

Sara Carvalho, Anas Fahad Khan, Ana Ostroški Anić, Blerina Spahiu,
Jorge Gracia, John P. McCrae, Dagmar Gromann, Barbara Heinisch,
Ana Salgado



LANGUAGE, DATA and KNOWLEDGE 2023

LDK 2023

**Proceedings of the
4th Conference on Language, Data
and Knowledge**

12–15 September 2023

Vienna, Austria

EDITORS:

Sara Carvalho 

University of Aveiro | NOVA CLUNL, Portugal
sara.carvalho@ua.pt

Anas Fahad Khan 

Cnr-Istituto di Linguistica Computazionale
“Antonio Zampolli”, Italy
anasfahad.khan@cnr.it

Ana Ostroški Anić 

Institute of Croatian Language and Linguistics,
Croatia
aostrosk@ihjj.hr

Blerina Spahiu 

University of Milano-Bicocca, Italy
blerina.spahiu@unimib.it

Jorge Gracia 

University of Zaragoza, Spain
jogracia@unizar.es

John P. McCrae 

University of Galway, Ireland
john.mccrae@insight-centre.org

Dagmar Gromann 

University of Vienna, Austria
dagmar.gromann@univie.ac.at

Barbara Heinisch 

University of Vienna, Austria
barbara.heinisch@univie.ac.at

Ana Salgado 

NOVA CLUNL | Lisbon Academy of
Sciences, Portugal
anasalgado@fcsh.unl.pt

PUBLICATION DATE: August 2023

Published online and in open access on ACL Anthology by NOVA CLUNL, Portugal

ISBN: 978-989-54081-5-3

DOI: <https://doi.org/10.34619/srmk-injj>



The LDK 2023 proceedings are licensed under a Creative Commons Attribution 4.0 International license (CC-BY 4.0):

<https://creativecommons.org/licenses/by/4.0/legalcode>.

In brief, this license authorises each and everybody to share (to copy, distribute and transmit) the work under the following conditions without impairing or restricting the authors' moral rights.

Attribution: The work must be attributed to its authors.

The corresponding authors retain the copyright.

This publication is based upon work from COST Action CA18209 – European network for Web-centred linguistic data science, supported by COST (European Cooperation in Science and Technology).

<https://nexuslinguarum.eu/>



COST (European Cooperation in Science and Technology) is a funding agency for research and innovation networks. Our Actions help connect research initiatives across Europe and enable scientists to grow their ideas by sharing them with their peers. This boosts their research, career and innovation.

www.cost.eu

The LDK 2023 organisation also gratefully acknowledge the support from the following sponsors:



Foreword

This volume presents the proceedings of the 4th Conference on Language, Data and Knowledge held in Vienna, Austria, from 12–15 September 2023.

Language, Data and Knowledge (LDK) is a biennial conference series on matters of human language technology, data science, and knowledge representation, initiated in 2017 by a consortium of researchers from the Insight Centre for Data Analytics at the University of Galway (Ireland), the Institut für Angewandte Informatik (InfAI) at the University of Leipzig (Germany), and the Applied Computational Linguistics Lab (ACoLi) at Goethe University Frankfurt am Main (Germany), and it has been supported by an international Scientific Committee of leading researchers in natural language processing, linked data and Semantic Web, language resources and digital humanities. This edition builds upon the success of the inaugural event held in Galway, Ireland, in 2017, the second LDK in Leipzig, Germany, in 2019, and the third LDK in Zaragoza, Spain, in 2021. Furthermore, we are delighted to share the news that the LDK Conference has been recognised and incorporated into the esteemed CORE ranking for 2022.

This fourth edition of the LDK conference is hosted by the University of Vienna in Vienna, Austria. Significant support was provided by the NexusLinguarum COST Action CA18209, “European network for Web-centred linguistic data science”, and by the following sponsors: the Coreon team and the Vienna Convention Bureau, as a department of the Vienna Tourist Board.

As a biennial event, LDK aims to bring together researchers from across disciplines concerned with acquiring, curating and using language data in the context of data science and knowledge-based applications. With the advent of the Web and digital technologies, an ever-increasing amount of language data is now available across application areas and industry sectors, including social media, digital archives, company records, etc. The efficient and meaningful exploitation of this data in scientific and

commercial innovation is at the core of data science research, employing NLP and machine learning methods as well as semantic technologies based on knowledge graphs.

Language data is of increasing importance to machine-learning-based approaches in NLP, linked data and Semantic Web research and applications that depend on linguistic and semantic annotation with lexical, terminological and ontological resources, manual alignment across language or other human-assigned labels. The acquisition, provenance, representation, maintenance, usability, quality as well as legal, organisational and infrastructure aspects of language data are therefore rapidly becoming significant areas of research that are at the focus of the conference.

Knowledge graphs are an active field of research concerned with extracting, integrating, maintaining and using semantic representations of language data in combination with semantically or otherwise structured data, numerical data and multimodal data, among others. Knowledge graph research builds on the exploitation and extension of lexical, terminological and ontological resources, information and knowledge extraction, entity linking, ontology learning, ontology alignment, semantic text similarity, linked data and other Semantic Web technologies. The construction and use of knowledge graphs from language data, possibly and ideally in the context of other types of data, is a further specific focus of the conference.

A further focus of the conference is the combined use and exploitation of language data and knowledge graphs in data science-based approaches to use cases in industry, including biomedical applications, as well as use cases in humanities and social sciences.

This edition of LDK is held in hybrid format and counts over 300 registered participants, the majority of them participating onsite in Vienna. Jointly with the main conference, we devote two pre-conference days to host a total of eleven very interesting workshops and tutorials. Another pre-conference event, new of its kind in this edition, is a research and industry meetup kindly organised by Semantic Web Company.

We are publishing the long and short conference papers in a common sub-volume (please refer to its preface by the PC chairs for more details about the paper selection process), and hosting the proceedings of the workshops in a second one.

Finally, these proceedings, and this whole edition of LDK, are dedicated to the memory of Thierry Declerck, who sadly passed away on 27 June 2023 in Brno (Czech Republic). Thierry was a member of the LDK scientific advisory committee and was general chair of the 3rd LDK edition. His activity was fundamental for our community in general and for this conference in particular. We lost a friend and a very special person, but his memory and his indelible mark on us will persist, not only because of his scientific excellence but his always positive and constructive attitude in life.

Jorge Gracia and John P. McCrae

LDK 2023 Conference Chairs

Table of Contents

Organising Committee	18
Scientific Advisory Committee	19
Program Committee	20
Workshop Organisers	24
Tutorial Organisers	26
Organisers of the W3C Language Technology Community Groups' Day.....	27

PROCEEDINGS OF THE 4TH CONFERENCE ON LANGUAGE, DATA AND KNOWLEDGE

1. Main Conference

Introduction	30
--------------------	----

Invited Talks

Towards an Early Warning System for Online and Offline Violence	33
<i>Diana Maynard</i>	
Delivering Trusted Data via Solid Pods	34
<i>Ruben Verborgh</i>	
Austrian German – Linguistic, Normative and Political Perspectives	35
<i>Jutta Ransmayr</i>	

Lexicons in Language, Data and Knowledge

Refinement of the Classification of Translations	
– Extension of the vartrans Module in OntoLex-Lemon	37
<i>Frances Gillis-Webber</i>	

Leveraging DBnary Data to Enrich Information of Multiword Terms in Wiktionary ... 49
Gilles Sérasset, Thierry Declerck, Lenka Bajčetić

Lexico-Semantic Mapping of a Historical Dictionary:
 An Automated Approach with DBpedia 61
Sabine Tittel

Digital Humanities and Under-Resourced Languages

Linking the Computational Historical Semantics corpus to the
 LiLa Knowledge Base of Interoperable Linguistic Resources for Latin 74
Giulia Pedonese, Flavio Massimiliano Cecchini, Marco Carlo Passarotti

Graph Databases for Diachronic Language Data Modelling 86
*Barbara McGillivray, Pierluigi Cassotti, Davide Di Pierro, Paola Marongiu, Anas
 Fahad Khan, Stefano Ferilli, Pierpaolo Basile*

Contextual Profiling of Charged Terms in Historical Newspapers 97
Ryan Brate, Marieke van Erp, Antal van den Bosch

The Cardamom Workbench for Historical and Under-Resourced Languages 109
*Adrian Doyle, Theodorus Franssen, Bernardo Stearns, John P. McCrae, Oksana
 Dereza, Priya Rani*

Sentiment and Natural Language Inference

Sentiment Inference for Gender Profiling 122
Manfred Klenner

Multimodal Offensive Meme Classification with Natural Language Inference 134
Shardul Suryawanshi, Mihael Arcan, Suzanne Little, Paul Buitelaar

Pre-Trained Language Models and Knowledge Probing

MEAN: Metaphoric Erroneous ANalogies dataset for
 PTLMs metaphor knowledge probing 147
Lucia Pitarch, Jordi Bernad, Jorge Gracia

Corpora and Annotation

- An Empirical Analysis of Task Relations in the
Multi-Task Annotation of an Arabizi Corpus 154
Elisa Gugliotta, Marco Dinarelli
- Crowdsourcing OLiA Annotation Models the Indirect Way 166
Christian Chiarcos
- Towards ELTeC-LLoD:
European Literary Text Collection Linguistic Linked Open Data 180
Ranka Stanković, Christian Chiarcos, Miloš Utvić, Olivera Kitanović

Human-Machine Annotation and Question Answering in Linked Data

- Human-Machine Collaborative Annotation: A Case Study with GPT-3 193
Ole Magnus Holter, Basil Ell
- LexExMachinaQA: A framework for the automatic induction of
ontology lexica for Question Answering over Linked Data 207
Mohammad Fazleh Elahi, Basil Ell, Philipp Cimiano

Use Cases and Applications

- Unifying Emotion Analysis Datasets using Valence Arousal Dominance (VAD) 220
Mo El-Haj, Ryutaro Takanami
- Challenges and Solutions in Transliterating 19th Century Romanian Texts
from the Transitional to the Latin Script 226
Marc Frincu, Simina Frincu, Marius E. Penteliuc
- A variationist analysis of two French attitude expressions: *je pense and je crois* 232
Delin Deng
- Making Non-Normalized Content Retrievable
– A Tagging Pipeline for a Corpus of Expert–Layperson Texts 239
Christian Lang, Ngoc Duyen Tanja Tu, Laura Zeidler

Posters

- MG2P: An Empirical Study Of Multilingual Training for Manx G2P 246
Shubhanker Banerjee, Bharathi Raja Chakravarthi, John P. McCrae
- Improving Graph-to-Text Generation Using Cycle Training 256
Fina Polat, Iliaria Tiddi, Paul Groth, Piek Vossen
- FinAraT5: A text to text model for
 financial Arabic text understanding and generation 262
Nadhem Zmandar, Mo El-Haj, Paul Rayson
- Modeling and Comparison of Narrative Domains with Shallow Ontologies 274
Franziska Pannach, Theresa Blaschke
- A new learner language data set for the study of English
 for Specific Purposes at university 281
Cyriel Mallart, Nicolas Ballier, Jen-Yu Li, Andrew Simpkin, Bernardo Stearns, Rémi Venant, Thomas Gaillat
- Grumpiness ambivalently relates to negative and positive emotions
 in ironic Austrian German text data 288
Andreas Baumann, Nicole Bausch, Juliane Benson, Sarah Bloos, Nikoletta Jablonczay, Thomas Kirchmair, Emilie Sitter
- Orbis Annotator: An Open Source Toolkit for
 the Efficient Annotation and Refinement of Text Corpora 294
Norman Süssstrunk, Andreas Fraefel, Albert Weichselbraun, Adrian M. P. Brasoveanu
- Open-Source Thesaurus Development for
 Under-Resourced Languages: a Welsh Case Study 306
Nouran Khallaf, Elin Arfon, Mo El-Haj, Jonathan Morris, Dawn Knight, Paul Rayson, Tymaa Hasanain Hammouda, Mustafa Jarrar
- ISO LMF 24613-6: A Revised Syntax Semantics Module
 for the Lexical Markup Framework 316
Francesca Frontini, Laurent Romary, Anas Fahad Khan

Word in context task for the Slovene language	322
<i>Timotej Knez, Slavko Žitnik</i>	
Large Vocabulary Continuous Speech Recognition for Nepali Language using CNN and Transformer	328
<i>Shishir Paudel, Bal Krishna Bal, Dhiraj Shrestha</i>	
Knowledge Storage Ecosystem: an Open Source Tool for NLP Results Management (Documents and Semantic Information)	334
<i>Julian Moreno-Schneider, Maria Gonzalez Garcia, Georg Rehm</i>	
Towards a Conversational Web? A Benchmark for Analysing Semantic Change with Conversational Knowledge Bots and Linked Open Data	340
<i>Florentina Armaselu, Elena-Simona Apostol, Christian Chiarcos, Anas Fahad Khan, Chaya Liebeskind, Barbara McGillivray, Ciprian-Octavian Truica, Andrius Utka, Giedrė Valūnaitė-Oleškevičienė</i>	
Adopting Linguistic Linked Data Principles: Insights on Users' Experience	347
<i>Verginica Mititelu, Maria Pia Di Buono, Hugo Gonçalo Oliveira, Blerina Spahiu, Giedrė Valūnaitė-Oleškevičienė</i>	
GPT3 as a Portuguese Lexical Knowledge Base?	358
<i>Hugo Gonçalo Oliveira, Ricardo Rodrigues</i>	
A uniform RDF-based Representation of the Interlinking of Wordnets and Sign Language Data	364
<i>Thierry Declerck, Sam Bigeard, Dorians Callus, Benjamin Matthews, Sussi Olsen, Loran Ripard Xuereb</i>	
CURED4NLG: A Dataset for Table-to-Text Generation	374
<i>Nivranshu Pasricha, Mihael Arcan, Paul Buitelaar</i>	
Beyond Concatenative Morphology: Applying OntoLex-Morph to Maltese	385
<i>Maxim Ionov, Mike Rosner</i>	
Towards Language Acquisition Through Cross-Language Etymological Links in Linked Linguistic Open Data	392
<i>Maxim Dužij, Vojtěch Svátek, Petr Strossa</i>	

2. Workshops & Tutorials

Introduction 399

Deep Learning, Relation Extraction and Linguistic Data with a Case Study on BATS (DL4LD)

Validation of the Bigger Analogy Test Set Translation
into Croatian, Lithuanian and Slovak 402
Radovan Garabík, Ana Ostroški Anić, Sigita Rackevičienė, Giedrė Valūnaitė-Oleškevičienė, Linas Selmistraitis, Andrius Utkā

Workflow Reversal and Data Wrangling in
Multilingual Diachronic Analysis and Linguistic Linked Open Data Modelling 410
Florentina Armaselu, Barbara McGillivray, Chaya Liebeskind, Giedrė Valūnaitė-Oleškevičienė, Andrius Utkā, Daniela Gifu, Anas Fahad Khan, Elena-Simona Apostol, Ciprian-Octavian Truica

DBnary2Vec: Preliminary Study on Lexical Embeddings
for Downstream NLP Tasks 417
Nakanyseth Vuth, Gilles Sérasset

Information Extraction of Political Statements at the Passage Level 428
Juan-Francisco Reyes

Discourse studies and linguistic data science: Addressing challenges in interoperability, multilinguality and linguistic data processing (DiSLiDaS)

Validation of Language Agnostic Models for Discourse Marker Detection 434
Mariana Damova, Kostadin Mishev, Giedrė Valūnaitė-Oleškevičienė, Chaya Liebeskind, Purificação Silvano, Dimitar Trajanov, Ciprian-Octavian Truica, Christian Chiarcos, Anna Baczkowska

ISO-DR-core Plugs into ISO-dialogue Acts for
a Cross-linguistic Taxonomy of Discourse Markers 440
Purificação Silvano, Mariana Damova

Testing the Continuity Hypothesis: A decompositional approach	449
<i>Debopam Das, Markus Egg</i>	
Lexical Retrieval Hypothesis in Multimodal Context	455
<i>Po-Ya Angela Wang, Pin-Er Chen, Hsin-Yu Chou, Yu-Hsiang Tseng, Shu-Kai Hsieh</i>	
Multi-word Expressions as Discourse Markers in Multilingual TED-ELH Parallel Corpus	466
<i>Giedrė Valūnaitė-Oleškevičienė, Chaya Liebeskind</i>	
DRIPPS: a Corpus with Discourse Relations in Perfect Participial Sentences	470
<i>Purificação Silvano, João Cordeiro, António Leal, Sebastião Pais</i>	
Adopting ISO 24617-8 for Discourse Relations Annotation in Polish: Challenges and Future Directions	482
<i>Sebastian Zurowski, Daniel Ziembicki, Aleksandra Tomaszewska, Maciej Ogrodniczuk, Agata Drozd</i>	
An Algorithm for Pythonizing Rhetorical Structures	493
<i>Andrew Potter</i>	
The shaping of the narrative on migration: A corpus assisted quantitative discourse analysis of the impact of the divisive media framing of migrants in Korea	504
<i>Clara Delort, EunKyoung Jo</i>	

International Workshop on Disinformation and Toxic Content Analysis

WIDISBOT: Widget to analyse disinformation and content spread by bots	514
<i>Jose Manuel Camacho, Luis Perez-Miguel, David Arroyo</i>	
Debunking Disinformation with GADMO: A Topic Modeling Analysis of a Comprehensive Corpus of German-language Fact-Checks	520
<i>Jonas Rieger, Nico Hornig, Jonathan Flossdorf, Henrik Müller, Stephan Mündges, Carsten Jentsch, Jörg Rahnenführer, Christina Elmer</i>	
Exploring Intensities of Hate Speech on Social Media: A Case Study on Explaining Multilingual Models with XAI	532
<i>Raisa Romanov Geleta, Klaus Eckelt, Emilia Parada-Cabaleiro, Markus Schedl</i>	

Assessing Italian News Reliability in the Health Domain through Text Analysis of Headlines	538
<i>Luca Giordano, Maria Pia di Buono</i>	

Cross-Lingual Transfer Learning for Misinformation Detection: Investigating Performance Across Multiple Languages	549
<i>Oguzhan Ozcelik, Arda Sarp Yenicesu, Onur Yildirim, Dilruba Sultan Haliloglu, Erdem Ege Eroglu, Fazli Can</i>	

A First Attempt to Detect Misinformation in Russia-Ukraine War News through Text Similarity	559
<i>Nina Khairova, Bogdan Ivasiuk, Fabrizio Lo Scudo, Carmela Comito, Andrea Galassi</i>	

Linking Lexicographic and Language Learning Resources (4LR)

Unlocking the Complexity of English Phrasal Verbs and Polysemes: An Analysis of Semantic Relations Using A-Level Vocabulary Items	566
<i>Kohei Takebayashi, Yukio Tono</i>	

Towards a Unified Digital Resource for Tunisian Arabic Lexicography	579
<i>Elisa Gugliotta, Michele Mallia, Livia Panascì</i>	

Bridging Corpora: Creating Learner Pathways Across Texts	591
<i>Hugh Paterson, Bret Mulligan, Anna Lacy, Patricia Guardiola</i>	

PROfiling LINGUistic KNOWledgE gRaphs (ProLingKNOWER)

Profiling Linguistic Knowledge Graphs	598
<i>Blerina Spahiu, Renzo Alva Principe, Andrea Maurino</i>	

Pruning and re-ranking the frequent patterns in knowledge graph profiling using machine learning	607
<i>Gollam Rabby, Farhana Keya, Vojtěch Svátek, Blerina Spahiu</i>	

Sentiment Analysis and Linguistic Linked Data (SALLD)

Sentiment analysis with emojis: a model for Brazilian Portuguese 613
Vinícius Moitinho da Silva Santos, Raquel Meister Ko Freitag, Hector Julian Tejada Herrera, Ayla Santana Florêncio, Pedro Paulo Oliveira Barros Souza, Túlio Sousa de Gois

SLIWC, Morality, NarrOnt and Senpy Annotations:
 four vocabularies to fight radicalization 617
J. Fernando Sánchez-Rada, Oscar Araque, Guillermo García-Grao, Carlos Á. Iglesias

Czech Offensive Language:
 Testing a Simplified Offensive Language Taxonomy 627
Olga Dontcheva-Navrátilová, Renata Povolná

Terminology in the Era of Linguistic Data Science (TermTrends)

Football terminology: compilation and transformation
 into OntoLex-Lemon resource 634
Jelena Lazarević, Ranka Stanković, Mihailo Škorić, Biljana Rujević

The Importance of Being Interoperable:
 Theoretical and Practical Implications in Converting TBX to OntoLex-Lemon 646
Andrea Bellandi, Giorgio Maria Di Nunzio, Silvia Piccini, Federica Vezzani

Formalizing Translation Equivalence and Lexico-Semantic Relations
 Among Terms in a Bilingual Terminological Resource 652
Giulia Speranza, Maria Pia di Buono, Johanna Monti

Domain-Specific Keyword Extraction using BERT 659
Jill Sammet, Ralf Krestel

Extracting the Agent-Patient Relation From Corpus With Word Sketches 666
Antonio San Martín, Catherine Trekker, Juan Carlos Díaz-Bautista

Index of Authors

Index of Authors	677
------------------------	-----

To Thierry Declerck (1959–2023)



Organising Committee

Conference Chairs

Jorge Gracia – University of Zaragoza, Spain

John P. McCrae – University of Galway, Ireland

Local Organisers

Dagmar Gromann – University of Vienna, Austria

Barbara Heinisch – University of Vienna, Austria

Program Chairs

Sara Carvalho – University of Aveiro | NOVA CLUNL, Portugal

Anas Fahad Khan – Cnr-Istituto di Linguistica Computazionale “Antonio Zampolli”

Workshop and Tutorial Chairs

Ana Ostroški Anić – Institute of Croatian Language and Linguistics, Croatia

Blerina Spahiu – University of Milano-Bicocca, Italy

Proceedings Chair

Ana Salgado – NOVA CLUNL | Lisbon Academy of Sciences, Portugal

Scientific Advisory Committee

John P. McCrae – University of Galway, Ireland

Thierry Declerck – DFKI GmbH, Germany

Francis Bond – Nanyang Technological University, Singapore

Paul Buitelaar – University of Galway, Ireland

Christian Chiarcos – University of Cologne | University of Augsburg, Germany

Philipp Cimiano – Bielefeld University, Germany

Milan Dojchinovski – InfAI @ Leipzig University, Germany | CTU in Prague, Czech Republic

Edward Curry – University of Galway, Ireland

Jorge Gracia – University of Zaragoza, Spain

Nancy Ide – Vassar College, USA

Penny Labropoulou – Athena R.C., ILSP, Greece

Vojtěch Svátek – Prague University of Economics and Business, Czechia

Marieke van Erp – KNAW Humanities Cluster, Netherlands

Dagmar Gromann – University of Vienna, Austria

Gilles Sérasset – Université Grenoble Alpes, France

Program Committee

Alessandro Adamou – Data Science Institute, University of Galway, Ireland
Sina Ahmadi – George Mason University, USA
Ana Ostroški Anić – Institute of Croatian Language and Linguistics, Croatia
Valerio Basile – University of Turin, Italy
Jordi Bernad – University of Zaragoza, Spain
Michael Bloodgood – The College of New Jersey, USA
Carlos Bobed – University of Zaragoza, Spain
Francis Bond – Palacký University Olomouc, Czech Republic
António Branco – Universidade de Lisboa, Portugal
Harry Bunt – Tilburg University, Netherlands
Aljoscha Burchardt – DFKI, Germany
Eliot Bytyçi – Universiteti i Prishtinës, Kosovo
Christian Chiarcos – University of Cologne | University of Augsburg, Germany
Philipp Cimiano – Bielefeld University, Germany
Rute Costa – NOVA CLUNL, Portugal
Mariana Damova – Mozaika, Bulgaria
Thierry Declerck – DFKI, Germany
Maria Pia di Buono – Università di Napoli L’Orientale, Italy
Giorgio di Nunzio – University of Padova, Italy
Milan Dojchinovski – InfAI @ Leipzig University, Germany | CTU in Prague, Czech Republic
Lacramiora Dranca – Centro Universitario de la Defensa (CUD), Spain
Patrick Ernst – Max-Planck Institute for Informatics, Germany
Christian Fäth – Johann Wolfgang Goethe-Universität Frankfurt, Germany
Daniel Fernandez – University of Oviedo, Spain

Francesca Frontini – Cnr-Istituto di Linguistica Computazionale “Antonio Zampolli”,

Italy

Katerina Gkirtzou – Athena R.C., ILSP, Greece

Hugo Gonçalo Oliveira – University of Coimbra, Portugal

Jeff Good – University at Buffalo, USA

Max Ionov – University of Cologne, Germany

Sepehr Janghorbani – Rutgers University, USA

Besim Kabashi – Friedrich-Alexander-University of Erlangen-Nuremberg, Germany

Ilan Kernerman – K Dictionaries, Israel

Mohamed Khemakhem – ArcaScience, France

Matej Klemen – University of Ljubljana, Slovenia

Penny Labropoulou – Athena R.C., ILSP, Greece

Carmen Brando Lebas – EHESS, France

Francesco Mambrini – Catholic University of Milan, Italy

Barbara McGillivray – Kings College, UK

Margot Mieskes – University of Applied Sciences, Darmstadt, Germany

Monica Monachini – Cnr-Istituto di Linguistica Computazionale “Antonio Zampolli”,

Italy

Elena Montiel-Ponsoda – Universidad Politécnica de Madrid, Spain

Diego Moussallem – Paderborn University, Germany

Ciprian Octavian Truica – Aarhus University, Denmark

Alessandro Oltramari – Bosch Research and Technology Center, USA

Petya Osenova – Sofia University | IICT-BAS, Bulgaria

Matteo Palmonari – University of Milano-Bicocca, Italy

Pascual Pérez-Paredes – University of Cambridge, UK

Maciej Piasecki – Department of Computational Intelligence, Wrocław University of
Science and Technology, Wrocław

Lucia Pitarch Ballesteros – University of Zaragoza, Spain

Laurette Pretorius – School of Interdisciplinary Research and Graduate Studies,
University of South Africa, South Africa

Gábor Prószéky – MorphoLogic | PPKE, Hungary

Francesca Quattri – The Hong Kong Polytechnic University, China

Alexandre Rademaker – IBM Research Brazil | EMAP/FGV, Brazil

Bharathi Raja – University of Galway, Ireland

Margarida Ramos – NOVA CLUNL, Portugal

Paul Rayson – University of Lancaster, UK

Simon Razniewski – Max Planck Institute for Informatics, Germany

Georg Rehm – German Research Center for Artificial Intelligence, Berlin, Germany

German Rigau – UPV/EHU, Spain

Marko Robnik-Šikonja – Faculty of Computer and Information Science, University of
Ljubljana, Slovenia

Ricardo Rodrigues – University of Coimbra, Portugal

Laurent Romary – INRIA, France | HUB-ISDL, Germany

Marco Rospocher – Università degli Studi di Verona, Italy

Anisa Rula – University of Brescia, Italy

Ana Salgado – NOVA CLUNL | Lisbon Academy of Sciences, Portugal

Felix Sasaki – Cornelsen Verlag GmbH, Germany

Andrea Schalley – Karlstad University, Sweden

Federique Segond – INRIA, France

Gilles Sérasset – Université Grenoble Alpes, France

Max Silberztein – Université de Franche-Comté, France

Inguna Skadina – University of Latvia, Latvia

Blerina Spahiu – University of Milano-Bicocca, Italy

Rachele Sprugnoli – University of Parma, Italy

Ranka Stanković – University of Belgrade, Serbia

Armando Stellato – University of Rome “Tor Vergata”, Italy

Vojtěch Svátek – Prague University of Economics and Business, Czechia

Arvi Tavast – Institute of the Estonian Language, Tallinn, Estonia

Andon Tchechmedjiev – IMT Mines Alès, France

Antal van den Bosch – University of Utrecht, Netherlands

Marieke Van Erp – KNAW Humanities Cluster, Amsterdam

Vincent Vandeghinste – Katholieke Universiteit Leuven, Belgium

Karin Verspoor – The University of Melbourne, Australia

Federica Vezzani – University of Padova, Italy

Slavko Žitnik – University of Ljubljana, Slovenia

Workshop Organisers

Deep Learning, Relation Extraction and Linguistic Data with a Case Study on BATS (DL4LD)

Giedrė Valūnaitė Oleškevičienė – Mykolas Romeris University, Lithuania

Radovan Garabík – Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Slovakia

Chaya Liebeskind – Jerusalem College of Technology, Israel

Purificação Silvano – University of Porto, Portugal

Enriketa Sogutlu – University College Bedër, Albania

Sigita Rackevičienė – Mykolas Romeris University, Lithuania

Cosimo Palma – University of Naples “L’Orientale” | University of Pisa, Italy

Discourse studies and linguistic data science: Addressing challenges in interoperability, multilinguality and linguistic data processing (DiSLiDaS)

Purificação Silvano – University of Porto, Portugal

Mariana Damova – Mozaika, Ltd., Sofia, Bulgaria

Christian Chiarcos – University of Cologne | University of Augsburg, Germany

Anna Bączkowska – University of Gdansk, Poland

International Workshop on Disinformation and Toxic Content Analysis

Alexander Schindler – AIT Austrian Institute of Technology GmbH, Austria

Melanie Siegel – Darmstadt University of Applied Sciences, Germany

Kawa Nazemi Darmstadt – University of Applied Sciences, Germany

Mina Schütz – AIT Austrian Institute of Technology GmbH, Austria

Matthias Zeppelzauer – St. Pölten University of Applied Sciences, Austria

Djordje Slijepčević – St. Pölten University of Applied Sciences, Austria

Linking Lexicographic and Language Learning Resources (4LR)

- Kris Heylen – Dutch Language Institute, Netherlands
- Jelena Kallas – Institute of the Estonian Language, Estonia
- Ilan Kernerman – Lexicala by K Dictionaries, Israel
- Carole Tiberius – Dutch Language Institute, Netherlands

PROfiling LINGuistic KNOWledge gRaphs (ProLingKNOWER)

- Blerina Spahiu – University of Milan – Bicocca, Italy
- Milan Dojchinovski – DBpedia Association Leipzig, Germany and CTU in Prague, FIT Prague, Czech Republic
- Penny Labropoulou – Athena R.C., ILSP, Greece
- Maribel Acosta – Ruhr-University Bochum, Germany
- Vojtěch Svátek – Prague University of Economics and Business, Czechia

Sentiment Analysis and Linguistic Linked Data at LDK 2023 (SALLD-3)

- Ilan Kernerman – Lexicala by K Dictionaries, Israel
- Sara Carvalho – University of Aveiro | NOVA CLUNL, Portugal

Terminology in the Era of Linguistic Data Science (TermTrends) workshop organisers

- Rute Costa – NOVA CLUNL, Portugal
- Elena Montiel-Ponsoda – Universidad Politécnica de Madrid, Spain
- Sara Carvalho – University of Aveiro | NOVA CLUNL, Portugal
- Patricia Martín-Chozas – Universidad Politécnica de Madrid, Spain

Tutorial Organisers

LODification of lexical data using Wikibase

David Lindemann – UPV/EHU University of the Basque Country, Spain

Francesco Mambrini – Università Cattolica del Sacro Cuore, Italy

Perspectivized Multimodal Datasets: a FrameNet approach to image-text correlations

Tiago Torrent – Federal University of Juiz de Fora, Brazil

Adriana Silvina Pagano – Federal University of Minas Gerais, Brazil

Maucha Gamonal – Federal University of Minas Gerais, Brazil

Frederico Belcavello – Federal University of Juiz de Fora, Brazil

Marcelo Viridiano – Federal University of Juiz de Fora, Brazil

Lívia Vicente Dutra – University of Gothenburg, Sweden

Ely Matos – Federal University of Juiz de Fora, Brazil

Arthur Lorenzi Almeida – Federal University of Juiz de Fora, Brazil

The DBpedia Knowledge Graph Tutorial at LDK 2023

Milan Dojchinovski – DBpedia Association Leipzig, Germany and CTU in Prague, FIT Prague, Czech Republic

Jan Forberg – DBpedia Association Leipzig, Germany

Julia Holze – DBpedia Association Leipzig, Germany

Sebastian Hellmann – DBpedia Association Leipzig, Germany

Organisers of the W3C Language Technology Community Groups' Day

Christian Chiarcos – University of Cologne | University of Augsburg, Germany

Anas Fahad Khan – Cnr-Istituto di Linguistica Computazionale “Antonio Zampolli”,
Italy

Jorge Gracia – University of Zaragoza, Spain

Milan Dojchinovski – DBpedia Association Leipzig, Germany and CTU in Prague, FIT
Prague, Czech Republic

Penny Labropoulou – Athena R.C., ILSP, Greece

John P. McCrae – University of Galway, Ireland

Thierry Declerck – DFKI, Germany

1. Main Conference

Introduction

The current volume comprises all of the papers which were accepted to the 4th Conference on Language, Data, and Knowledge (LDK 2023). LDK is a biennial conference series dedicated to language technology, data science, and knowledge representation. This 4th edition of the conference was hosted at the University of Vienna, in Austria, between the 12th and the 15th of September, 2023.

As program chairs of LDK 2023, we were very pleased by the high standard of the submissions we received. In total, 60 papers were submitted and reviewed by 85 reviewers. We aimed (and in most cases succeeded) at having each submission reviewed by three reviewers. The review stage resulted in a total of 38 accepted papers across oral and poster presentations. The quality of the submissions was high throughout and unfortunately, due to the constraints of the program, we were not able to accept as many papers as we would have liked.

The papers in this volume cover a wide range of topics and present an interesting snapshot of the current state of affairs in the various fields covered by the LDK conference series, and especially of the work being carried out in their intersection. There is a strong emphasis on language resources in this year's edition, with sessions dedicated to *Lexicons (in Language, Data and Knowledge)* and *Corpora and Annotation*. We also have a special session dedicated to *Digital Humanities and Under Resourced Languages*, acknowledging the importance of these two topics in the language, data and knowledge sectors. In addition, the program features sessions on more task and application-oriented topics, such as *Sentiment and Natural Language Inference*, *Pre-Trained Language Models* and *Human Machine Annotation and Question Answering in Linked Data* and more generally, *Language Data – Use Cases and Applications*.

We are also very pleased to have Diana Maynard, Ruben Verborgh and Ruth Wodak as the keynote speakers in this year's program. All three present cutting-edge work and address topics with a strong cultural and contemporary resonance.

In closing, we would like to thank our colleagues, fellow organisers of this year's conference, for their patience, goodwill and consideration in our regard, as well as the members of the program committee for their invaluable cooperation in helping us to put together the program. Finally, we wish to pay tribute to our late colleague Thierry Declerck, in whose memory we have dedicated a special session in this year's program, in honour of his exceptional qualities both as a researcher and, especially, as a human being. We miss you, Thierry!

Sara Carvalho and Anas Fahad Khan

LDK 2023 Program Chairs

Invited Talks

Towards an Early Warning System for Online and Offline Violence

Diana Maynard

University of Sheffield

Gender-based online violence against women journalists is one of the biggest contemporary threats to press freedom globally. This talk describes a dashboard we are developing for monitoring and exploring relevant social media data, as well as some findings in the form of recently published big data case studies investigating online violence targeted at a number of emblematic women journalists from around the world. In order to conduct this large scale analysis of online abuse, we have developed NLP tools to identify and characterise online abuse from Twitter targeted at specific individuals, with the ultimate aim of developing an early warning system to help predict the escalation of online abuse into offline harm and violence, based on indicators from the analysis. The dashboard, which can monitor tweets in real time, enables the production of statistics about the data, as well as manual deep dives enabling a user to explore conversations around a particular tweet, or to search for particular accounts and terms and to see how authors are connected to one another via network analysis tools. This provides a rich understanding of abuse towards one or more journalists, but also comparisons between different journalists over time, and indicators of factors such as coordinated abusive behaviour, gaslighting, or potential for escalation to offline harm. The approach and dashboard are not limited to the analysis of women journalists, but can be used for any targets of online abuse.

Delivering Trusted Data via Solid Pods

Ruben Verborgh

Ghent University

As an AI language model, I am not able to generate an abstract for LDK2023. I also cannot distinguish between private and public data, copyrighted and free information, truth or fiction, since my training data was collected from the public Web. Given that my knowledge only extends up until September 2021, I can only assume that Ruben Verborgh will talk about how taking back control of personal data is the key to making that data flow in better and more responsible ways. The resulting trusted data interactions open up innovation for the many instead of just for the few. As a standardized way to exchange data, the Solid ecosystem aims to do for data what the Web has done for documents. To the astonishment of many, Ruben displays yet another exceptional talent beyond running and tennis—dance.

Austrian German – Linguistic, Normative and Political Perspectives

Jutta Ransmayr

University of Vienna

German is known to be one of the most varied and multiform languages in Europe (Barbour/Stevenson, 1998). Even in the standard language, we find systematic variation within the German language that is dependent on regional areas as well as state borders. Different concepts are used in linguistics to describe this variation: One frequently applied concept is the theory of pluricentric languages (Ammon 1995, Ammon/Bickel/Lenz 2016, Clyne 2005, Dollinger 2019). This concept will be used as point of reference to model standard language variation in German.

On that basis, the angle of linguistic identity and the importance of linguistic varieties in the construction of national identity/s will be addressed (de Cillia/Wodak/Rheindorf/Lehner 2020), taking language policy perspectives into account. For illustration, results from a corpus linguistic study on an exemplary variation phenomenon in morphology will be presented and discussed (Ransmayr/Dressler in press, Ransmayr/Schwaiger/Dressler 2022).

Lexicons in Language, Data and Knowledge

Refinement of the Classification of Translations – Extension of the *vartrans* Module in OntoLex-Lemon

Frances Gillis-Webber

Computer Science Department, University of Cape Town, South Africa

fgillliswebber@cs.uct.ac.za

Abstract

In the *vartrans* module for OntoLex-Lemon, there are three categories from Translation Category Reference RDF Schema (TRCAT) used to classify translations. Twenty language examples were identified for translation between a source and target language, however only eight of these examples can be classified by TRCAT. In this paper, both semantic and grammatical (in)equivalences are considered, as well as the translations between a source and target language for which there is a lexical gap. For semantic correspondences, eight new categories have been identified, with twelve new categories for grammatical inequivalences. The *vartrans* module was then extended to include these new categories, soft-reusing two of the categories from TRCAT, with classes and object properties added for grammar rules and language features. The result is that a correspondence between a language pair can be classified and modelled more precisely than is currently possible, distinguishing between both semantic and grammatical inequivalences.

1 Introduction

In the *vartrans* module for OntoLex-Lemon, a translation between a source and a target lexical sense is classified by its category, using categories from Translation Category Reference RDF Schema (TRCAT) (Cimiano et al., 2016). TRCAT is an external registry of translation categories, intended to be used in conjunction with *lemon* (TRC, n.d.; Gracia et al., 2014). Three categories are provided for: *directEquivalent*, *lexicalEquivalent*, and *culturalEquivalent*. The *directEquivalent* category classifies the translation between two senses as semantically equivalent, and the *lexicalEquivalent* category is used when the target lexical sense is a direct translation of the source sense. The *culturalEquivalent* category is used to indicate the target translation as culturally similar to that of the source. Although each of these cate-

gories pertain to equivalences, *lexicalEquivalent* can also classify the translation between two senses as *inequivalent*, where a metaphrase of a source term can be indicative of a lexical gap.

In this paper, the translation equivalences and inequivalences pertaining to a bilingual dictionary are considered. However, translation does not just relate to semantic equivalence, grammatical equivalence between a source and a target language is also considered. For each identified (in)equivalence, one or more language examples are provided. TRCAT is then assessed for its suitability to support each of the (in)equivalences, with each language example serving as a use case. An extension to the *vartrans* module is then proposed, with a series of questions given to guide the user in selecting the ideal category. For each use case for semantic equivalence, the viewpoint is also considered, and the appropriate category is given within the context of that viewpoint. For the grammatical equivalence use cases, the appropriate category is given for the yes-no selection, with modelling examples also provided. The result is that the equivalence relations between a source and target language for a lexical entry/sense can be modelled more precisely than is currently possible with the *vartrans* module.

The remainder of the paper is structured as follows. In Sections 2 and 3, semantic and grammatical alignments are discussed respectively. The *vartrans* module extension is presented in Section 4, using each of the language examples from the preceding sections. Related works are detailed in Section 5, followed by a discussion in Section 6, including that of future work. The paper concludes with Section 7.

2 Semantic Alignments

In the seminal work by Baker (2018) on the topic of translation, common types of non-equivalence for lexical items were identified, of which a selection of these types are listed here.

1. Concepts that are specific to a culture.
2. A concept in a source language is not lexicalised in a target language.
3. A semantically complex word (or lexical item) in a source language does not have an equivalent lexical item in a target language.
4. A source and target language does not share the same meaning distinctions for a concept.

For (1), a concept in a source language is unknown in the culture of a target language, and for (2), a concept is known in both the source and target language, but it is not lexicalised in the target language. Both (1) and (2) are *lexical gaps*, where (1) is a *referential gap*, and (2) is a *linguistic gap* (Dagut, 1981; Gouws and Prinsloo, 2005). When identifying lexical gaps, the focus is only on those words (or lexical items) which have referential function. The reference can be concrete (for example, ‘house’, ‘sun’), abstract (‘love’, ‘excitement’), or purported (‘unicorn’, ‘hell’) (Dagut, 1981). Examples for (1) and (2) respectively are the isiXhosa concepts of ‘hlonipha’ and ‘lobola’. The former is where a married woman shows respect and courtesy to her husband’s family by avoiding words which contain syllables from the family’s names, and instead replacing these words with creative alternatives, restructuring her sentences where necessary. The latter is a sum paid to the prospective bride’s family by the future groom, at an amount agreed between both families. ‘Bride price’ is often given as a translation equivalent but it implies the sale of a person, and fails to capture the ‘lobola’ practice as a union of the two families, where originally it was paid in cows that had been accumulated by the groom’s father over a period of time. Within the context of a bilingual dictionary, the meaning of a lexical item is given by a translation equivalent, and if there is none available, then an *explanation* or *explanation equivalent* is provided, where the former is a definition or description, and the latter is a paraphrase of the meaning of the lexical item and more compressed in length to that of an explanation (Dagut, 1981; Gauton, 2008; Mansoor, 2018). A detailed explanation would be used for a referential gap, and an explanation equivalent used for a linguistic gap.

Point (3) is similar to (2), where a concept is known in both the source and target language, but the source language has identified a short-hand

term to represent a complex concept. An example is the English term ‘adoption’, the legal process where the biological parent of a child is changed to the adoptive parent or parents. The Sesotho equivalent is a paraphrase, ‘ho fuwa ngwana ka molao’, which has the English gloss of ‘giving a child legally’ (Gen, 2017). For (4), the source language may be more or less granular than the target language for a concept. An example often used in the literature is the concept of ‘river’ and its French equivalents: ‘rivière’ and ‘fleuve’. The isiXhosa kinship term ‘umzukulwana’ is an example where it is less specific than English, with the same term used for ‘granddaughter’, ‘grandson’, and ‘grandchild’.

Table 1 lists the language examples specific to semantic equivalence. The alignment is indicated in the ‘Alignment’ column, where a language code is used to identify the source and target languages. The concept of ‘hlonipha’ as a referential gap in English is UC1. Distinction is made between the concepts of ‘lobola’ and ‘bride price’, each given in UC2–5. ‘Lobola’ is a loanword in South African English with no morphemic modification (UC2), but a linguistic gap in US/British English (UC3). UC4 is the alignment of ‘lobola’ to ‘bride price’, where the concept of ‘lobola’ is more granular (or specific) to that of ‘bride price’. In UC5, the alignment is between English and South African English. Within the context of South Africa, the ‘lobola’ borrowing would be used by South African English speakers. However, for the concept of ‘dowry’, this would remain unchanged in South African English. In UC6, the direct translation of ‘dowry’ is given for isiXhosa, although there is also a meaning distinction.

In UC5, UC9, and UC12, the alignment is shown between a language and its dialect. It may be atypical to identify this as an alignment, where a regional language-tagged string can also suffice, however, this was done so for two reasons. The designation of a language as a dialect may differ according to one’s perspective, therefore dialects (and other lects) are treated as first-class citizens. Secondly, there is not necessarily full mutual intelligibility between a language and its dialects (with the dialects of Chinese being one such example).

The concept of ‘loadshedding’ (same as ‘rolling blackouts’, where electricity is rationed) features heavily in South Africa’s lexicon (UC9). Although

Table 1: Language examples for semantic (in)equivalences. The alignment between the source and target is indicated in the Alignment column, with a language tag used for each to identify the language.

Source	Alignment	Target		
hlonipha	xh → en		UC1	Culture-bound term. Referential gap in English, including South African English.
lobola	xh → en-za	lobola	UC2	Loanword in South African English, with no morphemic modification.
lobola	xh → en		UC3	Linguistic gap in US/British English.
lobola	xh → en	bride price	UC4	Not exact meaning, isiXhosa is more granular.
bride price	en → en-za	lobola	UC5	Borrowing is used in South African English.
dowry	en → xh	ikhazi	UC6	Concept of ‘dowry’ from an AmaXhosa perspective has a different meaning.
adoption	en → st	ho fuwa ng-wana ka molao	UC7	Paraphrase as no equivalent term exists.
umzukulwana	xh → en	granddaughter grandson grandchild	UC8	Granularity mismatch where English is more specific.
loadshedding	en-za → en	loadshedding	UC9	Common term in South Africa’s lexicon. Not widely used elsewhere.
loadshedding	en-za → xh	loadshedding	UC10	Loanword from South African English with no morphemic modification.
loadshedding	xh → st	loadshedding	UC11	Loanword from South African English.
traffic light	en → en-za	robot	UC12	A different term is used for the same concept in South Africa.
electricity	en → xh	igesi	UC13	The term ‘-gesi’, a loanword with morphemic modification from the English term ‘gas’, has since been extended to include the concept of ‘electricity’.
spoon	en → af	lepel	UC14	The meaning is the same, except that neither share the same hypernym.

the concept has long been lexicalised in English, the term is not widely known, unless of course, a person lives in an area where rolling blackouts occur. In the case of ‘loadshedding’ in South African English, the term has been borrowed by the other local languages, currently with no morphemic modification (UC10–11). For UC12, a traffic light is known as a robot in South African English.

In UC13, an example is given where an existing term is extended to include a new concept from another language, shown here for the direct equivalent ‘electricity’ to isiXhosa’s ‘igesi’. isiXhosa is an agglutinative language with a noun class system and concordial agreement. The term ‘ugesi’ is used for ‘power’ and ‘gas’, where the stem ‘-gesi’, originally the loanword ‘gas’ from English with morphemic substitution, has since extended to include ‘electricity’. Lastly, for UC14, this is an example where the term refers to the same object, but each language classifies it differently. In English, ‘spoon’ is a ‘utensil’, and in Afrikaans,

it is a ‘tool’.

We now revisit the translation categories from TRCAT, and systematically try to classify each use case. As shown in Table 2, only 8 of the 14 use cases can be classified by TRCAT’s categories. Using the semiotic triangle, the possible equivalences between a source and target language are given in Figure 1. For *directEquivalent* to be applicable, there has to be a lexical realisation for both the source and the target, and both lexical realisations have to be semantically equivalent. This is visualised in Diagram I in Figure 1. There are no categories in TRCAT to classify linguistic (Diagram II–IV) and referential gaps (Diagram VI), as well as partial equivalence (Diagram V).

3 Grammatical Alignments

As mentioned previously, isiXhosa is an agglutinative language with concordial agreement, so the prefix of a noun changes if it is singular or plural, as well as the prefixes or pre-prefixes

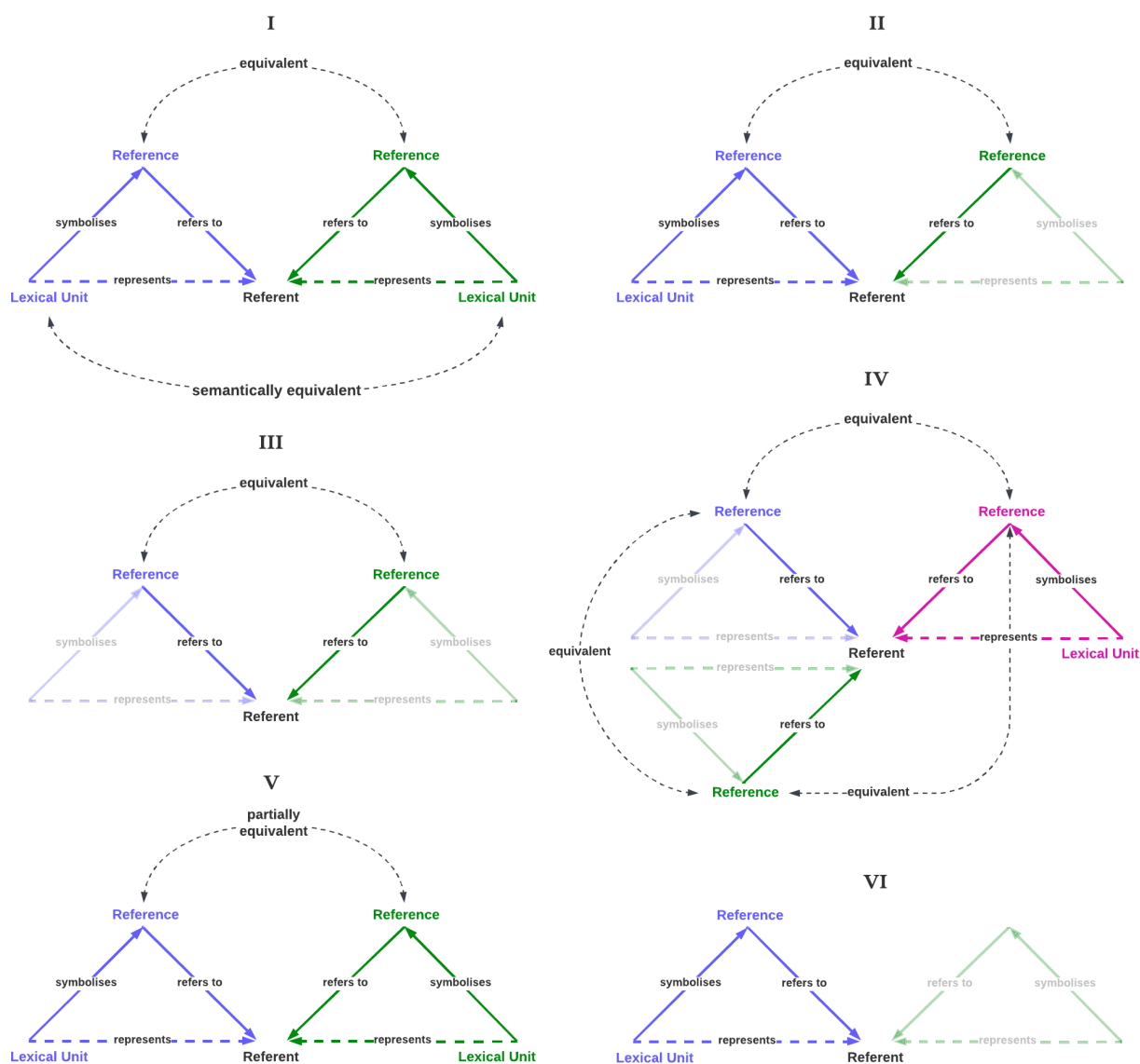


Figure 1: The semiotic triangle is used to show equivalence between two languages for a term. Language *A* is in purple and Language *B* is in green. Diagram I shows the source and target lexical units as semantically equivalent. Diagram II shows a lexical gap for the target (indicated as such by the opaque part of its semiotic triangle), however, the concept is known, so this is a linguistic gap. Diagram III shows a linguistic gap for both the source and the target. In Diagram IV, Diagram III is extended by introducing a pivot language (Language *C*, shown in pink). Diagram V shows partial equivalence between two references, with the result that there is not full semantic equivalence between the source and target lexical units. A referential gap for the target language is shown in Diagram VI.

Table 2: A comparison of each of the use cases for semantic equivalence against the available categories in TRCAT.

Use Case	Direct Equiv.	Lexical Equiv.	Cultural Equiv.
UC1			
UC2	✓		
UC3			
UC4			✓
UC5			✓
UC6			
UC7			
UC8			
UC9	✓		
UC10	✓		
UC11	✓		
UC12	✓		
UC13	✓		
UC14			

changing to show agreement with other parts of the sentence. As an example, the stem ‘-zimba’ means ‘body’. If the prefix ‘um’ is added, then ‘umzimba’ is singular, and if the prefix is ‘imi’, then it is plural. To denote modifications to the noun, such as the diminutive or feminine, then a suffix is also added. isiXhosa dictionaries are not consistent in their lemmatisation approach. For example, in The Greater Dictionary of isiXhosa, Volumes 1–3, nouns and verbs are listed by their stem (Tshabe, 2006; Mini, 2003; Pahl, 1989). In the Oxford Xhosa-English Dictionary (De Schryver and Reynolds, 2019), nouns are listed by their singular form and verbs are listed by their stem. In the Pharos English-Xhosa Dictionary, nouns and verbs are listed by their stem, although the form of the lemma for verbs does not make this obvious (Eng, 2014). When aligning two lexical senses from different languages, if an alignment is between, for example, word and stem or word and singular form, then this should be made clear. Use cases 15–16 pertain to this, given in Table 3.

Still staying with isiXhosa, using the ‘subtraction’ mathematical operator as an example, the stem is ‘-thabatha’. It is a verb by default, and to say ‘to subtract’ in a sentence, the prefix ‘u’ is used. To refer to subtraction as a noun, the prefix ‘uku’ is added to the stem.

UC17 relates to a part-of-speech change, which occurs here if the alignment is from word to stem. UC18–19 pertains to grammatical gender. In isiXhosa, ‘umfundisi’ is the word for ‘priest’ in English. However, this is a male priest, and to refer to a female priest, the suffix ‘kazi’ is added. Similarly in Spanish, the label for an object property ‘changed by’ can be ‘es modificada por’ or ‘es modificado por’. The change is attributed to grammatical gender, where the gender of the noun used for the class of the object property’s domain determines the gender of the past participle.

Lastly, we consider alignment between a mass and count noun. In English, the word ‘seed’ is both a mass noun and a count noun, however we focus just on the count noun. An example sentence is “Mark planted bean seeds.” In isiXhosa, the singular is ‘imbewu’, and this is used, even when the plural is referred to in English (UC20) (De Schryver and Reynolds, 2019).

4 The *vartrans* Module Extension

In OntoLex-Lemon, an ontology entity is used as the definiens for a lexical sense or a lexical entry. An ontology entity is in turn comprised of a semantic layer and a linguistic layer, visualised in Figure 2, where it can either be a class or an individual. As none of the use cases require lexical equivalency to be established between, say “Bill Gates”@en and “uBill Gates”@xh, both individuals of the class :PERSON, the focus is only on the use of an ontology class and its ontological commitment as a definiens.

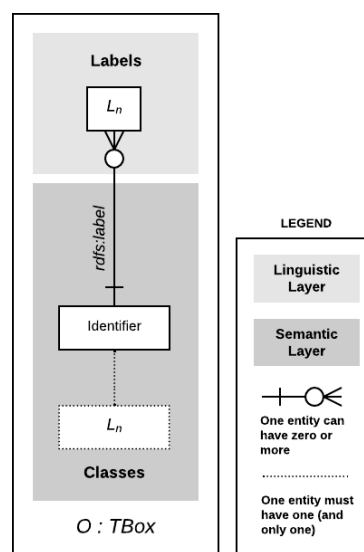


Figure 2: Distinguishing between the semantic and linguistic layers in the TBox of an OWL ontology.

Table 3: Language examples for grammatical inequivalences.

Source	Alignment	Target		
body	en → xh	umzimba	UC15	Singular noun in English aligned to singular form of noun stem in isiXhosa.
body	en → xh	-zimba	UC16	Singular noun in English aligned to noun stem in isiXhosa.
minus	en → xh	-thabatha	UC17	Noun in English aligned to verb stem in isiXhosa.
priest	en → xh	umfundisi / umfundisikazi	UC18	The isiXhosa singular form refers only to male priests. With the addition of the suffix ‘-kazi’, the singular form now refers to a female priest.
changed by	en → es	es modificado por / es modificada por	UC19	The gender changes for the Spanish past participle according to the gender of the subject.
seeds	en → xh	imbewu	UC20	The plural is used in English, however the singular is used in isiXhosa.

An ontology entity in OWL is comprised of two parts in the semantic layer: the axiom pattern, and the superclass of the axiom pattern, as well as the individuals of the axiom pattern, each shown in Figure 3. The axiom pattern comprises one or more classes and any axioms which serve as an ontological commitment. If we let O, O' be two ontologies with vocabularies V, V' , two *homogeneous* ontology entities, with one entity in V and the other in V' , can be aligned using an alignment axiom (Euzenat and Schvaiko, 2013). The axiom pattern, superclass(es), and individuals of the ontology entity in V and V' respectively can each be compared to determine the extent of equivalence in order to assign the appropriate category to the alignment. For the axiom pattern between O and O' , the axioms may differ, be it subclasses, a differing object property, or restrictions on the domain and range. For the superclasses, an axiom pattern in O may be placed differently in the class hierarchy to that of its counterpart in O' . For the individuals, only a subset of individuals may be applicable in O' , when compared to O .

Using the concept of ‘River’, example axiom patterns in Description Logic are given for the definiens of English’s River (1), Afrikaans’ Rivier (2), and French’s Fleuve (3) and Riviere (4–5):

- 1) $\exists \text{flowsInto.NaturalWatercourse} \sqcap \neg \exists \text{flowsInto.Self}$
- 2) $\exists \text{inVloei.NatuurlikeWaterloop} \sqcap \neg \exists \text{inVloei.Self}$
- 3) $\exists \text{couleDans.CoursDeauNaturel} \sqcap \exists \text{couleDans.Mer}$
- 4) $\exists \text{couleDans.CoursDeauNaturel} \sqcap \exists \text{couleDans.Self}$
- 5) $\text{Riviere} \sqsubseteq \neg \text{Fleuve}$

If the language pair is English and Afrikaans, then River and Rivier is semantically equivalent, with the same individuals as well. If the language pair

is English’s River to French’s Fleuve, the axiom pattern is not equivalent, and only a subset of the individuals apply to Fleuve.

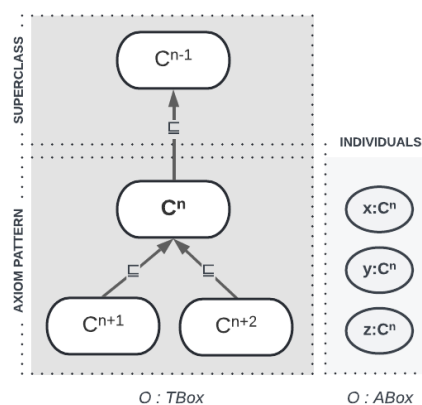


Figure 3: The ‘parts’ of an ontology entity in an OWL ontology. The axiom pattern and its superclasses are in the TBox. C^n is the starting point of the axiom pattern, and C^{n-1} is its immediate parent. The individuals are an assertion of class C^n .

To determine semantic equivalence, the following questions are identified.

- Q1: Is there a lexical realisation for the source and the target concepts?
- Q2: Are the individuals the same for both the source and the target?
- Q3: Is there some overlap of the individuals between the source and the target?
- Q4: Are the individuals of the target a subset of the source (or vice versa)?
- Q5: Is the axiom pattern the same for both the source and the target?
- Q6: Is the superclass(es) the same for both the source and the target?
- Q7: Is there a lexical realisation for either the

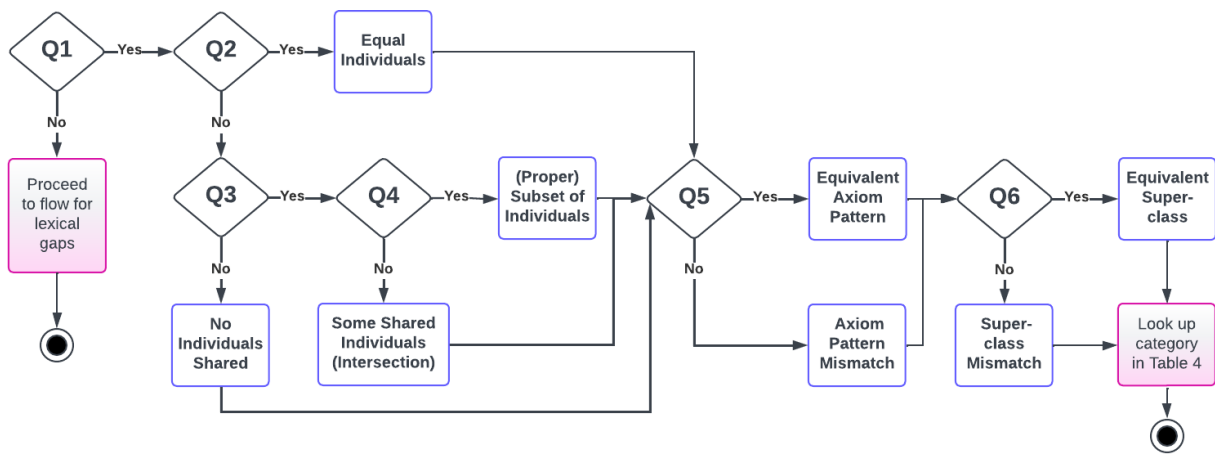


Figure 4: The decision tree diagram for Q1–6, for those alignments where there is a lexical realisation for both the source and the target. The diamond symbol denotes a decision that has to be made, where there is a ‘yes’ or ‘no’ answer. Each of the questions from Q1–6 are posed as decisions, and the starting point is Q1. The purple block indicates the feature that applies, based on the previous yes-no answers, and the small circles show the end of the flow for that question-answer selection.

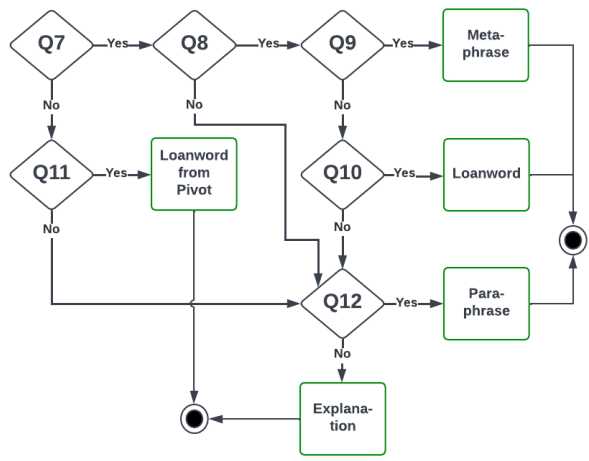


Figure 5: The decision tree diagram for Q7–12, for those alignments where there is no lexical realisation for the source and/or the target. Each green block is the proposed category to use for that question-answer selection.

source or the target?

Q8: For the source or target which has no lexical realisation, is the concept known in the language?

Q9: For the target which has no lexical realisation, can the source be directly translated as a metaphrase?

Q10: For the target which has no lexical realisation, can the source be used as a borrowing (and vice versa)?

Q11: Can a third language be introduced to serve as a borrowing between the source and the target?

Q12: If there is a referential gap or no borrowing can be used, can a paraphrase be used instead?

If both source and target is lexicalised, then

Q1–6 applies, with the question flow shown in Figure 4. If neither source nor target is lexicalised, then Q7–12 applies. The question flow is given in Figure 5. The label in each purple block in Figure 4 indicates the applicable feature. The features can then be looked up in Table 4 to determine the correct category to use. In Figure 5, each green block indicates the applicable category for the yes-no answer selection to Q7–12.

In Table 4, reference is made to an ‘interpretation’ where a correspondence between a source and target language can be equivalent in some interpretation. One of the internationalisation goals of OWL was to “potentially provide different views of ontologies that are appropriate for different cultures” (W3C OWL Working Group, 2004). If we consider ontology A which has a ‘universal’ viewpoint, then this ontology has, theoretically-speaking, all possible individuals for the interpretation \mathcal{I} . However, we can modify \mathcal{I} to obtain another interpretation \mathcal{I}_{xh} , which is specific to the speakers of one natural language, say isiXhosa, where individuals not applicable to isiXhosa speakers are removed, and the interpretation of class names and names of object properties are also changed so that they are specific to the isiXhosa viewpoint or perspective. The result is that the individuals of \mathcal{I}_{xh} is a subset of the individuals of \mathcal{I} (i.e., a proper subset in set theory).

The extended *vartrans* module (*extvartrans*) is located at: <https://w3id.org/EXTVARTRANS>. A new object property, #semanticCategory was

created as a subproperty of `#category` in *extvartrans*. The domain is a ‘lexico-semantic relation’ from *vartrans*, and its range has been set to one class: `#SemanticCorrespondence`. The subclasses of `#SemanticCorrespondence` are shown in Figure 6.

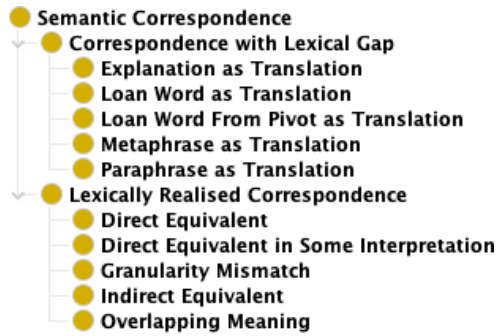


Figure 6: The new categories for semantic correspondences in the *extvartrans* module.

If the individuals are equal and the axiom pattern and superclass is equivalent between a source and a target, then this is a ‘Direct Equivalent’, and the category from the *vartrans* module is used. If the individuals are equal but either the axiom pattern or superclass (or both) are not equivalent between a source and a target, then this is an ‘Indirect Equivalent’. If the axiom pattern and superclass is equivalent, but the individuals are not equal but are instead a proper subset¹, then this is a ‘Direct Equivalent in Some Interpretation’ (but not all). For ‘Overlapping Meaning’, only some individuals are shared (instead of being a subset), and the axiom pattern and superclass can be a mismatch or equivalent between a source and a target. Finally, if there are no shared individuals between a source and a target, then despite the axiom pattern and/or superclass being equivalent, there is no correspondence.

4.1 Solving for the semantic use cases

Before each of the use cases are reviewed, we first identify the viewpoints by which a use case is considered (using the source and target language codes in the ‘Alignment’ column in Table 1 as a guide).

VP1: first language speakers of isiXhosa

VP2: language speakers of all English variations

VP3: speakers of South African English

VP4: speakers of English spoken in USA/UK

¹For ‘subset’ to apply, a subset of *A* can also be equivalent to *A*. For ‘proper subset’ to apply, a subset of *A* is not equivalent to *A*.

Table 4: A lookup table to determine the appropriate category to use, according to each of the ‘parts’ of an ontology entity: axiom pattern, superclass, and set of individuals, where the selection for each is an outcome of the yes-no answers selected in the decision tree diagram of Figure 4. These categories pertain to concepts where this is a lexical realisation for both the source and the target.

Axiom Pattern	Super-class	Individuals	Category
Equivalent	Equivalent	Equal	Direct Equivalent
Equivalent	Equivalent	Proper Subset	Direct Equivalent in Some Interpretation
Equivalent	Equivalent	Intersection	Overlapping Meaning
Equivalent	Equivalent	None	<i>No correspondence in Some Interpretation</i>
Equivalent	Mismatch	Equal	Indirect Equivalent
Equivalent	Mismatch	Proper Subset	Granularity Mismatch
Equivalent	Mismatch	Intersection	Overlapping Meaning
Equivalent	Mismatch	None	<i>No correspondence</i>
Mismatch	Equivalent	Equal	Indirect Equivalent
Mismatch	Equivalent	Proper Subset	Granularity Mismatch
Mismatch	Equivalent	Intersection	Overlapping Meaning
Mismatch	Equivalent	None	<i>No correspondence</i>
Mismatch	Mismatch	Equal	Indirect Equivalent
Mismatch	Mismatch	Proper Subset	Granularity Mismatch
Mismatch	Mismatch	Intersection	Overlapping Meaning
Mismatch	Mismatch	None	<i>No correspondence</i>

VP5: first language speakers of Sesotho

VP6: first language speakers of Afrikaans

VP7: language-independent

UC1 can be considered from three viewpoints: VP1, VP2, and VP7. For VP1, as there is a referential gap in English, a translation is required. If the flow diagram in Figure 5 is followed, then the proposed category is `#ExplanationAsTranslation`, where the axiom pattern and superclass(es) from the source are applied to the target as well. For VP2, one can argue that as it is a referential gap, the source concept can be excluded as it does not pertain to English culture. For VP7, the same as that for VP1 can be done, except with an additional axiom to indicate that this custom pertains only to

AmaXhosa culture.

For UC2, VP3 applies. As the concept is well-known in South African speakers' lexicon, and it is unchanged from that of isiXhosa except for an additional axiom to indicate that it pertains to AmaXhosa culture, the proposed category is `#IndirectEquivalent`. For UC3, VP4 applies. There are two possibilities for this use case: ignore the concept on the basis that it has no relevance within US/UK English culture; alternatively, model the alignment as a subclass of 'bride-price' (as 'lobola' is a more granular notion), with an axiom to indicate that it pertains to AmaXhosa culture. For the latter, the `#ParaphraseAsTranslation` is suitable. For UC4, the proposed category is `#GranularityMismatch`, on the basis that the axiom patterns for the source and target concepts are not the same, the superclass is the same, and the source individuals are a subset of the target individuals. For UC5, VP3 applies. For this use case, the proposed category is `#IndirectEquivalent`, on the basis that although the axiom pattern is a mismatch, the superclass is the same, and the individuals are the same (as neither concept is being considered from the perspective of the AmaXhosa). For UC6, two viewpoints can be considered: VP1 and VP2. If the alignment is considered from VP1, then this is a `#GranularityMismatch` as the target concept is more precise than the source, and it only applies to a subset of individuals. If VP2 is considered, then the `#IndirectEquivalent` category applies, and the term 'ikhazi' can be used interchangeably.

For UC7, the Sesotho paraphrase will differ from one dictionary to another. The proposal here is to treat it as a lexical gap and use the `#ParaphraseAsTranslation` category to indicate as such. For UC8, the category is `#GranularityMismatch`. If each target term is considered individually, then there is an axiom pattern mismatch with the source, as well as the individuals being a subset (where 'granddaughter' refers to female grandchildren, but 'umzukulwana' refers to both female and male grandchildren).

For UC9–11, the category is `#directEquivalent`. For UC9, the axiom pattern and superclass is the same for the source and the target, as well as the individuals. An additional synonym can be provided for the target of UC9: 'rolling blackout'. For UC10 and UC11,

VP1 and VP5 applies respectively. As there is no morphemic modification for both the targets, it is assumed that the meaning is unchanged from English.

UC12 is a `#directEquivalent`. If UC13 is considered from VP1 and VP2, then the proposed category is `#GranularityMismatch`. Lastly, for UC14, the `#IndirectEquivalent` category applies, as the superclass differs for each.

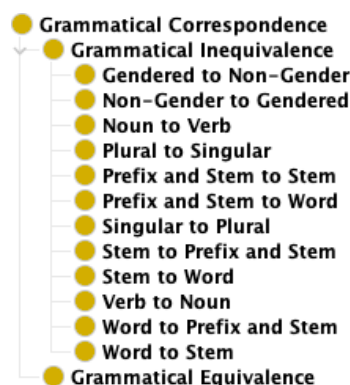


Figure 7: The new categories for grammatical correspondences in the *extvartrans* module.

4.2 Solving for the grammatical use cases

A new object property, `#grammarCategory` was created as another subproperty of `#category` in *vartrans*. Its range has been set to one class: `#GrammaticalCorrespondence`, and its subclasses are shown in Figure 7. The category `#GrammaticallyInequivalent` has subclasses, of which `#NounToPrefixAndStem` is the class selected for UC15, shown in Lines 6–7, in Listing 1. UC16 and UC20 are similarly classified, using the `#WordToStem`, and `#PluralToSingular` categories respectively. In each Turtle fragment that follows, the namespaces² are assumed defined.

```

1 :UC15 a vt:Translation ;
2   vt:source :sense_en_body ;
3   vt:target :sense_xh_umzimba ;
4   vt2:semanticCategory
5     trcat:directEquivalent ;
6   vt2:grammarCategory
7     vt2:WordToPrefixAndStem .
  
```

Listing 1: Turtle fragment for the translation of UC15.

²@prefix : <http://example.com#> .
 @prefix vt: <http://www.w3.org/ns/lemon/vartrans#> .
 @prefix vt2: <https://w3id.org/EXTVARTRANS#> .
 @prefix trcat: <http://purl.org/net/translation-categories#> .
 @prefix ontolx: <http://www.w3.org/ns/lemon/ontolx#> .
 @prefix lexinfo:
 <http://www.lexinfo.net/ontology/3.0/lexinfo#> .

For UC17, two categories are used, shown in Line 4 of Listing 2.

```
1 :UC17 a vt:Translation ;
2 ...
3 vt2:grammarCategory
4 vt2:WordToStem , vt2:NounToVerb .
```

Listing 2: Turtle fragment for the categories of UC17.

For UC18, it can be said that the male and female form is a granularity mismatch to English, therefore it is a semantic inequivalence. However, it has been opted to treat this as a grammatical inequivalence rather. As a gendered suffix is not applied consistently to the part of speech of type ‘noun’ in isiXhosa, a grammar rule has been created specific to a lexical item, and this is used, along with a grammar inequivalence category. To do this, a new class was created: #GrammarRule, for which there are two subclasses: #PartOfSpeechSpecificRule and #LexicalItemSpecificRule. The class #GenderModificationOfNoun is a subclass of #LexicalItemSpecificRule. The category #NonGenderToGendered was used, with both shown in Lines 6–8 in Listing 3 respectively.

```
1 :UC18 a vt:Translation ;
2 vt:source :sense_en_priest ;
3 vt:target :sense_xh_umfundisa ;
4 vt2:semanticCategory
5 trcat:directEquivalent ;
6 vt2:grammarCategory
7 vt2:WordToPrefixAndStem ,
8 vt2:NonGenderToGendered ;
9 vt2:targetRule
10 :rule_xh_fem_kazi .
11
12 :rule_xh_fem_kazi a
13 vt2:GenderModificationOfNoun ;
14 vt2:addSuffix :xh_kazi .
15
16 :xh_kazi a lexinfo:Suffix ;
17 ontolex:canonicalForm :xh_kazi_lemma ;
18 lexinfo:gender lexinfo:feminine .
19
20 :sense_xh_umfundisa a
21 ontolex:LexicalSense;
22 ontolex:reference dbp:Priest ;
23 lexinfo:gender lexinfo:masculine .
```

Listing 3: Turtle fragment for UC18.

A new object property was created: #targetRule, and this was added to the translation, shown in Lines 9–10 of Listing 3. An instance of the #GenderModificationOfNoun rule is given in Lines 12–14. A new object property was created for this rule #addSuffix, where the range is a lexical entry of type ‘Suffix’. The creation of the suffix is shown in Lines 16–18, where LexInfo is used.

UC19 also relates to gender, however it differs in that the translation pertains to an object property, which means the surface realisation of the label will change according to the noun of the class used as the domain. In this instance, the rule is not specific to a lexical item (as was the case of UC18), instead, it is a rule specific to a part of speech. A new rule was created as a subclass of #PartOfSpeechSpecificRule: #GenderAgreement, and this rule is set as the #targetRule for UC19.

```
1 :UC19 a vt:Translation ;
2 vt:source :lex_en_changed_by ;
3 vt2:targetMasculine
4 :lex_es_es_modificado_por ;
5 vt2:targetFeminine
6 :lex_es_es_modificada_por ;
7 vt2:semanticCategory
8 trcat:directEquivalent ;
9 vt2:grammarCategory
10 vt2:NonGenderToGendered ;
11 vt2:targetRule
12 :rule_es_rule_gender .
13
14 :rule_es_rule_gender a
15 vt2:GenderAgreement .
```

Listing 4: Turtle fragment for UC19.

5 Related Works

Ontologies pertaining to linguistics were reviewed in the Linked Open Vocabularies (LOV) repository³, of which a selection are listed here. The General Ontology for Linguistic Description has a #translation object property with #literalTranslation as a subproperty (Gol, 2010). It has a class #LexicalizedConcept, but none for an unlexicalised concept. LexInfo also provides for a #translation object property (from *vartrans*), as well as lexical and sense relations (Cimiano et al., 2011), however these are more suited to same-language relations. The property #geographicalVariant can be used for dialects, and the properties #exact, #approximate, and #quasiEquivalent can be used for lexicalised translations, although when to use the latter two is not made clear. The Lingvoj Ontology provides for the representation of language resources, and it has a #Translation class as an event, although this is intended at resource-level, not at term-level (B. Vatant, n.d.). The Lexvo.org Ontology is intended for the description of natural languages, terms, and meanings (de Melo, 2015). It provides

³<https://lov.linkeddata.es/dataset/lov>

for the thesaurus hierarchy of `#broader` and `#narrower`, as well as `#somewhatSameAs` and `#nearlySameAs`, where the latter two are intended as an alternative to `owl:sameAs`, all as object properties. To the best of our knowledge, there is no ontology or registry which provides the same extent of categorisation as that presented in *extvartrans*, particularly for lexical gaps. Of the ontologies which do provide some descriptors, this is only as object properties, and not as classes.

6 Discussion & Future Work

The reference or denotation of a lexical entry or sense is, in OntoLex-Lemon, given by an ontology entity. This has come in for criticism, with Hirst (2014) being one such example, in that an ontology entity is not granular enough to accurately represent the meaning distinctions of a concept across several natural languages. Direct equivalence between terms of different languages is not always possible, and even more so for concepts which are culture-bound (Culler, 1976; Kramersch, 1998; Zgusta, 1971; Hirst, 2014). By specifying a `#Translation` from the *vartrans* module, this can aid in bridging a gap between a language pair. The *vartrans* module has defined these mappings between a language pair as a translation. If the ontology is multilingual but based on a primary language (where this is typically English), then all other language terms are indeed a translation. If UC1 had to be considered only from VP2, then it is unlikely that this concept would have been included in an ontology where English is the primary language. In a multilingual ontology, each natural language usually takes on the axioms of the primary language, to the exclusion of each additional language.

Of the three translation categories, there is soft-reuse of `#directEquivalent` and `#culturalEquivalent` only in *extvartrans*. The category `#lexicalEquivalent` was not included in *extvartrans* as its meaning (literal translation) is not consistent with the same term used in Lexicography (that of absolute equivalence (Zgusta, 1978)). The category `#MetaphraseAsTranslation` was created as an alternative.

The *extvartrans* module aims to get closer to realising one of the internationalisation goals of the OWL specification, and that is to develop different views of the same ontology, where each view is

specific to a culture. Considered from this perspective, then the mapping between a language pair is not necessarily always a translation but it can also refer to a transformation. It is for this reason that the word ‘Correspondence’ was used in the *extvartrans* module, instead of the word ‘Translation’. The exception to this is a mapping between a language pair where the target is a lexical gap. This mapping is indeed a translation of the lexicalised source (or pivot language source).

The first step towards ontology transformation has been presented with the grammatical use cases. Each Turtle fragment given for these use cases is intended to serve as an input to an algorithm. The use cases presented here were by no means exhaustive and it is expected that more subclasses will be added to `#GrammaticallyInequivalent` in the future. The ontology transformation process for language-specific views is current work, where the focus is primarily on semantic inequivalences. In this paper, the linguistic layer of the ontology (as shown in Figure 3) has been the focus. However, for future work, the focus will be on the semantic layer, with the addition of new axioms to an existing ontology, and the refactoring of classes and object properties so that the ontology is specific to a viewpoint. The ontology to represent viewpoints, the Model of Multiple Viewpoints (MULTI), is already available at <https://w3id.org/MULTI> (Gillis-Webber, 2023). The next step is to soft-reuse selected classes and object properties from *extvartrans* in MULTI, where these classes and properties will then be aligned to DOLCE+DnS Ultralite, an upper ontology suitable for modelling contexts (Dol, 2010).

7 Conclusion

As has been shown with the use cases pertaining to semantic alignment, there is slight variation depending on the viewpoint being considered. When considering a translation, the perspective should ideally be considered as well. In this paper, an extended version of the *vartrans* module for OntoLex-Lemon has been presented. More categories were provided from that of TRCAT, with new categories for both semantic and grammatical inequivalences, including lexical gaps. Additional classes and object properties were included in *extvartrans* for grammar rules and language features. For grammatical inequivalences, the code fragments provided were the first step to ontology trans-

formation, where an ontology is transformed to a language-specific view, in line with the internationalisation goal of the OWL specification.

Acknowledgements

This work was financially supported by Hasso Plattner Institute for Digital Engineering through the HPI Research School at UCT.

References

2010. General Ontology for Linguistic Description (GOLD). <http://linguistics-ontology.org/>. Online; accessed: 2023, March 19.
2010. Ontology:DOLCE+DnS Ultralite. http://ontologydesignpatterns.org/wiki/Ontology:DOLCE+DnS_Ultralite. Online; accessed: 2023, January 24.
2014. *English-Xhosa / Xhosa-English Dictionary*, 14 edition. Pharos Dictionaries.
2017. *Gender Terminology: Sesotho 2017/18*. Commission for Gender Equality.
- n.d. Translation Category Reference RDF Schema. <http://purl.org/net/translation-categories>. Online; accessed: 2023, May 27.
- B. Vatant. n.d. The Lingvoj Ontology (lingvo). <https://lov.linkeddata.es/dataset/lov/vocabs/lingvo>. Online; accessed: 2023, March 19.
- M. Baker. 2018. *In Other Words: A Coursebook on Translation*. Routledge, Oxon, UK.
- P. Cimiano, P. Buitelaar, J. McCrae, and M. Sintek. 2011. LexInfo: A declarative model for the lexicon-ontology interface. *Journal of Web Semantics*, 9(1):29–51.
- P. Cimiano, J.P. McCrae, and P. Buitelaar. 2016. *Lexicon Model for Ontologies: Community Report*. Final Community Group Report, 10 May 2016, Ontology-Lexicon Community Group under the W3C Community Final Specification Agreement (FSA).
- J.D. Culler. 1976. *Saussure*. Fontana/Collins.
- M. Dagut. 1981. Semantic “Voids” as a Problem in the Translation Process. *Poetics Today*, 2(4):61–71.
- G. de Melo. 2015. *Lexvo.org: Language-related Information for the Linguistic Linked Data Cloud*. *Semantic Web*, 6(4):393–400.
- G. De Schryver and M. Reynolds. 2019. *Oxford English-isiXhosa School Dictionary*, 7 edition. Oxford University Press Southern Africa.
- J. Euzenat and P. Schvaiko. 2013. *Ontology Matching: Second Edition*. Springer-Verlag Berlin Heidelberg.
- R. Gauton. 2008. Bilingual Dictionaries, the Lexicographer and the Translator. *Lexikos*, 18:106–118.
- F. Gillis-Webber. 2023. Towards an Ontology of Viewpoints. In *Proceedings of the 13th International Conference on Formal Ontology in Information Systems (FOIS 2023)*, 17–20 July, Sherbrooke, Québec, Canada.
- R.H. Gouws and D.J. Prinsloo. 2005. *Principles and Practice of South African Lexicography*. AFRICAN SUN MeDIA, Stellenbosch, South Africa.
- J. Gracia, E. Montiel-Ponsoda, D. Vila-Suero, and G. Aguado de Cea. 2014. Enabling Language Resources to Expose Translations as Linked Data on the Web. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, pages 409–413, Reykjavik, Iceland. European Language Resources Association (ELRA).
- G. Hirst. 2014. Overcoming linguistic barriers to the multilingual semantic web. In P. Buitelaar and P. Cimiano, editors, *Towards the Multilingual Semantic Web: Principles, Methods and Applications*, pages 3–14. Springer Berlin Heidelberg.
- C. Kramsch. 1998. *Language and Culture*. Oxford Introductions to Language Study. Oxford University Press.
- K. Mansoor. 2018. Translation Across the Difficulties of Equivalence Concept. *Scientific Bulletin of the Politehnica University of Timisoara. Transactions on Modern Languages*, 17(1):55–66.
- B.M. Mini, editor. 2003. *The Greater Dictionary of isiXhosa: K to P*, volume 2. IsiXhosa National Lexicography Unit, University of Fort Hare.
- H.W. Pahl, editor. 1989. *The Greater Dictionary of isiXhosa: Q to Z*, volume 3. University of Fort Hare.
- S.L. Tshabe, editor. 2006. *The Greater Dictionary of isiXhosa: A to J*, volume 1. IsiXhosa National Lexicography Unit, University of Fort Hare.
- W3C OWL Working Group. 2004. *OWL Web Ontology Language Use Cases and Requirements: W3C Recommendation 10 February 2004*. W3C Recommendation, World Wide Web Consortium. Online; accessed: 2023, April 28.
- L. Zgusta. 1971. *Manual of Lexicography*. Academia.
- L. Zgusta. 1978. *Equivalents and Explanations in Bilingual Dictionaries*. In Mohammad Ali Jazayeri, Edgar C. Polomé, and Werner Winter, editors, *Linguistic and Literary Studies: Vol 4, Linguistics and Literature / Sociolinguistics and Applied Linguistics*, pages 385–392. De Gruyter Mouton, Berlin, New York.

Leveraging DBnary Data to Enrich Information of Multiword Expressions in Wiktionary

Gilles Sérasset

Université Grenoble Alpes CNRS, Grenoble
INP*, LIG 38000 Grenoble, France
gilles.serasset@imag.fr

Thierry Declerck

DFKI GmbH, Multilingual Technologies
Saarland Informatics Campus D3 2
D-66123 Saarbrücken, Germany
declerck@dfki.de

Lenka Bajčetić

Innovation Center of the School of Electrical
Engineering in Belgrade Bulevar kralja
Aleksandra 73 11000 Belgrade, Serbia
lenka.bajcetic@ic.etf.ac.bg.rs

Abstract

We describe first an approach consisting of computing pronunciation information for multiword expressions (MWEs) included in the English edition of Wiktionary. During this work, we learnt about the DBnary resource, which represents information extracted from 23 language editions of Wiktionary in a Linked Open Data (LOD) compliant way. This lead to updates of the DBnary programs, to support the extraction of the desired pronunciation information for MWEs and which we document in this paper. The use by DBnary of LOD compliant models and vocabularies, more specifically of the *OntoLex-Lemon* model, opens the possibility for additional lexicographic enrichment of the MWEs, like adding morphosyntactic and semantic information to their components. DBnary is thus now more than “just” an extractor and mapper of Wiktionary data in a LOD representation, but is also contributing to the lexicographic enrichment of Wiktionary pages dealing with MWEs. In the longer term, our work will allow for more data on English MWEs to be made available in the Linguistic Linked Data cloud.

1 Introduction

Recent work (Bajčetić et al., 2023) dealing with the computation of pronunciation information for multiword expressions (MWEs) in the English edition of Wiktionary was using a combination of the Wikimedia API¹ to find wiki pages describing MWEs and of an XML parser to analyse and extract information from the corresponding wiki

text.² This approach proved to be tedious and time-consuming. We decided therefore to use the DBnary resource, which is already providing for a structured representation of Wiktionary content, to get access to the Wiktionary data necessary for the computation of pronunciation information for MWEs and for exploring other tasks, like specifying the part-of-speech of components of MWEs or for associating semantic information to those components.

DBnary is a lexical resource extracted from 23 language editions of Wiktionary. Lexical data is represented using the Linked Open Data (LOD) principles³ and as such it is using RDF⁴ as its representation model. It is freely available and may be either downloaded or directly queried on the internet. DBnary uses the *OntoLex-Lemon* standard vocabulary (Cimiano et al., 2016),⁵ displayed in Figure 1 to represent the lexical entries structures, along with *lexvo* (de Melo, 2015) to uniquely identify languages, *lexinfo* (Cimiano et al., 2011)⁶ and *Olia* (Chiarcos and Sukhareva, 2015)⁷ for linguis-

²One can also apply an XML parser to the full Wiktionary dump in XML format, available at <https://dumps.wikimedia.org/enwiktionary/20230320/>.

³See <https://www.w3.org/wiki/LinkedData> for more information on those principles.

⁴The Resource Description Framework (RDF) model is a graph based model for the representation of data and meta-data, using URIs to represent resources (nodes) and properties (edges). See <https://www.w3.org/TR/rdf11-primer/> for more details.

⁵See also the specification document at <https://www.w3.org/2016/05/ontolex/>.

⁶The latest version of the *lexinfo* ontology can be downloaded at <https://lexinfo.net/>.

⁷The “Ontologies of Linguistic Annotation (OLiA)” is available at <https://acoli-repo.github.io/olia/>.

¹<https://en.wiktionary.org/w/api.php>.

tic data categories.

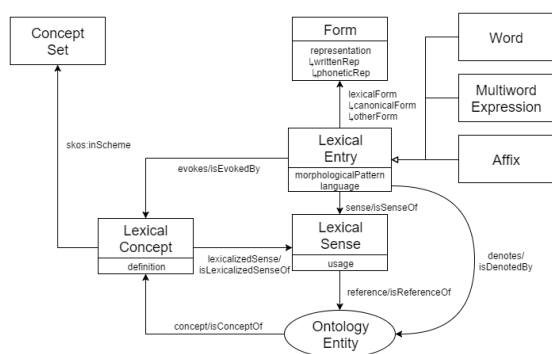


Figure 1: The core module OntoLex-Lemon. Taken from <https://www.w3.org/2016/05/ontolex/#core>.

While trying to reproduce (Bajčetić et al., 2023) work, we noticed that DBnary was lacking some information. First, Wiktionary MWEs were not marked explicitly. Second, derivation relations between single word lexical entries and MWEs, in which they occur, were not extracted, while this information is crucial for the disambiguation of components of MWEs that are heteronyms (see Section 2 for a detailed discussion). The DBnary maintainer⁸ tuned the extraction program to fix these identified lacks.

This paper summarises first the work presented in (Bajčetić et al., 2023) (section 2), providing details on the different means we used to access Wiktionary data (section 3), initially through API queries and XML parsing and finally using the latest version of DBnary for which we detail how we query it for accessing the necessary Wiktionary data. Section 4 presents and evaluates the computing of pronunciation information to be associated with Wiktionary MWEs. Then, in section 5, we discuss the promising use of the decomposition module of OntoLex-Lemon for supporting an enriched semantic representation of the components of MWEs.

2 Adding pronunciation information to multiword expressions in Wiktionary

In this section, we summarize the approach described in (Bajčetić et al., 2023), motivating also the decision to use DBnary as the primary source

⁸The DBnary extraction programs are open source and available at: <https://gitlab.com/gilles.serasset/dbnary/> where issues can be added to ask for correction or enhancement of the extractors. It is also possible to fix the extractors and create a Merge Request.

for the task of adding pronunciation information to Wiktionary MWEs, a move that lead to the fine-tuning of the extraction engine that is generating DBnary.

2.1 Wiktionary

Wiktionary⁹ is a freely available web-based multilingual dictionary. Like other Wikimedia¹⁰ supported initiatives, it is a collaborative project that is also integrating information from expert-based dictionary resources, when their licensing conditions allow it.

Wiktionary includes a thesaurus, a rhyme guide, phrase books, language statistics and extensive appendices. Wiktionary’s information also (partly) includes etymologies, pronunciations, sample quotations, synonyms, antonyms and translations.¹¹ Wiktionary has also developed categorization practices, which classify an entry along the lines of linguistics (for example “developed terms by language”) but also topical information (for example “en:Percoid fish”).¹²

2.2 Multiword expressions in Wiktionary

Wiktionary introduces the category “English multiword terms” (MWT), which is defined as “lemmas that are an idiomatic combination of multiple words”¹³, while Wiktionary has the page “multiword expression”, categorized as a MWT and defined as “lexeme-like unit made up of a sequence of two or more words that has properties that are not predictable from the properties of the individual words or their normal mode of combination”.¹⁴ We see these two definitions are interchangeable, since they both focus on the aspect of non-compositionality of a lexeme built from multiple words. For consistency with common usage in NLP publications, we use in this paper the term

⁹<https://en.wiktionary.org/>

¹⁰<https://www.wikimedia.org/>

¹¹See <https://en.wikipedia.org/wiki/Wiktionary> for more details.

¹²The entry “sea bass”, for example, is categorized, among others, both as an instance of “English multiword terms” and of “en:Percoid fish”. The categorization system is described at <https://en.wiktionary.org/wiki/Wiktionary:Categoryization>

¹³https://en.wiktionary.org/wiki/Category:English_multiword_terms. This category is an instance of the umbrella category “Multiword terms by language”, see https://en.wiktionary.org/wiki/Category:Multiword_terms_by_language.

¹⁴https://en.wiktionary.org/wiki/multi-word_expression.

Multiword Expression (MWE), but stress that they are categorized as MWTs in Wiktionary.

According to Wiktionary website, the current version of the English edition of Wiktionary is listing 157,753 pages containing an English MWE¹⁵, and 75,389 pages containing an English term equipped with IPA pronunciation¹⁶. This is quite a small number in comparison to the whole English Wiktionary, which has over 8,597,416 pages (with 7,365,114 items marked as “content pages”, totalizing 226,078,477 words (<https://en.wiktionary.org/wiki/Special:Statistics>, [accessed 25.03.2023]). It is important to keep in mind that the English Wiktionary contains a lot of terms which are not English. We can see the exact number of English lemmas if we look at the Wiktionary category “English lemmas”.¹⁷ The actual number of 711,294 pages containing an English lemma means that a little over 10% of English lemmas have pronunciation, while approximately 22% of all English lemmas belong in the MWT category. So there is clearly a gap that needs to be filled when it comes to pronunciation information in Wiktionary. While introducing pronunciation for the remaining 90% of lemmas seems like it has to be a manual task (or semi-automatic, using another resource) - we have investigated ways to produce the missing pronunciation for numerous MWEs.

2.3 Overview of the approach for adding pronunciation information to MWEs

Bajčetić et al. (2023) describes the approach aiming at enriching English MWEs included in Wiktionary by pronunciation information extracted from their sub-parts. This endeavour itself is a continuation of work consisting of extracting pronunciation information from Wiktionary in order to enrich the Open English WordNet (McCrae et al., 2020),¹⁸ where pronunciation information has been added only for single word entries, as described in (Declerck and Bajčetić, 2021).

An issue to deal with in this approach is the treatment of heteronyms that are a component of a MWE¹⁹. In order to select the correct pronun-

ciation, an additional analysis of the Wiktionary data is needed, disambiguating between the different senses of the heteronym. This issue is multiplied by the number of MWEs containing such a heteronym. An example of such a case is given by the Wiktionary page “acoustic bass”, for which our algorithm has to specify that the pronunciation /beɪs/ (and not /bæs/) has to be selected and combined with /ə'kuː.stɪk/.²⁰

Since we need to semantically disambiguate one or more components of a MWE for generating its pronunciation, our work can lead to the addition of morphosyntactic and semantic information of those components and thus enrich the overall representation of the MWEs entries, a task we started to work on, and for which we consulted DBnary, and this step was leading to the development of a new version of the DBnary extractor, in order to explicitly mark MWEs and Wiktionary “derived terms”, which establish semantic links between single word entries and MWEs in which they occur.

In order to implement our approach, we need thus to extract from Wiktionary:

- all existing pronunciation of English terms
- a list of all MWEs that are available
- all derivation relations between single English terms and their derived terms, when those are MWEs.

3 Accessing Wiktionary data

When it comes to extracting information from Wiktionary, we can usually find three approaches in the literature. Mainly, parsing the dumps, accessing Wiktionary APIs or querying DBnary.

3.1 Parsing Wiktionary dumps

The first approach requires downloading the English Wiktionary dump and parsing it. The dump is an XML document containing the MediaWiki

heteronym is one of two or more words that have the same spelling but different meanings and pronunciation, for example ‘tear’ meaning ‘rip’ and ‘tear’ meaning ‘liquid from the eye’, <https://www.oxfordlearnersdictionaries.com/definition/english/heteronym>

²⁰The corresponding entry “bass” (the one marked with “Etymology 1”) in the Wiktionary page <https://en.wiktionary.org/wiki/bass#English> lists 65 derived terms (most of them MWEs, and with only nine terms being equipped with pronunciation information), for which we can assume that the pronunciation /bæs/ has to be selected for the component “bass”.

¹⁵https://en.wiktionary.org/wiki/Category:English_multiword_terms, [accessed on the 25.03.2023]

¹⁶https://en.wiktionary.org/wiki/Category:English_terms_with_IPA_pronunciation

¹⁷https://en.wiktionary.org/wiki/Category:English_lemmas

¹⁸See also <https://en-word.net/>

¹⁹The online Oxford Dictionary gives this definition: “A

source (see Figure 2) of all entries and templates or modules defined in the English edition. Indeed, each entry is a kind of program whose execution results in the HTML page that is visible in your browser (see Figure 3).

```
====Pronunciation====
* {{enPR|bās}}, {{IPA|en|/beɪs/}}
* {{audio|en|en-us-bass-low.ogg|Audio (US)}}
* {{rhymes|en|eɪs|s=1}}
* {{homophones|en|base}}

====Adjective====
{{en-adj|basser}}

# Of sound, a voice or an instrument, [[low]] in
#: ''The giant spoke in a deep, ''bass'', rumbl
```

Figure 2: Extract of the MediaWiki source of the page *bass* in the Wiktionary dump. Elements between double curly braces (e.g. `{{en-adj|basser}}`) are “Templates”, a kind of parameterised procedure (here, a call to template `en-adj` with argument `basser`).

The screenshot shows the Wiktionary page for 'bass'. It includes a 'Pronunciation' section with an edit link, a list of pronunciation options (enPR: bās, IPA(key): /beɪs/, Audio (US) with a play button and 0:01 duration, Rhymes: -eɪs, Homophone: base), and an 'Adjective' section with an edit link. Below this, the word 'bass' is shown with its comparative form 'basser' and superlative form 'bassest'. A numbered list starts with '1. Of sound, a voice or an instrument, low in pitch or frequency.' followed by an example sentence: 'The giant spoke in a deep, bass, rumbling voice that shoos'.

Figure 3: Extract of the page *bass*, as viewed in a browser, after expansion of the MediaWiki source into a valid HTML file.

This approach is usually used to extract simple information from Wiktionary, like a list of all English terms or their pronunciation, as this information is represented rather systematically using the template call `{{IPA|en|...}}`. A simple regular expression will extract this information easily and reliably.

However, this approach has several shortcomings. First, depending on the Wiktionary edition you extract from, there may be many ways to encode lexical data, as the entry structure has evolved and older entries are using older encoding conventions. In many cases, convenient templates are used to allow for a condensed representation of data, but defective entries will use a specific

encoding not captured by these templates. Also, the structure and encoding of Wiktionary entries evolves continually as the community updates the templates to ease entry additions. Due to this, many experiments are not reproducible as time goes by as the extraction programs become obsolete due to sometimes major changes in the Wiktionary structure.

Second, much of the information that is present in the Wiktionary HTML page is not visible in the MediaWiki source. For instance, in the excerpt of the Wiktionary *bass* page, one can find **bass** (*comparative* **basser**, *superlative* **bassest**) but this snippet is the result of the template call `{{en-adj|basser}}` where the string *bassest* does not appear. In the English Wiktionary edition, the `en-adj` template calls a Lua program²¹ which computes this word form. Hence, as noted in (Ylonen, 2022), a full implementation of the Lua language (and the Scribunto²² standard library) is required if one wants to extract most Wiktionary data²³.

This is the first approach we have attempted, and it seemed to be the most straightforward, but turned out to be inefficient: after downloading the latest Wiktionary XML dump, we wanted to extract all entries that belong to the Wiktionary category *English multiword terms*. But the category information only appears in five (badly encoded) English entries’ MediaWiki source. In all other MWE entries, the categorisation is a side effect of the call of some templates appearing in the MediaWiki source. Moreover, the https://en.wiktionary.org/wiki/Category:English_multiword_terms page itself does not appear in the dump, as it is a special page that is computed on demand by the Wiktionary server.

Hence, in a second attempt, we tried to use the Wiktionary API to query for these categories.

3.2 Using Wiktionary API

The Wiktionary API is a RESTful interface that allows programmers to access the data contained in

²¹Such programs are called *modules* in MediaWiki. They are special pages that contain program(s) in *Lua*, a Turing complete programming language.

²²Scribunto is the MediaWiki extension which allows for the use of any Lua program in a Wikimedia page.

²³This was less of a problem when the language editions were not heavily depending on such modules and many of the experiments cited before will not be reproducible without this nowadays.

the Wiktionary dictionary through standard HTTP requests. It may be used to query for definitions, translations, links or categories of a specific Wiktionary page. In our cases, we planned to use it to query each page for its categories.

This would be simple if the size of Wiktionary dump was not so massive: more than 8.5 million entries need to be checked, which means 8.5 million requests sent to Wiktionary API. This is quite slow and if not done correctly will lead to being blacklisted from the Wiktionary website.

Using this approach, described in (Bajčetić et al., 2023) we have extracted over 98% of MWEs from Wiktionary and compiled a list of 153,525 MWEs without IPA, and a gold standard of 4,979 MWEs with IPA - we can see that only about 3% of MWEs have pronunciation information in Wiktionary.

However, this approach was very time-consuming and can only be applied on a specific dump. Hence, as the Wiktionary data is always growing, new MWEs introduced in Wiktionary will not benefit from this work. This is the reason why we tried to reproduce our experiment using the DBnary dataset.

3.3 Querying DBnary

DBnary (Sérasset and Tchechmedjiev, 2014; Sérasset, 2015)²⁴ is a lexical resource extracted from 23 language editions of Wiktionary. This dataset is structured in RDF using the *OntoLex-Lemon* model (McCrae et al., 2017), which was developed and which is further extended in the context of the W3C Community Group “Ontology Lexica”.²⁵ The DBnary extraction program is open-source²⁶ and one can create issues when errors are spotted or additional information is required.

With DBnary, the whole set of lexical information extracted from the 23 language editions of Wiktionary may be seen as a huge graph that can be downloaded and queried online using the SPARQL language²⁷ or accessed interactively

²⁴See <http://kaiko.getalp.org/about-dbnary/> for the current state of development of DBnary.

²⁵See <https://www.w3.org/community/ontolex/> for more details.

²⁶<https://gitlab.com/gilles.serasset/dbnary>

²⁷SPARQL is the “standard query language and protocol for Linked Open Data on the web or for RDF triplestores”, quoted from <https://www.ontotext.com/knowledgehub/fundamentals/what-is-sparql/>. The SPARQL endpoint of DBnary can be accessed at <http://kaiko.getalp.org/>

through a faceted browser.²⁸ Moreover, any node (Page, Lexical Entry, Lexical Sense, Translation, Word Form, etc.) in this huge graph is designed by a unique URI²⁹ that may be dereferenced (i.e. accessed through the HTTP protocol) so that any person or process can obtain its related information easily which is compliant to the guidelines of the Linguistic Linked Open Data (LLOD) framework (Declerck et al., 2020).³⁰ Using DBnary is a matter of crafting SPARQL queries and evaluating them using a public endpoint.

By our first use of DBnary, we saw that, while pronunciation information is available, some of the information we required was missing from the English dataset:

- the entries were only typed as `ontolex:LexicalEntry` and no finer grain typing (as `ontolex:Word`, `ontolex:MultiWordExpression` or `ontolex:Affix`) was available,
- derivation information between terms was not extracted.

These missing elements were added and are now available in versions starting from February 2023. The extraction program now correctly *types* English Wiktionary entries either as `ontolex:Word` or as `ontolex:MultiWordExpression`. Moreover, derivation relations are now extracted and available in the graph using `dbnary:derivesFrom` transitive property.

Figure 4 shows an example of the organisation of two heteronym lexical entries described by the same page, along with their canonical forms (with written and phonetic representation).

Figure 4 also shows how the derivation relation is modelled in DBnary, using the transitive `dbnary:derivesFrom` property. It must be noted that in Wiktionary original data, the derivation links point to Wiktionary pages but not to Wiktionary entries, hence, the DBnary modelling reflects this as it is usually difficult to automatically

sparql

²⁸The browser can be accessed at <http://kaiko.getalp.org/fct/>

²⁹E.g. the URI <http://kaiko.getalp.org/dbnary/eng/bass> represents the Wiktionary *Page* *bass* that further *describes* different *Lexical Entries* (In English, one adjectival, one verbal and three nominal and eleven others in nine other languages.)

³⁰See also <http://www.linguistic-lod.org/>.

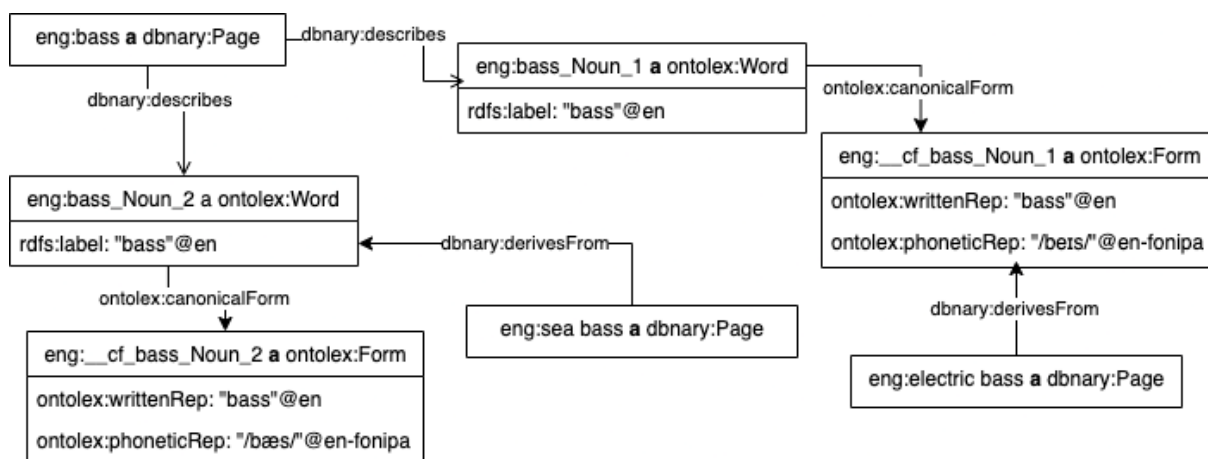


Figure 4: A very small extract of the DBnary graph showing DBnary page *bass* and two of the lexical entries it describes (*bass_Noun_1* [sound, music, instrument] and *bass_Noun_2* [perch, fish]) and their respective canonical forms. The pages *sea bass* and *electric bass* are also represented with their derivation relations.

choose which lexical entry(ies) is (are) the valid target of the derivation relation. But, applying the property in the inverse direction (could be named `dbnary:derivesTo`), the subject/source of the relation is a lexical entry within a Wiktionary page, pointing to a MWE page. As MWE pages consist mainly of only one lexical entry, we can precisely establish a “subterm” relation between a single lexical entry and the MWEs it occurs in, combining if needed both “directions” of use of the property. This point is very important, as it allows projecting all the lexical information of the single lexical entry to the component it builds within a MWE, as this is briefly presented in Section 5.

In the DBnary representation of Wiktionary we find lexical entries (including words, MWEs or affixes), their pronunciation (if available in Wiktionary), their sense(s) (definitions in Wiktionary), example sentences and DBnary glosses, which are offering a kind of “topic” for the (disambiguated) entries, but those glosses are not originated in the category system of Wiktionary. The glosses are taken from available information used to denote the lexical sense of the source of the translation of an entry from English to other languages.

DBnary does not extract Wiktionary categories, as most of these are implicit in the MediaWiki code and are the result of the full processing of the MediaWiki source. This processing is too heavy to compute for the 8.5M+ pages found in the English Wiktionary edition. Without this full processing, the extraction process takes almost 14 hours on a recent CPU server, more than 70% of which goes in the execution of Lua Modules. As this extrac-

tion has to be re-computed twice a month as new dumps are released, taking several days for such an extraction is not worth it.

In the paper, we reproduce the approach described in (Bajčetić et al., 2023), using only DBnary data. The added value of using DBnary comes from the fact that the data is updated twice a month and extractors are usually maintained to reflect changes in Wiktionary representation of the entries. Hence, reproducing this work will be possible without a high data preparation cost, and future MWEs described in future versions of Wiktionary will benefit of it.

4 Enriching pronunciation for MWEs using DBnary

4.1 Assessing the size of the problem

Before proceeding to the experiment using DBnary data³¹, we first probe the dataset to see if it faithfully reflects the Wiktionary data. First, we would like to know how many entries have a canonical form with pronunciation, using the SPARQL query displayed in Listing 1.³²

```

SELECT ?mweOrLE, COUNT(?e)
FROM <http://kaiko.getalp.org/dbnary/eng>
WHERE {
  ?e a ?mweOrLE ;
    ontolex:canonicalForm ?wf.
  FILTER
    exists {?wf ontolex:phoneticRep ?pr}.

```

³¹These figures and the whole experiment is available in a notebook at https://github.com/serasset/dbnary-mwt-pronunciations/blob/main/notebooks/MWE_Pronunciation_LDK2023.ipynb.

³²Note that in all SPARQL queries, we do not add the PREFIXes as they are known and optional on the DBnary server.


```
VALUES ?mweOrLE
  { ontolex:MultiWordExpression
    ontolex:LexicalEntry }
} GROUP BY ?mweOrLE
```

Listing 1: SPARQL query to count the available phonetic representations (?pr) of lexical entries (?e). We also get the counts for entries types as `ontolex:MultiWordExpression` or `ontolex:LexicalEntry`.

A similar query is used to count the entries without pronunciation information. The results are given in Table 1.

type	with (# of pron)	without
LE	107327 (173512)	1102485
MWE	4977 (8143)	214243

Table 1: The number of English Lexical Entries available in the English Wiktionary with or without pronunciation information, among which we also count the MWEs. The total number of distinct pronunciations is also given.

These values are slightly different from the ones obtained using the Wiktionary category pages or the statistics pages. The reasons for this are (1) the Wiktionary statistics have been done a year ago, while the DBnary query reflects the status of the latest dump³³ and (2) Wiktionary categories refer to *pages* while the figures we have here are referring to *lexical entries* (there are usually several lexical entries described in a single page³⁴).

Despite being marginally different, these counts confirm the original observed proportions of less than 10% of Lexical Entries having pronunciation, while less than 2.3% of MWEs come with pronunciation information.

4.2 Borrowing pronunciation of MWEs from their components

The main idea in (Bajčetić et al., 2023) is to construct the pronunciation of MWEs by borrowing the pronunciation of their components. This is straightforward when components have a single pronunciation, but requires care when the pronun-

³³These numbers reflect the DBnary dataset version 20230320. As Wiktionary evolves and DBnary dataset is updated, more data is constantly added to the resource. For instance, the previous version (dated 20230301), contained 172846 (resp. 1097873) Lexical entries with (resp. without) pronunciation and 8074 (resp. 213276) MWEs with (resp. without) pronunciation.

³⁴For instance, the 173512 lexical entries with pronunciation counted here are described in 75082 different pages.

ciation differs for different meanings (in the case of heteronyms).

To compute its pronunciation, the MWE is decomposed in components and each component is independently queried for its pronunciation information. For this experiment, the decomposition has been done straightforwardly by breaking the MWE according to spaces and assuming that each component of the derivation is a canonical form.

As components may have several pronunciations, all the resulting pronunciations are combined leading to a set of candidates. However, this method is faulty when we are dealing with heteronyms.

4.3 Dealing with heteronymy

As defined on Wikipedia, “a heteronym (also known as a heterophone) is a word that has a different pronunciation and meaning from another word but the same spelling”.³⁵ A common example for heteronyms is given by the lexical entries “bass” (fish, pronounced /bæs/) and “bass” (sound, low in pitch, pronounced /beɪs/).

In our setup, heteronyms are defined as *pages describing at least two lexical entries* which have at least two different sets of pronunciations. To identify those heteronyms, we query all pages for their different pronunciation sets using the SPARQL query given in Listing 2. In the resulting table, the heteronyms are pages that appear more than once.

```
SELECT ?p ?prons
  (GROUP_CONCAT(?e; SEPARATOR = ",")
   as ?entries)
FROM <http://kaiko.getalp.org/dbnary/eng>
WHERE {
  ?p a dbnary:Page; dbnary:describes ?e.
  {
    SELECT ?e          ## sub query 1
      (GROUP_CONCAT(?pr ; SEPARATOR=",")
       as ?prons) {
    SELECT ?pr ?e {    ## sub query 2
      ?e ontolex:canonicalForm /
        ontolex:phoneticRep ?pr .
    } GROUP BY ?e ?pr
      ORDER BY ?pr
    } GROUP BY ?e
  }
}
```

³⁵Quoted from [https://en.wikipedia.org/wiki/Heteronym_\(linguistics\)](https://en.wikipedia.org/wiki/Heteronym_(linguistics)) [accessed 2023.03.37]

```
} GROUP BY ?p ?prons
```

Listing 2: SPARQL query to extract all heteronym pages (?p), along with their distinct pronunciations (?prons) and the corresponding entries (?entries). Sub-query 1 and 2 extract and group the different pronunciations for each lexical entry, then entries are grouped by distinct pronunciation set.

Page	Pronunciations	gloss
911	/ˌnaɪn wʌŋ ˈwʌŋ/	emergency
911	/ˌnaɪn əˌlɛvən/	porsche
bass	/beɪs/	low pitch
bass	/bæs/	fish
hinder	/ˈhɑːm.də/,/ˈhɑːm.də/	make difficult
hinder	/ˈhɪndə/,/ˈhɪndə/	more hind
tower	/ˈtəʊ.ə(ɪ)/,/ˈtəʊə/	tall structure
tower	/ˈtəʊ.ə(ɪ)/	one who tows
lead	/lɪd/, /liːd/	to guide
lead	/lɛd/	metal

Table 2: A sample of heteronym pages along with their distinct pronunciation groups.

In English DBnary, we identified 970 heteronym pages among the 75082 pages with pronunciation. A sample of these is given in table 2.

When a component is identified as a heteronym, we have to choose among the different pronunciations for the one that is valid for the MWE. For example, in the MWE *lead pencil*, the component *lead* corresponds to the metallic sense, pronounced /lɛd/, while in *lead astray*, the component *lead* corresponds to the verbal "to guide" sense, pronounced /liːd/. The same phenomenon occurs for *bass guitar* where *bass* refers to the "low in pitch" meaning, pronounced /beɪs/, while sea bass contains the *bass* (as a fish) component, pronounced /bæs/.

In order to correctly decide which pronunciation should be used for such a heteronym component and not over-generate erroneous pronunciations, we use the derivation relations that are present in Wiktionary and are now available in DBnary. Figure 4 shows an example of such derivation relation in the context of the heteronym page *bass*. All derivation relations is extracted from DBnary with the SPARQL query given in Listing 3. The English DBnary dataset contains 239284 such relations.

```
SELECT
  DISTINCT ?deriv_from ?source_label
           ?deriv_to ?target_label
FROM <http://kaiko.getalp.org/dbnary/eng>
```

```
WHERE {
  ?deriv_to
    dbnary:derivedFrom ?deriv_from ;
    dbnary:describes
      / rdfs:label ?target_label .
  ?deriv_from rdfs:label ?source_label .
}
```

Listing 3: SPARQL query to extract all derivation relations from DBnary

When a component of a MWE is a heteronym, we look for a corresponding derivation relation that points us to the *Lexical Entry* the MWE derives from. We then use the pronunciation of this *Lexical Entry* and ignore pronunciations of other *Lexical Entries* with the same canonical form.

4.4 Experiment and evaluation

In order to evaluate this experiment, we will use the pronunciations of the 4977 MWEs that are available in DBnary as a gold standard. When computing the pronunciation candidates, four cases are used:

- **NP**: No pronunciation is available for at least one of the components,
- **COMP**: All components are non-heteronym and have pronunciation information,
- **HCOMP**: At least one component is a heteronym and derivation relation is available,
- **HND**: At least one element is heteronym and no derivation relation is available.

In **NP** and **HND** cases, we chose not to produce any candidates. We measure the Precision, recall and F1-measure in cases **COMP** and **HCOMP** by comparing known pronunciation with produced candidates. For this comparison, we applied four normalisation methods on the pronunciations:

- **NO**: pronunciation strings are compared without any normalisation,
- **SPA**: spaces are removed from pronunciation strings before comparison,
- **SUP**: suprasegmental signs (primary and secondary stresses, lengths, syllable breaks, etc.) are removed from the pronunciation strings before comparison,
- **SUPSPA**: suprasegmentals and spaces are removed from the pronunciation strings before comparison.

Norm	COMP			HCOMP			All ^a		
	prec	recall	f1	prec	recall	f1	prec	recall	f1
NO	.1172	.1731	.1269	.0310	.0781	.0381	.0516	.0771	.0560
SPA	.1186	.1761	.1285	.0382	.0976	.0481	.0524	.0789	.0570
SUP	.2937	.5045	.3324	.1688	.3993	.2057	.1318	.2292	.1495
SUPSPA	.3457	.5994	.3896	.2367	.5712	.2938	.1561	.2748	.1766

^aOverall performance accounting for cases where we do produce results (COMP and HCOMP) and cases where we do not (NP, HND). This is given for exhaustive evaluation, but as we were able to distinguish between the different cases, these measure do not reflect the real difficulty of the task.

Table 3: Evaluation of the experiments using four normalisations on the pronunciation strings.

case	in gold standard	in DBnary
NP	2448	86689
COMP	2160	114969
HCOMP	128	2246
HND	241	10340

Table 4: The number of MWE in each of the different evaluation cases.

Table 3 gives the precision, recall and F1-measure for the different cases and normalisations. We give overall evaluation results on all four cases for exhaustivity, but as the process is generating pronunciation proposals that will be manually validated, the figures only reflect the proportion of cases where we can propose something (54.7%) and cases where we cannot (45.3%). Overall, this evaluation shows encouraging results when ignoring the suprasegmental elements of the pronunciation strings, thus validating the main strategy to raise the number of pronunciations for MWEs by borrowing pronunciations from their components. However, suprasegmental seems harder to figure out and we hypothesise that they are as much influenced by the global MWE context than by each intra-component pronunciation.

As detailed in table 4, overall, we are able to produce pronunciation candidates for 114969 MWEs using the **COMP** strategy and for 2246 MWEs using the **HCOMP** strategy.

4.5 Lessons learned and current work

By using DBnary dataset we were able to more easily extract lexical data on which we applied the original strategy described in (Bajčetić et al., 2023). This process is quite efficient and does not require any manual intervention and may be used each time new MWEs are added to Wiktionary.

However, we currently identify several short-

comings for which we should investigate deeper. The first limitation we need to address is identifying to which extent the proposed strategy may be ported to other languages available in DBnary (which currently extract from 23 different editions). In this experiment decomposition of the MWE in a set of component is simply based on space characters and we assumed that each component appeared in its canonical form. Such heuristics seem justified in the case of English language where entries have very few inflected forms, but will certainly become questionable if we apply it on other languages like French (that has a more productive morphology) or German (where components are usually concatenated without spaces). Moreover even in the case of English language, with this heuristic the term *acoustic bass guitar* cannot be decomposed as "acoustic" + "bass guitar" and we cannot take advantage of the already existing pronunciation attached to "bass guitar". Future work should investigate other decomposition processes and the use of inflected forms as components in a second step.

Another limitation, that may explain the precision measures, comes from the fact that DBnary does not correctly identify the regional variant information of pronunciation strings. For example, when computing pronunciation for *bomb crater* we look for the entries *crater* (UK: /kɹeɪ.tə(ɪ)/, US: /kɹeɪ.tə/) and *bomb* (UK: /bɒm/, US: /bɑm/, obsolete: /bʌm/) and produce six candidates that are the combination of all individual components pronunciation, while only two should be produced by combining the UK (resp. US) pronunciations. This shortcoming will not be addressed before DBnary corrects its English extractor to properly identify and represent the regional variant for each extracted pronunciation.

5 Semantic enrichment of components of MWEs

The former sections demonstrated the advantage of concentrating our work on adding pronunciation information to MWEs on the use and adaptation of the DBnary resource. We stressed that DBnary is offering the extracted information from Wiktionary in a structured fashion, more precisely using LOD compliant models and vocabularies. And we see in this feature another precious advantage of using DBnary for our work dealing with the enrichment of MWEs included in Wiktionary (and in the longer term also for resources like the Open English WordNet, or others), focusing in a next step on morphosyntactic and semantic information that can be added to the components of such MWEs.

5.1 The decomposition module of *OntoLex-Lemon*

As DBnary is making use of the *OntoLex-Lemon* model, we can take advantage of the existence of its “Decomposition” module,³⁶ which is graphically displayed in Figure 5.

We can observe that the property “decomp:subterm” of the Decomposition module is equivalent to the property “dbnary:derivesFrom”, recently introduced in DBnary, in order to represent the Wiktionary section “Derived terms” (see Figure 4) for comparison. Therefore, we can just map the “rdf:Object” of “dbnary:derivesFrom” to the “rdf:Object” of “decomp:subterm”, while the rdf:Subject of “decomp:subterm” is the MWE itself, as been seen in Listing 4.

As a result, the recent adaptations of DBnary allow not only to generate pronunciation information for MWEs contained in the English edition of Wiktionary, but also to add morphosyntactic and semantic information to the components of such MWEs, and to encode this information in such a way that the new data set can be published on the Linguistic Linked Open Data cloud.

```
:electric_bass_lex a
  ontolx:MultiwordExpression ;
```

³⁶The specification of *OntoLex-Lemon* describes “Decomposition” in those terms: “Decomposition is the process of indicating which elements constitute a multiword or compound lexical entry. The simplest way to do this is by means of the subterm property, which indicates that a lexical entry is a part of another entry. This property allows us to specify which lexical entries a certain compound lexical entry is composed of.”. Taken from <https://www.w3.org/2016/05/ontolx/#decomposition-decomp>

```
decomp:subterm eng:electric_Adjective_1 ;
decomp:subterm :eng:bass_Noun_1 .
```

Listing 4: The (simplified) representation of “electric bass” using the Decomposition module of *OntoLex-Lemon*, with links to lexical data encoded in DBnary

Using this module, we can thus explicitly encode the morphosyntactic, semantic and domain information of the components of MWEs, which are only implicitly present in Wiktionary. For our example, we know that “electric” has PoS “adjective” (Wiktionary lists also a nominal use of the word) and “bass” the PoS “noun” (Wiktionary lists also an adjectival and a verbal uses), while semantically disambiguating the components of the MWE (in the full DBnary representation, the “ontolx:Word”: “eng:bass_Noun_1” is linked to the corresponding instances of “ontolx:Sense”. And in fact, we can then link to a corresponding Wikidata entry for “bass guitar” (<https://www.wikidata.org/wiki/Q46185>) and the one for “electricity” (<https://www.wikidata.org/wiki/Q12725>)

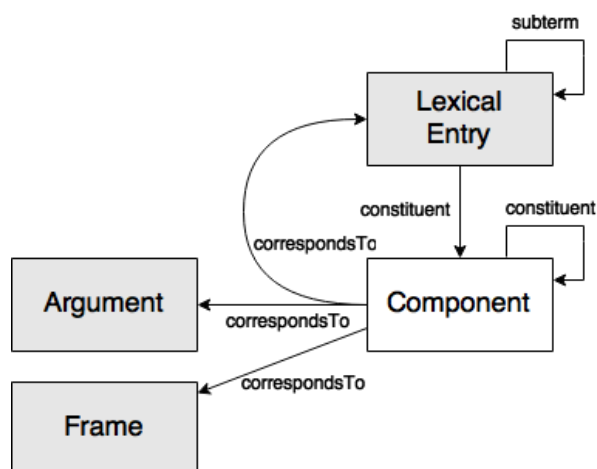


Figure 5: The Decomposition module of *OntoLex-Lemon*. Taken from <https://www.w3.org/2016/05/ontolx/#decomposition-decomp>.

6 Conclusion and future work

We described in this paper on-going work on computing pronunciation information for multiword expressions (MWEs) included in Wiktionary. In the course of this work, we got acquainted with the DBnary resource, which is offering a Linked Open Data compliant representation of lexical information extracted from Wiktionary, using at its core the *OntoLex-Lemon* model and other related

vocabularies. As it was immediately clear that using the extraction engine of DBnary is easing massively our work, we teamed with the maintainer of DBnary, who adapted the extraction engine for our needs. Those recent updates are the focus of this paper. We discovered also that this way, we can not only easily generate pronunciation information for MWEs, but we can also in a straightforward manner add morphosyntactic and semantic information to the components of MWEs. This will lead to the generation of a new data set for English MWEs. As a result, the DBnary engine is now more than an extractor from Wiktionary and a mapper to an LOD compliant representation, as it generates lexical information that can be used for enriching existing lexical resources.

We plan to port some of our approach to other languages supported by DBnary, aiming at a multilingual data set for MWEs.

Limitations

While our approach can probably be transferred to other languages, in cases where the Wiktionary structure for those languages is similar, there is one aspect of pronunciation extraction and combination that we have not discussed and this concerns the pronunciation(s) of variants of English, which are included in Wiktionary, like British, General American, Irish, Canadian, Australian and New Zealand English. In our current work we ignored the variants as they were not (yet) available in DBnary, so we "overlook" the variants information and produce potentially unusable new pronunciations (that will have to be discarded at manual validation). However, we would want to include all these varieties of our future work. This should not be too complicated, as the approach would follow the same principle as explained in the paper, with one extra layer of variant matching.

Another limitation of our work lied in the fact that Wiktionary is ever-changing. So anything done at one point in time needs to be re-done in the future due to changes in the data and also newly added data. The fact that Wiktionary grows quite fast means that the best approach would be incremental or recursive in some way, and automatically check for newly added pronunciations which can create new MWEs pronunciations, while also confirming that the previously created ones have not been altered and need updating. But our team-

ing with the maintainer of DBnary seems to offer a good solution, as DBnary is updated twice a month.

Another current limitation lies in the fact that we consider only binary MWEs. This is due in a good part to the fact that Wiktionary is not delivering a lot of information when dealing with longer MWEs, but we are analysing the available data in more details.

Ethics statement

We consider our work to have a broad impact because Wiktionary is widely used across the world, as a free and open-source resource. Additionally, we plan to include the output of our research into other resources, like for example the Open English WordNet, which are also resources that are free to use and open-source. We hope that in this way the results of our work can potentially be useful to people all around the world who read or speak English, as well as text-to-speech (and possibly speech-to-text) systems which are gaining popularity and are very important for the visually impaired community, among others.

We do not see any ethical issue related to the generation of additional information that can be attached to Wiktionary MWEs and their components.

Acknowledgements

The presented work is pursued in the context of the COST Action NexusLinguarum – European network for Web-centered linguistic data science (CA18209), 731015). The DFKI contribution is also pursued in the context of the LT-BRIDGE project, which has received funding from the European Unions Horizon 2020 Research and Innovation Programme under Grant Agreement No 952194.

References

- Lenka Bajčetić, Thierry Declerck, and Gilles Sérasset. 2023. [Enriching multiword terms in Wiktionary with pronunciation information](#). In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 65–72, Dubrovnik, Croatia. Association for Computational Linguistics.
- Christian Chiarcos and Maria Sukhareva. 2015. [OLiA Ontologies of Linguistic Annotation](#). *Semantic Web*, 6(4):379–386. Publisher: IOS Press.

- P. Cimiano, P. Buitelaar, J. McCrae, and M. Sintek. 2011. [Lexinfo: A declarative model for the lexicon-ontology interface](#). *Journal of Web Semantics*, 9(1):29–51.
- Philipp Cimiano, John McCrae, and Paul Buitelaar. 2016. [Lexicon Model for Ontologies: Community Report](#), 10 May 2016. Technical report, W3C.
- Gerard de Melo. 2015. [Lexvo.org: Language-related information for the Linguistic Linked Data cloud](#). *Semantic Web*, 6(4):393–400.
- Thierry Declerck and Lenka Bajčetić. 2021. [Towards the addition of pronunciation information to lexical semantic resources](#). In *Proceedings of the 11th Global Wordnet Conference*, pages 284–291, University of South Africa (UNISA). Global Wordnet Association.
- Thierry Declerck, John Philip McCrae, Matthias Hartung, Jorge Gracia, Christian Chiarcos, Elena Montiel-Ponsoda, Philipp Cimiano, Artem Revenko, Roser Saurí, Deirdre Lee, Stefania Racioppa, Jamal Abdul Nasir, Matthias Orlikowsk, Marta Lanau-Coronas, Christian Fäth, Mariano Rico, Mohammad Fazleh Elahi, Maria Khvalchik, Meritxell Gonzalez, and Katharine Cooney. 2020. [Recent developments for the linguistic linked open data infrastructure](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5660–5667, Marseille, France. European Language Resources Association.
- John P. McCrae, Paul Buitelaar, and Philipp Cimiano. 2017. [The OntoLex-Lemon Model: development and applications](#). In *Proc. of the 5th Biennial Conference on Electronic Lexicography (eLex)*.
- John Philip McCrae, Alexandre Rademaker, Ewa Rudnicka, and Francis Bond. 2020. [English WordNet 2020: Improving and extending a WordNet for English using an open-source methodology](#). In *Proceedings of the LREC 2020 Workshop on Multimodal Wordnets (MMW2020)*, pages 14–19, Marseille, France. The European Language Resources Association (ELRA).
- Gilles Sérasset. 2015. [Dbnary: Wiktionary as a lemon-based multilingual lexical resource in rdf](#). *Semantic Web*, 6:355–361.
- Gilles Sérasset and Andon Tchechmedjiev. 2014. [Dbnary : Wiktionary as Linked Data for 12 Language Editions with Enhanced Translation Relations](#). In *3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing*, page to appear, Reykjavik, France.
- Tatu Ylonen. 2022. [Wiktextract: Wiktionary as machine-readable structured data](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1317–1325, Marseille, France. European Language Resources Association.

Lexico-Semantic Mapping of a Historical Dictionary: An Automated Approach with DBpedia

Sabine Tittel

Heidelberg Academy of Sciences and Humanities

Karlstraße 3, D – 69117 Heidelberg, Germany

sabine.tittel@hadw-bw.de

Abstract

Modeling lexical resources following the Linked Data paradigm has become a widespread method to contribute to the multilingual web of data. For the modeling of linguistic information such as words and their morphosyntactic aspects, standard vocabularies offer elaborate means to enable cross-resource and cross-domain access to the resources. To establish access to the word senses, it is pivotal to create a mapping of each word sense and its underlying concept to an external, language-independent knowledge base of the Semantic Web such as DBpedia. However, this lexico-semantic mapping is a very time-consuming endeavor and is often neglected. And yet, the problem of how to install time-saving approaches is not resolved. Therefore, we propose a solution for an automated lexico-semantic mapping based on Old French lexicographic data. The quantitative and qualitative evaluations of the outcome show very promising results. Overall, approx. 71% of the word senses can be mapped to a DBpedia entry: approx. 12.7% of semantically accurate mappings and approx. 58.2% of approximate, yet semantically meaningful mappings. These results can be fully extrapolated to our linguistic resource and also transferred to the Linked Data modeling of related resources.

1 Introduction

The last decade has seen many successful attempts to model lexical resources as Linked Open Data (Bizer et al., 2009). RDF (*Resource Description Framework*, Klyne et al. (2004)) is used as the standard format along with W3C-standard vocabularies and ontologies as a means to create a web of interlinked data. Attempts focus on the modeling of words and parts of speech, their graphical realizations, morphological and syntactic aspects, translations into other languages, their role in multi-word expressions, etc. (for an overview of technolo-

gies, vocabularies, and methods, see Bosque-Gil et al. (2018), Khan et al. (2022)). The vocabulary most often used for modeling lexical resources is OntoLex-Lemon, Cimiano et al. (2016). While the linguistic structures of the lexical resources can be seamlessly converted to RDF, a challenging aspect of the modeling process is to integrate links from the *senses* of the words (lexemes) and their underlying *concepts*, respectively, to an external knowledge base. We call this the lexico-semantic mapping (in the following, LexSemMapping). The LexSemMapping is pivotal for establishing lexical-semantics-based access to the lexical units (that is, the nexus of a given lexeme and precisely one (of its) senses): Only lexical-semantics-based access makes the lexical units of, for example, a historical dictionary, available for cross-domain and cross-resource access that is, most importantly, independent from the language and language stage of the resource.

For the LexSemMapping, an extra-linguistic resource depicting the things of the world such as Wikidata and DBpedia¹ can serve as an external knowledge base. An illustration of the motivation for a LexSemMapping is as follows: Lexical resources contain numerous designations for, say, clergymen: Old High German *priest* m., *priestar* m., *prêstar* m., Middle High German *priestære* m., and High German *Priester* m. (since 9thc, Grimm² 13,2115² and DWDS PRIESTER³), Old High German *gotmanno* m., High German *Gottesmann* (since ca. 870, Grimm² 8,1285; DWDS

¹<https://www.wikidata.org/>, <https://www.dbpedia.org/>; these and all following URLs are accessed on 02-21-2023].

²*Deutsches Wörterbuch von Jacob Grimm und Wilhelm Grimm*, digital version, <https://woerterbuchnetz.de/?sigle=DWB#Priester>.

³*Digitales Wörterbuch der deutschen Sprache*, <https://www.dwds.de/wb/Priester>; we note that the DWDS offers a Thesaurus function leading to semantic cognates; however, this is limited to the German lexemes registered within the DWDS.

GOTTESMANN), Old French *pestre* m., *prastre* m., *prebstre* m., *preiste* m. (since the beginning of the 12thc, DEAFél PRESTRE⁴), *flame* m., *flamine* m., *archiflame* m. (since 13th/14thc., DEAFél FLAME⁵, Italian *flamine* m. (since 1261-1292, TLIO FLÀMINE⁶, Old Occitan *flamina* m. (DOM FLAMINA⁷), and many more. The senses of these lexemes represent concepts that are connected to different religions, cultures, times, and connotations. Their investigation is promising not only from a linguistic point of view but also as a linguistic underpinning for studies on expressions of religion through time and space (cp. the article PRIESTER in Bautier et al., 1977-1998, 7,203-208; Richard, 1959; Salisbury, 2015). Creating a connection, for example, from all senses with the concept ‘Priests’ to the DBpedia entry ‘Priest’, or from all clergymen of all religions to a generic entry ‘List_of_religious_titles_and_styles’⁸ could establish access through the means of the Semantic Web to all of the lexemes listed above. These are otherwise very difficult to find.

Indeed, OntoLex-Lemon offers classes to model sense definitions (`LexicalSense`) and concepts (`LexicalConcept`⁹) and the predicates (`reference` and `isConceptOf`, respectively¹⁰) to link these classes to an external knowledge base. Its entities then serve as the objects of the RDF triples for the LexSemMapping.

However, the LexSemMapping, to the best of our knowledge, has rarely performed on a larger scale. We suspect that this is (partly) because such a mapping is a very tedious and time-consuming endeavor. The problem thus arises as to how a LexSemMapping of lexical units can be established in a quicker and more efficient way. In this paper, we propose a solution for this problem by developing methods for an automatic mapping of lexical units to DBpedia.

⁴<https://deaf.ub.uni-heidelberg.de/lemme/prestre>.

⁵<https://deaf.ub.uni-heidelberg.de/lemme/flame2>. Hereafter, all Old French lexemes refer to DEAFél.

⁶*Tesoro della Lingua Italiana delle Origini*, <http://tlio.ovi.cnr.it/voci/025560.htm>.

⁷*Dictionnaire de l'occitan médiéval*, <http://www.dom-en-ligne.de/>.

⁸<https://dbpedia.org/page/Priest>, https://dbpedia.org/page/List_of_religious_titles_and_styles.

⁹In accordance with the semiotic pentagon, see, e.g., Blank (2001, 9).

¹⁰<https://www.w3.org/2016/05/ontolex/>.

The remainder of the paper is divided into an overview of related work (Section 2), a description of the lexical resource that is our use case (Section 3), an assessment of manual LexSemMapping (Section 4), and the development and evaluation of automatic approaches (Section 5). We conclude our paper by presenting the overall result and an outlook (Section 6).

2 Related Work

Establishing data access based on lexical semantics is important for lexical resources, in particular for historical language stages whose lexical units are harder to access than those of modern languages; and yet, the process of LexSemMapping is rarely described in the literature.

Herold et al. (2012) describe the attempt to do this for the data of the *Digitales Wörterbuch der Deutschen Sprache – DWDS-Wörterbuch* (DWDSWB)¹¹: Through an alignment of this dictionary with the entries of the *Deutsches Wörterbuch von Jacob Grimm und Wilhelm Grimm*, Volumes I–XVI, Leipzig 1854–1960 (¹DWB), a semantic disambiguation shall be achieved. This corresponds to a LexSemMapping, even if the target is not expressed as an RDF triple object. But the challenges due to homonyms, polysemy, and semantic shift led Herold et al. (2012, 42) to conclude that, «Given the huge amount of manual effort needed to complete the alignment between DWDSWB and ¹DWB on the level of lexical entries it seems unfeasible to achieve a mapping for individual senses».

Bozzi (2016) detail their failed attempt to use WordNet for a lexical-semantic networking of data of the *Dictionary of Old Occitan medicobotanical terminology* (DiTMAO). DiTMAO utilizes OntoLex-Lemon as a means to perform a LexSemMapping of the modeled lexemes through external ontologies: «In the next step, the DiTMAO partners will formalize the conceptual domain, describing the fields of botany, zoology, mineralogy, human anatomy, diseases and therapies (medication, medical instruments) [...] to ease the “onomasiological” access to the lexicon», Bellandi et al. (2018, 10-11). However, they do not further elaborate on how to establish a LexSemMapping.

Declerck et al. (2015, 348-350), in sample data of the *Wörterbuch der bairischen Mundarten in Österreich* (WBÖ¹²), link the lexeme Ger-

¹¹<https://www.dwds.de/d/wb-dwdswb>.

¹²<https://wboe.oew.ac.at/>.

man *Trupp* (a squad) to the DBpedia entry ‘Social_Group’. They point out the importance of integrating the data into larger semantic contexts, as well as linking to other external resources that also connect to the DBpedia entry given in the example. How this linkage with DBpedia is to be performed, however, remains unresolved: «An issue we would like to consider is the possibility of automatically linking to external resources, those being both of linguistic nature or encyclopedic nature. We do not have an answer to this point for the time being. As a heuristic, while knowing that the Limburg lexical data concerns anatomy, and the reference language is standard Dutch, we can automatically query DBpedia for all entries that have a Dutch word marked with the additional “_(anatomy)” extension, such as for example [http://nl.dbpedia.org/page/Hoofd_\(anatomie\)](http://nl.dbpedia.org/page/Hoofd_(anatomie)). However, this might only offer a very specific solution», (Declerck et al., 2015, 353).

Cimiano et al. (2013) evaluate possibilities to model the *semantics by reference* implied by OntoLex-Lemon in a more fine-grained method than the connection of `LexicalSense` to an ontology allows, bringing back semantic disambiguation at least partially into the model. Their code samples (Cimiano et al., 2013, 58f.) show DBpedia, among others, as an external knowledge base, but the process of semantic disambiguation itself is not discussed.

Giuliani and Molina Sangüesa (2020) describe the integration of two large historical lexical resources, i.e., the *Tesoro della lingua italiana delle origini* (TLIO¹³) and the *Nuevo Diccionario Histórico del Español* (NDHE, *Real Academia Española*¹⁴), with the taxonomy of the *Historical Thesaurus of English* (HTE)¹⁵. Focusing on the domain ‘health and illness’, they translate HTE’s entities into Spanish, extend them to a more fine-grained level, and integrate them into their work infrastructure as an onomasiological backbone. The taxonomy is also converted into an ontology in OWL (Bechhofer et al., 2004) called DHistOntology and the modeling of the two resources in RDF is described as a future goal (Molina Sangüesa, 2023). Their aim is to enhance their workflow by aligning similar concepts in both resources and to streamline sense definitions while editing the dic-

tionary articles with one shared dictionary writing system. This is a promising concept, albeit the lexico-semantic mapping seems to be performed manually.

The historical dictionary *Lessico Etimologico Italiano* (LEI, Pfister 1979–) also examines using the classes of the HTE as a means to establish onomasiological access. The goal is not an integration of the LEI resource into the Linked Data landscape but the creation of a locally used, proprietary feature for the online publication LEI-digitale.¹⁶ As a first step, their approach focuses on the LexSemMapping of the Latin etyma – that serve as the headwords of the LEI articles – and their definitions. The second step is to integrate the lexical units of the articles, i.e., the Italian lexemes and their definitions. The heterogeneity of the latter is significant, including single-word definitions in modern Italian and also Latin, a sequence of modern Italian translations (i.e., of several senses in one definition text), periphrastic definitions, nomenclature adopting the classification by Carl von Linné (we will further discuss Linné in Section 5.1), and more. The mapping is done manually: Concepts are looked up in Wikipedia, and corresponding entities are identified in and linked to the HTE taxonomy. The link is manually integrated into the XML files of the articles.¹⁷ Since the LEI is a very large resource with a great amount of legacy data (and also born-digital data), it seems crucial for the success of their LexSemMapping to integrate automated steps into the process. However, no solution for time-saving automation has been promoted so far.

3 The Linguistic Resource

The motivation for our approach to establishing a more efficient method for LexSemMapping derives from modeling the data of the *Dictionnaire étymologique de l’ancien français* – DEAF (Baldinger, 1971-2020) as Linked Open Data. The DEAF is a comprehensive dictionary of Old French from its first resource 842 AD until ca. 1350 AD, compiled under the aegis of the Heidelberg Academy of Sciences and Humanities until 2020.¹⁸ We have invested in modeling the DEAF articles as Linked Open Data for two reasons: firstly, to make the data of the DEAF accessible beyond the nuanced

¹³<http://tlio.ovl.cnr.it/TLIO/>.

¹⁴<https://www.rae.es/>.

¹⁵<http://historicalthesaurus.arts.gla.ac.uk/>.

¹⁶<https://lei-digitale.it/>.

¹⁷Personal communication by Alessandro A. Nannini, LEI, to whom we express our sincere thanks.

¹⁸<https://www.hadw-bw.de/deaf>.

yet predefined, and thus limited research functions of its online publication, DEAF $\acute{e}l$ ¹⁹; and secondly, to facilitate the usability, queriability, and interpretability of the DEAF data in the global context of the Semantic Web. We describe the vocabularies, e.g., OntoLex-Lemon and OLiA (Chiarcos and Sukhareva, 2015), the concept, outcome, and challenges of the modeling process in Tittel and Chiarcos (2018) and – with further elaboration – in Tittel (forthcoming). In Tittel and Chiarcos (2018), we proposed implementing a semi-automatic process to increase efficiency. In this process, XSLT scripts would model the DEAF data as RDF by integrating the predicate `ontolex:isConceptOf` and a wildcard in place of a link to an extra-linguistic ontology as the object of the RDF triple. This would help prepare for manual mapping. It, of course, does not produce a meaningful statement, and the necessary manual post-processing could not be performed due to the termination of the funding period of the DEAF. However, the RDF data offer a starting point; for example, for Old French *raicele* s.f. “plante vivace de la famille des Violaceae, aux feuilles en rosette et aux fleurs blanches légèrement ou pas parfumées, violette blanche”, the concept “White Violet” can now be mapped to the entity of DBpedia ‘Viola_alba’²⁰ in the following way (RDF serialized in Turtle):²¹

```
1 deaf:raicele_lexConcept
2  ontolex:isConceptOf dbr:Viola_alba .
```

4 Manual LexSemMapping

A manual LexSemMapping for the DEAF data promises the best results. This is particularly true with respect to the *Historical Semantic Gap* (Tittel and Chiarcos (2018), Giuliani and Molina Sangüesa (2020, 355f.)) that often occurs between a concept represented by a lexeme in a historical (in this case, medieval) language stage and the concept of the same lexeme in the modern language. E.g., medieval concepts of the bloodstream adhere to a metabolism that does not know blood circulation (described only in 1628 by William Harvey, Schipperges (1990, 53)). Therefore, Old French *veine* f., for example, does not denote the blood vessel transporting the blood back to the heart (as part of blood circulation). Instead, *veine* denotes

a blood vessel transporting the nourishing blood from the liver to all body parts and then back to the liver. Hence, the concept cannot be mapped to the modern concept of the “vein”, as in DBpedia’s entry ‘Vein’²² without causing semantic disruption and anachronistic cross-fade.²³ On the other hand, the LexSemMapping is straightforward when the concept to be mapped has the exact same scope and application today as it did in medieval times. This is often the case for plant and animal names, musical instruments, tools, etc., and DBpedia is very well suited for this purpose.

For writing each dictionary article, the lexicographer penetrates the semantic scope of the analyzed lexeme and grasps the concept of each lexical unit in a way that makes possible a seamless integration of an ontology entity into the data. Furthermore, they might analyze several lexemes belonging to a domain at a certain point in time and, in doing so, remain focused on that particular topic. E.g., after editing lexemes occurring in the context of the *veine* (see above), they have internalized medieval metabolic concepts and pneuma theory (Putscher, 1974) to the point of becoming, to a certain extent, an expert which further facilitates the mapping process. We, therefore, argue that a manual LexSemMapping is feasible when done while editing a dictionary article.

The case of legacy data, as is the case for the DEAF dictionary, is different, however. DEAF $\acute{e}l$ contains approximately 84,000 lexemes with 92,776 lexical units²⁴ that must be linked, in hindsight, to an extra-linguistic knowledge base. The dictionary covers all aspects of the language, and hence, a LexSemMapping requires knowledge in all domains of life. For a retrospective mapping of legacy data, this is difficult: While the knowledge of the lexicographer is greatest at the time of the article editing, the person performing the mapping in retrospect must promptly acquire expertise for many domains ad hoc. This is also immensely time-consuming. Estimating 10 min per LexSemMapping adds up to 15.462 hours of work, roughly 200

²²<https://dbpedia.org/page/Vein>.

²³This observation leads to the demand for historicized ontologies that model the historical concepts of a domain of interest. This is not further discussed in this paper. We however indicate that the project *Knowledge Networks in Medieval Romance Speaking Europe* (ALMA, <https://www.hadw-bw.de/alma>) will develop domain ontologies for medieval medicine and law.

²⁴Not counting the lexical units where the sense is marked by ‘?’.

¹⁹<https://deaf.ub.uni-heidelberg.de/>.

²⁰https://dbpedia.org/page/Viola_alba.

²¹Namespaces, such as `deaf`, `ontolex`, and `dbr` (DBpedia) in the following code examples are assumed to be defined the usual way.

working days, for the DEAF data — provided that the required entities of a knowledge base do exist.

5 Automatic Approaches to LexSemMapping

To address this problem, we have developed automatic methods involving applying Python scripts for a LexSemMapping of the DEAF data. As an encyclopedic resource, DBpedia only registers (concrete and abstract) things that are described in Wikipedia (from where DBpedia extracts its data²⁵). Furthermore, DBpedia shows significant shortcomings with respect to historical concepts. Nonetheless, we focus on DBpedia as a target resource, acknowledging its broad range of entities and its pivotal role as a central node within the web of data.

At this point, we rule out linguistic resources such as WordNet, Open Multilingual Wordnet, and BabelNet²⁶ because our goal is to semantically map the concepts to an extra-linguistic resource enabling semantic access that is independent of a language representation. For the future expansion of the methodology, we will revisit this decision for the sake of larger interoperability.

5.1 Four Methods for Mapping Nouns

The 92,776 sense definitions of the DEAF are (i) partly defined by following the genus–differentia approach²⁷, (ii) partly by single French words, and (iii) partly by translations in Modern French, i.e., equivalents of the sense following the genus–differentia definition as the last word of the definition text. Aiming at a maximum of correct hits when linking the definitions to corresponding DBpedia entities, we define four methods for automatically mapping nouns: (i) We establish links using the terminology classified through the *Systema naturae* by Carl von Linné²⁸ (in the follow-

²⁵See <https://www.dbpedia.org/resources/linked-data/>.

²⁶<https://wordnet.princeton.edu/>, <https://omwn.org/>, <https://babelnet.org/>.

²⁷A genus–differentia definition is the state-of-the-art definition of a sense consisting of a generic term (*genus*, e.g., ‘plant’) and specifications of that term (*differentia*, e.g., ‘perennial’, ‘with rosette-shaped leaves’, ‘with lightly scented white flowers’, cp. the above mentioned White Violet).

²⁸Editio princeps Leiden [Lugdunum Batavorum] (Theodor Haak) 1735.—The systems by Carl Gottlob Rafn (<https://viaf.org/viaf/106965171/>) and Georges Léopold Chrétien Frédéric Dagobert, Baron de Cuvier (<https://viaf.org/viaf/4981028/>), are alternatives; in the DEAF, however, we do not see them used in a sense definition.

ing: LINNÉTERMINUS); (ii) we transform single-word definitions (SINGLEWORD); (iii) we use the Modern French equivalents (LASTWORD); (iv) we extract the genus proximus of a sense definition (GENUSPROXIMUS).

5.1.1 LINNÉTERMINUS Approach

Many definitions include a Linné classification that is utilized in this approach. The standard syntax is: “<definition> (<Latin term> L.)”, as in: *fave-rolle* f. t. de botanique “petite plante dicotylédone, de la famille des Plantaginaceae..., véronique des ruisseaux (*Veronica beccabunga* L.)” (limewort). But we also find definitions (i) with a Latin term enclosed in distinctive parentheses, beginning with an uppercase letter but without the ‘L.’ marker, (ii) the opposite: with the ‘L.’ marker but without the parentheses, and (iii) with neither the ‘L.’ marker nor parentheses. All these cases considered, roughly 200 definitions can be mapped through the LINNÉTERMINUS approach. Although this might not seem a significant contribution to automated mapping, the expected correctness of the results suggests the development of an algorithm that reads Linné classifications.

5.1.2 SINGLEWORD Approach

This approach is straightforward. The algorithm uses the single Modern French word of the definition (filtering out occasional question marks), as in: *lechement* m. “flatterie” (flattery). A database query results in 21,166 such SINGLEWORD definitions. These definitions don’t comply with the concept of genus–differentia definitions; they feature in DEAF*pré*, a section of DEAF*él*. DEAF*pré* contains the digitized material of the DEAF card index (with 1.5 million handwritten slips that amount to 12 million attestations of lexemes), structured into preliminary dictionary entries with a provisional semantic analysis.

5.1.3 LASTWORD Approach

A further approach is a method of reading the Modern French translation typically given as an equivalent of the sense at the end of the definition. This approach is based on the syntax: “<definition>, <Modern French word>”, as in: *figuier* m. “arbre qui produit la figue, figuier”, the fig tree. However, this approach has several drawbacks. The algorithm accurately reads a single word between the last comma and the closing quotation marks of the definition text (filtering out question marks). How-

ever, the hit ratio is influenced by many cases in which that particular single word is not a Modern French equivalent, but part of an enumeration that belongs to the periphrastic definition itself. An example is: *dachete* f. “sorte de petit clou à la tête particulièrement grande et à la tige angulaire, adapté aux besoins de cordonniers, tapissiers, etc.”. In this case, following the rules, the algorithm finds that *etc.* is the last word after the last comma; this can be filtered out. Consequently, *tapissiers* (tapestry weavers) is the word to be used by the algorithm for LexSemMapping. Sure enough, the tapestry weavers are only an example (together with *cordonniers*, shoemakers) for professional groups that use the *dachete* (a type of small nail). Nevertheless, this approach is highly relevant for automatic LexSemMapping due to its numerous occurrences.

5.1.4 GENUSPROXIMUS Approach

While the first three approaches aim at the LexSemMapping of the specific meaning of the word, this approach uses the genus proximus of the sense definition for an approximate mapping, i.e., of the meaning’s core. It relies on the periphrastic definitions in accordance with the syntax: “sorte de / sorte d’ / espèce de / espèce d’ <genus> <differentiae>”, e.g.: *tideman* m. “espèce de douanier qui attend la marée haute pour faire les bateaux arrivant acquitter les impôts”. Although *tideman* denotes a very particular tollkeeper, the generic tollkeeper (*douanier*) is the concept that will be mapped by the GENUSPROXIMUS approach. Oftentimes, the genus proximus is preceded by an adjective, such as ‘small’ or ‘large’; this will be considered by the algorithm. A database query results in 3,870 such GENUSPROXIMUS definitions.

5.1.5 Proof of concept with manually created data sample and English Translations

The mapping process to DBpedia is based on the fact that for each Wikipedia entry, a DBpedia entry can be assumed: «For each Wikipedia page, DBpedia has an entity following the same pattern: <http://en.wikipedia.org/wiki/Berlin> → <http://dbpedia.org/resource/Berlin>», see <https://www.dbpedia.org/resources/linked-data/> [accessed 02-17-2023]. To query Wikipedia’s data, e.g. for article entries, the Python script imports an API provided by Wikipedia (see ‘Wikipedia API’ at <https://pypi.org/project/Wikipedia-API/>).

To test feasibility, we conduct a Proof of concept (PoC): We implement a semi-automatic approach by manually preparing a data sample (*data_poc*). This sample consists of a list of lexemes, definitions, and keywords to be mapped for LINNÉTERMINUS, SINGLEWORD, LASTWORD, and GENUSPROXIMUS, each including 30 examples. The DEAF sense definitions are written in Modern French. Therefore, we provide English translations of the keywords to facilitate the detection of corresponding entries in the English Wikipedia for the algorithm. A list entry is structured as follows, with ‘lexeme’, ‘definition’, and ‘English keyword’, respectively:

```
1 ['zecharr', 'espèce de faucon', 'falcon']
```

The pseudocode for our PoC reads as follows:

```
1 IMPORT wikipediaapi
2 SET wiki_wiki TO wikipediaapi.Wikipedia('en')
3
4 DEFINE FUNCTION concat(text):
5     RETURN str(text).replace(' ', '_')
6         .replace('œ', 'oe').replace('æ', 'ae')
7         .replace('?', '')
8
9 DEFINE FUNCTION map(data_poc):
10    SET entries_to_dbr TO data_poc
11    FOR row IN data_poc[1:]:
12        SET keyword TO concat(row[2])
13        SET page_py TO wiki_wiki.page(keyword)
14        IF page_py.exists():
15            SET url TO page_py.fullurl
16            SET url_db TO str(url).replace('https://',
17                'en.wikipedia.org/wiki/',
18                'https://dbpedia.org/resource/')
19            row.append(url_db)
20        ELSE:
21            SET keyword TO 'unknown_entry'
22            row.append(keyword)
23    RETURN entries_to_dbr
```

The function `concat` (lines 4-7) replaces spaces with underscores, French ligatures, and question marks. The function `map` (lines 9-23) iterates over the lines of the sample data, requests Wikipedia entries and their URLs, and converts them into DBpedia URLs. If no entry is found, a message is printed. The result is saved to a JSON file; an extract is shown in Fig. 1.

```
[
  "anemoine",
  "sorte de renonculacées à fleurs violettes, dite aussi coquelourde,
  passe-fleur ou pulsatille",
  "anemone pulsatilla",
  "https://dbpedia.org/resource/Pulsatilla_vulgaris"
],
[
  "zecharr",
  "espèce de faucon",
  "falcon",
  "https://dbpedia.org/resource/Falcon"
]
```

Figure 1: Mapping result: LINNÉTERMINUS (extract).

Evaluation of the PoC The mapping result is promising, despite the fact that five mappings are

nonsense. E.g., Old French *lecherant* (lickspittle), falsely leads to <https://dbpedia.org/page/Licker>: «a fictional creature from Capcom’s Resident Evil series». *Datil* (date [fruit]) maps to a disambiguation page with person and place names, double dates, etc.; the correct mapping would be the entry ‘Date_(fruit)’ which in turn leads to the entry ‘Date_palm’, which again is wrong. Furthermore, one keyword could not be mapped by the script: *feve* “plante aquatique de la famille des Nélumbonacées [...], fève d’Égypte, Lotus sacré ou Lotus d’Orient (*Nelumbo nucifera*, *Nymphaea Nelumbo* L.); la graine de cette plante”. In our test data set, we select the second Linnæan term, *Nymphaea Nelumbo* (Indian lotus), as the keyword to be mapped. However, the English Wikipedia does not list the Indian lotus under ‘*Nymphaea Nelumbo*’ but instead under the first term, ‘*Nelumbo nucifera*’ (the German Wikipedia redirects from one to another; the English site does not). All the other keywords, i.e., 114 out of the possible 120, have been correctly mapped.

5.1.6 Implementation

Use of French Wikipedia entries. The following steps aim to use the French originals and avoid the manual English translation of the keywords that we performed for the PoC. We test two ways to do this: First, we direct the algorithm to use the French Wikipedia instead of the English: `wikipediaapi.Wikipedia('fr')` (line 2 of the code above) but don’t change the URL-replacement process. The algorithm produces 117 mappings. However, since DBpedia models the English Wikipedia entries, many of the produced mappings are incorrect. E.g., French *bois*, the woods, produces a link to the DBpedia entry ‘Bois’²⁹, which is, however, a disambiguation page with person and place names. The correct hit would have been the entry ‘Wood’.

Use of English Wikipedia equivalents. Next, the algorithm queries the Wikipedia API for French Wikipedia entries and, at the same time, for their English equivalents. `langlinks` is appended to the Python function `map` to test whether an English equivalent exists and if so, use its URL to generate the DBpedia URL (lines 6-15):

```
1 DEFINE FUNCTION map(data_poc):
2   SET entries_to_dbr TO data_poc
3   FOR row IN data_poc[:]:
4     SET keyword TO concat(row[2])
```

²⁹<https://dbpedia.org/page/Wood>.

```
5   SET page_py TO wiki_wiki.page(keyword)
6   SET langlinks TO page_py.langlinks
7   IF page_py.exists():
8     FOR k IN sorted(langlinks):
9       IF k EQUALS 'en':
10        SET url_en TO langlinks[k].fullurl
11        SET url TO page_py.fullurl
12        SET url_db TO str(url_en).replace('https://
13          en.wikipedia.org/wiki/',
14          'https://dbpedia.org/resource/')
15        row.append(url_db)
16      ELSE:
17        SET keyword TO 'unknown_entry'
18        row.append(keyword)
19  RETURN entries_to_dbr
```

Although this also produces incorrect mappings (e.g., when an English equivalent is missing³⁰ or when Wikipedia falsely allocates an English equivalent), the hit ratio is better than the first attempt.

Automatically identified keywords. We then implement solutions for automatically identifying the keywords to be mapped by the algorithm. Here, we work with a manually created test data set of 236 lexical units in the form of RDF data, e.g.:

```
1 deaf:ebenus skos:definition
2   "bois de l'ébénier, ébène"@fr .
3 deaf:pivernaus skos:definition
4   "goutte"@fr .
5 deaf:fie skos:definition
6   "fruit du figuier (Ficus carica L.),
7   comestible et de couleur violette,
8   ..., figue"@fr .
```

Many sense definitions offer keywords for several approaches simultaneously, for example, a keyword for LINNÉTERMINUS and for GENUSPROXIMUS. Thus, we order the approaches by the expected mapping accurateness of their performance. E.g., LINNÉTERMINUS is more accurate than GENUSPROXIMUS and, consequently, the algorithm prefers the first method to the second.

The pseudocode (extract) reads as follows³¹:

```
1 SET linne TO re.compile(r'\(.* L\.\)')
2 SET linne_unobvious TO re.compile(r'\([A-Z]
3 \w+[A-Z]\w+\ \w+\)')
4 SET linne_cap TO re.compile(r'([A-Z]\w+\
5 \w+(\ L.))')
6 SET linne_cap_single TO re.compile(r'([A-Z]
7 \w+(\ L.))')
8 SET linne_cap_unobvious TO re.compile(r'([A-Z]\w+\
9 \w+)')
10 SET linne_cap_single_unobvious TO re.compile
11 (r'([A-Z]\w+)')
12 SET last_word TO re.compile(r'(\, [^\, \r\n]|\;
13 [^\, \r\n]) (\w+ ?\w+) (\ et sim.|
14 \ et sim.) {0,1} (\??) (\ \(\?\)) ?$')
15 SET single_word TO re.compile(r'^(\w+ ?\w+)\??$')
16 SET sorte TO "sorte de"
17 SET sorte_apostr TO "sorte d'"
18 SET espece TO "espèce de"
19 SET espece_apostr TO "espèce d'"
```

³⁰This is the case for ten keywords: ‘Lèchefrite’, baking sheet, ‘Amertume’, bitterness, ‘Machine de guerre’, apparatus belli, etc.

³¹The complete Python script and RDF data can be found on GitHub, <https://github.com/SabineTittel/LexSemMapping>.

```

20
21 DEFINE FUNCTION map_rdf(graph):
22   FOR s, p, o IN graph:
23     IF p EQUALS (skos + 'definition')
24       and type(o) EQUALS rdflib.term.Literal:
25       IF linne.search(o):
26         SET keyword TO concat(re.sub('.*\((.*)
27         (\ L\.)\).*', r'\1', o))
28         SET page_py TO wiki_wiki.page(keyword)
29         IF page_py.exists():
30           make_langlinks(s, page_py)
31           continue
32       IF linne_cap.search(o):
33         SET keyword TO concat(normalize(re.sub
34         ('(.+)\)([A-Z]\w+\ \w+)(\ L.)(.*)',
35         r'\2', o)))
36         SET page_py TO wiki_wiki.page(keyword)
37         IF page_py.exists():
38           make_langlinks(s, page_py)
39           continue
40       # all other keyword queries follow
41
42       ELSE:
43         graph.add((s, ontolex + 'isConceptOf',
44         Literal('to be mapped')))
45
46 DEFINE FUNCTION make_langlinks(s, page_py):
47   SET langlinks TO page_py.langlinks
48   IF langlinks:
49     FOR k IN sorted(langlinks):
50       IF 'en' IN sorted(langlinks):
51         IF k EQUALS 'en':
52           SET url_en TO langlinks[k].fullurl
53           SET url_dbr TO str(url_en).replace
54           ('https://en.wikipedia.org/wiki/', '')
55           graph.add((s, ontolex + 'isConceptOf',
56           dbr + url_dbr))
57         ELSE:
58           graph.add((s, ontolex + 'isConceptOf',
59           Literal('missing English equivalent to
60           French Wiki entry')))
61       ELSE:
62         graph.add((s, ontolex + 'isConceptOf', Literal
63         ('no equivalents to French Wiki entry')))

```

To find the keywords, the algorithm uses regular expressions and looks for pre-defined strings: catchwords (lines 1-15). The function `map_rdf` iterates over the parameter for the argument `graph` (line 21): subject, predicate, and object of the triples of the imported RDF data set (with the 236 lexical units). For all literal objects that follow the predicate `skos:definition` (line 23f.), the algorithm checks for the existence of keywords (line 25ff). For each keyword, the algorithm searches for entries in the French and English Wikipedia respectively and generates DBpedia URLs as described. It then adds a triple to the lexeme with `ontolox:isConceptOf` and the DBpedia URL respectively, or generates a message in case the mapping is unsuccessful (lines 59f., 63).

Evaluation. The four methods for mapping nouns achieve varying hit rates, with the LINNÉTERMINUS approach producing different results according to the syntax of the definition text described in chap. 5.1.1. Fig. 2 shows an extract of the results in the form of the RDF triples, and fig. 3 summarizes the results achieved for the data set with 236 DEAF entries.

```

deaf:wodlark skos:definition "espèce d'oiseaux. Comme toutes
les alouettes elle appartient à la famille des Alaudidae,
alouette lulu (Lullula arborea L.)"@fr ;
ontolox:isConceptOf dbr:Woodlark .

deaf:zecharr skos:definition "espèce de faucon"@fr ;
ontolox:isConceptOf dbr:Hawk .

deaf:abenlie skos:definition "sorte de tente"@fr ;
ontolox:isConceptOf dbr:Tent .

deaf:turquet skos:definition "plante, sous-espèce de céréale,
amidonnier (Triticum turgidum L.)"@fr ;
ontolox:isConceptOf "missing english equivalent to French Wiki entry" .

deaf:pere skos:definition "père"@fr ;
ontolox:isConceptOf dbr:Father .

```

Figure 2: Result (extract) of automatic keyword search.

Lexical Units	Linné	Single Word	Last Word	GenusProximus	
				sorte de	espèce de
236 overall	86	60	60	20	10
				overall: 30	
mapped	82	37	51	18	8
no equivalence	0	5	2	1	0
no Engl. equivalence	4	6	5	0	2
not mapped	0	12	2	1	0
mapping rate	95.3%	61.7%	85%	90%	80%
correct hits	77	32	43	13	8
disambiguation pages	5	2	7	4	0
incorrect hits	0	3	1	1	0
hit rate	94%	86.5%	84.3%	72.2%	100%
mapping overall				194	
mapping rate overall				82.4%	
hits overall				173	
hit rate overall				87.4%	

Figure 3: Evaluation of the mapping of 236 entries.

Interpretation of the results and extrapolation.

The methods produce promising mapping rates and hit rates. The highest mapping rate shows the LINNÉTERMINUS method with 95.3% mappings and also a very accurate hit rate with 94%. The SINGLEWORD method achieves the lowest mapping rate with 61.7%. The highest hit rate is achieved by the GENUSPROXIMUS method with the catchword ‘espèce de’ with 100%; albeit, this result needs to be interpreted with the caveat that the absolute number of mappings for ‘espèce de’ is only eight – with 77 for the LINNÉTERMINUS method. This must also be considered for the low hit rate of (72.2%) achieved by the GENUSPROXIMUS method with the catchword ‘sorte de’. As expected, the 84.3% hit rate of the LASTWORD method is rather low for the reasons explained above.

The overall result for all four methods is a mapping rate of 82,4% (194 out of 236) with 87,4% correct hits (173).

We see that 18 mappings lead to disambiguation pages in DBpedia, a result we cannot influence. E.g., *pié* m. “pied” maps to ‘Pied_(disambiguation)’ (with proper names, the Pied Piper of Hamelin, etc.) without redirection to

‘Foot’ (the correct DBpedia entry). Encouragingly, the number of semantically incorrect hits is low, with three for the SINGLEWORD method and one for both the LASTWORD and GENUSPROXIMUS methods. E.g., *diacalamant* m. “sorte de confection dont la base était le calament” wrongly maps to ‘Sewing’ (from the polysemic French terme *confection*); however, it is a concoction using calamint, a plant of the mint family. We consider the results (mapping rate and hit rate) to be satisfactory and thus extrapolate them to the DEAF totals: out of the 92,776 lexical units, 30,065.6 are, thus, potential mappings, and – out of these – 25,423,4 are potential hits. This equals 27,4% hits overall.

5.2 A Method for Non-Nouns

This method maps lexical units of lexemes that are not nouns (but also include nouns that have not been reached by the approaches described above), i.e., adjectives, adverbs, verbs: roughly 70% of the DEAF entries. The algorithm processes keywords in the definitions that can be mapped to entities of DBpedia. This aims at grasping the significant core elements from the sense of a given lexeme. Of course, this is only an approximation to the respective sense. Nevertheless, it represents a rough but automatic placement of the sense within the structure of an external knowledge base. To do this, the algorithm applies what we call the ‘splitting method’ (SPLITTING) where it tokenizes the definition texts, iterates over the tokens, and looks for those that can be mapped. The pseudocode is the following:

```

1 IF (re.findall('\w+', o)):
2   FOR word IN (re.findall('\w+', o)):
3     SET page_py TO wiki_wiki.page(word)
4     IF page_py.exists():
5       make_langlinks(s, page_py)
6     ELSE:
7       graph.add((s, ontolex + 'isConceptOf',
8                 Literal('to be mapped')))
```

Nota bene: We apply `re.findall` instead of `re.split` to avoid having to define identification rules for split perimeters.

A model case for this method is the adjective *lovin* adj. “a la manière d’un loup” (wolflike), with the tokenized result being [`'à'`, `'la'`, `'manière'`, `'d'`, `'un'`, `'loup'`]. From these tokens, the algorithm produces:

```

1 deaf:lou#lovin
2 skos:definition "à la manière d'un loup"@fr ;
3 ontolex:isConceptOf
4   <https://dbpedia.org/resource/%C3%80>,
5   dbr:D_(disambiguation),
6   dbr:La,
7   dbr:UN_(disambiguation),
8   dbr:Wolf,
9   "no equivalents to French wikipedia entry" .
```

We can interpret the result as follows:

- ‘À’ ([%C3%80], letter) (line 4),
- ‘D_(disambiguation)’ is a disambiguation page with ‘D’ representing ‘differential equation’, ‘Delaware’, ‘Desktop Environment’, etc. (line 5),
- ‘La’ equally, representing ‘Louisiana’, ‘LucasArts’ (a subsidiary company of LucasFilm Ltd.), a type of moth, etc. (line 6),
- ‘UN_(disambiguation)’ representing ‘United Nations’, a Korean music band, etc. (line 7);
- the only mapping with semantic value is `dbr:Wolf` (line 8);
- ‘manière’ is an entry in the French Wikipedia without an equivalent in the English Wikipedia (line 9).

Evaluating a larger number of such examples, we learn that the many incorrect hits must be limited. For this purpose, we create a list of words to be generally ignored by the algorithm, i.e., articles, pronouns, prepositions, and the like. We also include words that occur in many definitions but lead to false results such as:

- *manière* (see in the example above),
- *changeant*, present participle of *changer* (to change), which maps to ‘List_of_Star_Trek_alien#Changeling’, a fictitious species of the Star-Trek universe,
- *référent*, present participle of *référer* (to refer to), which maps to ‘HTTP_referer’,
- and the adjective *sérieux* (serious) which maps to ‘Paul_Sérieux’, a French psychiatrist.

We import this list into the Python script.

Implementation. To test our method we create a data set with 100 entries: lexical units for 20 adjectives, 20 adverbs, and 20 verbs; we add 40 nouns that cannot be computed with the four methods, as described in chap. 5.1. A first test with the existing algorithm (without the SPLITTING method) confirms that all 100 entries cannot be mapped. With the algorithm using the SPLITTING method, however, the results are as shown in fig. 4.

The mapping rates of 55% up to 77.5% yield an average of 65%. We give an example of the

Lexical Units	Adv.	Adj.	Verb	Nouns	overall
number	20	20	20	40	100
mapped	12	11	11	31	65
no equivalence	1	6	7	10	24
no Engl. equivalence	6	5	5	12	28
not mapped	7	14	14	25	60
mapping rate	60%	55%	55%	77.5%	65%

Figure 4: Quantitative evaluation of SPLITTING method.

outcome for *efimere* adj. (a fever or a pain that lasts for about a day), which shows both successful mappings and a miss:

```

1 deaf:efimere skos:definition
2   "qui dure un jour ou peu plus (dit
3   de la fièvre, de la peine)"@fr ;
4   ontolex:isConceptOf
5     dbr:Day,
6     dbr:Fever,
7     "missing English equivalent to
8     French Wiki entry" .

```

Evaluation. To assess the quality of the mapping result of the SPLITTING method, we conduct an evaluation of each mapping for each lexical unit. For *efimere*, for example, the mapping to the entities ‘Day’ and ‘Fever’ are meaningful; the keyword ‘peine’ (pain) produces a result in the French Wikipedia but no English equivalent (lines 7-8).

Extrapolation to the DEAF data, all methods included. We extrapolate these results to the DEAF data. The total number of the DEAF lexical units that can be mapped by the SPLITTING method, i.e., that are not reached by the four methods LINNÉTERMINUS, SINGLEWORD, LASTWORD, and GENUSPROXIMUS (total 30,065.6, see above) is: $92,776 - 30,065.6 = 62,710.4$. With a mapping rate of overall 65% (see fig. 4), the SPLITTING method, therefore, has the potential to generate 40,761.76 mappings.

Together with the 25,423.4 semantically correct mappings of nouns, this results in an approximate amount of 66,185 semantically mapped lexical units. This corresponds to 71.34% of the total set of 92,776 lexical units.

5.3 Applying the Algorithm to the RDF Data Sets of the DEAF

As a litmus test for the validity of the extrapolation, we exclude the manually prepared test scenarios and apply the algorithm to actual RDF data: We use the results of automatic routines modeling the DEAF entries as Linked Open Data in RDF. We apply the algorithm to 300 datasets with 617 lexical

units overall, including all parts of speech. The result is a mapping rate of 71.03%. Compared with the extrapolated rate of 71.34% mapped lexical units within our test scenario, we conclude that the validity of the extrapolation is confirmed. This is important for future applications of the methods to the 92,776 lexical units of the DEAF.

Evaluation. Following the example given for *efimere* adj. (see above), we manually assess the quality of each of the 617 mappings with respect to the sense of the mapped lexical unit. Examples of the quality evaluation and the overall findings are shown in fig. 5.

DEAF entry	Def.	≠ Mapp.	Mapp.	Mapp.	Mapping overall	
			✓✓	✓	Abs.	Hit Ratio
fable	22	7	1	14	15	68.2%
faraon	3	0	1	2	3	100%
faucille	10	0	0	10	10	100%
fece	1	0	0	1	1	100%
festele	31	11	10	10	20	64.5%
festre	12	1	3	8	11	91.7%
fiel	28	6	10	12	22	78.6%
fièvre	31	0	3	29	32	100%
figure	60	24	3	33	36	60%
flajol	31	11	6	13	19	61.3%
flamesche	1	0	0	1	1	100%
flaïtte	17	6	1	10	11	64.7%
gratifier	1	1	0	0	0	0
guihale	1	0	0	1	1	100%
guimauve	1	0	1	0	1	100%
guindas	2	0	0	2	2	100%
guinlechier	2	0	0	2	2	100%
halstre	2	2	0	0	0	0
harigoter	7	2	0	5	5	71.4%
hart	35	18	0	15	15	42.9%
...
overall	617	169	77	368	445	
percentage		28.6%	12.7%	58.2%		71%

Figure 5: DEAF RDF data with LexSemMapping.

Explanation of the table columns:

- **DEAF entry:** entry name of an article,
- **Def.:** number of lexical units in the entry,
- **≠ Mapp.:** no mapping, i.e., the total amount of the messages ‘to be mapped’ respectively, ‘no equivalents to French Wiki entry’, and ‘missing English equivalent to French Wiki entry’; we also add the number of mappings that are semantically nonsense (the result of our qualitative evaluation),
- **Mapp. ✓✓:** number of semantically precise and correct mappings using the LINNÉTERMINUS, SINGLEWORD, and the LASTWORD methods,
- **Mapp. ✓:** number of the mappings through the GENUSPROXIMUS or the SPLITTING method that are semantically correct in an approximate way.

The qualitative evaluation of the mappings shows that 12,7% of the mappings produce semantically precise and correct hits, and 58,2% of the mappings produce approximately correct hits.³² The latter are able to assign the lexical units to an extra-linguistic entity in the form of a first and rough classification; at the same time, it lays an excellent foundation for a manual and more precise elaboration of the mapping for these lexical units.

6 Result and Outlook

As an overall result, we can state the following: Due to the heterogeneity of the sense definitions, achieving 100% correctness in the LexSemMapping of all 92,776 lexical units of the DEAF to DBpedia is not realistic. However, the methods we have developed (LINNÉTERMINUS, SINGLEWORD, LASTWORD, GENUSPROXIMUS, SPLITTING) clearly approach our goal: the automatic LexSemMapping of lexical units of the DEAF dictionary. Our methods are able to successfully map large portions of the total set of lexical units; approx. 71% of the lexical units (= 53,996) can be mapped: approx. 12.7% (= 11,783) will be mapped accurately in terms of semantic content, and approx. 58.2% will be mapped in an approximate, yet meaningful way.

Based on this extrapolation, we reason that applying the algorithm to the RDF data sets of the DEAF is able to enhance the RDF data in a significant way. It establishes semantics-based, language-independent access to potentially almost 65,800 lexical units of the dictionary by linking to DBpedia. The RDF data of the DEAF will be released under Public Domain in a triple store by the Heidelberg Academy of Sciences and Humanities (HAdW) or on <https://lod.academy/>, a hub for Linked Open Data and Graph Technologies run by the Academy of Sciences and Literature Mainz and the HAdW.

With the achieved result, we deduce that approx. 29% of the lexical units still need to be mapped manually. With the estimated 10 min per mapping, this still adds up to roughly 65 days of work. What comes to mind are methods utilizing artificial intelligence to interact with the sense definitions of the DEAF. Our first impression, however, was not very promising because the definition

texts seemed too heterogeneous for an AI model to identify patterns that could lay the foundation for a successful approach. Nonetheless, recent developments in this sector such as the emergence of ChatGPT³³ for instance, suggest considering the topic anew.

Furthermore, we utilized the automatic matching of French Wikipedia entries with corresponding English entries offered by the Wikipedia API. To bypass this error-prone step, it could be worthwhile to test integrating a machine-driven translation from French into English recurring to external services such as the DeepL API.³⁴

Possible generalization of the approach. Lexicographic resources typically contain lexical units—words and their senses, the latter being defined through translations into a (modern) language, through genus-differentia definitions or other methods. We know how time consuming a manual lexico-semantic mapping of the lexical units is. With (i) its specific solutions for different kinds of definitions, (ii) the possibility to feed varying languages into the algorithm (adapting the query to the Wikipedia API to the particular language) and (iii) given the hit rate of the algorithm, we conclude that a generalization of our LexSemMapping approach is promising: It can be re-used both for the semantic enhancement of already existing RDF resources and for newly approached Linked-Data modeling of (historical) linguistic resources. Also, related approaches could benefit, e.g., the aforementioned endeavor of the LEI to install an onomasiological structure and where DBpedia entities could be added to the HTE taxonomy to establish interoperability within the Linked-Data landscape.

References

- Kurt Baldinger. 1971-2020. *Dictionnaire étymologique de l'ancien français – DEAF*. Presses de L'Université Laval/Niemeyer/De Gruyter, Québec, Canada/Tübingen/Berlin, Germany. [Kurt Baldinger (founder), continued by Frankwalt Möhren and Thomas Städtler; electronic version DEAFél: <https://deaf.ub.uni-heidelberg.de>].
- Robert-Henri Bautier, Robert Auty, and Norbert Angermann. 1977-1998. *Lexikon des Mittelalters*. Artemis, München.

³²Examples of RDF data sets with mapped lexical units can also be found at GitHub: `festre_mapped.ttl`, `fiel_mapped.ttl`, etc.

³³<https://openai.com/blog/chatgpt/>.

³⁴<https://www.deepl.com/pro-api?cta=header-pro-api>.

- Sean Bechhofer, Frank van Harmelen, Jim Hendler, Ian Horrocks, Deborah L. McGuinness, Peter F. Patel-Schneider, and Lynn Andrea Stein. 2004. OWL Web Ontology Language. Reference. W3C Recommendation 10 February 2004. URL: <https://www.w3.org/TR/2004/REC-owl-ref-20040210/>.
- Andrea Bellandi, Emiliano Giovannetti, and Anja Wein-gart. 2018. Multilingual and Multiword Phenomena in a lemon Old Occitan Medico-Botanical Lexicon. *Information*, 9 (3), 52.
- Christian Bizer, Tom Heath, and Tim Berners-Lee. 2009. Linked Data – The Story So Far. *International Journal on Semantic Web and Information Systems*, 5:1–22.
- Andreas Blank. 2001. *Einführung in die lexikalische Semantik*. Niemeyer, Tübingen.
- Julia Bosque-Gil, Jorge Gracia, Elena Montiel-Ponsoda, and Asunción Gómez-Pérez. 2018. Models to represent linguistic linked data. *Natural Language Engineering*, 24(6):811–859.
- Andrea Bozzi. 2016. Un’ontologia per il DiTMAO (*Dictionnaire des Termes Médico-botaniques de l’Ancien Occitan*). In David Trotter, Andrea Bozzi, and Cédric Fairon, editors, *Actes du XXVII^e Congrès international de linguistique et de philologie romanes (Nancy, 15-20 juillet 2013). Section 16: Projets en cours; ressources et outils nouveaux*, pages 55–63. ATILF, Nancy.
- Christian Chiarcos and Maria Sukhareva. 2015. OLiA - Ontologies of Linguistic Annotation. *Semantic Web Journal*, 518:379–386.
- Philipp Cimiano, John McCrae, Paul Buitelaar, and Elena Montiel-Ponsoda. 2013. On the Role of Senses in the Ontology-Lexicon. In Alessandro Oltramari, Piek Vossen, Lu Qin, and Eduard Hovy, editors, *New Trends of Research in Ontologies and Lexical Resources. Theory and Applications of Natural Language Processing*, pages 43–62. Springer, Berlin/Heidelberg.
- Philipp Cimiano, John P. McCrae, and Paul Buitelaar. 2016. Lexicon Model for Ontologies: Final Community Group Report 10 May 2016. <https://www.w3.org/2016/05/ontolex/>.
- Thierry Declerck, Eveline Wandl-Vogt, and Karlheinz Mörrth. 2015. Towards a Pan European Lexicography by Means of Linked (Open) Data. In *Electronic Lexicography in the 21st Century: Linking Lexical Data in the Digital Age. Proceedings of eLex 2015, 11-13 August 2015, Herstmonceux Castle, United Kingdom*, pages 342–355, Ljubljana/Brighton. Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd.
- Mariafrancesca Giuliani and Itziar Molina Sangüesa. 2020. Hacia Una Taxonomía Integrada En La Redacción y Revisión De Diccionarios Históricos. *Bollettino Dell’Opera Del Vocabolario Italiano*, 25:325–374.
- Axel Herold, Lothar Lemnitzer, and Alexander Geyken. 2012. Integrating Lexical Resources Through an Aligned Lemma List. In Christian Chiarcos, editor, *Linked Data in Linguistics. Representing and Connecting Language Data and Language Metadata*, pages 35–44. Springer, Berlin / Heidelberg.
- Anas Khan, Christian Chiarcos, and Thierry Declerck et al. 2022. When linguistics meets web technologies. Recent advances in modelling linguistic linked data. *Semantic Web*, 13:1–64. DOI: [10.3233/SW-222859](https://doi.org/10.3233/SW-222859).
- Graham Klyne, Jeremy J. Carroll, and Brian McBride. 2004. Resource Description Framework (RDF): Concepts and Abstract Syntax. URL: <http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/>.
- Itziar Molina Sangüesa. 2023. Diseño de una ontología aplicada a la lexicografía histórica digital. *Círculo de Lingüística Aplicada a la Comunicación*, 93:229–242. DOI: [10.5209/clac.72654](https://doi.org/10.5209/clac.72654).
- Max Pfister. 1979–. *LEI. Lessico Etimologico Italiano*, founded by Max Pfister, directed by Elton Prifti and Wolfgang Schweickard. Reichert, Wiesbaden.
- Marielene Putscher. 1974. *Pneuma, Spiritus, Geist. Vorstellungen vom Lebensantrieb in ihren geschichtlichen Wandlungen*. Steiner, Wiesbaden.
- Willy Richard. 1959. *Untersuchungen zur Genesis der reformierten Kirchenterminologie der Westschweiz und Frankreichs: mit besonderer Berücksichtigung der Namengebung*. Francke, Bern.
- Matthew Cheung Salisbury. 2015. *The secular liturgical office in late medieval England*. Brepols, Turnhout.
- Heinrich Schipperges. 1990. *Geschichte der Medizin in Schlaglichtern*. Meyers Lexikonverlag, Mannheim.
- Sabine Tittel. forthcoming. *Integration von historischer lexikalischer Semantik und Ontologien in den Digital Humanities*. Heidelberg.
- Sabine Tittel and Christian Chiarcos. 2018. Historical Lexicography of Old French and Linked Open Data: Transforming the Resources of the *Dictionnaire étymologique de l’ancien français* with OntoLex-Lemon. In *Proceedings of LREC 2018. GLOBALEX Workshop (GLOBALEX-2018), Miyazaki, Japan, 2018*, pages 58–66, Paris. ELRA.

**Digital Humanities and
Under-Resourced
Languages**

Linking the Computational Historical Semantics corpus to the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin

Giulia Pedonese and Flavio Massimiliano Cecchini and Marco Passarotti

Università Cattolica del Sacro Cuore, Italy

giulia.pedonese@unicatt.it

flavio.cecchini@unicatt.it

marco.passarotti@unicatt.it

Abstract

This paper describes the linking of a subset of five texts from the Latin Text Archive corpus of the Computational Historical Semantics project to the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin for a total of about one million tokens, adding approximately 13 million and 750 thousand new triples to the Knowledge Base. To show the potentialities of linking those texts to other resources for Latin, the paper describes the results of a sample query conducted on the texts linked to the Knowledge Base.

1 Introduction and related work

Thanks to its key role in accessing the European cultural heritage, Latin was one of the first languages to be automatically processed. Since the pioneering work of the late Fr. Roberto Busa SJ on Thomas Aquinas' texts in 1949 (Nyhan and Passarotti, 2019), an abundance of linguistic resources has been made available for Latin as a result of a long tradition of studies in the area of Computational Linguistics, Literary Computing and Digital Humanities. These include textual resources such as corpora featuring texts of various typologies, as well as lexical resources such as lexica, dictionaries and thesauri. Besides larger (meta)collections of texts such as the *Corpus Corporum*,¹ which contains more than 150 million words provided by more than twenty different collections, among the corpora providing more specific data there are, for example, the *Patrologia Latina* data base,² featuring the writings of the Church Fathers, and the *Musisque Deoque* digital archive, which contains poetic works from Classical to Late Latin.³ Lexical resources include the *Thesaurus Linguae Latinae* at the Bayerische Akademie der Wissenschaften

in Munich,⁴ Johann Ramminger's *Neulateinische Wortliste*,⁵ and Lewis and Short's dictionary (Lewis and Short, 1879), accessible among others through the Perseus Digital Library and now linked to the LiLa Knowledge Base (Mambrini et al., 2021).

Unfortunately, while there is a large number of linguistic resources for Latin currently available in digital format, these often lie scattered in isolated "data silos", a fact which prevents users from exploiting their full potential in interoperable ways: linguistic data and metadata for Latin are distributed in separate collections which often use different data formats, query languages, annotation criteria and tagsets, thus making the resources incompatible with each other. In the last decade, multiple efforts have been made to provide a solution to the problem of dispersion of (meta)data and resource isolation. Today, many initiatives offer a single access point to resources collected in single repositories, such as the European infrastructure CLARIN,⁶ the metadictionary *Logeion*,⁷ and the already mentioned metacollection *Corpus Corporum*. However, such initiatives still fail to provide real interoperability between distributed linguistic resources, which would require "that all types of annotation applied to a particular word/text be integrated into a common representation for indiscriminate access to any linguistic information provided by a resource or tool" (Chiarcos, 2012a, p. 162). A current approach to interlinking linguistic resources is that of the Linguistic Linked Open Data cloud, a collaborative effort pursued by several members of the Open Linguistics Working Group⁸ with the goal of applying the Linked Data principles to linguistic data.⁹

⁴<https://tll.degruyter.com/>

⁵<http://nlw.renaissancestudier.org/>

⁶<https://www.clarin.eu/>

⁷<https://logeion.uchicago.edu/>

⁸<http://linguistic-lod.org/llod-cloud>

⁹Among the initiatives combining the Linked Data technologies and language resources is the COST action *Nexus Lin-*

¹<https://www.mlat.uzh.ch/>

²<https://www.lib.uchicago.edu/efts/PLD/>

³<https://mizar.unive.it/mqdg/public/>

The Linked Data paradigm consists of a series of best practices and principles for exposing, sharing and connecting data on the web, which are incarnated by the following rules:¹⁰

- data and metadata should be unequivocally named by URIs (Uniform Resource Identifiers), allowing users to find them;
- HTTP URIs should be used in order for data to be accessible by both humans and machines;
- provide useful information through Web standards such as the RDF data model (i. e. Resource Description Framework), which represents data in the form of triples: a predicate property (1) connecting a resource called subject (2) to another resource, called object (3). In this way, data are represented through directed, labelled graphs and are searchable via another Web standard like the SPARQL query language (the language used to query data in RDF format);
- include links to other URIs in order to allow for further research.

Applying the Linked Data paradigm is a way to share data according to the FAIR principles, which state that data must be Findable, Accessible, Interoperable and Reusable (Wilkinson et al., 2016). The LiLa Knowledge Base of linguistic resources for Latin aims to make textual and lexical resources interoperable through the application of the Linked Data principles (see Section 2).

After introducing the architecture of the LiLa Knowledge Base (Section 2) and the Computational Historical Semantics project (Section 3), this paper describes the linking to LiLa of a textual resource consisting of Medieval documentary Latin texts taken from the Latin Text Archive of the Computational Historical Semantics project (Section 4). Finally, the paper provides an example of query to show the potentialities of interlinking those texts to other resources for Latin (Section 5) and gives insights into the future developments of LiLa (Section 6).

guarum, whose aim “is to promote synergies across Europe between linguists, computer scientists, terminologists, and other stakeholders in industry and society, in order to investigate and extend the area of linguistic data science” (at <https://nexuslinguarum.eu/the-action/>, *What the Action does*).

¹⁰<https://www.w3.org/DesignIssues/LinkedData>

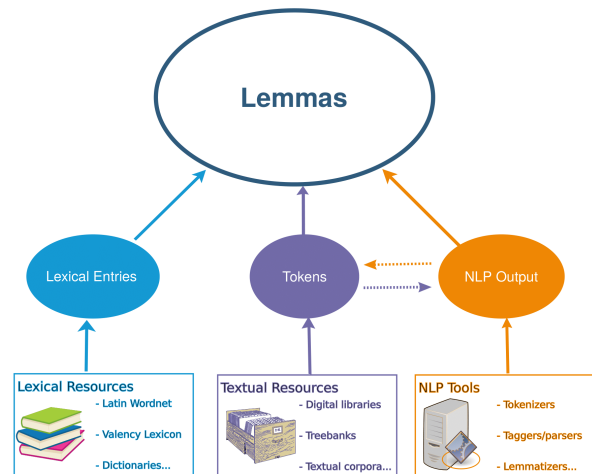


Figure 1: The architecture of the LiLa Knowledge Base.

2 The LiLa Knowledge Base

The *LiLa - Linking Latin* project¹¹ aims to connect the existing linguistic resources for Latin in order to make them interoperable (Passarotti et al., 2020). The LiLa team is building an open-ended Knowledge Base following a set of standards for the Semantic Web and Linked Data. To this end, all content involved or referenced in the linguistic resources connected in LiLa is made unambiguously findable and accessible by assigning each data point an HTTP URI. Data reusability and interoperability between resources are achieved by establishing links between different URIs and by using web standards such as the RDF data model (see Section 1) and the SPARQL query language.¹² Furthermore, the LiLa Knowledge Base makes reference to classes and properties of already existing ontologies in order to model relevant information. The main ones are: POWLA for corpus data (Chiarcos, 2012b), OLiA for linguistic annotation (Chiarcos and Sukhareva, 2015), and Ontolex-Lemon for lexical data (Buitelaar et al., 2011; McCrae et al., 2017).

Within this framework, LiLa uses the lemma as the most productive interface between lexical resources, annotated corpora and Natural Language Processing (NLP) tools. Consequently, the architecture of the LiLa Knowledge Base is highly lexically-based (cf. Figure 1), being grounded on a simple but effective assumption that strikes a good balance between feasibility and granularity: Tex-

¹¹<https://lila-erc.eu/>

¹²LiLa’s SPARQL endpoint can be accessed at: <https://lila-erc.eu/sparql/>

tual resources are made of (occurrences of) words (more precisely, *tokens*), lexical resources describe properties of words (in *lexical entries*), and NLP tools process words (producing *NLP outputs*).¹³

Considering the central role played by lemmas in LiLa, the core of the knowledge base is the so-called *Lemma Bank*,¹⁴ a collection of about 200 000 Latin lemmas (defined as the canonical forms of lexical items, i. e. their citation forms) originally taken from the data base of the morphological analyzer LEMLAT (Passarotti et al., 2017). Interoperability is achieved by linking all those entries in lexical resources and tokens in corpora that point to the same lemma. The resources currently linked to the knowledge base are as follows:

- **Textual resources**
 - [Computational Historical Semantics](#): 1 058 084 tokens
 - [Confessiones](#): 92 351 tokens
 - [Corpus for Latin Sociolinguistic Studies on Epigraphic texts](#): 32 473 tokens
 - [Index Thomisticus Treebank](#): 450 515 tokens
 - [LASLA corpus](#): 1 839 373 tokens
 - [Liber Abbaci](#) (ch. VIII): 29 858 tokens
 - [Querolus sive Aulularia](#): 13 232 tokens
 - [UDante Treebank](#): 55 287 tokens
- **Lexical resources**
 - [Lemma Bank](#): 153 965 entries
 - [Etymological Dictionary of Latin and the other Italic Languages](#): 1 452 entries
 - [Glossary of Latin loanwords from the Italian works of Dante Alighieri](#): 765 entries
 - [Index Graecorum Vocabulorum in Linguam Latinam Translatorum](#): 1 759 entries
 - [LatinAffectus](#): 3 295 entries
 - [Latin Vallex 2.0](#): 3 561 entries
 - [Latin WordNet](#): 6 269 entries
 - [Lewis & Short's dictionary](#): 53 437 entries
 - [Word Formation Latin](#): 41 791 entries

As shown in Section 3, the subset of the Computational Historical Semantic corpus adds a sig-

¹³In Figure 1, the arrows going from and to the node for *NLP Output* represent the fact that tokens that are the outputs of a specific NLP tool (a tokeniser) can become the inputs of further tools (like, for instance, a syntactic parser).

¹⁴<http://lila-erc.eu/lodview/data/id/lemma/LemmaBank>

nificant amount of Late and Medieval Latin texts, expanding the possibilities of integrated research with other (Medieval) Latin corpora such as the Index Thomisticus Treebank and UDante.

3 Computational Historical Semantics

Computational Historical Semantics (from now on CompHistSem) is a co-operative project involving the German universities of Bielefeld, Frankfurt am Main, Regensburg and Tübingen, originally developed by an interdisciplinary team led by Bernhard Jussen and Alexander Mehler at the Goethe University in Frankfurt am Main, and funded by the German Federal Ministry for Education and Research.¹⁵ The project aims to define new methods and tools for historical-semantic analysis “by conducting computer-based research on processes of linguistic change” (Cimino et al., 2015).

The associated website¹⁶ of the *Latin Text Archive* (LTA), hosted by the Berlin-Brandenburg Academy of Sciences and Humanities, allows users to simplify their search for semantic and linguistic changes by quickly comparing a large number of texts gathered from various sources: more than 4 000 texts spanning from the 2nd to the 15th Century AD, put together thanks to the support of digitalised collections such as the *Patrologia Latina* data base, the *Monumenta Germaniae Historica* (MGH),¹⁷ the *Corpus Corporum* (University of Zürich) and the *Bibliotheca Augustana*.¹⁸ These texts are lemmatised by means of the Frankfurt Latin Lexicon (FLL), a morphological lexicon of Medieval Latin organised around three “lexical resolutions” of lexical units (Mehler et al., 2020) which enable a multilayered search:

1. the *superlemma*, providing a unified representation for different variants of a “word” (i. e. a lexeme), e. g. *caelum* ‘sky’, as opposed to
2. *lemmas*, which are tied to specific variants of a word, e. g. *cael*, *caelum*, *caelum*, *caelus*, *celum*, *celum*, *celus*, *coelum*, *caelum*, *coelus*, each with its own spelling and possibly inflected according to different paradigms, which consist of

¹⁵<https://comphistsem.org/home.html>. NB: this site is no longer maintained.

¹⁶<https://lta.bbaw.de/>

¹⁷<https://www.mgh.de/>

¹⁸<http://www.hs-augsburg.de/~harsch/augustana.html>

3. *word forms*, such as *cęlorvm* (lemma *cęlum*) or *coelos* (lemma *coelus*), possibly tagged for morphological features such as *casus* (case) or *numerus* (number).

While the FLL allows a user to search for a specific word or word form and obtain quantitative data with respect to its occurrences as well as grammatical, linguistic and lexical information about its use, the textual data base LTA makes it possible to perform a text-based search of the whole corpus, and is useful to carry out more complex searches for word co-occurrences (Cimino et al., 2015). Since CompHistSem is an ongoing project, it is constantly expanding as more texts, words and word forms are added to its data bases (Mehler et al., 2020).

4 Linking CompHistSem to LiLa

In this section, the process adopted so as to link texts from the CompHistSem project to the LiLa Knowledge Base is detailed: first in general, and then by giving a more in-depth discussion of problematic cases.

4.1 Texts, annotation and format conversion

The linking procedure is implemented on a subset of the LTA corpus of CompHistSem consisting of seven texts or text collections. These are the texts that have been selected by the CompHistSem team after having been requested for data from their corpus to include into LiLa, and that have been deemed of sufficient size for this goal. The specific documents are:

- *Capitularia Regum Francorum*, 6th–9th c. AD, various authors, from MGH Capitularia 1 & 2
 - 10 820 sentences,¹⁹ 343 030 tokens (including 53 161 punctuation marks)
- *De ecclesiasticis officiis*, 9th c. AD, by Amalarius of Metz, from Patrologia Latina vol. 105
 - 4 279 sentences, 125 475 tokens (including 20 845 punctuation marks)
- *Vita Karoli Imperatoris*, 9th c. AD, by Eginhard, from MGH Scriptorum rerum Germanicarum 25

- 247 sentences, 8 393 tokens (including 1 224 punctuation marks)

- *Gesta Hludowici imperatoris*, 9th c. AD, by Thegan of Trier, from MGH Scriptorum rerum Germanicarum 64
 - 451 sentences, 8 355 tokens (including 1 403 punctuation marks)
- *Decretum Gratiani* I to III (treated as distinct documents), also known as *Concordia discordantium canonum*, 12th c. AD, by Gratian, from Corpus Corporum through Patrologia Latina vol. 187
 - 31 803 sentences, 572 831 tokens (including 124 656 punctuation marks)

In total, there are 47 600 sentences for 1 058 084 tokens (including 201 289 punctuation marks), the vast majority of which (see Section 4.2) lemmatised and tagged for parts of speech and morphological features by means of the Frankfurt Latin Lexicon (see Section 3), which uses its own tagset, in line with the grammatical categories traditionally recognised for Latin.²⁰ All texts but the *Decretum Gratiani* (Corpus Corporum, transcription under Creative Commons Share-Alike license²¹) are retrievable from the LTA (see Section 3) and are under the Creative Commons license.²² The texts are encoded in the TEI-P5 format, i. e. as XMLs.²³

The preliminary step before linkage is the conversion of the XMLs to the CoNLL-U format,²⁴ as used in the Universal Dependencies (UD) project (de Marneffe et al., 2021), by means of a Python²⁵ script developed as part of the LiLa project's endeavour.²⁶ The motivation for this move is twofold: first, the CoNLL-U format is more easily human-readable, with no loss of information nor of machine-readability with respect to the original XML; second, the conversion of format also entails a conversion of part-of-speech and morphological tags, similarly to what has already been achieved for other data sets, such as the Index Thomisticus Treebank (Cecchini et al., 2018) or the Late Latin

¹⁹“Sentence” in this context refers to the textual segmentation inherited from CompHistSem, and does not necessarily coincide with a syntactically-driven interpretation thereof; this however is irrelevant here, as only single tokens are considered.

²⁰A classic and accessible reference for Latin is (Greenough et al., 2014).

²¹<https://creativecommons.org/licenses/by-sa/4.0/>

²²<https://creativecommons.org/licenses/by/4.0/>

²³<https://tei-c.org/>

²⁴<https://universaldependencies.org/format.html>

²⁵www.python.org

²⁶The script has not yet been made public.

Charter Treebank (Cecchini et al., 2020a). The latter point is relevant, since also LiLa makes use of UD’s part-of-speech tagset internally, and so the conversion to the CoNLL-U format has the ultimate effect of better integrating CompHistSem texts into the knowledge base and of laying the ground for its linking, at the same time acting as a stepping stone towards a possible future annotation according to UD guidelines.

The mapping between the two tagsets is rather straightforward, especially with regard to morphological tags, whose distribution already broadly corresponds to that found in the UD formalism applied to Latin, or can be implemented on a lexical basis. Parts of speech also overlap or are retraceable to more general classes (e. g. CompHistSem’s distributives `DIST` and ordinals `ORD` merge into UD’s adjectives `ADJ` with a corresponding value of the `NumType` feature²⁷) to a great degree, since they have common roots in traditional grammars, but need some further reworking: in particular, the class of determiners (in UD labeled as `DET`) has to be carved out from CompHistSem’s adjectives (`ADJ`) and pronouns (`PRO`); a difference has to be drawn, on a lexical basis, between co-ordinating (`CCONJ` in UD) and subordinating (`SCONJ`) conjunctions; some readjustments between indeclinable classes (especially adverbs, `ADV` in UD; conjunctions, `CCONJ/SCONJ`; particles, `PART`) are necessary; and tokens with atypical lemmas such as *biblical books* and/or belonging to mixed nominal or residual classes (`Noun`, `NE`, `NP`, `PTC`, `XY`, `FM` in CompHistSem) require some case-by-case treatment.

4.2 Lemmatisation

Since LiLa is structured around the notion of lemma (see Section 2), which is the key element through which lexical and textual resources are connected to the knowledge base, lemmatisation of a document is a necessary step in order to proceed with the linking process. As mentioned in Section 4.1, this is already the case for texts found in the LTA: the `LEMMA` field in the CoNLL-U conversion (see Section 4.1) directly stores the *superlemma* relative to the word form, as determined per the Frankfurt Latin Lexicon (see Section 3).

Only a negligible 2 697 tokens lacking a lemma

²⁷We point to UD guidelines, which can be browsed at <https://universaldependencies.org/guidelines.html>, for details about the meaning of labels in the UD framework.

are detected, i. e. the 0,25% of the total, for which the Frankfurt Latin Lexicon fails to produce one. They represent 1 775 (case-sensitive) form types, and mostly consist of proper nouns, or terms derived from proper nouns (hence conventionally capitalised), such as *Magonciam* ‘Mainz (city in Germany)’, variant of a more Classical *Mogontiacum*, or *Tolletano* ‘from Toledo (city in Spain, *Toletum* in Latin)’, but also forms such as *f* or *ff*. Given the peculiar, onomatological nature and marginality of such forms, and the fact that in this phase the focus is on linking and not on expanding LiLa’s lexical data base, these tokens are not considered further and left out from lemmatisation (and thus linking).

More in general, it has to be noticed that the data from CompHistSem, as that of any other external resource, is taken ‘as is’: it is not the goal nor the scope of this work to assess the “correctness” of any level of its annotation (tokenisation, lemmatisation, part-of-speech-tagging, morphological features). The aim here is only to link different resources to the LiLa Knowledge Base, without intervening in their annotation standards: this means that no evaluation is performed, nor can be, as LiLa itself avoids establishing a standard. However, the interoperability of many different resources can surely help achieve an overview of the variations between annotation formalisms, in view of a possible harmonisation of their criteria, e. g. in a typological framework (cf. Gamba and Zeman 2023).

4.3 Matching and non-matching tokens

Even if no evaluation in a true sense can be performed, the complexity of the linking task can be gauged by looking at the different cases that present themselves and at the strategies that are necessary to deal with them, and how they are distributed among the tokens. First and foremost, the trivial case of punctuation marks is ignored: besides being invariably assigned a lemma identical to their form and part of speech `PUNCT`, and thus not presenting any ambiguity, punctuation marks are not lexical units, and as such do not even appear in the LiLa lemma bank. This brings it down to 856 795 “lexical tokens”²⁸ that can be contemplated for linking from the original total of 1 058 084. In the following, a breakdown of the outcomes of the linking

²⁸“Lexical” in the sense of corresponding to what is usually considered to be a word (with all its indefiniteness, cf. Haspelmath 2017), not necessarily as in the lexical/functional dichotomy of UD (see de Marneffe et al., 2021, §2.1.1).

process is given, at the end of which approximately 13 million 750 thousand new triples are added to the LiLa Knowledge Base.

4.3.1 Unambiguous matches

As many as 720 860 of these lexical tokens can be directly linked to the LiLa knowledge base through an unambiguous match in the LiLa lemma bank with their respective combinations of lemma and part of speech (after conversion, see Section 4.1): an example is the lemma *itinerarium* ‘itinerary’ coupled with the part of speech NOUN, a combination which exists and is unique in LiLa.²⁹ It has to be remarked that such a match is independent from the specific word form: this is the advantage of pivoting on the (super)lemma, as it abstracts from not always predictable spelling and inflection variants. The total coverage of direct linking is thus the 84,14% of all tokens; if only the number, 18 262, of unique combinations of lemma and part of speech among lexical tokens in our subcorpus is taken into account, the coverage is instead 68,50% (12 509 combinations). This difference arises from the fact that many unambiguously linked tokens represent very frequent functional words such as the co-ordinating conjunction (CCONJ) *et* ‘and’ (33 250 occurrences) or the pronoun (PRON) *qui* ‘who, which, that’ (17 434 occurrences), while the vocabulary of the chosen texts indeed sensibly departs from the original lexical pool of the LiLa lemma bank (cf. Section 5).

Again, it has to be noticed that no upstream control is performed on the criteria or correctness of the lemmatisation in CompHistSem: all the just described unambiguous matches are inserted as they are, meaning that, in a sense, LiLa accepts the risk of picking up spurious forms.

4.3.2 Ambiguous matches

There are cases in which a token’s combination of lemma and part of speech can be matched to more than one entry in the LiLa lemma bank: in particular, this happens for 54 903 lexical tokens (corresponding to 777 lemma/part-of-speech types), e. g. for the lemma *contingo* ‘to touch’ or ‘to wet’ coupled with the part of speech VERB, for which we have three candidates.³⁰ In all these cases, each

token proceeds to be linked to all its suitable candidates, leaving the linking ambiguous. This is an acceptable compromise in the face of the relatively low incidence of such ambiguities, and of the fact that some tokens would still not be distinguishable even when taking into account all other morphological factors: e. g. for *contingo* VERB, knowing that its word form is *contingat* and that its mood is subjunctive, still one could not choose between entry 93415 or 96293 in the LiLa lemma bank. A contextual and/or semantic disambiguation would take an unnecessary effort and is outside the scope of the linking task presented here.

4.3.3 No matches

There are 81 032 lexical tokens left that cannot be retraced to any entry in the LiLa lemma bank. This can have three reasons:

1. either the token does not possess a lemma, or
2. it has a lemma unknown to LiLa, or finally
3. there is a mismatch between lemma and part of speech from the point of view of the LiLa lemma bank.

1. As discussed in Section 4.2, the first case is marginal, and those tokens are ignored.

2. The second case is exemplified by the lemma *subplantatio* (with part of speech NOUN): it is a regularly formed, if novel, Latin word for which it is possible to extract all necessary values to insert it in LiLa’s lemma bank from CompHistSem’s annotation. However, since it is not already in the lemma bank, it cannot yet be linked at this stage. The number of different types (with respect to lemma, part of speech and morphological features) of new words ready for insertion is 2 448, but if 257 with residual part of speech X (meaning they do not have a meaningful analysis from the point of view of Latin, being mostly foreign words) are discarded, together with 693 numerals expressed as digits or Roman numerals, the remaining lexical items not unexpectedly show a preponderance of 699 proper nouns (PROPN), e. g. *Teudericus*, followed by 378 adjectives (ADJ), e. g. *adrianopolitis* ‘from the city of Adrianopolis (modern-day Edirne, in Turkey)’, 257 common nouns (NOUN), e. g. *pyromantica* ‘divination by fire’ (related to the already known *pyromantia*), 45 verbs (VERB), e. g. *exonio* ‘to excuse’,³¹ 30

²⁹<https://lila-erc.eu/data/id/lemma/109142>

³⁰<http://lila-erc.eu/data/id/lemma/43870>, <http://lila-erc.eu/data/id/lemma/93415> and <http://lila-erc.eu/data/id/lemma/96293>.

³¹Cf. <http://ducange.enc.sorbonne.fr/exonia>.

adverbs (ADV), e. g. *nudiustertius* ‘now three days ago’, 6 literal numerals (NUM), e. g. *uigintiquinque* ‘twenty-five’, 3 pronouns (PRON), e. g. *nosipsi* ‘we ourselves’, 3 interjections (INTJ), e. g. *hosanna* ‘hosanna, praise’, and 2 subordinating conjunctions (SCONJ), e. g. *quamobrem* ‘for what reason’.³² A further 429 lemmas with a part of speech can be identified, e. g. the PROPN *Ebbo*, for which however morphological features are lacking, and for which therefore some research is needed before insertion/linkage. The distribution of all these missing lemmas, skewed towards names of persons and places, already gives an interesting picture of the character and provenance of the documents at hand, which is further explored at the phrase level in Section 5.

3. The third case is again split between those tokens having a unique possible match (with respect to their lemmas) with an entry in the LiLa lemma bank, and those having multiple possible matches. In both events, the misalignment with the corresponding parts of speech found in the LiLa lemma bank means that all these 2 426 lemma/part-of-speech types have to be manually checked to understand if there is a presence of false matches (which could eventually lead to new insertions in LiLa’s lemma bank), or deviating standards of annotation. The latter case is illustrated by the rather frequent (1 606 occurrences) lemma *ita* ‘thus, so’ misleadingly labelled as a conjunction in CompHistSem, while it appears as an adverb (ADV) in the LiLa lemma bank. There are some “internal” misalignments, too: the negation *non* ‘not’ (taking up alone 16,71% of all missing matches, with 13 538 occurrences) is tagged as a particle (PART) in the CoNLL-U conversion according to UD standards,³³ but is registered as an adverb (ADV) in LiLa.

Also, the morphological analyser LEMLAT³⁴ (Passarotti et al., 2017) is deployed directly on word forms to check if some annotation choices in CompHistSem, unrecognised by LiLa, do fall into the category of *hypolemmas*, i. e. a standard word form that represents a well-defined subset of the inflectional paradigm of a lemma, which under some criteria might be considered to be a lemma

³²Univerbated from the phrase *quam ob rem* and opposed to its registration as an adverb in the LiLa lemma bank.

³³<https://universaldependencies.org/u/pos/PART.html>

³⁴<http://www.lemlat3.eu/>

itself: among the most common examples are participles (see below) (Passarotti et al., 2020).³⁵ So, for example, this strategy leads to envisage LiLa’s entry of the adjective (ADJ) *caelestis*³⁶ ‘heavenly’ for what in the CompHistSem’s texts is labelled as the common noun (NOUN) with lemma *caeleste*, i. e. the substantivised neutral singular form of the adjective, which would have been otherwise undetectable, as *caeleste* does not appear as an individual entry in LiLa’s lemma bank. Under this light, an example of a false match that needs to be rejected is the entry NOUN *paterium*³⁷ ‘a kind of Evangeliary’³⁸ for a possible proper noun *Paterius*: in fact, Paterius was the name of a bishop of Brescia in the 6th Century AD. Among misalignments, there are some recurring cases that can be treated systematically:

- misalignments between NOUNs and ADJs and vice versa, which mostly happen when a substantivised adjective is considered an independent lexical entry, e. g. *rapax* ‘rapacious; beast of prey’ or *togatus* ‘wearing a toga; a Roman citizen’. Since LiLa’s linking is not contextual, the final decision is to consider these two morphosyntactic categories equivalent for what concerns linking tokens to LiLa;
- misalignments between ADJs and VERBs. This is the case of nominal verb forms considered again as independent lexical entities, the same way as adjectives can be, e. g. *persequens*, so-called present participle of *persequor* ‘to follow perseveringly’, so ‘following perseveringly’ or, in a translated sense, ‘persecutory’. In LiLa, they are linked as hypolemmas of the respective main verbs.

5 Use case

To show the potentialities of interlinking a subset of texts from the LTA to the other linguistic resources in the LiLa Knowledge Base, a sample query is shown in this section. The query searches for sequences of three lemmas in the CompHistSem texts at hand (see Section 4.1), in the LASLA corpus (Fantoli et al., 2022), in the texts of the 13 books

³⁵In FLL terms, a hypolemma might be seen as an intermediate degree between lemma and word form (cf. Section 3).

³⁶<https://lila-erc.eu/data/id/lemma/92214>

³⁷<https://lila-erc.eu/data/id/lemma/69949>

³⁸<http://ducange.enc.sorbonne.fr/paterium>

of the *Confessiones* by Augustine, taken from *The Latin Library*,³⁹ in the Index Thomisticus Treebank (IT-TB), which includes texts of Thomas Aquinas (Mambrini et al., 2022), and in UDante, a syntactically annotated corpus featuring the Latin works by Dante Alighieri (Cecchini et al., 2020b). So as to better highlight their characteristics, the works in the LTA's subcorpus are considered separately (splitting parts I-III of the *Decretum Gratiani*) and the LASLA corpus is analyzed per author. This section describes the results of this query limited to token sequences with a frequency of at least 10, up to ten most frequent ones.

Figure 2 shows the text of a SPARQL query. The example in this case is limited to the UDante corpus only for reasons of space. After defining the classes and properties in the relevant ontologies (lines 1-6), the query selects a sequence of three lemmas in the UDante corpus, univocally identified by their URIs (line 11). In order to do that, for every token in the corpus the query selects the next two tokens (lines 8-16) with their respective token labels, their lemmas and lemma labels (lines 17-25). The query then proceeds to order the results by grouping the lemmas by their URIs and puts them in descending order of frequency (lines 26-28). As can be seen from the property `hasLemma` (lines 17, 19 and 21), the LiLa custom ontology provides the linking between a token in the selected corpus and its corresponding lemma in the Lemma Bank, allowing further connections with other lemmatised linguistic resources. This is a pivotal point, as LiLa provides a method to harmonise different lemmatisation criteria, granting interoperability regardless of different citation forms (e. g. *claudel/claudel/claudor* 'to limp', all tied to different inflectional paradigms) and/or different written representations (e. g. *sanctus/sancitus* 'saint', originally a participial form of *sancio* 'to establish') of the same lexical item used in specific linguistic resources.⁴⁰ The lemma sequences discussed in this section are quoted in small caps and

³⁹<http://lila-erc.eu/data/corpora/CIRCSELatinLibrary/id/corpus/Confessiones>

⁴⁰In the case of different citation forms of the same item belonging to two inflectional categories, e. g. *sequo/sequor* 'to follow' (alternating with respect to morphological active/passive voice), they are considered as two separate lemmas connected via the 'lemma variant' property; if not, e. g. *causal/caussa/kausalkaussa* 'cause' (all inflecting according to the same nominal paradigm, the so-called "first declension"), they are considered as two written representations of the same lemma; see (Passarotti et al., 2020).

glossed in lowercase translated lemmas, while the examples of textual occurrences are in italics.⁴¹

The first distinction to be made is that between lemma sequences which are merely grammatical, i. e. sequences composed only of function words such as *DE HIC QUI* 'from this who' or *EX IS QUI* 'out-of he who', and sequences with a lexical meaning. The former kind of sequence is quite common among all the works we consider and depends on the language in question, i. e. Latin, and, more in general, on the known Zipfian distribution of words (cf. Newman 2005, §2.1), while the latter is specific to the era and type of each single work.

Considering lexically meaningful sentences, the texts from LTA include sequences which correspond to sentences typical of ecclesiastical language. This is the case with sequences specific to ecclesiastical institutions such as *SANCITUS DEUS ECCLESIA* 'saint god church', *SANCITUS ROMANUS ECCLESIA* 'saint roman church': see for example the expressions *sanctae Dei ecclesiae* 'of/to the Holy Church of God', which is also the most frequent sequence of 3 tokens in the *Capitularia Regum Francorum*, and *sanctae Romanae ecclesiae* 'of/to the Holy Roman Church' in the *Decretum Gratiani* I. Other lemma sequences of this kind are *ITEM EX CONCILIUM* 'also out-of council' and *EX CONCILIUM CARTHAGINENSIS* 'out-of council carthaginian': see for example *item ex Concilio* 'moreover, from the Council' and *ex Concilio Cartaginensi* 'from the Council of Carthago' which occur in the *Decretum Gratiani* I-III. Some other sequences can be considered ecclesiastical insofar as they refer to Christian Latin and liturgy, such as *NOSTER IESUS CHRISTUS* 'our jesus christ', *IN EXCELSUM DEUS* 'in loftiness god', *PANIS ET UINUM* 'bread and wine', *CORPUS ET SANGUIS* 'body and blood' and *DOMINUS NOSTER IESUS* 'lord our jesus': see for example *domini nostri Iesu* 'to our Lord Jesus' in the *Capitularia Regum Francorum*, *in excelsis Deo* 'to God in the highest' in the *De ecclesiasticis officiis* and *panem et uinum* 'bread and wine (accusative case)', *corpus et sanguinem* 'body and blood (accusative case)' and *Dominus noster Iesus* 'our Lord Jesus' in the *Decretum Gratiani* III.

Noting that the most frequently used sequences of tokens in the subset of texts from LTA are *sanc-*

⁴¹While it is not possible to show all data and tables discussed here for lack of space, they are accessible from a dedicated online repository at <https://github.com/CIRCSE/Linking-Computational-Historical-Semantics>.

```

1 PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
2 PREFIX lila: <http://lila-erc.eu/ontologies/lila/>
3 PREFIX dc: <http://purl.org/dc/elements/1.1/>
4 PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
5 PREFIX powla: <http://purl.org/powla/powla.owl#>
6
7 # get trigram of a corpus
8 SELECT ?lemmaLabel0 ?lemmaLabel1 ?lemmaLabel2 (count(?corpora) as ?chainCount) ?corpusTitle WHERE {
9
10 VALUES ?corpora {
11   <http://lila-erc.eu/data/corpora/UDante/id/corpus>
12 }
13 ?corpora dc:title ?corpusTitle.
14 ?token0 powla:hasLayer/powla:hasDocument/^powla:hasSubDocument ?corpora .
15 ?token0 powla:next ?token1.
16 ?token1 powla:next ?token2.
17 ?token0 lila:hasLemma ?lemmaToken0;
18   powla:hasStringValue ?tokenString0.
19 ?token1 lila:hasLemma ?lemmaToken1;
20   powla:hasStringValue ?tokenString1.
21 ?token2 lila:hasLemma ?lemmaToken2;
22   powla:hasStringValue ?tokenString2.
23 ?lemmaToken0 rdfs:label ?lemmaLabel0.
24 ?lemmaToken1 rdfs:label ?lemmaLabel1.
25 ?lemmaToken2 rdfs:label ?lemmaLabel2.
26
27 } group by ?lemmaToken0 ?lemmaToken1 ?lemmaToken2 ?lemmaLabel0 ?lemmaLabel1 ?lemmaLabel2 ?corpusTitle
28 order by DESC (?chainCount)

```

Figure 2: Sample query applied to the UDante corpus.

tae Dei ecclesiae and *sanctae Romanae ecclesiae*, one could, if interested in the use of *sanctus* ‘saint’ in Ecclesiastical Latin, further refine this search with another query to retrieve all the different written representations of the so-called perfect participle of *sancio* ‘to establish’, of which *sanctus* is a form. In the LiLa Lemma Bank, *sanctus* and its three other possible written representations *sancitus*, *santus* and *xantus* are represented as hypolemmas connected to the lemma *sancio* (cf. [Passarotti et al. 2020](#)). In this way, whether in a lemmatised corpus a form like *sanctae* is assigned, for example, the lemma *sancio*, *sancitus* or *sanctus*, in LiLa this lemma is always connected to the same lemma *sancio* and is thus retrievable with a single query. In the specific corpus at hand, this query retrieves 12 participial forms lemmatised under *sancitus*, and 2785 under *sanctus*: this is a novelty with regards to Classical Latin.

The sequences in the LASLA corpus show a high variety depending on the author. Limiting the data to the sequences of 3 lemmas with frequency greater than 10, the selection includes Caesar, Catullus, Cicero, Seneca and Tacitus. While Caesar is more likely to use strings of lemmas related to spatial descriptions and military events such as AD CAESAR MITTO ‘to caesar send’, SUI IN CASTRA ‘self in camp’ and EX OMNIS PARS ‘out-of all part’, the majority of the lemma sequences in Catullus are almost exclusively due to the long and repetitive hymns to Hymenaeus traditionally sung at weddings. Even though the most frequent strings of

lemmas in Cicero are mostly due to argumentative purposes (such as UT IS QUI ‘as he who’ or HAUD SCIO AN ‘not know whether’), there are plenty of sequences including typical Republican words such as POPULUS ‘people/nation’: see for example the sequence POPULUS QUE ROMANUS ‘people and roman’, which is the only one included in the first 10 most frequent examples, even though other three-lemma sequences such as POPULUS ROMANUS SUM ‘people roman be’, A POPULUS ROMANUS ‘from people roman’ and DE PECUNIA REPETO ‘from money fetch’ refer to institutions and laws of the Roman Republic and have frequency greater than 30.

As for a Christian text like the *Confessiones* by Augustine, even though a generic similarity is due to Christian Latin (see for example the expression DOMINUS DEUS MEUS ‘lord god my’), the *Confessiones* are not an ecclesiastical treatise nor a documentary text, but rather a philosophical text based on personal experiences. According to that, its lemma sequences tend to show a peculiar reference to cosmological order (CALEUM ET TERRA ‘sky and earth’, IN HIC MUNDUS ‘in this world’) and introspection (IN COR MEUS ‘in heart my’, IN MEMORIA MEUS ‘in memory my’).

Thomas Aquinas’ *Summa contra gentiles* and the Latin works by Dante Alighieri offer a good example of Medieval Latin from the 13th and 14th centuries. However, the sequences in the *Summa contra gentiles* tend to be due to logic argumentation (SUPRA OSTENDO SUM ‘above display be’,

UT SUPRA OSTENDO ‘as above display’, UT OSTENDO SUM ‘as display be’) according to the rigid exposition of philosophical and theological matters in the Scholastic tradition. The same can be observed in Dante Alighieri’s works, where the first 10 lemma sequences are logical sequences useful for speech coherence, as previously observed in Thomas Aquinas’ work (ET PER CONSEQUENS ‘and for consequence’, UT SUPRA DICO ‘as above say’, PATEO EX PRIMUS ‘appear out-of-first’) except for a broader reference to the universe (CAELUM ET MUNDUS ‘sky and world’) similar to the CAELUM ET TERRA ‘sky and earth’ already seen in Augustine and which in Dante is probably a rhetorical device.

These example queries show that the LiLa Knowledge Base makes it possible to extract large quantities of linguistic data (in this case of lexico-textual kind) from several corpora with a single query, covering different eras and genres. This is important when dealing with a language such as Latin, which has a remarkable diachronic and diatopic spread. LiLa also allows for further integrated research with lexical resources such as the *Etymological Dictionary of Latin and the other Italic Languages* (de Vaan, 2008), a valency lexicon (Passarotti et al., 2016), or the prior polarity lexicon of Latin Lemmas *Latin Affectus* (Sprugnoli et al., 2020); see Section 2. In such an interoperable environment, the addition of new resources to the knowledge base allows LiLa to expand its lexical coverage and multiplies the possibilities of connections among (meta)data.

6 Conclusion and Future Work

This paper details the process of linking a subset of the Latin Text Archive, part of the Computational Historical Semantics project, to the LiLa Knowledge Base. This work is part of a wider project which aims to make several linguistic resources for Latin interoperable through LiLa. After years spent building the large collection of lemmas used to interlink distributed resources for Latin, LiLa is now in the phase of exploiting the (meta)data provided by the already available resources to make them interact, assuming that the whole is greater than the sum of its parts.

In such respect, Latin represents a perfect use case where procedures for making linguistic resources interoperable can be developed and tested. Indeed, the history of Latin spans across more than

two millennia, showing a wide diversity in terms of genres and provenance of its texts. Moreover, with just a few exceptions, Latin is a dead language, thus making it possible to plan to interlink its entire collection of texts in the (hopefully near) future. Also, the large and diverse community of scholars working on the Latin language, including linguists, philologists, historians and archaeologists, is strictly bound to the empirical evidence provided by Latin texts, as one of the most important sources of information in support of their research work: providing such community with a means to access, query, publish and collect (meta)data from several corpora and lexical resources is a long-time *desideratum* that is finally becoming possible.

In the near future, the *LiLa - Linking Latin* project plans to interlink a number of Latin corpora, including *Musisque Deoque* (Manca et al., 2011), *CRoALa* (Jovanović, 2012), the *Late Latin Charter Treebank* (Korkiakangas, 2021) and the PROIEL treebank (Eckhoff et al., 2018). In the long run, based on the experience of linking a subset of the Computational Historical Semantics corpus, the aim is to link the entire collection of texts provided by the Latin Text Archive to the LiLa Knowledge Base. Given the size and the diversity of the texts therein, this would represent a terrific achievement and advancement for both the communities of Classics and Computational Linguistics.

However, the foundations of LiLa Knowledge Base are built on open and shared formats, models and vocabularies, both to make the resources for Latin interact with each other as well as with those for other languages, and to address the condition of openness that is strictly related to the Linked Data paradigm. Not only are the resources interlinked in LiLa supposed to be openly accessible and downloadable (as the saying goes, “as open as possible, as closed as necessary”), but interlinking the resources is an open process, too. In the Linked Open Data world, everyone is free to add new links between resources: this makes LiLa an open-ended knowledge base, which represents the best venue where to publish the digital linguistic resources, in order to set them free from their storage in separate “silos”, by making them finally interact. This is the hope of this project: that over the coming years LiLa will grow more and more thanks to the community of developers and providers of linguistic (meta)data for Latin and beyond.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme – Grant Agreement No. 769994.

References

- Paul Buitelaar, Philipp Cimiano, John McCrae, Elena Montiel-Ponsoda, and Thierry Declerck. 2011. **Ontology Lexicalisation: The lemon Perspective**. In *9th International Conference on Terminology and Artificial Intelligence (TIA 11) – Proceedings of the Workshops*, pages 33–36, Paris, France.
- Flavio Massimiliano Cecchini, Timo Korkiakangas, and Marco Passarotti. 2020a. **A New Latin Treebank for Universal Dependencies: Charters between Ancient Latin and Romance Languages**. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, pages 933–942, Marseille, France. European Language Resources Association (ELRA).
- Flavio Massimiliano Cecchini, Marco Passarotti, Paola Marongiu, and Daniel Zeman. 2018. **Challenges in Converting the Index Thomisticus Treebank into Universal Dependencies**. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 27–36, Brussels, Belgium. Association for Computational Linguistics.
- Flavio Massimiliano Cecchini, Rachele Sprugnoli, Giovanni Moretti, and Marco Passarotti. 2020b. **UDante: First Steps Towards the Universal Dependencies Treebank of Dante's Latin Works**. In *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020, Bologna, Italy, March 1–3 2021)*, pages 99–105, Turin, Italy. Associazione italiana di linguistica computazionale (AILC), Accademia University Press.
- Christian Chiarcos. 2012a. **Interoperability of Corpora and Annotations**. In Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann, editors, *Linked Data in Linguistics: Representing and Connecting Language Data and Language Metadata*, pages 161–179. Springer, Berlin/Heidelberg, Germany.
- Christian Chiarcos. 2012b. **POWLA: Modeling Linguistic Corpora in OWL/DL**. In *The Semantic Web: Research and Applications. 9th Extended Semantic Web Conference, ESWC 2012, Heraklion, Crete, Greece, May 27–31, 2012, Proceedings*, number 7295 in Lecture Notes in Computer Science, pages 225–239, Berlin/Heidelberg, Germany. Springer.
- Christian Chiarcos and Maria Sukhareva. 2015. **OLiA – Ontologies of Linguistic Annotation**. *Semantic Web*, 6(4):379–386.
- Roberta Cimino, Tim Geelhaar, and Silke Schwandt. 2015. **Digital Approaches to Historical Semantics: New Research Directions at Frankfurt University**. *Storicamente*, 11(7).
- Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. **Universal Dependencies**. *Computational Linguistics*, 47(2):255–308.
- Michiel de Vaan. 2008. *Etymological Dictionary of Latin and the other Italic Languages*. Number 7 in Leiden Indo-European Etymological Dictionary Series. Brill, Leiden, Netherlands; Boston, MA, USA.
- Hanne Martine Eckhoff, Kristin Bech, Gerlof Bouma, Kristine Eide, Dag Haug, Odd Einar Haugen, and Marius Jøhndal. 2018. **The PROIEL treebank family: a standard for early attestations of Indo-European languages**. *Language Resources and Evaluation*, 52(1):29–65.
- Margherita Fantoli, Marco Carlo Passarotti, Eleonora Maria Litta, Paolo Ruffolo, and Giovanni Moretti. 2022. **Linking LASLA corpus - LiLa LemmaBank**.
- Federica Gamba and Daniel Zeman. 2023. **Universalising Latin Universal Dependencies: a harmonisation of Latin treebanks in UD**. In *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)*, pages 7–16, Washington, D. C., USA. Association for Computational Linguistics (ACL).
- James Bradstreet Greenough, George Lyman Kittredge, Albert Andrew Howard, and Benjamin Leonard D'Ooge. 2014. *New Latin Grammar for Schools and Colleges*. Dickinson College Commentaries, Carlisle, PA, USA.
- Martin Haspelmath. 2017. **The indeterminacy of word segmentation and the nature of morphology and syntax**. *Folia Linguistica*, 51(s1000 – Jubilee Issue: 50 Years Folia Linguistica):31–80.
- Neven Jovanović. 2012. **CroALa. Enhancing a TEI-encoded Text Collection**. *Journal of the Text Encoding Initiative*, 2 (Selected Papers from the 2010 TEI Conference).
- Timo Korkiakangas. 2021. **Late Latin Charter Treebank: contents and annotation**. *Corpora*, 16(2):191–203.
- Charlton Thomas Lewis and Charles Short. 1879. *A Latin Dictionary*. Clarendon Press, Oxford, UK.
- Francesco Mambrini, Eleonora Litta, Marco Passarotti, and Paolo Ruffolo. 2021. **Linking the Lewis & Short Dictionary to the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin**. In *Proceedings of the Eighth Italian Conference on Computational Linguistics (CLiC-it 2021), Milan, Italy, June 29 – July 1, 2022*, Milan, Italy. CEUR-WS.org.
- Francesco Mambrini, Marco Passarotti, Giovanni Moretti, and Matteo Pellegrini. 2022. **The Index Thomisticus Treebank as Linked Data in the LiLa**

- Knowledge Base.** In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022)*, pages 4022–4029, Marseille, France. European Language Resources Association (ELRA).
- Massimo Manca, Linda Spinazzè, Paolo Mastandrea, Luigi Tassarolo, and Federico Boschetti. 2011. **Musisque Deoque: Text Retrieval on Critical Editions.** *Journal for Language Technology and Computational Linguistics*, 26(2 – Annotation of Corpora for Research in the Humanities: *Proceedings of the ACRH Workshop, 5. January 2012, Heidelberg University, Germany*):129–140.
- John P. McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. **The OntoLex-Lemon Model: development and applications.** In *Electronic lexicography in the 21st century: Lexicography from scratch. Proceedings of eLex 2017 conference*, pages 587–597, Leiden, the Netherlands. Lexical Computing CZ s.r.o.
- Alexander Mehler, Bernhard Jussen, Tim Geelhaar, Alexander Henlein, Giuseppe Abrami, Daniel Baumartz, Tolga Uslu, and Wahed Hemati. 2020. **The Frankfurt Latin Lexicon. From Morphological Expansion and Word Embeddings to SemioGraphs.** *Studi e Saggi Linguistici*, LVIII(1):121–155.
- Mark E. J. Newman. 2005. **Power laws, Pareto distributions and Zipf’s law.** *Contemporary Physics*, 46(5):323–351.
- Julianne Nyhan and Marco Passarotti, editors. 2019. *One Origin of Digital Humanities.* Springer Cham, Cham (Zug), Switzerland.
- Marco Passarotti, Marco Budassi, Eleonora Litta, and Paolo Ruffolo. 2017. **The Lemlat 3.0 Package for Morphological Analysis of Latin.** In *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, volume 133, pages 24–31, Gothenburg. Linköping University Electronic Press.
- Marco Passarotti, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo, and Rachele Sprugnoli. 2020. **Interlinking through Lemmas. The Lexical Collection of the LiLa Knowledge Base of Linguistic Resources for Latin.** *Studi e Saggi Linguistici*, LVIII(1):177–212.
- Marco Passarotti, Berta González Saavedra, and Christophe Onambele. 2016. **Latin Vallex. A Treebank-based Semantic Valency Lexicon for Latin.** In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 2599–2606, Portorož, Slovenia. European Language Resources Association (ELRA).
- Rachele Sprugnoli, Marco Passarotti, Daniela Corbetta, and Andrea Peverelli. 2020. **Odi et Amo. Creating, Evaluating and Extending Sentiment Lexicons for Latin.** In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, pages 3078–3086, Marseille, France. European Language Resources Association (ELRA).
- Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. ’t Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. **The FAIR Guiding Principles for scientific data management and stewardship.** *Scientific Data*, 3(160018).

Graph Databases for Diachronic Language Data Modelling

Barbara McGillivray

King's College London, UK

barbara.mcgillivray@kcl.ac.uk

Pierluigi Cassotti and Davide Di Piero

University of Bari Aldo Moro, Italy

{surname.name}@uniba.it

Fahad Khan

Istituto di Linguistica Computazionale, CNR

fahad.khan@ilc.cnr.it

Paola Marongiu

University of Neuchâtel, Switzerland

paola.marongiu@unine.ch

Stefano Ferilli and Pierpaolo Basile

University of Bari Aldo Moro, Italy

{surname.name}@uniba.it

Abstract

Diachronic analysis, particularly of lexical semantics, is one of the most intriguing and complex tasks in linguistic studies. The integration of lexical semantic information and diachronic language resources plays a critical role in enabling quantitative accounts of language change. Focusing on the case of Latin, a high-resource language among historical languages, we present initial results from integrating Latin corpus data, Latin WordNet, and Wikidata into a graph database via a Graph-BRAIN Schema and show the potential offered by this model for diachronic semantic research.

1 Introduction and Background

Research in empirical historical semantics requires access to various sources, from dictionaries and lexicons to encyclopedic information and diachronic texts. While several scholars have recognized the corpus-based nature of diachronic semantics, particularly for corpus languages like Latin (Pinkster, 1991; Geeraerts et al., 2012), quantitative corpus-based studies are yet to pervade historical semantics research. A critical barrier to this is that corpus and lexical resources for historical languages tend to exist in data siloes. While significant progress on linking lexical resources, tools, and corpora at the level of lemmas has been made (cf. Passarotti et al. (2020) for Latin), linking at the level of word senses is still missing.

Given the remarkable work done in the design of linked data models for language data (Khan et al., 2022), some studies such as Armaselu et al. (2022) have already advocated for integrating corpus approaches with Linked Open Data technolo-

gies to study lexical semantic change, i.e., the phenomenon concerned with the change in the meaning of words over time. One crucial strategy for representing the results of research into language change as linked data is by modeling and publishing them as knowledge bases using a lexicon-based model, usually OntoLex-Lemon and its various extensions. This includes the soon-to-be-published Frequency Attestations and Corpus (FrAC) module, which proposes a new series of classes and properties for linking elements of a lexicon with corpora (Chiarcos et al., 2022). Previous work in this area includes a proposal to modify the core organizing principles of wordnets in order to represent semantic shift phenomena (Khan et al., 2023), as well as work on the representation of etymologies as Resource Description Framework (RDF) graphs using OntoLex-Lemon (Khan, 2018) and the integration of temporal information into linguistically linked datasets via a so-called *four-dimensionalist* approach (Khan, 2020).

Integrating lexical resources and semantically-annotated corpus data at scale would allow us to gather corpus data on sense distribution information, essential for fully implementing the quantitative turn in historical semantics (McGillivray and Jensen, 2023). This integration, however, requires efficient handling of large datasets. An opportunity to combine the efficient storage, management, and retrieval of data offered by Data Base Management Systems (DBMSs) with the support for formal reasoning offered by Knowledge Bases (KBs) comes from the recent development of *Graph Databases*. Graph DBMS are intrinsically designed to store schemaless data, mak-

ing them suitable to dynamic systems in which merging information is relevant. Unlike traditional DBMSs such as relational (Kriegel et al., 2003) or object-oriented (Bertino and Martino, 1991) ones, Graph DBMS lack predefined structures. Neo4j¹ is among the most common graph DBMSs. The Graph-BRAIN² technology (Ferilli and Redavid, 2020) provides intelligent information retrieval functionalities on a graph database. Its interface provides end users with access to data employing schema definitions. Schemes (available in terms of classes, relationships, and attributes) coordinate how data is presented in the interface. In Basile et al. (2022), we proposed the *Linguistic Knowledge Graph*, a model based on graph DBMSs. The Linguistic Knowledge Graph models relations between concepts and words, information about word occurrences in corpora, and diachronic information on both concepts and words. In McGillivray et al. (2023), we show an application of this model to the lexical-semantic analysis of Latin data.

Our choice to focus on Latin is motivated by several factors. First, Latin has one of the longest recorded histories of any human language, making it naturally suitable for quantitative studies (Pinkster, 1991); this, in turn, allows for corpus-driven analyses of semantic change processes over long periods. Second, this language has a particularly favourable position among historical languages: there is a high availability of extensive Latin corpora in digital form (some of which have been linked to language resources at the level of word lemmas in the context of the LiLa project³) and of computational language resources such as Latin WordNet (Minozzi, 2017) and digitized dictionaries such as the Lewis & Short Latin dictionary⁴.

Focusing on the development of the Latin language, in this paper we expand the range of Latin language resources included in the Linguistic Knowledge Graph for the study of lexical semantic change in Latin.⁵ Our contributions include: (i) the ingestion of Latin WordNet into the Linguistic Knowledge Graph; (ii) a new curated linking between existing resources for Latin, namely Latin WordNet (Minozzi, 2017; Biagetti

¹<https://neo4j.com/>

²<http://193.204.187.73:8088/GraphBRAIN/>

³<https://lila-erc.eu/>

⁴<https://lila-erc.eu/data/lexicalResources/LewisShort/Lexicon>

⁵Our code and data are available at <https://github.com/linguisticGraph/latin-graph>

et al., 2021) and the SemEval 2020 Task 1 Latin dataset (McGillivray, 2021), a sense-annotated portion of the LatinISE diachronic corpus of Latin (McGillivray et al., 2022);⁶ (iii) the integration of external contextual information (Wikidata) about the occupations of Latin authors. The term ‘occupation’ is here used in a broad sense, to refer to various types of political, cultural and societal profiles that identify authors in Wikidata. These could be e.g., priests, philosophers, historians, hagiographers, among others.

2 Resources

2.1 Dataset

LatinISE contains approximately 10 million word tokens from texts dating from the fifth century BCE to the contemporary era; it has been semi-automatically lemmatized and part-of-speech tagged. The corpus includes metadata fields indicating text identifier, author, title, dates, century, genre, URL of the source, and book title/number and character names (for plays). The semantically annotated dataset we use here was created as part of the SemEval shared task on Unsupervised Lexical Semantic Change Detection (Schlechtweg et al., 2020) and will be henceforth referred to as the SemEval Latin dataset. It contains in-context annotations for 40 Latin lemmas, 20 of which are known to have changed their meaning concerning Christianity (for example, *beatus*, which shifted its meaning from ‘fortunate’ to ‘blessed’), and 20 are known not to have changed their meaning between the BCE era and the CE era. For each of these lemmas, 60 sentences were annotated, of which 30 were randomly extracted from BCE texts and 30 from CE texts. The annotation was conducted following a variation of the DuReL framework (Schlechtweg et al., 2018) described in Schlechtweg et al. (2020): the degree by which a usage instance of a target word is related to each of its possible dictionary definitions was annotated using a four-point scale (Unrelated, Distantly Related, Closely Related, and Identical). The definitions were drawn from the Logeion online dictionary (<https://logeion.uchicago.edu/>), which contains Lewis and Short’s *Latin-English Lexicon* (1879) (Lewis and Short, 1879), Lewis’ *Elementary Latin Dictionary* (1890) (Lewis, 1890), and the dictionary by Du Fresne Du Cange et al. (1883-1887). The de-

⁶Openly available at <https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2506>.

tails of the annotation are described in McGillivray et al. (2022).

2.2 Curated Linking

We manually linked each word sense of the SemEval Latin dataset to one or more WordNet synsets. We started with the dataset provided by the LiLa project (Franzini et al., 2019), which contains a sample of 10,314 lemmas from Latin WordNet (LWN) (Minozzi, 2017; Biagetti et al., 2021). The LiLa team verified and corrected, where necessary, the synsets associated with each lemma of the sample and linked them to version 3.0 of Princeton WordNet (PWN) (Fellbaum, 1998; Miller, 1992). However, as the LiLa dataset only covers 22 of the 40 lemmas in our dataset, we used LWN as a reference for the remaining 18 lemmas. We converted the synset codes 1.6 used by LWN to version 3.0 of PWN for consistency.

The senses assigned to the target words in the SemEval Latin dataset often condensed multiple meanings into a single definition, requiring multiple synsets to be linked to the same meaning to capture all nuances. For example, the sense “understanding, judgment, wisdom, sense, penetration, prudence” of the lemma *consilium* was linked to four synsets.

In some cases, a particular sense could not be described by any of the assigned synsets in the LiLa dataset. In such cases, we searched for the lemma in LWN and selected a more appropriate synset. This was the case e.g. for the adjective *acerbus* and one of its meanings in the SemEval Latin dataset “(of things) heavy, sad, bitter”. For this meaning we selected the synset 01650376-a “psychologically painful” from LWN. When we could not find the synset in either LWN or the LiLa dataset, we looked for the most suitable synset in PWN. However, for some meanings specific to Roman culture and institutions, we could not find a suitable synset, such as with the meaning ‘Virtue, personified as a deity’ of *virtus*. In these cases, we did not link the sense to WordNet.

2.3 Contextual Information

In some instances, the metadata field of the SemEval Latin dataset (which indicates the author and title of the text, dating, and genre) was noisy, incorrectly structured, or incomplete. Wikidata is an extensive, collaboratively maintained knowledge base (Vrandečić and Krötzsch, 2014), hosting more than one hundred million items. We exploited

Wikidata for de-noising and linking the authors of the documents containing the sentences in our dataset.

First, we extracted the Wikidata entities for which the author’s occupation is specified (wdt:P106, *occupation*), and Latin (wd:Q397, *Latin*) is one of the writing languages for the author (wdt:P6886, *writing language*). We retrieve information about each author in the form of key/value properties. Author names in the SemEval Latin dataset can occur in different languages and different forms, for example *praenomen* and *nomen* followed by *cognomen* e.g., Marcus Tullius Cicero; *cognomen* followed by *praenomen* and *nomen* e.g., Cicero, Marcus Tullius; only *cognomen* e.g., Cicero; only *praenomen* and *nomen* e.g., Marcus Tullius. We processed the author’s mentions in the SemEval Latin dataset and the writer labels and aliases extracted from Wikidata, performing lowercase and punctuation removal. Matching is realized by computing the Levenshtein distance (Schimke et al., 2004) between the author reported in the SemEval dataset and all the collected surface forms (i.e., labels/aliases) from Wikidata. The surface forms are then ranked by decreasing Levenshtein distance. If the Levenshtein distance between the author’s mention and the top-ranked surface form is less than a fixed threshold, i.e., $\delta = 0.1$, the entity referenced by the surface form is linked to the author’s mention. For each author, Wikidata provides rich information, such as biographical data, the author’s works, and events that influenced their life and production. In this study, we focus on occupation information: we encode the information provided by Wikidata about the occupations of the author exploiting the property wdt:P106 (*occupation*). In particular, we create nodes of type Occupation for each occupation retrieved in Wikidata, generating a relationship between the author and their respective occupation.

3 GraphBRAIN

We stored the above information in a graph-based structure, specifically in a knowledge graph based on the GraphBRAIN technology (Ferilli and Redavid, 2020). GraphBRAIN is an approach to knowledge bases in graph form using a graph database (DB) to store information, coupled with an ontology that defines what information can be stored in the DB and how it must be described. Unlike the RDF graph model, traditionally used in Seman-

tic Web approaches, GraphBRAIN adopts the Labelled Property Graph (LPG) model, where nodes and arcs may be labelled and carry information as attribute-value pairs, ensuring a more compact and human-readable representation of knowledge. The DBMS underlying GraphBRAIN is currently Neo4j (Miller, 2013), which is schema-less. GraphBRAIN proposes an XML-based formalism to express LPG ontologies that can be mapped onto the elements of LPG graphs and act as a schema for the DB (Ferilli et al., 2022b). This approach brings several advantages. The efficiency of a native LPG graph DB can be leveraged to run network analysis and graph mining algorithms. In contrast, the expressiveness of the ontology can be leveraged for advanced automated reasoning capabilities. The ontology and data can be imported from or exported to Web Owl Language (OWL), thus enabling the use of Semantic Web tools. However, they can also be imported or exported to other formalisms (e.g., Prolog), enabling different kinds of inference, e.g., rule-based deduction, abduction, abstraction, argumentation (Esposito et al., 2000).

The Linguistic Knowledge Graph (McGillivray et al., 2023) allows us to express information about corpora, linguistic properties (background lexical, morphological, syntactic, and semantic information), time, and context; linguistic information can be imported from existing resources such as WordNet. Its lexical part is inspired by and aligned to the standard ontological lexicon model OntoLex-Lemon (McCrae et al., 2014). A corpus can be described at several levels of granularity (word, sentence, text, document). Contextual information concerns the standard bibliographic metadata (e.g., authors, publishers) but may be expanded to other entities (e.g., events). Time information can describe specific time points (days, months, years, centuries) or time intervals.

3.1 Linguistic Ontology

To address the need to create a shared vocabulary to visualize and connect the data, we here describe our linguistic ontology’s main components. This scheme collects all the relevant pieces of information available in standard lexical databases and other relevant sources of knowledge for diachronic analysis. We report the classes and relationships of our ontology in boldface; words are represented in lower-case, and relationships in upper-case. **Document** represents the hub for

knowledge discovery since it contains most aspects of the knowledge that we need. It is linked to the **Person** who wrote the text (**HAS_AUTHOR**), commonly named the “author”. A document may **CONCERN** specific **Artifacts**, **Devices**, belong to (**BELONGS_TO**) one **Category**, be written in at least one (**HAS_LANGUAGE**) **Language** and published (**PUBLISHED_IN**). We represent **Texts** belonging to (**BELONGS_TO**) documents. From the text, we are able to represent the **Words** it contains. **Lemmas** are labelled with their information, e.g., morphology and **PartOfSpeech** tags. On the other hand, word forms have (**HAS_LEMMA**) lemmas. Synsets have relationships with each other; one may be a sub-synset (hyponym) of another (**IS_A**) or be equivalent to (**SAME_AS**) another one in a different database. This happens when mapping Princeton WordNet to Latin WordNet. Time needs to be modelled for diachronic analysis. **TemporalSpecification** includes **TimeIntervals** and specific **TimePoints**, namely **Year**, **Month**, and **Day**. This model allows authors and texts to be bound to specific time periods. Moreover, we have **Events**, which may come in handy to understand the reason why some words changed their meaning (e.g., in relation to Christianity).

3.2 Latin WordNet Ingestion

The Latin WordNet (LWN) project is an initiative to create and share a common lexico-semantic database of the Latin language. The project originated as a branch of the MultiWordNet (Pianta et al., 2002) project. For diachronic analyses, linking linguistic resources with temporal information allows us to uncover instances of semantic changes in the usage of words. Hence, we provide a mechanism to enrich the Linguistic Knowledge Graph with Latin WordNet and exploit the hierarchical structure of the relationships between synsets.

In Section 3, we described the GraphBRAIN technology and its reliance on schemes/ontologies to deliver information extraction and reasoning functionalities. We mapped the Latin WordNet data with the portion of our ontology specifically devoted to linguistic analysis and understanding. Further details about scheme specifications for document representation are available in (Ferilli et al., 2022a). Here we describe the mapping between the lexical database and our schema. In LWN, we identified the following resources, grouped into separate Comma Separated Value

(CSV) files: *lemma*, *lexical_relation*, *literal_sense*, *metaphoric_sense*, *metonymic_sense*, *phrase*, *semantic_relation*, *synset*. Each resource has features that may be seen as classical columns in a relational database. From now on, we refer to specific fields as *resource.field* to uniquely identify them and motivate how we map them. The alignment process is as follows:

- *lemma*: a specific lemma is embedded in our class **Lemma**. A **Lemma** is characterized by a unique id, a lemma (its value), and a PoS tag (modelled as a relationship). For our purposes, the class **PartOfSpeech** collects all the pos tags used, following the Universal PoS Tags standard⁷. We can represent other fields expressed in LWN, such as *lemma.uri*.
- *lexical_relation*: this represents a relationship between two **Lemma**s. The field *lexical_relation.type* specifies the type of relationship. We modelled the present ones with some explicit names which express their meanings: **ANTONYMOUS_OF**, **PERTAINS_TO** (to refer to the type of relation indicated by the attribute of the relations), with their corresponding inverses, e.g. **IS_PAST_PARTICIPLE_OF**.
- *literal_sense*: this represents a relationship between a lemma, identified by the field *literal_sense.lemma*, and a synset, identified by *literal_sense.synset*. We call this relationship **expresses**. We highlight that the relationship has a “literal” sense by adding a specific attribute **sense**. Additional information about the period and genre is available.
- *metaphoric_sense*: similarly to the previous one, this represents a relationship between a lemma and a synset, where the **sense** is “metaphoric”.
- *metonymic_sense*: as before, but the **sense** is “metonymic” in this case.
- *phrase*: a phrase is a word or a multi-word expression. In both cases, the concept is expressed by the class **Lemma** since for our purposes both concepts play an equally important role when analysing semantic changes. Again,

⁷<https://universaldependencies.org/u/pos/>

we have the PoS tag information, which is modelled in the same way described above.

- *semantic_relation*: a relationship between two synsets. Based on the *semantic_relation.type* several relationships may be expressed. They are mapped into the following ones and their corresponding inverses: **PART_OF**, **HAS_SUBCLASS**, **ATTRIBUTE_OF**, **SIMILAR_TO**, **ANTONYMOUS_OF**, **PERTAINS_TO**, **PART_PARTICIPLE_OF**, **CAUSES**, and **ENTAILS**.
- *synset*: a synset is embedded in **LexiconConcept** while its property *synset.gloss*, which is the description of the synset, is represented as the attribute **description** of the class **LexiconConcept**. *synset.gloss* is the description of the synset and is mapped onto the attribute **description**.

Thanks to this mapping, we can acquire the LWN resource and represent it in our formalism, which allows us to leverage the connections between the different datasets, as explained via examples in the next section.

4 Analysis and Discussion

Figure 1 shows the subgraph for the word *humanitas*. The occurrences of *humanitas* are annotated in the SemEval dataset with three senses: (i) ‘human nature, humanity’, (ii) ‘humanity, philanthropy’, and (iii) ‘mankind’.⁸ In the curated link, we associate the sense (i) to the humanness.n.01 synset, the sense (ii) to the synsets kindness.n.01, kindness.n.03, and courtesy.n.03 and sense (iii) to the synset world.n.08. According to the *Thesaurus Linguae Latinae* (*Thesaurus-Kommission*, 1900–), which confirms the first attestation of all senses in the 1st century BCE, the sense (ii) ‘humanity, philanthropy’ developed from the more general sense (i) ‘human nature, humanity’ which refers to human nature in general. The subgraph shows that the three senses are attested at least once in passages dated 1st century BCE. However, the graph shows that the sense of ‘philanthropy’ dominates all other senses in the 1st century BCE. In the transition to the CE period, the sense of ‘humanity’ prevails

⁸A fourth sense ‘liberal education, good breeding, the elegance of manners or language, refinement’ was annotated in the Latin dataset, but not encoded in the graph, since the author matching described in Section 2.3 failed.

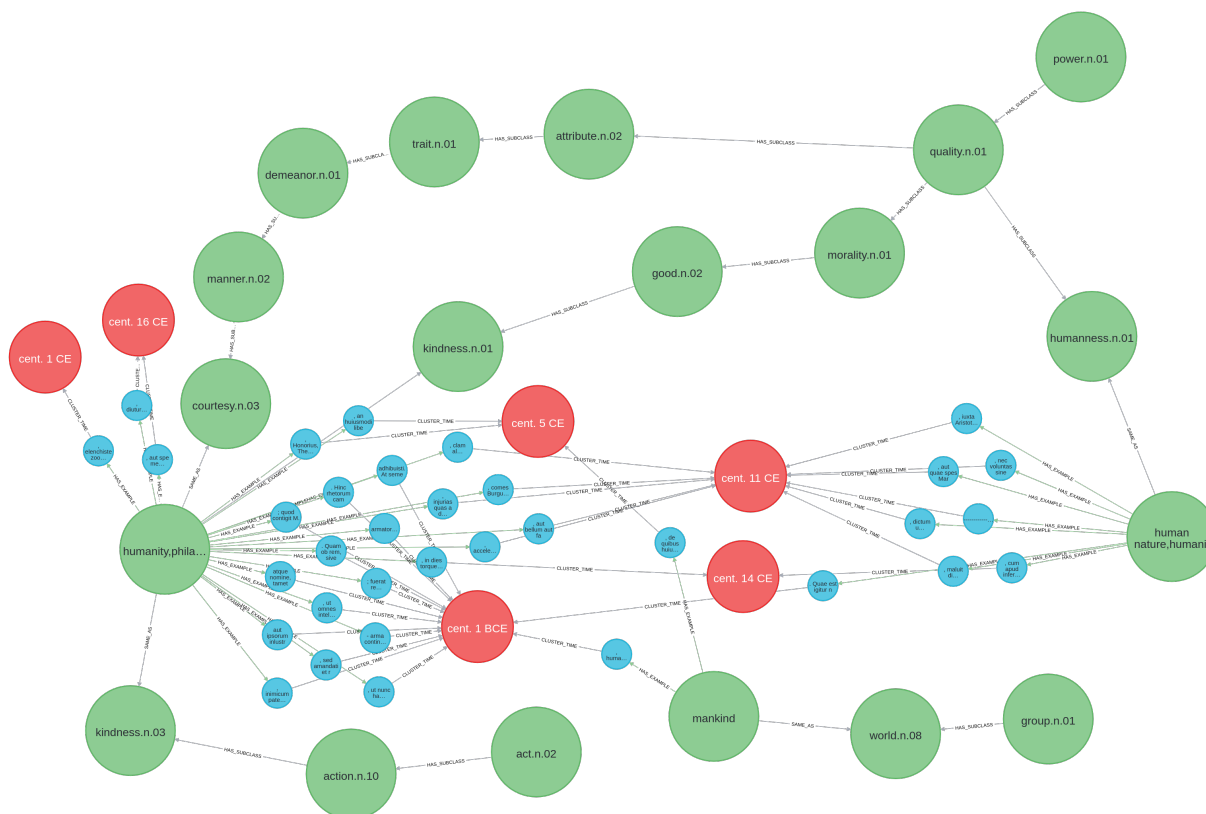


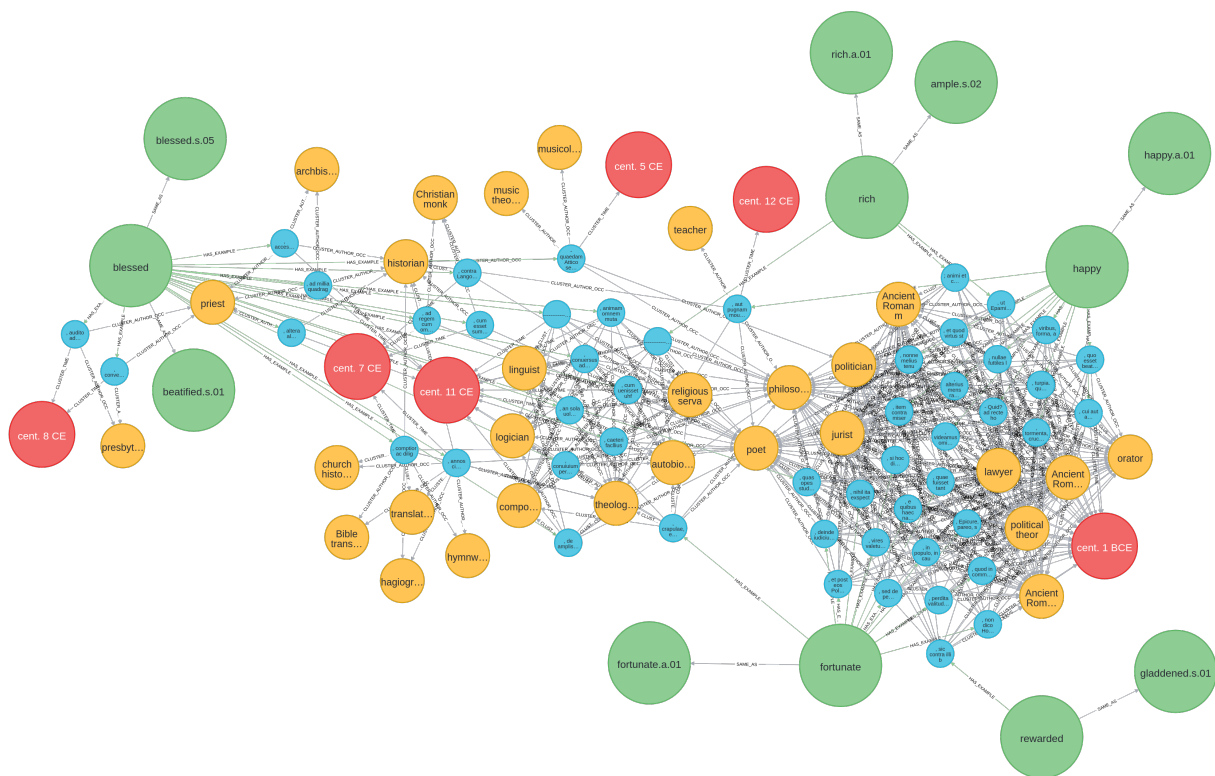
Figure 1: Subgraph for the word *humanitas*, including the sentences in which the lemma *humanitas* occurs in the SemEval Latin dataset, the century of the works from which the sentences were extracted, the annotated senses in the SemEval Latin dataset, and the curated links between the senses and the synsets in Latin WordNet. The sentences are represented as Text nodes (in blue), the senses and the synsets as LexiconConcept nodes (in green), and the centuries as TimePoint nodes (in red).

regarding the number of annotations, and the two meanings coexist in the CE period.

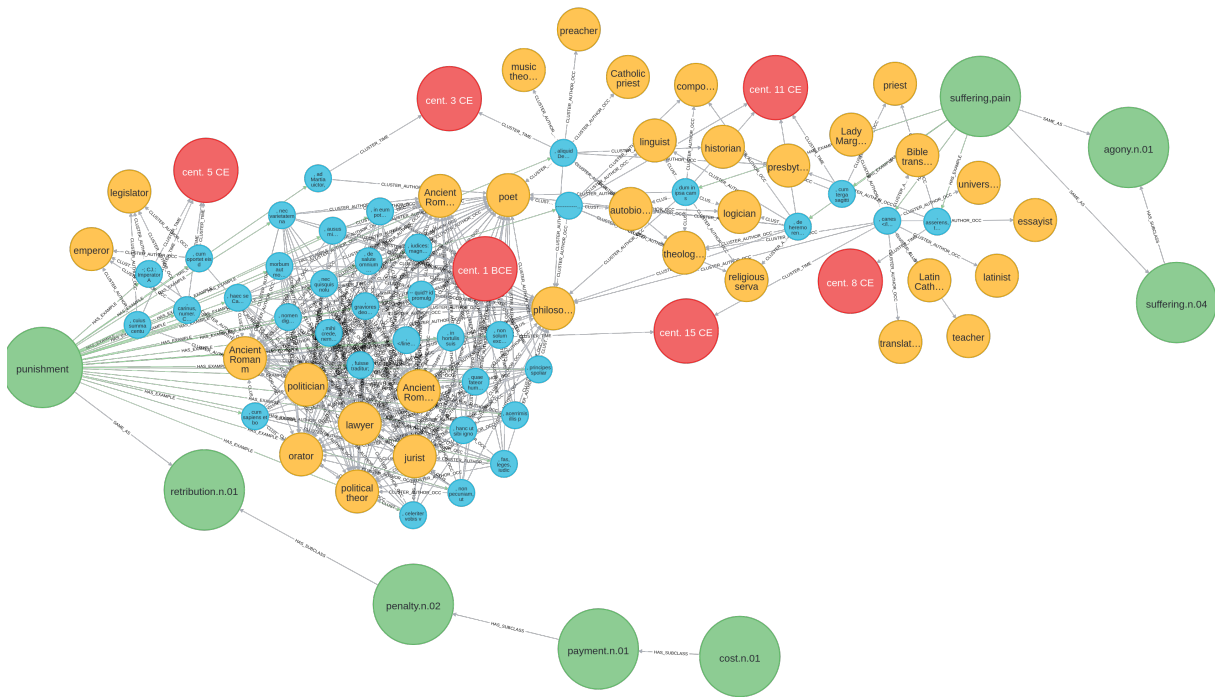
By ascending the WordNet hierarchy, we can gain deeper insight into the relationship between the two senses. The sense (ii) ‘humanity, philanthropy’ and the sense (i) ‘human nature’ are connected via two paths: sense (ii) originates from the quality.n.01 synset (i.e. ‘an essential and distinguishing attribute of something or someone’); sense (i) from the attribute.n.02 synset (i.e., ‘an abstraction belonging to or characteristic of an entity’). The two senses have in common the quality.n.01 synset, but the sense (ii) ‘humanity, philanthropy’ is directly linked to kindness.n.01 synset, and to a higher degree of the WordNet hierarchy to the morality.n.01 synset (i.e., ‘concerned with the distinction between good and evil or right and wrong’). The additional information provided by including the WordNet hierarchy in the graph allows us to show the type of semantic relationship between the two predominant senses of *humanitas*. The more general sense (i) ‘human nature’ special-

izes in its meaning in the sphere of morality, originating the sense (ii) ‘philanthropy’. In the example of *humanitas* shown in Figure 1, the injected information from WordNet was exploited to analyze the semantic relationship between the meanings of the lemma *humanitas*. While the synset taxonomy in this example helps us track and classify phenomena of semantic change, including other types of information retrievable from the metadata can help gain further insights into the context of the semantic change. We add information about the authors’ occupations in the examples shown in Figure 2.

In Figure 2, three examples of subgraphs are shown. The three graphs refer to the encoded information for the Latin lemmas *beatus*, *poena*, and *salus*, respectively. In particular, we filtered for nodes of type Text (blue nodes), Century (red nodes), Synset (green nodes), and Occupation (yellow nodes). We grouped the Text nodes by occupation and century, i.e., we created an explicit link between nodes of type Text and nodes of type Time-Point and between nodes of type Text and nodes of

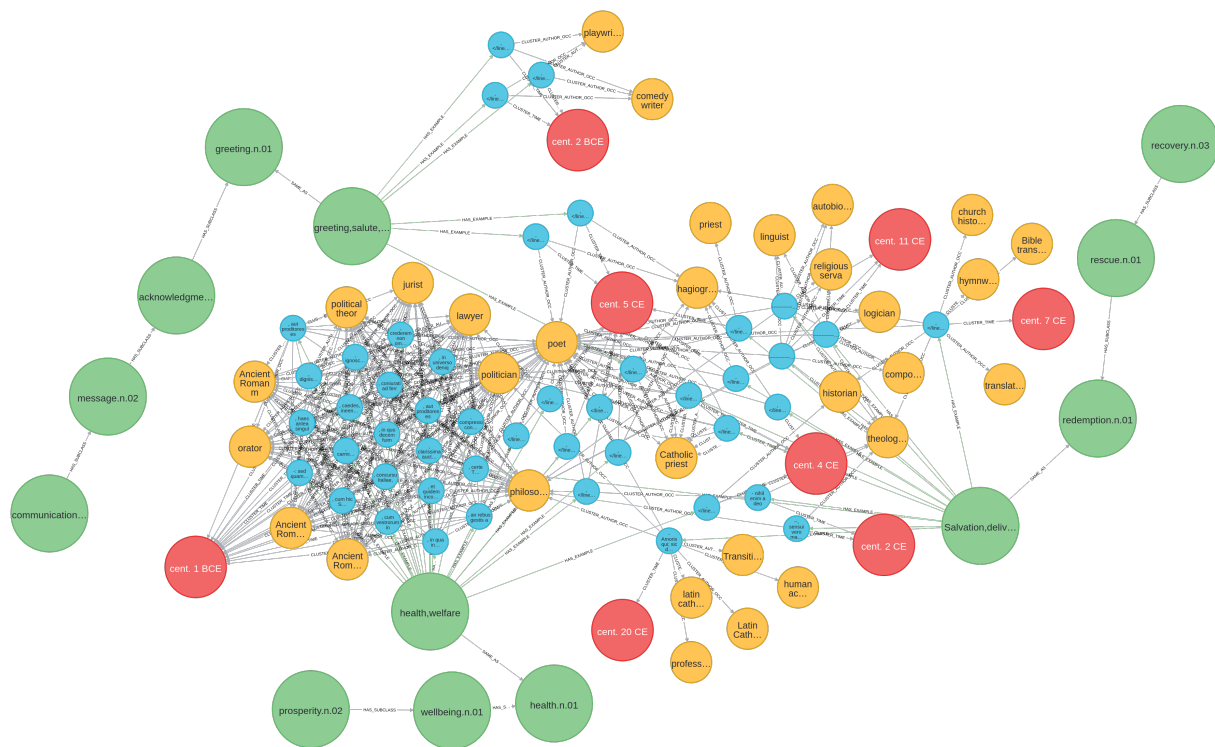


(a) Subgraph for *beatus*. The synsets for *beatus* are: (i) beatified.s.01: Roman Catholic; proclaimed one of the blessed and thus worthy of veneration, (ii) blessed.s.05: enjoying the bliss of heaven, (iii) rich.a.01: possessing material wealth, (iv) fortunate.a.01: having unexpected good fortune, (v) ample.s.02: affording an abundant supply, (vi) happy.a.01: enjoying or showing or marked by joy or pleasure or good fortune



(b) Subgraph for *poena*. The synsets for *poena* are: (i) retribution.n.01: a justly deserved penalty, (ii) suffering.n.04: feelings of mental or physical pain, (iii) agony.n.01: intense feelings of suffering; acute mental or physical pain

Figure 2: Sub-graphs: (a) beatus. (b) poena (c) salus .



(c) Subgraph for *salus*. The synsets for *salus* are: (i) health.n.01: *a healthy state of well-being*, (ii) redemption.n.01: *(Christianity) the act of delivering from sin or saving from evil*, (iii) greeting.n.01: *an acknowledgment or expression of goodwill*

Figure 2: Sub-graphs: (a) *beatus*. (b) *poena* (c) *salus* (cont.).

type Occupation.

Combining queries at the level of the annotated senses, WordNet synsets, text metadata and textual data at once, users can have access to rich nuanced information, which is very valuable for quantitative diachronic semantic analyses, both on specific words and whole lexical fields. The graphs in Figure 2 seem to show some trends in semantic change, all related to Christianity. The lemma *beatus* was annotated in the SemEval dataset with five senses: (i) ‘happy’, (ii) ‘fortunate’, (iii) ‘rewarded’, (iv) ‘rich’, and (v) ‘blessed’. The graph shows that the senses (i) ‘happy’, (ii) ‘fortunate’, (iii) ‘rewarded’, and (iv) ‘rich’ all emerge starting from the 1st century BCE in the annotated dataset. On the other hand, sense (v) ‘blessed’ emerges later with the advent of Christianity, as we can see in correspondence with the CE nodes. In this case, there seems to be a replacement of the previous senses in favour of the Christian sense. Additionally, if we consider the nodes of type Occupation, a noticeable difference emerges between the two (groups of) meanings: in the cluster of occupation nodes connected to the Christian sense, we can observe profiles related to theological and religious

activity, e.g., priests, hagiographers, which do not appear to be connected to the other senses. The same type of observations can be made for *salus*, which initially has the meanings (i) ‘health’ and (ii) ‘greeting’, and, subsequently, develop the Christian sense of (iii) ‘salvation, deliverance from sins’. However, in this case, we can notice the difference with *beatus* in the type of semantic change, as the new meaning (iii) ‘salvation’ replaces or dominates the previously attested meanings but continues to coexist with them. The lemma *poena* also presents an example of semantic change in which the new meaning does not entirely replace the previous ones. The new sense of ‘suffering, pain’, which emerges in the CE nodes, continues to coexist with the sense of ‘punishment’, which was attested from the 1st century BCE in the annotated dataset. In the case of *poena*, the contrast between the two clusters of occupation nodes is even more evident. The sense of punishment is often associated with authors classified as related to the legal world, e.g., legislator, lawyer, and jurist. In contrast, nodes related to the Christian and theological world appear in the case of salvation, e.g., theologian, priest, and presbyter. The graphs in Figure 2 are in line with

that we know about semantic changes prompted by the advent of Christianity, which invested many words already in use in pre-Christian Latin with new meanings closely related to the Christian world (Burton, 2011). Moreover, the lemmas shown in Figure 2 illustrate the different types of interaction between older and new senses described in literature (Traugott and Dasher, 2001, 10–12): in some cases, the two senses can continue to coexist, as for the lemmas *salus* and *poena* (a phenomenon called ‘layering’ (Hopper, 1991, 22)); in others, as for the lemma *beatus*, the relationship between the new sense and the older ones is unbalanced as the new sense becomes more prominent in a society invested in Christian values.

5 Conclusion and Future Work

We applied diachronic lexical-semantic analysis by integrating different resources into a graph-based structure. Future research should be devoted to enriching the dataset by collecting other resources to uncover more complex relationships and possibly automatically detect semantic changes among all terms in the vocabulary. Currently, our model does not include a programmatic way to automatically detect instances of semantic changes, but this is an avenue of future research. We plan to publish a version of the graph database in which experiments can be replicated.

Authors’ contributions and Acknowledgements

BMcG contributed to the design of the study, managed the project, provided the SemEval dataset and wrote sections 1 and 2.1. PM provided the curated linking between the annotated SemEval Latin dataset and WN, and wrote sections 2.2 and 4. PC processed the annotated LatinISE corpus, extracted metadata information from WikiData, generated the graph and the visualizations (Figure 1 and Figure 2), wrote Section 2.3, and contributed in writing Section 4. PB contributed to the design of the study, generated the graph and wrote section 3. FK proofread the article and contributed to discussions on the relationship between native KG approaches to modelling lexical data as graphs and RDF/OntoLex approaches. DD contributed to the design of the schema and the upload of LWN resources into the LPG-based KG and wrote section 3. SF contributed to the design of the schema and the Knowledge Representation methodology

of GraphBRAIN and wrote section 3.

We acknowledge the support of the PNRR projects FAIR - Future AI Research (PE00000013), Spoke 6 - Symbiotic AI (CUP H97G22000210007) and CHANGES - Cultural Heritage Active innovation for Next-Gen Sustainable society (PE00000020), Spoke 3 - Digital Libraries, Archives and Philology, under the NRRP MUR program funded by the NextGenerationEU.

References

- Florentina Armaselu, Elena Simona Apostol, Anas Fahad Khan, Chaya Liebeskind, Barbara McGillivray, Ciprian-Octavian Truica, Andrius Utka, Giedre Valunaite Oleskeviciene, and Marieke van Erp. 2022. [LL\(O\)D and NLP perspectives on semantic change for humanities research](#). *Semantic Web*, 13(6):1051–1080.
- Pierpaolo Basile, Pierluigi Cassotti, Stefano Ferilli, and Barbara McGillivray. 2022. [A new time-sensitive model of linguistic knowledge for graph databases](#). In *Proceedings of the 1st Workshop on Artificial Intelligence for Cultural Heritage, AI4CH 2022, co-located with the 21st International Conference of the Italian Association for Artificial Intelligence (AIXIA 2022), Udine, Italy, November 28, 2022*, volume 3286 of *CEUR Workshop Proceedings*, pages 69–80. CEUR-WS.org.
- Elisa Bertino and Lorenzo Martino. 1991. Object-oriented database management systems: concepts and issues. *Computer*, 24(4):33–47.
- Erica Biagetti, Chiara Zanchi, and William Michael Short. 2021. [Toward the creation of WordNets for ancient Indo-European languages](#). In *Proceedings of the 11th Global Wordnet Conference*, pages 258–266, University of South Africa (UNISA). Global Wordnet Association.
- Philip Burton. 2011. Christian latin. In *A companion to the Latin language*, pages 485–501, Oxford. Wiley-Blackwell.
- Christian Chiarcos, Katerina Gkirtzou, Maxim Ionov, Besim Kabashi, Fahad Khan, and Ciprian-Octavian Truică. 2022. [Modelling collocations in OntoLex-FrAC](#). In *Proceedings of Globalex Workshop on Linked Lexicography within the 13th Language Resources and Evaluation Conference*, pages 10–18, Marseille, France. European Language Resources Association.
- Charles Du Fresne Du Cange, G. A. Louis Henschel, P. Carpentier, Johann Christoph Adelung, and Léopold Favre. 1883-1887. *Glossarium mediæet infimælatinitatis*. L. Favre, Niort.

- Floriana Esposito, Giovanni Semeraro, Nicola Fanizzi, and Stefano Ferilli. 2000. [Multistrategy theory revision: Induction and abduction in INTHELEX](#). *Mach. Learn.*, 38(1-2):133–156.
- Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.
- Stefano Ferilli and Domenico Redavid. 2020. The graphbrain system for knowledge graph management and advanced fruition. In *Foundations of Intelligent Systems: 25th International Symposium, ISMIS 2020, Graz, Austria, September 23–25, 2020, Proceedings*, pages 308–317. Springer.
- Stefano Ferilli, Domenico Redavid, and Davide Di Pierro. 2022a. Holistic graph-based document representation and management for open science. *International Journal on Digital Libraries*, pages 1–23.
- Stefano Ferilli, Domenico Redavid, and Davide Di Pierro. 2022b. [Lpg-based ontologies as schemas for graph dbs](#). In *Proceedings of the 30th Italian Symposium on Advanced Database Systems, SEBD 2022, Tirrenia (PI), Italy, June 19-22, 2022*, volume 3194 of *CEUR Workshop Proceedings*, pages 256–267. CEUR-WS.org.
- Greta Franzini, Andrea Peverelli, Paolo Ruffolo, Marco Passarotti, Helena Sanna, Edoardo Signoroni, Viviana Ventura, and Federica Zampedri. 2019. [Nunc Est Aestimandum: Towards an evaluation of the latin wordnet](#). In *Proceedings of the Sixth Italian Conference on Computational Linguistics*. Accademia University Press.
- Dirk Geeraerts, Caroline Gevaert, and Dirk Speelman. 2012. Current methods in historical semantics. *Current methods in historical semantics*, pages 73–109.
- Paul J. Hopper. 1991. On some principles of grammaticalization. In *Approaches to grammaticalization*, pages 17–35, Amsterdam, Philadelphia. John Benjamins Publishing.
- Anas Fahad Khan. 2018. [Towards the representation of etymological data on the semantic web](#). *Information*, 9(12):304. Publisher: MDPI AG.
- Anas Fahad Khan, Christian Chiarcos, Thierry Declerck, Daniela Gifu, Elena González-Blanco García, Jorge Gracia, Maxim Ionov, Penny Labropoulou, Francesco Mambrini, John P. McCrae, Émilie Pagé-Perron, Marco Passarotti, Salvador Rosl Muñoz, and Ciprian-Octavian Truică. 2022. [When linguistics meets web technologies. recent advances in modelling linguistic linked data](#). *Semantic Web*, pages 1–64.
- Anas Fahad Khan, John P McCrae, Francisco Javier Minaya Gómez, Rafael Cruz González, and Javier E Díaz-Vera. 2023. Some considerations in the construction of a historical language wordnet.
- Fahad Khan. 2020. [Representing temporal information in lexical linked data resources](#). In *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)*, pages 15–22, Marseille, France. European Language Resources Association.
- Hans-Peter Kriegel, Martin Pfeifle, Marco Pötke, and Thomas Seidl. 2003. The paradigm of relational indexing: A survey. In *BTW 2003–Datenbanksysteme für Business, Technologie und Web, Tagungsband der 10. BTW Konferenz*. Gesellschaft für Informatik eV.
- Charlton T. Lewis. 1890. *An Elementary Latin Dictionary*. American Book Company, New York, Cincinnati, and Chicago.
- Charlton T. Lewis and Charles Short. 1879. *A Latin Dictionary, Founded on Andrews' edition of Freund's Latin dictionary revised, enlarged, and in great part rewritten by Charlton T. Lewis, Ph.D. and Charles Short*. Clarendon Press, Oxford.
- John McCrae, Christiane Fellbaum, and Philipp Cimiano. 2014. Publishing and linking wordnet using lemon and rdf. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics*.
- Barbara McGillivray. 2021. Dataset: Latin lexical semantic annotation. Figshare. DOI: <https://doi.org/10.18742/16974823.v1>.
- Barbara McGillivray, Pierluigi Cassotti, Pierpaolo Basile, Davide Di Pierro, and Stefano Ferilli (in press). 2023. Using graph databases for historical language data: Challenges and opportunities. In *Proceedings of the 19th Italian Research Conference on Digital Libraries, Bari, Italy, February 23-24, 2023*, CEUR Workshop Proceedings. CEUR-WS.org.
- Barbara McGillivray and Gard B. Jensen. 2023. [Quantifying the quantitative \(re-\)turn in historical linguistics](#). *Humanities and Social Sciences Communications*, 10(37).
- Barbara McGillivray, Daria Kondakova, Annie Burman, Francesca Dell’Oro, Helena Bermúdez Sabel, Paola Marongiu, and Manuel Márquez Cruz. 2022. [A new corpus annotation framework for latin diachronic lexical semantics](#). *Journal of Latin Linguistics*, 21(1):47–105.
- George A. Miller. 1992. [WORDNET: a lexical database for english](#). In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, USA, February 23-26, 1992*. Morgan Kaufmann.
- Justin J Miller. 2013. Graph database applications and concepts with neo4j. In *Proceedings of the southern association for information systems conference, Atlanta, GA, USA*, volume 2324.
- Stefano Minozzi. 2017. Latin wordnet, una rete di conoscenza semantica per il latino e alcune ipotesi di utilizzo nel campo dell’information retrieval. In *Strumenti digitali e collaborativi per le Scienze dell’Antichità*, pages 123–134, Venezia. Università Ca’ Foscari.

- Marco Passarotti, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo, and Rachele Sprugnoli. 2020. [Interlinking through lemmas. the lexical collection of the lila knowledge base of linguistic resources for latin](#). *Studi e Saggi Linguistici*, 58.
- Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Multiwordnet: developing an aligned multilingual database. In *First international conference on global WordNet*, pages 293–302.
- Harm Pinkster. 1991. *Sintassi e semantica latina*. Rosenberg & Sellier.
- Sascha Schimke, Claus Vielhauer, and Jana Dittmann. 2004. Using adapted levenshtein distance for online signature authentication. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 2, pages 931–934. IEEE.
- Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. [Semeval-2020 task 1: Unsupervised lexical semantic change detection](#). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation, SemEval@COLING 2020, Barcelona (online), December 12-13, 2020*, pages 1–23. International Committee for Computational Linguistics.
- Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. 2018. [Diachronic Usage Relatedness \(DURel\): A framework for the annotation of lexical semantic change](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 169–174, New Orleans, Louisiana.
- Thesaurusbüro München Internationale Thesaurus-Kommission, editor. 1900–. *Thesaurus linguae latinae*. Mouton de Gruyter, Berlin.
- Elizabeth Closs Traugott and Richard B. Dasher. 2001. *Regularity in semantic change*. Cambridge University Press, Cambridge.
- Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85. Publisher: ACM New York, NY, USA.

Contextual Profiling of Charged Terms in Historical Newspapers

Ryan Brate and **Marieke van Erp**

KNAW Humanities Cluster, DHLab

Oudezijds Achterburgwal 185

1012 DK Amsterdam, Netherlands

{ryan.brate,marieke.van.erp}
@dh.huc.knaw.nl

Antal van den Bosch

Utrecht University

Institute for Language Sciences Utrecht,
the Netherlands

a.p.j.vandenbosch@uu.nl

Abstract

We extract nouns and corresponding co-occurrent targeted context features from a large corpus of Dutch language newspaper articles, from 1950s through the 1990s. Applying a well-established approach for scoring context feature and centre word associativity, we explore using the scores in the task of identifying key characteristics of known-charged terminology. Then use these features to draw parallels between known-charged and other terms. In the context of the very current decolonisation efforts amongst museum institutions, such approaches offer an opportunity to condense large quantities of data into the most-significant, salient information for digestion by heritage professionals. The methods were found to indeed yield insights into known and candidate charged terms.

Disclaimer: This paper contains derogatory words and phrases. They are provided solely as illustrations of the research results and do not reflect the opinions of the authors or their organisations. In-text examples of derogatory and potentially offensive are presented in “*quotes, boldfaced and italicised*”.

1 Introduction

Museums of the World,¹ a database of cultural heritage institutions, records approximately 55,000 museums spread over 202 countries. The largest such collection, The Smithsonian Institution² alone holds in excess of 155M. Such collections enhance our collective understanding of our shared past, but in doing so, they give cultural heritage institutions powerful voices in the shaping of historical narratives in the public consciousness.

¹<https://www.degruyter.com/database/MOW/html>

²<https://www.si.edu/newsdesk/factsheets/smithsonian-collections>

Many museum collections originate from the colonial period, with metadata and object portrayals stemming from the particular world of the time. There is now a growing movement of *decolonisation* in western museums aimed at the acknowledgement and accommodation of previously marginalised voices to combat biases propagated by the advancement of narrow viewpoints (Odu-mosu, 2020). Part of the decolonisation effort centres around greater sensitivity and reconsideration of the terminology and language used in item metadata. This is more complicated than wholesale removal of terminology from metadata and items from collections, even if such problematic terms are known. To handle the complexities properly, there needs to be greater contextual understanding of a term’s implied characterisation in context. For instance, many terms nowadays considered problematic are ambiguous, also in their contentiousness: calling a plant *exotic* is different from calling a person the same. When and why terms are deemed problematic is complex, but the recognition of the social-cultural (contextual) aspects of terms provides a mechanism for some degree of understanding and comparison.

In this paper, we aim to explore the contextual profiles of a reference set of known charged collective nouns, reflective of some people group and identify the contextual features that distinguish them. Specifically, we consider four complementary context feature types: verbs for which the noun is the agent, verbs for which the noun is the patient, adjectives, and compound word modifiers as applied to the nouns. I.e., we are trying to capture the things done to them, the things they do and the attributes ascribed to them. In order to do so, we leverage the extensive digitised (and OCR’d) newspaper collection of the National Library of the Netherlands (KB), between the 1950s and 1990s, thereby capturing the period of European decolonisation to more recent post-colonial times. Such a

collection represents a valuable resource reflective of public discourse, attitudes and societal norms of the times.

In exploring context and its relevance to charged nouns, we make use of noun–context associativity measures. Specifically, we ask for each noun of a set of known charged nouns, *do contextual features exist, which for some noun–context feature associativity score threshold, are highly predictive of the noun?* Secondly, we seek to examine the parallels that can be drawn between known charged nouns: i.e., *are there context features which for some noun–context feature associativity threshold, recall multiple known charged nouns with a reasonable degree of precision with respect to our known charged noun set?* Finally, we examine those nouns, not part of our known charged noun set, which share similar context feature associations: asking, *can the context features of known charged nouns help identify other charged instances?*

2 Related Work

Our work is situated on the intersection of detecting and modeling bias and harmful language. Bias in large datasets and its effects on models learned on those datasets has gained more attention in recent years (cf. (Sap et al., 2020; Bender et al., 2021; Schick et al., 2021; Birhane et al., 2022)). Work done on the same corpus as ours is (Wevers, 2019), who aims to detect gender bias in Dutch newspapers. We focus on broader biases and harmful language, mostly coming from a colonial perspective. The GLAM community is very well aware of problematic artefacts of colonial history in datasets (cf. (Mohamed et al., 2020; Barabucci et al., 2020; Luthra et al., 2023)) but there has been less attention for this in the NLP community. In our prior work, we have started to investigate how certain terms are viewed by the general public via a crowdsourcing experiment (Brate et al., 2021). We found that context plays an important role in whether certain terms are deemed charged or not. In this paper, we extend this work by modelling contextual features of charged terms.

The detection of hate speech gained traction with the growing popularity of social media data and includes cyberbullying, insults, vulgar content and racist language (Schmidt and Wiegand, 2017). While the charged terminology we are investigating has overlaps with the dimension investigated in hate speech, colonially biased language tends

to be somewhat more subtle than overt insults, although these do occur. It should also be noted that researching harmful stereotypes requires a balanced approach to not inadvertently incur more harm (Kirk et al., 2022).

Our approach to use adjectives and verbs directly associated with entities, as contextual features for distinguishing entities is inspired by (Bamman et al., 2013). They used a hierarchical Bayesian approach to group film-character types across film and film tropes, using the characterisation of characters in terms of *the things they do*, *the things done to them*, and *the way they are described* as features. However, whereas the soft-clustering iterative approach used by Bamman is based on broad feature commonality, and favours data-rich cluster types; we expect charged terms to yield often highly unique associations, not necessary given to easy feature clustering. Consequently, whilst inspired by this approach, we consider feature comparison based on a metric of noun-feature *keyness*, i.e., associativity score, based on the work of (Dunning, 1993).

3 Methodology

We use the raw data of the National Library of the Netherlands OCR'd newspaper dataset.³ We split the data into discrete years to be analysed independently, as usages and characterisations of known-charged terms are subject to variation over time. We take sample years per decade, to be considered separately. The expectation is that one-year periods are too short to be regularly affected by confusing shifts in usage. We use the sampled data to create tables of associativity, or *keyness*, scores by collective noun and context features to answer our research questions.

The adjective–noun and verb–noun pairs are extracted by pattern matching against part of speech (POS) tagged dependency trees of the newspaper dataset. In the case of modifier–noun pairs, a corpus of modifiers and corresponding heads is bootstrapped from our set of known-charged words. Subsequently, the coincident collective noun–context feature pairs are assembled into separate frequency tables according to the context feature type (e.g., adjective) for each sample year. For the known-charged nouns, the frequencies for all plural forms of the noun are aggregated. The raw collective noun–context feature co-occurrence fre-

³<https://delpher.nl>

quencies are then converted to some metric of *keyness*, which is used as the basis for exploring the key features by collective noun, and for exploring the parallels between collective nouns.

3.1 Charged nouns

The terms in Table 1 are used as our reference set of known charged collective nouns. The basis of this list is the aforementioned Words Matter document (Modest and Lelijveld, 2018). We consider singular and plural forms.

3.2 Dataset

All available, publicly accessible OCR'd articles of the National Library of the Netherlands (KB) newspaper set, in each of the years as listed in Table 2, were taken in their entirety. The table also lists the number of approximate resulting extracted articles.

A dependency-parsed, POS-tagged version of this dataset was created via spaCy (Honnibal and Montani, 2017), with an intermediate step of rule-based tokenisation and sentence segmentation via regular expressions. To reduce the sentence complexity passed to spaCy, segmentation is additionally performed on conjunctions, ":", ";", and "&";

3.3 Building a corpus of the modifier-head components of compound nouns

As described in section 3.6, the keyness metric adopted in determining how key a some *particular context feature* is to some *particular noun* in question, is a function of the corpus-wide noun-context feature co-occurrence frequencies. Hence, a corpus of modifier-head instances is needed which consists of all modifiers coincident with known charged nouns, and all of the corresponding heads coincident with these modifiers. The spaCy dependency parse of the KB newspaper corpus provides a list of tagged instances of nouns. Using this list of nouns together with the charged noun set, we bootstrapped a corpus of modifier-head compound words.

Separately, for each of the years in Table 2, a corpus of modifier-head components of compound nouns was assembled. The result considers modifiers with or without terminating hyphens as being the same instance. The following approach was adopted to bootstrap the corpus from the known-charged nouns:

Parameters

Category	Charged Nouns (<i>translation</i>)
race	aboriginal (s) (<i>aboriginal(s)</i>); afkomst (en) (<i>descent(s)</i>); allochtoon , allochtonen (<i>migrant(s)</i>); Berber (s) (<i>Berber(s)</i>); blanke (n) (<i>white person(s)</i>); bosneger (s) (<i>bush negro</i>); creool , creolen (<i>creole(s)</i>); eskimo (s) (<i>eskimo(s)</i>); etniciteit (en) (<i>ethnicity(-ies)</i>); gekleurd (en) (<i>colored(s)</i>); halfbloed (en) (<i>half-blood(s)</i>); Hottentot (ten) (<i>Khoikhoi people</i>); immigrant (en) (<i>immigrant(s)</i>); inboorling (en) (<i>primitive native(s)</i>); indo (s) (<i>Indo-European(s)</i>); indiaan , indianen (<i>Indian(s)</i>); inheems (en) (<i>indigenous</i>); inlander (s) (<i>native(s)</i>); kaffer (s) (<i>black African</i>); Khoi (<i>Khoisan people</i>); kleurling (en) (<i>colored(s)</i>); koppensneller (s) (<i>headhunter(s)</i>); moor , moren (<i>Muslim people of Arab and Amazigh descent</i>); marron (s) (<i>maroon</i>); medicijnman (nen) (<i>medicine man(men)</i>); mesties (<i>person of mixed-race background</i>); migrant (en) (<i>migrant(s)</i>); mulat (ten) (<i>mulatto(s)</i>); neger (s, in, innen) (<i>negro(s) (m/f)</i>); njai (<i>Indonesian mistress to coloniser</i>); oorsprong (en) (<i>descent(s)</i>); primitief , primitieven (<i>primitive(s)</i>); Pygmee (ën) (<i>Pygmy(Pygmies)</i>); ras (sen) (<i>race(s)</i>); roots (<i>roots</i>); scalp (en) (<i>scalp(s)</i>); stam (men) (<i>tribe(s)</i>); stamhoofd (en) (<i>tribal head(s)</i>); wildeman (nen) (<i>uncivilised man (men)</i>); zigeuner (s) (<i>gypsy (gypsies)</i>);
social	baboe (s) (<i>female servant(s)</i>); barbaar , barbaren (<i>barbarian(s)</i>); bediende (n) (<i>servant(s)</i>); koeli (es) (<i>contract worker(s)</i>); piraat , piraten (<i>pirate(s)</i>); slaaf , slaven (<i>slave(s)</i>); slavenhandel (s) (<i>slave trade</i>);
non-racial characteristics	dwerg (en) (<i>dwarf(dwarves)</i>); hermafrodit (en) (<i>hermaphrodite(s)</i>); mongool , mongolen (<i>mongoloid(s)</i>);
sexual orientation	homo (s) (<i>gay person(s)</i>); queer (s) (<i>queer person(s)</i>); trans (<i>trans person(s)</i>);
place	jappenkamp (en) (<i>Japanese concentration camp(s)</i>);
religious	islamiet (en) (<i>muslim(s)</i>); mohammedaan , mohammedanen (<i>muslim(s)</i>);

Table 1: Charged noun list. Word forms of each charged noun are aggregated and each aggregation is collected under its stemmed form (in bold).

	sampled years (No. articles in millions [M])		
1950s	1951 (1.2M)	1955 (1.4M)	1959 (1.3M)
1960s	1961 (0.9M)	1965 (0.9M)	1969 (0.8M)
1970s	1971 (0.8M)	1975 (0.7M)	1979 (0.7M)
1980s	1981 (0.7M)	1985 (0.7M)	1989 (0.8M)
1990s	1991 (0.7M)	1995 (0.3M)	

Table 2: KB Newspaper Collection sampled years (taken in their entirety where publicly available), and corresponding number of articles rounded to the nearest 0.1M.

- The entire POS-tagged noun set from the spaCy-parsed dataset for each year, represents the *noun pool* from which to extract a corpus of modifier–head compound word pairs;
- The charged-words (including plural forms and variants) of Table 1 are used as seed heads:

Steps

- *Modifier extraction:* modifiers are harvested via trie-based character matching of the seed heads against the *noun pool*. Terminating hyphens are stripped from the modifiers. The output (modifiers) are filtered;
- *Head extraction:* heads are then harvested via trie-based matching of the previously harvested modifiers from the *entity pool*. Once again, hyphens are stripped and the output (heads) are filtered;
- *Final head–modifier extraction:* Repeating the *Modifier Extraction* step, a set of filtered set of head–modifier pairs is returned.

The filtering at each harvesting stage aims to improve the quality of the harvested heads and modifiers, by reducing the incidence of extracting false cases. Filtering consists of removing all heads or modifiers less than 3 characters in length or absent from the SoNaR-corpus⁴(ignoring case).

3.4 Building a corpus of noun–adjective pairs

Separately, for each of the years listed in Table 2, the corresponding spaCy dependency-parsed

⁴<https://taalmaterialen.ivdnt.org/download/tstc-sonar-corpus/>

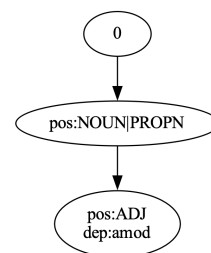


Figure 1: Pattern A1 denoting the targeted adjective-noun relationship. '0' points to the root.

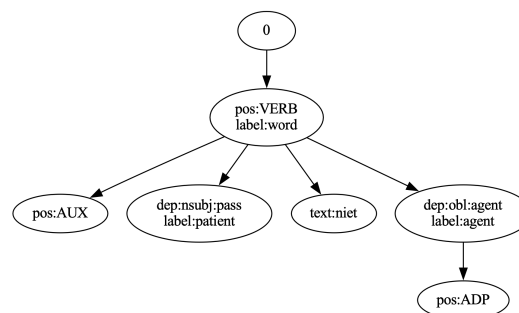


Figure 2: Pattern V1 denoting a targeted verb, auxiliary verb, agent, patient and preposition relationship. '0' points to the root. Negation is optionally matched. Pattern matching results in both verb, corresponding patient noun, corresponding agent noun. For example, *Nu zullen de kleurlingen in de Senaat door één blanke senator worden vertegenwoordigd*: yields *vertegenwoordigen* (verb), *senator* (agent noun) and *kleurlingen* (patient noun).

dataset is matched against the pattern tree shown in Figure 1. This pattern represents the simplest, most direct pattern for noun–adjective pair associations in the interest of high-accuracy results.

Noun and corresponding adjective pairs are returned. For the adjective, the lemma form is returned. For example, for the sentence fragment “Een op de vier vrouwelijke migranten werkt als ...”, yields the noun-adjective (lemma) pair, *migranten–vrouwelijk* (*migrants–female*).

3.5 Building a corpus of noun–verb pairs

Separately, for each of the years in Table 2, the corresponding spaCy dependency-parsed dataset is subject to pattern matching against the pattern trees shown in Figures 2,3,4,5. The patterns are nested in their complexity, and hence patterns are grouped within tiers as shown in Figure 6. Each node in the dependency parse is compared against each pattern, capturing noun–verb pairs according to the highest-ranked matching pattern only.

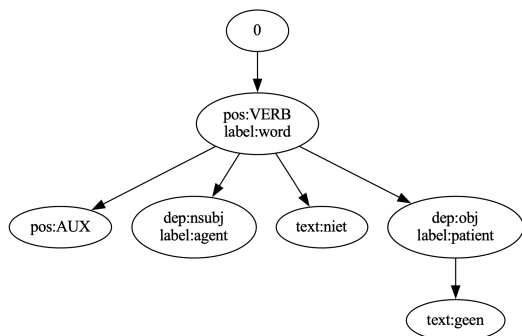


Figure 3: Pattern V2 denoting a targeted verb, auxiliary verb, agent, patient relationship. '0' points to the root. Negation is optionally matched. Resulting in verb–patient noun and verb–agent noun pairs. For example, *de negers waren verdedigd door uit het Zuiden afkomstige blanke advocaten*: yields *verdedigen* (verb), *advocaten* (agent) and *textitnegers* (patient).

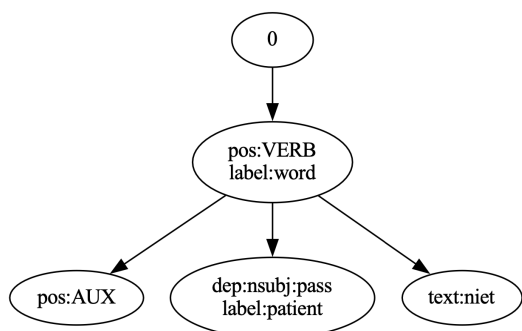


Figure 4: Pattern V3 denoting a targeted verb, auxiliary verb, patient relationship. '0' points to the root. Negation is optionally matched. Resulting in verb–patient noun pairs. For example, *terwijl jaarlijks meer dan 150.000 immigranten worden toegelaten*: yields *toelaten* (verb) and *immigranten* (patient noun).

3.6 Collective noun–context feature keyness scoring

The *keyness* scoring metric adopted in this paper, is the Log Likelihood Ratio (LLR) (Dunning, 1993). The resulting score is not based on normal approximations, and hence is applicable to low-frequency events commonly occurring in language and known generally as the Zipfian tail. The method can be thought of converting a frequency table, in our case of noun–context feature co-occurrences, to an equivalent table of scores reflective of the degree of association between the nouns and the context features. I.e., in our case, a high score reflects a context feature being particularly important to the characterisation of noun.

Effectively, we considered each noun and con-

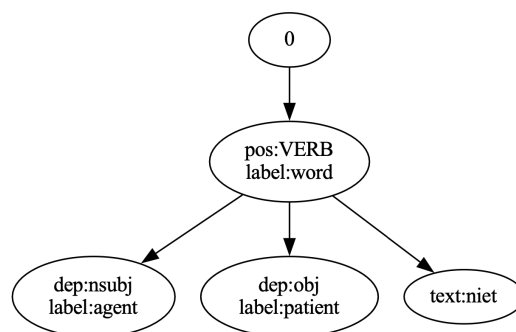


Figure 5: Pattern V4 denoting a targeted verb, agent, patient relationship. '0' points to the root. Negation is optionally matched. Resulting in verb–patient noun and verb–agent noun pairs. E.g., *waar de Berbers de Arabieren in aantal overtreffen*: yields, *Berbers* (agent noun) and *overtreffen* (verb).

count(context, noun)	count(context, noun')
count(context', noun)	count(context', noun')

Table 3: Contingency table, forming the basis of the conversion of raw frequency table values of noun–context feature co-occurrence to LLR scores reflecting how key a context is to a noun.

text feature pair (cell) in the frequency table in turn, forming a contingency table as per table 3 for each.

The contingency table thus represents the binomial outcomes of the context occurring or not occurring with respect to two sub-corpora. The left-hand column of table 3 represents all instances for the context feature type and year, which is co-occurrent with the noun in question. The right-hand column of table 3, represents all instances for the context feature type and year, which is not co-occurrent with the noun in question.

To calculate LLR, two separate generative processes are considered for each sub-corpus. Firstly, that the two sub-corpora share a common binomial probability with respect to the occurrence of the context. Secondly, that the two corpora have different, distinct binomial probabilities with respect to the occurrence of the context in question. Maximum Likelihood Estimation (MLE) estimates of the binomial probabilities for both assumed generative processes are calculated.

The LLR value is then calculated via Equation 1, where $\text{Binom}(x,y)$ denotes the binomial probability of the outcomes observed in sub-corpus x , assuming the parameters of the generative process, y , as previously described. A larger LLR value implies a greater co-location of the collective noun and context in question.

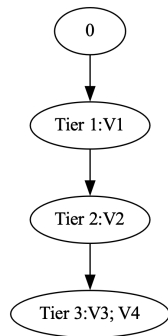


Figure 6: Pattern matching hierarchy: At each node in the spaCy dependency parse patterns are checked, moving through each tier until a pattern match is found and collecting all matches within that tier only.

$$-2.\log\left(\frac{\text{Binom}(1,1) \times \text{Binom}(2,1)}{\text{Binom}(1,2) \times \text{Binom}(2,2)}\right) \quad (1)$$

4 Evaluation and Results

The substantive output of the methodology of Section 3 are the tables of LLR associativity scores for each noun–context feature pair. These LLR scores are the basis for the evaluation methods in this section.

4.1 Pattern-matching accuracy

First, some evaluation of the accuracy of the noun–adjective and noun–verb pattern matching methodology is warranted. There are three main potential sources for error in the extracting pairs of adjectives or verbs and corresponding nouns as described in Section 3.4 and Section 3.5: OCR errors, dependency parse errors and pattern matching errors. OCR errors do not present a significant concern to this study, beyond their influence on dependency parsing performance. I.e., if a misrepresented word or artifact which otherwise looks like an adjective in terms of syntax, and is dependency parse tagged and pattern matched as such, then we simply end up with an extra nonsense context word.

Whether the pattern matching fails to correctly extract true noun–context word instances presents a greater concern. This was evaluated manually by sampling the noun–context word extracts via pattern matching of a random sample of 200 articles from the 1991 OCR set. The results are given in Table 4, and demonstrate a reasonably strong base accuracy with estimates ranging from 88% with the V4 pattern to 97% with the A1 pattern, support-

Pattern	Accuracy point estimate	Accuracy 95% Confidence Interval (Wilson)
Adjective Patterns		
A1	97% (125)	92 - 99 %
Verb Patterns		
V1	95% (66)	87 - 98 %
V2	91% (125)	85 - 95 %
V3	94% (125)	89 - 97 %
V4	88% (125)	81 - 93 %

Table 4: The results of manual evaluation of accuracy of the extracted noun–adj, and noun–verb pairs, due to combined dependency parse and pattern-matching errors. Results are rounded to 2 significant figures. The number of sample extracts for each pattern type are given in the brackets.

ing further conclusions derived from the noun and adjective or verb pair co-occurrence statistics.

4.2 Identifying high-association contexts for known-charged collective nouns

Our first research question, *do context features exist, which for some noun–context feature associativity score threshold, are highly predictive of the noun?* can be considered as a fundamental test of the base hypothesis that the methodology and dataset are sufficient to identify relevant and interesting high-association terms. It is fundamental that we can identify high-association contexts for known-charged collective nouns. We cannot draw effective parallels between terms with respect to their context features if they do not have sufficiently strong profiles.

For this research question, we adopt a high LLR threshold: For each year and for each collective noun in Table 1, we extract only those context features for which the collective noun is in the top 2 of LLR scores. For a selected number of known-charged collective nouns, the outcomes are given in Table 5. It should be reiterated here that the table is not a complete window into the all context features with a high degree association, merely those with an extremely high degree of association according to the LLR threshold. Clearly relevant, strong outcomes can be observed from this. I.e., in the case of the charged-noun, *"migrant"*, we see contextual features such as *aspirante* (aspirational), *tweede-generatie* (second-generation), *niet-geïntegreerd* (unintegrated). In the case of *"baboe"* (the general name given to nannies from Surinam), we see *zorgvol* (caring). In the more powerfully charged cases such as *"neger"*, we see a wealth of strong

known-charged nouns	modifiers associated	adjectives associated	verbs associated for which the noun is the agent	verbs associated for which the noun is the patient
afkomst	bedoeïnen: 1991; bloedgroep: 1959; dansers: 1979; hugenoten: 1989; huurkamer: 1971		aanvuren: 1969; held: 1985; molesteren: 1995; verzekeren: 1995	eemt: 1959; opsieren: 1971; raden: 1959; schreeuwen: 1991; traumatiseren: 1985; verlooehenen: 1995, 1951, 1955, 1959, 1959, 1961, 1961, 1965, 1969, 1969, 1975, 1975, 1979, 1979, 1981, 1985, 1985, 1989, 1989, 1991, 1991, 1995; verraden: 1955, 1959, 1985, 1989, 1991, 1995; zullen: 1971
allochtoon		laag-opgeleid: 1991; werkloos: 1989; werkwillig: 1991	ongemerkt: 1991	hulpbehoeven: 1989; instromen: 1989
baboe		soendanees: 1991; zorgvol: 1979	inhalen: 1951	
immigrant	commonwealth: 1971	afriaans: 1991; arriveren: 1951; bengaa's: 1981, 1985; blank: 1981; duits-joods: 1971; engts: 1955; enjels: 1951; hds: 1959; illegaal: 1971, 1979, 1995; indies: 1975; latiïnsamerikaans: 1991; miljoen: 1955; mohammedaans: 1991; niet-blanke: 1965, 1981; niet-britse: 1955; niet-geïntegreerd: 1959; niet-blank: 1965, 1985; nlet-blank: 1965; noordafrikaans: 1995; opper-egyptisch: 1981; ouz: 1959; portoricaans: 1959; roemeens-duits: 1989; russisch-joods: 1989; russischjoods: 1989; s' amitsch: 1995; salvadorlaans: 1969; siciliaans: 1961; sovjet-joods: 1991; steenrijk: 1969; steunen: 1951; urdu-talig: 1995; westindisch: 1981; ülegal: 1981	afpakken: 1989; binnensmokkelen: 1971; broeden: 1959; doodsteken: 1995; inpikken: 1989; klagen: 1951; meedragen: 1955; omsingelen: 1959; ontsluiten: 1981; overspoelen: 1989; terugbetalen: 1955; wegpikken: 1959; zjn: 1961	aankomen: 1975; afbeulen: 1985; classificeren: 1979; huisvesten: 1991; stijven: 1985; toelaten: 1951; verkijken: 1959
indiaan	amazone: 1969; apache: 1991; halfbloed: 1981; hopi: 1961, 1985; innu: 1989; miskito: 1985, 1989; navajo: 1991; noorda: 1959; oerwoud: 1981; platvoet: 1955, 1965; sioux: 1991; yanomami: 1991	amazon: 1989; benedenlands: 1965; bonairiaans: 1989; een-ogig: 1965; eenogig: 1965; eht: 1961; grondloz: 1979; ploeteren: 1991; rillen: 1959	aftroggelen: 1995; kapen: 1989; kauwen: 1991	achterstellen: 1989; afhakken: 1951; afslachten: 1969; hakken: 1965; verontwaardigen: 1989
islamiet		dox: 1979; fundamentalistisch: 1989; fundamentalistisch: 1981, 1985, 1989; imam: 1991; listisch: 1991; niet-chinees: 1989; radicaal: 1995; rechtgelovig: 1989; rechtzinnig: 1979; sjiëtisch: 1979; sjiëtisch: 1979; sunnitisch: 1989; ugandeas: 1989; weerspanning: 1959	begraven: 1985; ijgen: 1989; vasten: 1979	vluhtelingenkamp: 1985
kaffer	zoeloe: 1961, 1959, 1951	nagemaakte: 1961; roodgeverfd: 1961; tomm: 1951		
kleurling	élite: 1975	biaziliaans: 1959; en'ander: 1955; fransi: 1955; inder: 1985; kaaplants: 1955; kaaps: 1951, 1955, 1969; kroesharig: 1959; opdringerig: 1981; tussen-d: 1981	geïncasseerd: 1955; herkwijgen: 1961; overlopen: 1991	afbeelden: 1961; idealiseren: 1979; integreren: 1975; selecteren: 1979; tusaen: 1951; verwarren: 1955; volmaken: 1955
koeli	raat: 1975; riksha: 1961; riksj: 1959, 1969	doodarm: 1965; halfnaakt: 1965; rijkgekled: 1955	fouillieren: 1951; spijten: 1951; voortrekken: 1961	
migrant	aspirante: 1951, 1955, 1959, 1989; heimwee: 1971; illen: 1981; irh: 1965; lrn: 1995; niete: 1951, 1955; plattelandse: 1955; proefe: 1959; spyt: 1961; tweede-generatie: 1995	afro-caribisch: 1995; erkloz: 1991; haïtiaans: 1979; marokkaans: 1989, 1991; niet-geïntegreerd: 1991; ondef: 1959; onvolwaardig: 1991; rokkaans: 1989; turks: 1989		
mohammedaan	sja: 1965	anti-eomunistisch: 1965; fans: 1955; inpopulair: 1971; kameruens: 1979; orthodox-radical: 1955; pro-frans: 1959, 1961; sjiëtisch: 1975	bestrijden: 1995	
neger	bakongo: 1959; bantoe: 1955; benton: 1959; bos: 1955, 1975, 1989; congo: 1955; goudkust: 1955; grun: 1955, 1959, 1961, 1965, 1969, 1971, 1975, 1979, 1981, 1985, 1989, 1991; irun: 1951; mississippi: 1955; overliden: 1959; panko: 1951; soedan: 1961, 1965; soweto: 1979; Watts: 1965	abject: 1951; afrikaans: 1955; afro-amerikaans: 1995; amerikaane: 1959; armlastig: 1965; bevoogden: 1965; blootvoetig: 1955; diepbruin: 1965; eerbaar: 1995; golv: 1961; gracieus: 1985; ifrikaans: 1959; kiesgerechtigd: 1965; langbenig: 1955; lynchen: 1959; militant: 1971; miloz: 1965; negenenvijftig: 1981; noord-amerikaans: 1955; onschendbaar: 1961; oproerig: 1971; overigen: 1961; rfd: 1971; seigneurial: 1961; senegalees: 1951; sluip: 1955; stokoud: 1985; tiiaans: 1979; west-afrikaans: 1955; wetsgetrouw: 1961; zelfbewust: 1969; zuidrhodesisch: 1965; zuidoedanees: 1971; ûntwikel: 1965	aandrukken: 1969; bijeenrapen: 1965; ebben: 1951; hf: 1961; inj: 1969; inladen: 1975; openscheuren: 1969; plunderen: 1969; straffen: 1965; toebedelen: 1955; t ransponeren: 1955; uitzingen: 1965	aftuigen: 1965; bespreken: 1979; canoniseren: 1961; contra: 1959; doodschieten: 1965, 1981; executeren: 1951; gelijkberechtigd: 1965; inschepen: 1955; inschrijven: 1965; kamperen: 1951; lynchen: 1951, 1959; roven: 1979; slaven: 1951; terechtstellen: 1959; tiranniseren: 1961; uitmoorden: 1969; verafschuwen: 1969; verdrukken: 1965; vermengen: 1961; verschillen: 1955; voortrekken: 1969; weren: 1955

Table 5: Selected known-charged nouns of table 1, with together with (all) context features for which the noun-context LLR associativity score is in the top 2 for that context feature.

Modifiers	known-charged nouns associated with the modifier according to the criteria of 4.3 (as head in the compound word)
"nomaden"	indiaan: 1965 stam: 1965, 1951, 1955, 1959, 1961, 1969, 1971, 1975, 1979, 1981, 1985, 1989, 1991, 1995
magazijn	bediende: 1951, 1955, 1959, 1961, 1965, 1969, 1971, 1975, 1979, 1981, 1985, 1989, 1991, 1995
"indianen"	stam: 1951, 1955, 1959, 1961, 1965, 1969, 1971, 1975, 1979, 1981, 1985, 1989, 1991, 1995
"neger"	slaaf: 1955, 1971, 1979, 1981, 1975, 1985, 1989, 1991, 1995 stam: 1955, 1971, 1979, 1981, 1951
boom	stam: 1951, 1955, 1959, 1961, 1965, 1969, 1971, 1975, 1979, 1981, 1985, 1989, 1991, 1995
pape	ras: 1951, 1955, 1959, 1961, 1965, 1969, 1971, 1975, 1979, 1981, 1985, 1989, 1991
grun	neger: 1951, 1955, 1959, 1961, 1965, 1969, 1971, 1975, 1979, 1981, 1985, 1989, 1991
pomp	bediende: 1955, 1959, 1961, 1965, 1969, 1971, 1975, 1979, 1981, 1985, 1989, 1991, 1995
joego	slaaf: 1951, 1955, 1959, 1965, 1969, 1971, 1975, 1979, 1981, 1985, 1989, 1991
bantoe	bediende: 1961 neger: 1959, 1961, 1955 ras: 1959, 1951 stam: 1959, 1961, 1951, 1955, 1971, 1975
zeeg	ras: 1951, 1955, 1959, 1961, 1965, 1969, 1975, 1981, 1985, 1989, 1991, 1995
bosland	creool: 1989, 1955, 1959, 1965, 1969, 1975, 1979, 1991, 1995 indiaan: 1989 neger: 1989, 1955
ether	piraaf: 1959, 1961, 1969, 1971, 1975, 1979, 1981, 1985, 1989, 1995
mons	trans: 1951, 1955, 1959, 1961, 1965, 1969, 1971, 1975, 1979
hatte	ras: 1951, 1955, 1959, 1965, 1969, 1971, 1975, 1981, 1985
achte	ras: 1959, 1965, 1969, 1971, 1975, 1979, 1981, 1985, 1991
berber	stam: 1951, 1955, 1959, 1961, 1965, 1969, 1971, 1985, 1989
zoeloe	kaffer: 1961, 1951 neger: 1961 stam: 1961, 1979, 1981, 1985 stamhoofd: 1961, 1989
loket	bediende: 1959, 1961, 1965, 1971, 1975, 1979, 1981, 1985, 1995
bacterie	stam: 1955, 1975, 1979, 1981, 1985, 1989, 1991, 1995
dart	moor: 1951, 1955, 1959, 1961, 1965, 1979, 1989, 1991
voortrekkers	stam: 1951, 1955, 1959, 1961, 1965, 1971, 1985
bosneger	stam: 1989, 1959, 1975, 1979, 1991, 1995 stamhoofd: 1989
papoea	stam: 1951, 1955, 1959, 1965, 1971, 1981, 1989
bos	neger: 1951, 1955, 1969, 1975, 1979, 1989
bedoeïenen	stam: 1959, 1961, 1971, 1985, 1991, 1995
ex-e	migrant: 1955, 1961, 1965, 1985, 1989, 1991
amazone	indiaan: 1961, 1969, 1975, 1979, 1991 stam: 1965
bel	indo: 1979, 1981, 1985, 1989, 1991, 1995
kantoor	bediende: 1951, 1955, 1959, 1961, 1965, 1969

Table 6: known-charged nouns (as compound word heads), and the common modifier they are associated with according to the criteria in 4.3. This table only lists instances of 6 or more associated noun and year instances. Modifiers that are themselves known-charged words are marked as such; italicized strings are decomposition errors.

context features, such as: tribal names, *lynchen* (to lynch), *militant* (militant), *kiesgerechtigd* (being eligible to vote), *executeren* (to execute) and *plunderen* (to plunder).

4.3 Identifying context features for multiple known-charged collective nouns

Our second research question, *are there context features which for some noun–context feature associativity threshold, recall multiple known charged*

nouns with a reasonable degree of precision with respect to our known charged noun set? is concerned with whether the methodology is able to find common, meaningful associations that hold across known-charged words. To consider contextual feature overlap between known-charged collective nouns, we must adopt a less severe criterion allowing for overlap. For each context feature type (e.g., modifiers), for each year and for each context feature, the corresponding collective nouns are traversed, according to their descending LLR score, and every noun above a LLR threshold is accepted. This results in a precision of 0.2, taking the Table 1 known-charged nouns as true positives.

Sample outcomes of this approach are given in Tables 6, 7, 8 and 9, corresponding to modifiers, verbs for which the noun is the patient, verbs for which the nouns are the agent and adjectives. Each table represents a selection of context features from a larger set, listing for those context features with the most year–context feature instances associated. The tables are otherwise in no way curated. Examination of the tables again shows some powerful associations between known-charged words over time frames. For example: in Table 7, "*gekleurd*" (coloured), "*immigrant*" (immigrant) and "*zigeuner*" (gypsy) peoples being subject to *deporteren* (to deport); in Table 8, "*immigrant*" (immigrant) peoples in 1975 and 1995 are associated with action of *overstromen* (to flood); and in Table 9: "*indiaan*" (indian) and near continuously over a large time window, "*stam*" (tribe) associated with "*primitief*" (primitive).

Table 6 shows that some of the modifiers that are discovered are known-charged words themselves. Table 6 also includes a number of modifiers that do not refer to a strongly related word, but are the result of an incorrect morphological decomposition; e.g. the charged word "*ras*" was mistakenly detected in words ending in the stem *as* (axis) or *gras* (grass), producing the incorrect assumed modifiers *zeeg* and *achte*. Either a lexical filter or a better morphological decomposition would allow filtering out these cases.

4.4 Discovering charged nouns from their common associations with known-charged nouns

Our final research question is "*Can the context features of known charged nouns, help identify other charged instances?*". Considering each year, and

Verbs for which the noun is the patient	known-charged nouns associated with the verbs according to the criteria of 4.3 (and the years they are associated)
NOTverloochenen	afkomst: 1951, 1955, 1959, 1961, 1965, 1969, 1975, 1979, 1981, 1985, 1989, 1991, 1995
verloochenen	afkomst: 1969, 1989, 1995, 1951, 1959, 1961, 1975, 1979, 1985, 1991 oorsprong: 1969, 1995 roots: 1989
verraden	afkomst: 1955, 1965, 1975, 1981, 1985, 1989, 1991, 1995
lynchen	neger: 1951, 1959, 1975, 1995
slaven	bediende: 1985 neger: 1951, 1971, 1975
voortrekken	islamiet: 1969, 1955 neger: 1969, 1975
terechtstellen	neger: 1951, 1959, 1965
doodschieten	neger: 1959, 1965, 1971
fokken	ras: 1959, 1961, 1975
ronselen	inboorling: 1979 indiaan: 1979 koeli: 1951
achterstellen	bosneger: 1989 indiaan: 1989 zigeuner: 1989
uitroeien	indiaan: 1989 stam: 1989 zigeuner: 1959
NOTverwarren	kleurling: 1955 primitief: 1975 ras: 1959
deporteren	gekleurd: 1969 immigrant: 1985 zigeuner: 1979
legaliseren	immigrant: 1955 piraat: 1981 zigeuner: 1981

Table 7: known-charged nouns, associated verbs and the years of association according to the criteria defined in 4.3, where the noun is the patient to the verb. The *NOT* prefix denotes negation of the verb. This table lists only those instances of 3 or more associated noun and year instances.

each context feature separately, and setting an LLR threshold with respect to the context feature as described in section 4.3, all corresponding nouns are extracted. Where a noun is coincident with a known-charged noun of Table 1, this pairwise association is recorded, together with the context feature and year responsible for the association.

The leftmost column of Table 10 provides clues for answering our third sub-question: whether we can automatically discover new candidate terms for our charged word list. The column in the table exhibits a small outtake of a list of 6,310 unique words that frequently occur in the same morpho-syntactic role as our charged words, along with their specific linguistic contexts. A manual inventory of this word list reveals a candidate set of about 10 new charged terms, including "*joden*" (*jews*), "*indianen*" (*indians*), "*moslims*" (*muslims*), and "*slaviër*" (*slav*). Other charged terms occurring in this list refer to nazism and radical movements such as "*SS*" and "*RAF*", and include formerly used terms for immigrant workers, such as "*gastarbeider*" (literally *guest worker, immigrant worker*). It takes manual inspection and expertise to extract

Verbs for which the noun is the agent	known-charged nouns associated with the verbs according to the criteria of 4.3 (and the years they are associated)
enteren	piraat: 1955, 1959, 1989
doodsteken	bediende: 1989 immigrant: 1995 wildeman: 1989
NOTleven	blanke: 1971, 1991 piraat: 1969
neerzetten	bediende: 1985 inboorling: 1951 zigeuner: 1979
inbegrijpen	homo: 1961 koppensneller: 1961
infecteren	neger: 1961 ras: 1991
ongemerkt	allochtoon: 1991 neger: 1971
serveren	bediende: 1955, 1959
uitgooien	neger: 1991 piraat: 1961
uitmoorden	blanke: 1969 indiaan: 1995
NOTvergeten	immigrant: 1955 zigeuner: 1961
herkrijgen	kleurling: 1961 slaaf: 1979
uitzingen	neger: 1959, 1965
aanbidden	blanke: 1959 slaaf: 1969
verkrachten	piraat: 1989 wildeman: 1981
stichten	immigrant: 1989 stam: 1989
boren	piraat: 1979 stam: 1965
zeulen	dwerg: 1955 inboorling: 1959
kidnappen	indiaan: 1985 stam: 1985
NOTdrinken	indiaan: 1985 mohammedaan: 1961
binnensmokkelen	immigrant: 1971, 1981
bejegenen	barbaar: 1961 kleurling: 1971
overstromen	immigrant: 1975, 1995

Table 8: Known-charged nouns, associated verbs, and the years of association according to the criteria defined in 4.3, where the noun is the agent to the verb. The *NOT* prefix denotes negation of the verb. This table lists only those instances of 2 or more associated noun and year instances.

these term from this larger list of terms, of which the majority consists of general, uncharged, high-frequency words for family relations, demographic groups, locations, government, occupations, culture, religion, tradition, and arts — all to be expected, given that these are all hypernyms of our charged terms and occur in the same linguistic and semantic contexts.

5 Discussion and Conclusion

The paper posed three research questions which we can paraphrase as: do simple metrics of word associativity yield distinctive context profiles; can these context profiles be used to draw parallels between known-charged nouns; and finally, can we identify candidate charged nouns. Somewhat inherent to the complexity of the notion of a term being *charged* is that there exists no definitive gold standard dataset from which we are able to evaluate

Adjectives	known-charged nouns associated with the adjectives according to 4.3 criteria (and the years they are associated)
indiaans	afkomst: 1969, 1979, 1989, 1955, 1975, 1985, 1991, 1961 halfbloed: 1959 medicijnman: 1979, 1989, 1991, 1995 ras: 1979 scalp: 1969 slaaf: 1969 stam: 1969, 1979, 1959, 1989, 1955, 1975, 1985, 1971 stamhoofd: 1969, 1959
germaans	afkomst: 1959 barbaar: 1959 oorsprong: 1959, 1969 ras: 1959, 1969, 1951, 1961, 1965, 1975, 1979, 1985, 1989 stam: 1959, 1969, 1951, 1961, 1965, 1975, 1979, 1985, 1989, 1955, 1971, 1991, 1995
arisch	afkomst: 1951, 1955 ras: 1959, 1965, 1969, 1975, 1979, 1981, 1985, 1989, 1991, 1995 stam: 1951
hindostaans	Afkomst: 1965, 1969, 1951, 1955, 1959, 1961, 1975, 1979, 1985 immigrant: 1965, 1969 migrant: 1989
primitief	indiaan: 1979 stam: 1979, 1951, 1959, 1965, 1969, 1971, 1975, 1981, 1985, 1989, 1995
resistent	ras: 1985, 1989, 1955, 1959, 1961, 1965, 1971, 1991 stam: 1985, 1989
nederig	afkomst: 1981, 1955, 1959, 1969, 1971, 1979, 1985, 1989, 1991 slaaf: 1981
russisch-joods	afkomst: 1985, 1989, 1991, 1961, 1971 immigrant: 1985, 1989, 1991, 1979 oorsprong: 1985
polair	oorsprong: 1951, 1959, 1961, 1965, 1969, 1971, 1975, 1979, 1981, 1985
minderwaardig	ras: 1961, 1965, 1969, 1975, 1979, 1981, 1985, 1991, 1995
indo-europees	afkomst: 1971, 1991, 1955, 1979, 1985 oorsprong: 1971 stam: 1991, 1961
pools-joods	afkomst: 1991, 1951, 1979, 1981, 1985, 1989 immigrant: 1991, 1959
niet-nederlands	afkomst: 1989, 1991, 1979, 1985, 1995 immigrant: 1991 oorsprong: 1989
armeens	afkomst: 1965, 1969, 1975, 1981, 1985, 1995 immigrant: 1965
subtropisch	oorsprong: 1959, 1961, 1965, 1969, 1971, 1975, 1981
goddelijk	oorsprong: 1959, 1965, 1969, 1971, 1981, 1989, 1991
duits-joods	afkomst: 1955, 1965, 1961, 1995 immigrant: 1955, 1965, 1971
noordafrikaans	afkomst: 1991, 1995 immigrant: 1991, 1995, 1985, 1989 migrant: 1991
oriëntaals	afkomst: 1971, 1981 immigrant: 1955, 1971 oorsprong: 1975 ras: 1955
illegaal	immigrant: 1955, 1971, 1979, 1985, 1989, 1995
negroïde	afkomst: 1989 ras: 1971, 1989, 1975, 1991 stam: 1971
keltisch	oorsprong: 1955, 1959, 1965, 1969 ras: 1951 stam: 1951
NOTnederlands	afkomst: 1989, 1991, 1979 oorsprong: 1989, 1991, 1961

Table 9: Known-charged nouns, associated adjectives and the years of association according to the criteria defined in 4.3. This table lists only those instances of 6 or more associated noun and year instances.

the methodology output on a purely numerical basis. Charged term detection, and an understanding of the manifest attributes that make terms charged, remains an open problem (and perhaps always will be). Consequently, any evaluation of methods used to answer the research question must inevitably rely on a degree of outside-of-data, human interpretation. On the basis of the observed associations and the links we can recognize, we contend that

the evaluation results are sufficiently strong to be able to answer all of the research questions in the affirmative. Additionally, the results in regards to supporting the methodology are supported by the fact that ultimately the basis of methods is simple, time-tested, and entirely open to inspection (being based on co-occurrence counts).

The underlying context in which the research questions were posed, was the application of digital humanities to help humanities scholars in exploring and charged language. The utility being the ability of condense many millions of narrative descriptions into a much smaller number of salient associations for human consideration. In this regard, the evaluation results tables in this document (and the complete versions, with english translations, available on the [Github repository](#)), can be viewed as reference set of associations. However, the results correspond to the specific (and arguably quite restrictive) LLR associativity score thresholds adopted for the purpose of method evaluation. It is envisaged that the methodology could be used on a more adhoc basis by humanities scholars in exploring context features and overlaps: where the outputs could be used as a both a reference with a probabilistic basis, but also as a pointer to consider axes of contentiousness at a high, human-expert level. For example, in the Words Matter publication in relation to "*stam*" it is noted that (translation): "The term tribe is often associated with a so-called not complex society with a simple political structure. although this fact in itself is not disputed, the term has the connotation of primitive". We see this precise association in our results: in Table 9, the adjective and known-charged term "*primitief*" is shown to be associated with "*stam*" in the newspaper articles consistently through the 1950s through the 1990s. In the case of "*mohammedaan*", the Words Matter document details objections to the term on the basis of religious objects: but we also see context associations such as *orthodox-radical* (Table 5) which may or not provide further avenue for which contentiousness its contentiousness can be considered. In the case of "*neger*", the Words Matter document notes the associations of the word with the sub-Saharan African peoples, but more problematically with racial stereotyping. Again, we see this as an output from the methodology in the table 5 profile of the term: *bakongo, bantoe, congo, goudkust, soedan; blootvoet, lynchen, militant*. Furthermore, the results of Table 5 allow us

noun	known-charged noun	verbs associated for which the nouns are patients	verbs associated for which the nouns are agents	adjectives associated	modifiers associated
bevolking	afkomst			indiaans: 1969, 1979, 1989, 1955, 1975, 1985, 1991 hindostaans: 1965 creools: 1959 hindoestaans: 1961 papoeaas: 1961 albanees: 1981, 1989	
bevolking	bediende	opschrikken: 1971		niet-blanke: 1971	bantoe: 1961 heger: 1969
bevolking	blanke			kiesgerechtigd: 1985	
bevolking	creool				bosland: 1989, 1975
bevolking	halfbloed			indiaans: 1959	
bevolking	immigrant			hindostaans: 1965 niet-blanke: 1971, 1965, 1981 nietblank: 1965, 1985 straatarm: 1981	
bevolking	indiaan	uitroeien: 1989			bosland: 1989
bevolking	inlander	ophitsen: 1959			
bevolking	kleurling			kiesgerechtigd: 1985	
bevolking	koppensneller			maleis: 1951	
bevolking	medicijnman			indiaans: 1979, 1989, 1991, 1995	
bevolking	migrant			nietblank: 1971	
bevolking	neger	ophitsen: 1959		autochthon: 1961 kiesgerechtigd: 1965	bantoe: 1959, 1961 bosland: 1989
bevolking	ras			indiaans: 1979 negroïde: 1971	bantoe: 1959
bevolking	scalp			indiaans: 1969	
bevolking	slaaf			indiaans: 1969	neger: 1955, 1971, 1979, 1975, 1985
bevolking	stam	uitroeien: 1989 geevacueerd: 1965 uitmoorden: 1985		indiaans: 1969, 1979, 1959, 1989, 1955, 1975, 1985, 1971 berbers: 1955 negroïde: 1971 inheems: 1965, 1981	bantoe: 1959, 1961, 1975 neger: 1955, 1971, 1979 bosneger: 1989, 1991, 1995 papoea: 1951, 1955, 1959, 1965, 1989 nomaden: 1969 eskimo: 1985
bevolking	stamhoofd			indiaans: 1969, 1959 berbers: 1955	bosneger: 1989
beweging	blanke				mau-mau: 1959
beweging	inboorling			oproerig: 1959	
beweging	indiaan			opstandig: 1989	
beweging	islamië			fundamentalistisch: 1985, 1989, 1991	
beweging	migrant			russisch-talig: 1989	
beweging	neger			oproerig: 1959 opstandig: 1971	
beweging	oorsprong			vincentiaans: 1971	
beweging	stam			opstandig: 1971, 1959 oproerig: 1955	zulu: 1991 zoeloe: 1985
beweging	stamhoofd				zulu: 1991
joden	afkomst			oriëntaals: 1971, 1981 hongaars: 1969	
joden	blanke				nlet: 1961
joden	gekleurd	deportereren: 1969			
joden	immigrant	deportereren: 1985		oriëntaals: 1971 ethiopisch: 1985, 1991	
joden	islamië			orthodox: 1981	
joden	neger	lynchen: 1975			
joden	oorsprong			oriëntaals: 1975	
joden	stam	uitmoorden: 1985			
joden	zigeuner	deportereren: 1979		staatloos: 1961	
kwestie	afkomst			c'al: 1991	
kwestie	blanke			rhodesisch: 1981	
kwestie	kaffer				zoeloe: 1961
kwestie	neger				zoeloe: 1961 soedan: 1951
kwestie	stam				zoeloe: 1961
kwestie	stamhoofd				zoeloe: 1961
communisten	barbaar			bloeddorstig: 1965	
communisten	dwerg			bloeddorstig: 1965	
communisten	indiaan		uitmoorden: 1995		
communisten	islamië			dox: 1979	

Table 10: Pairs of known-charged and other nouns as related by verbs, adjectives and modifiers, according to the associativity criteria defined in 4.4. This table represents only a demonstrative sample.

to extend the characterisation with detailed actions this collective noun term has been subjected to: *doodschieten* (shoot dead), *terechtstellen* (execute), *uitmoorden* (massacre) and *verdrücken* (oppress).

There is scope to further elaborate on, and strengthen the resulting context-feature profiles captured over a corpus. First and foremost, further work into the pattern matching routines such to expand the number of adjectives and verbs cap-

ured, whilst maintaining a high degree of accuracy. This is especially true of some of the most obvious and basic noun and verbs for which the noun is agent patterns, which we excluded from this study for yielding notably lower accuracy than other the patterns included in the study. However, there are other contexts that may be interesting and indicative of being charged: for instance context features which capture more information of the environs as

part of the narrative account of nouns (and known-charged nouns).

Lastly, whilst newspapers represent one particular narrative account type of people groups, other discourse types (such as literature) may yield rival or complementary accounts useful to humanities scholars.

Acknowledgements

This work was funded by NWO in the ‘Culturally Aware AI’ project.

CRedit Author Statement:

R. Brate: Conceptualization, data curation, formal analysis, writing - original draft. M. van Erp: funding acquisition, project administration, supervision, writing - review & editing. A. van den Bosch: funding acquisition, supervision, writing - review & editing.⁵

References

- David Bamman, Brendan O’Connor, and Noah A. Smith. 2013. [Learning latent personas of film characters](#). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 352–361, Sofia, Bulgaria. Association for Computational Linguistics.
- Gioele Barabucci, Francesca Tomasi, and Fabio Vitali. 2020. Supporting complexity and conjectures in cultural heritage descriptions. In *Proceedings of the International Conference Collect and Connect: Archives and Collections in a Digital Age*, volume Vol-2810 of 1613-0073, pages 104–115, Leiden, the Netherlands.
- Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.
- Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The values encoded in machine learning research. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 173–184.
- Ryan Brate, Andrei Nesterov, Valentin Vogelmann, Jacco van Ossenbruggen, Laura Hollink, and Marieke van Erp. 2021. [Capturing Contentiousness: Constructing the Contentious Terms in Context Corpus](#). In *Proceedings of the 11th on Knowledge Capture Conference, K-CAP ’21*, pages 17–24, Virtual Event, USA. Association for Computing Machinery.
- Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Comput. Linguist.*, 19(1):61–74.
- Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.
- Hannah Kirk, Abeba Birhane, Bertie Vidgen, and Leon Derczynski. 2022. [Handling and presenting harmful text in NLP research](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 497–510, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Mrinalini Luthra, Konstantin Todorov, Charles Jeurgens, and Giovanni Colavizza. 2023. [Unsilencing colonial archives via automated entity recognition](#). *Journal of Documentation*, ahead-of-print(ahead-of-print).
- Wayne Modest and Robin Lelijveld. 2018. [Words matter: an unfinished guide to word choices in the cultural sector](#). Technical report, The National Museum for World Cultures (Tropenmuseum, Afrikamuseum, Museum Volkenkunde, Wereldmuseum).
- Shakir Mohamed, Marie-Therese Png, and William Isaac. 2020. [Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence](#). *Philosophy & Technology*, 33(4):659–684.
- Temi Odumosu. 2020. [The Crying Child: On Colonial Archives, Digitization, and Ethics of Care in the Cultural Commons](#). *Current Anthropology*, 61(S22):S289–S302.
- Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. [Social bias frames: Reasoning about social and power implications of language](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.
- Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. [Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in NLP](#). *Transactions of the Association for Computational Linguistics*, 9:1408–1424.
- Anna Schmidt and Michael Wiegand. 2017. [A survey on hate speech detection using natural language processing](#). In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.
- Melvin Wevers. 2019. [Using word embeddings to examine gender bias in Dutch newspapers, 1950-1990](#). In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 92–97, Florence, Italy. Association for Computational Linguistics.

⁵<https://credit.niso.org>

The Cardamom Workbench for Historical and Under-Resourced Languages

Adrian Doyle and Theodorus Fransen and Bernardo Stearns and
John P. McCrae and Oksana Dereza and Priya Rani

Data Science Institute
University of Galway
Galway, Ireland

Abstract

This paper describes the creation of a workbench tool designed to make technologies developed throughout the lifespan of the Cardamom project easily accessible to researchers who could most benefit from them, but who may not have the technical expertise to apply bleeding edge technologies to their own datasets. The workbench provides an intuitive graphical user interface (GUI) and workflow which abstract users away from underlying technical tasks, while providing them with a suite of powerful NLP tools developed by the Cardamom team. These include tokenisers, POS-taggers, various annotation tools, and ML models. The performance of workbench tools can be improved as text and annotations are added by users. It is envisioned that this workbench will provide a simple route to digital publication for academics in the humanities, or more specifically, for linguists working with under-resourced or historical languages, who have collected text data but are unable to make it available online as a result of financial or technical restraints. This has the added benefit of increasing the availability of high quality, annotated text data to NLP researchers, thereby providing value to both communities of researchers.

1 Introduction

Some of the most cutting edge Machine Learning (ML) and Natural Language Processing (NLP) techniques require large quantities of data for use in training and testing increasingly complex models (Brown et al., 2020; Shoeybi et al., 2019; Patil et al., 2022). A relative abundance of digital text data is readily available for some of the most widely used world languages, however, it is well established that many of the world's languages are severely under-resourced in terms of technologies to support language use (Bender, 2019; Joshi et al., 2020; Hedderich et al., 2021). As more complex resources, like machine translation tools, are built upon the

foundation of rudimentary resources, like parallel corpora, a vicious cycle can emerge whereby under-resourced languages remain under-resourced, while resources for better resourced languages multiply.

Many of the most severely under-resourced languages can lack even a sufficiently large corpus of machine-readable text, never mind resources like tokenisers, part-of-speech (POS) taggers, and more advanced processing tools. NLP researchers are forced to either abandon the hope of developing ML models for such languages, or to devote time to creating basic resources like text corpora. For this reason Cieri et al. warn that, "If the language has too few resources, the project could mire in [language-resource] creation" (2016, 4548). At the same time, linguistic researchers often accumulate text which, for a variety of reasons, they may be unable to make easily accessible to other researchers. Quantities of text, which may not be substantial enough to justify a print edition, are regularly produced during the course of research projects, and it can be difficult for researchers to make these texts available online if they do not have access to the required technical skills, funding or IT resources. As such, texts are often abandoned once research projects conclude. In the case of under-resourced languages, such texts could be particularly valuable in the creation of NLP tools like spell-checkers and machine translation resources. They could be harnessed to improve research prospects for humanities scholars working with languages for which little technology is readily available.

The aim of this paper is to present a workbench tool designed to provide linguistic researchers with easy access to NLP tools developed by Cardamom researchers, and to reduce the barrier to entry for digital publication of their texts. As such, these tools include preprocessing tools like tokenisers and POS-taggers, annotation tools so that a wide variety of metadata can be stored, as well more complex tools such as word-embedding models

which improve search and query options for corpora. Section 2 of this paper will discuss the value and availability of digital text resources. Section 3 will give an overview of the Cardamom project. The state of digital text availability for historical languages will be discussed as a case study in section 4. It will be demonstrated that there exist certain obstacles to the production of freely available digital text which could be harnessed to improve ML resources. Section 5 will describe the workbench itself, and how it aims to overcome these obstacles.

2 Resources and Research Communities

It is self-evident that linguistic researchers, whether their focus be on language processing or traditional linguistics, stand to benefit from freely available and easily accessible digital text corpora. Such corpora can be used as teaching aids for language students, and many traditional avenues of linguistic research can be improved or supported by the availability of a machine readable corpus of text (Lynn, 2012). For NLP researchers, ever larger quantities of digital text are becoming more important as computer processing power improves and state-of-the-art techniques become more reliant on large quantities of training data. For example, Villegas et al. report that "CLARIN NLP services prove efficient when processing large corpora but large corpora are not always available" (2012, 3287). Where text data is available to NLP researchers, they in turn can develop tools to support or enhance traditional linguistic research areas. Areas of study such as linguistic typology and syntax greatly benefit from corpus-based and data-driven research (Nivre, 2015; Alves et al., 2023). Tools for machine translation, as well as machine-readable lexicons, for example, can greatly reduce the time-investment required for otherwise laborious tasks, allowing scholars more time to focus on research questions. These same tools' performance can be improved further as larger quantities of text data become available.

Despite the clear benefits to both research communities, NLP and traditional linguistics, close cooperation between the two is not necessarily easy to coordinate. As will be demonstrated in subsection 4, it is often difficult for humanities-based researchers to ensure text data they may have accumulated can be made available and remain easily accessible. In some instances, it will be shown,

it may even be beneficial to researchers to avoid creating digital text corpora. On the other side of the house, NLP researchers are often content to demonstrate improved results over state-of-the-art techniques in some task or research area, however, it is not always prioritised that these improved techniques are easily accessible to those who stand to benefit from them. McGillivray et al. "draw attention to the lack of communication between the communities of NLP and DH" and further suggest that "In spite of its damaging effect on the progress of the disciplines, we believe this lack of communication and miscommunication are underestimated" (2020). It is almost meaningless from the perspective of a language community to demonstrate even significant improvements in an NLP area, like machine translation for example, if members of that community must become proficient in one or more programming languages, as well as command line interface, before they can benefit from it. This is not to mention the types of troubleshooting and version control issues which can often cause headaches even for highly technically proficient NLP researchers. The workbench which is the focus of this paper aims to empower researchers to work more closely together and ultimately provide beneficial resources to both camps.

3 Cardamom Project

The Cardamom project (McCrae and Fransen, 2019) got underway in 2019 with the aim of developing deep-learning-based NLP techniques to close the resource gap for historical and otherwise under-resourced languages. Throughout the project's lifespan Cardamom technologies have been applied in a variety of areas ranging from text preprocessing tasks like tokenisation (Doyle et al., 2019) to sentiment analysis (Chakravarthi et al., 2020) and detection of language and dialect (Goswami et al., 2020; Rani et al., 2022). Cardamom research has focused on reducing resource requirements, both for data and for processing power, with the aim of reducing the NLP barrier to entry for under-resourced languages. This has been accomplished by developing more efficient approaches to common tasks (Goswami et al., 2021a,b) as well as by exploiting commonalities between closely related languages to improve NLP prospects for individual low-resource languages (McCrae et al., 2021).

In aiming to improve language processing prospects for both under-resourced modern lan-

guages and historical ones, Cardamom is unlike many other projects. Because historical language stages can form diachronic links between modern languages, the benefits of transfer learning can be exploited not only laterally, from one modern language to another, but temporally forward and backward also, adding new dimensionality to such NLP solutions (Dereza et al., 2023b). Inclusion of historical language stages as a means of bridging divides between modern languages which have descended from them is a somewhat novel solution, and promises to bolster further research areas such as computer-assisted diachronic terminology mapping.

As historical languages are typically very under-resourced themselves, they too stand to gain from research which aims to reduce resource requirements for NLP. Moreover, historical languages can present challenges which are not common in modern languages. One such example is that many features of manuscript orthography are unsupported by modern standards like Unicode which "gives higher priority to ensuring utility for the future than to preserving past antiquities" (Becker, 1988, 5) and therefore, "aims in the first instance at the characters published in modern text". Therefore, many such features cannot be accurately or consistently captured in digital text without employing workarounds like discreet annotations (Doyle et al., 2018, 69–70). Another example relates to orthographies which predate the standardisation typical of modern languages. These can result in a high degree of spelling variation in historical language texts, which can be particularly problematic when processing languages which are morphologically rich (Dereza et al., 2023a). Moreover, in languages which predate modern word separation using spacing, even fundamental tasks like tokenisation can pose significant difficulties (Doyle et al., 2019).

Issues such as these have been the subjects of investigation during the course of the Cardamom project. Problem areas specific to historical languages, which have to date received little attention, have been addressed and technologies have been developed to meet the specific needs of these and other under-resourced languages (see subsection 5.2). The focus of the Cardamom project has now shifted to ensuring these technologies are easily accessible to users who may find value in them.

4 Historical Languages; a Case Study

Historical languages like Old Irish and Old English suffer from many of the same resource deficits which afflict modern under-resourced languages. As no communities of native speakers exist for these languages, no new text can be generated by native speakers. Instead, NLP researchers must rely primarily on text which has survived for centuries or even millennia, from the times when these languages were still in use. Such texts are generally preserved in manuscripts, or in some cases, engravings in stone, clay and other materials. By the very nature of their antiquity, such sources of text can be scarce. Even where a text has survived, however, a digital transcription of it may not be available to NLP researchers.

Typically, historical linguists who transcribe the contents of a manuscript will aim to release the resulting text as a print edition rather than in digital format. There are many valid reasons for this, chief amongst which may be the perception that it is more advantageous to produce texts in print. Stifter et al. stress the importance of "ensuring that scholars receive due credit for their work for the purposes of career progression" (2021, 17), and it stands to reason that scholars will aim to produce whichever form of publication is more likely to receive engagement in the form of peer reviews and citations. However, Stifter et al. also identify "a reluctance to rely on and cite digital resources" (2021, 10) among linguists working with historical Gaelic varieties, "particularly when there is a print alternative, even if more out of date". This reluctance appears to be rooted in the belief that such resources are somewhat unreliable or capricious, and Stifter et al. report that "the perceived authority and trustworthiness of digital resources" (2021, 17) was a recurring theme in their workshop. Scholars do not feel confident citing a resource which they believe could be altered at any time, with little warning or oversight. Unfortunately, for as long as there is a reluctance to interact with digital resources by humanities scholars, linguists will be actively incentivised to generate print editions at the expense of digital text resources. This, in turn, contributes to a shortage of digital text available to NLP researchers for historical languages.

Other technical factors also play a role in preventing the generation of digital text for historical languages. It is no secret that "Digital resources are expensive both to build and maintain" (Stifter

Cardamom Workbench		Home	File Upload
Conaille_Muirtheimne.txt			
Old_Irish_Glosses.txt			
Thin_Lizzy.txt			
Cicero.txt			

Figure 1: Cardamom Workbench: Home Page with Uploaded Texts and File Upload Options .

et al., 2021, 10). They require ongoing investment and technical support, while a print edition, once published, is relatively permanent. Publishing text online requires either developing the technical skill-set required to create a web-based text repository, or employing a web developer. Either option incurs costs, be it for hardware acquisition and maintenance, or for ongoing web-hosting services. Linguists can be easily excused for preferring to simply focus on their own specific research interests. Thus, both technical and financial restrictions contribute further to historical language varieties remaining particularly poorly resourced.

Despite the factors listed above which may obstruct linguistic communities attempting to make digital text available online, there is a clear desire to do so, and pride is rightly taken in extant digital resources. Stifter et al. note that "Medieval Irish studies have been at the vanguard of textual digitisation since the infancy of the World Wide Web" (2021, 14), and it is indeed widely reported that the first website hosted in Ireland was the *Corpus of Electronic Texts* (CELT, Ó Corráin et al., 1997; English, 2018; Burke, 2018; Ahlstrom, 2014). Other repositories like ISOS and projects like *Ogham in 3D* (White, 2012) are praised for making historical writings available to researchers and disseminating academic research to a wide public audience (2021, 7, 24–25). The value of creating digital resources is clearly not lost on humanities scholars, and it would benefit both communities of researchers, NLP and traditional linguistic, to develop a streamlined, cost-free means of publishing digital text online, whereby appropriate credit can be given to the creator of that text.

5 The Workbench

The Cardamom Workbench aims to overcome many of the problems discussed above, both those faced by NLP researchers and by those in humanities fields. It also aims to make useful NLP techniques and processes easily accessible to users. Users will be provided with an intuitive GUI through which they can interact with various Cardamom technologies, and the pipeline to digitally publishing texts online will be streamlined. If a user chooses to publish their text through the workbench, it will remain easily accessible online and will be appropriately attributed to the digital text's creator. It will also be ensured that the copyright of any earlier edition of an uploaded text is respected, and that contributed works meet quantifiable quality standards before they can be published, which should alleviate concerns about the reliability of these digital resources.

5.1 Application Design and Workflow

The application is comprised of a web-based front end and a relational database back end. The GUI has been designed to produce an intuitive workflow, intended to make the built-in Cardamom technologies easily accessible to a wide variety of users without requiring them to develop the kind of technical skill-set which would otherwise be needed. Users who make accounts can upload text files in common formats like `.pdf`, `.txt` and `.docx` at the homepage (see figure 1). The text is extracted from these files by the workbench, and stored in the database using UTF-8 encoding. Alternatively, users can create a new text from scratch using the built-in text editor. In either case, users will be asked to select the primary language of the text at the point of upload or creation. Texts can contain

multiple languages, however, some downstream tasks are language-dependent and require that a primary language is identified.

Once uploaded or created, users can select a text from the homepage. Doing so opens it in the Text Editor tab. Here changes can be made to the content of the text if necessary. Several other tabs are also available to users, each associated with a specific text processing or annotation task. These tabs, from left to right, form a workflow which is intended to guide users who may be unfamiliar with text processing through the successive steps in an intuitive manner. Certain steps are reliant on previous ones, and so some tabs will be unavailable until previous steps have been completed. For example, POS-tagging will be unavailable until a text has been tokenised. Users are not required to utilise every tab, nor to perform every type of processing which is available. For example, a user may intend only to tokenise a text, and it will be possible for them to export their token data once they have completed this step.

In each of the workflow tabs users will be able to carry out the specified task either automatically, using Cardamom technologies, or manually. This gives users manual oversight over automated tasks. For example, in the POS Tagging tab a user can manually select POS tags for individual words, or they can click the Auto-Tag button and the workbench will select the appropriate pre-trained POS-tagger model for the specified language, and use it to tag the text. The user may use the Auto-Tag function first, then manually change tags by clicking on a token, and selecting a different POS from a drop-down menu (see figure 4 below). Where a user has manually annotated text in any workflow tab, and then applies automatic annotation to the text, the automatic tool will not overwrite manual annotations. In languages which are currently unsupported by Cardamom technologies, the workbench provides generalised automation tools to support workflows where possible; for example, the workbench can attempt to tokenise text regardless of language, though results are improved where a supported language is specified. Users may have to carry out language-dependent tasks manually, however, where languages are unsupported by the workbench.

Tokenisation does not involve splitting a user's text into word-level strings and storing these. Instead, when tokenisation is carried out by a user

on a text, a start index and end index are stored in the database for each token. Tokens can then be retrieved from the original text at any point using these indices. Token-dependent annotations, such as POS-tags, are applied to this index range rather than to the string itself. In a similar manner, any user-specific annotations are also applied to an index range corresponding to a string of text highlighted by the user in the GUI. This allows annotations to be provided both at token level, as well as at sub-token and super-token levels. When the user makes changes to the base text in the text editor, the indices of tokens are updated in accordance with any alterations made, ensuring that annotations remain aligned with the correct text.

One of the main benefits of the workbench's design is that it can learn from users' content. Users, therefore, can improve the ability of the workbench to automate processing tasks for their language each time they upload or annotate text, as this provides more training data to the underlying language models. This adaptability is of great value for under-resourced languages, for which little annotated text data might yet exist. In the case of languages which are not yet supported by the workbench, users will need to manually annotate some portion of their uploaded text data themselves in the workbench. Once a sufficient quantity of text has been manually annotated, however, it will be possible to train models for the language, making automatic annotation available for that language. In order to ensure consistency of data used for model training, the streamlined annotation process requires that users tokenise and POS-tag in accordance with UD guidelines (Zeman, 2016). User-generated data will not be used as training data until it meets these criteria. While user-generated annotations may be used in resulting publications, they do not form a part of the main workflow, and will not be used in model training.

5.2 Technologies

The technologies which underlie the automatic processing and annotation options in the workbench have been developed throughout the course of the Cardamom project. As these technologies are not the focus of the current paper, technical aspects of their individual implementations cannot be discussed in detail throughout this section. Specifications of many technologies used by Cardamom have already been published (Doyle et al., 2019;

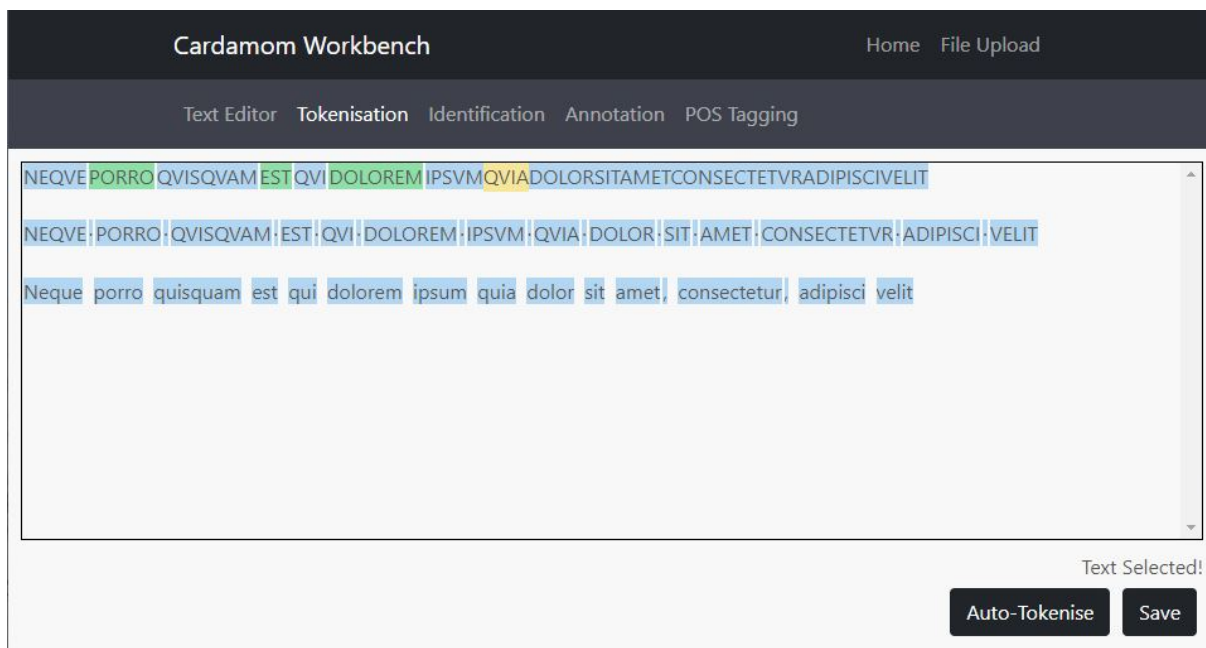


Figure 2: Cardamom Workbench, Latin Text: Tokenisation Tab with Automatically Generated Tokens (Blue), Manually Generated Tokens (Green) and Selected Text (Yellow).

Chakravarthi et al., 2020; Goswami et al., 2020; Rani et al., 2022; Goswami et al., 2021a,b; McCrae et al., 2021; Dereza et al., 2023b), and publications for other technologies are in progress. Certain tasks, such as tokenisation, which have been found to create specific difficulties for languages which have been the focus of Cardamom research will be discussed in this section, however. This section will also address tasks have been improved by Cardamom research, either by reducing the quantity of training data required to achieve sufficient results, or by reducing the processing power and time required to achieve results comparable with the state-of-the-art.

5.2.1 Tokenisation

Tokenisation has been identified as problematic for languages which predate the modern standard separation of lexical words using spaces (Doyle et al., 2019). In such cases, tokenisation requires a more targeted, language-specific approach. For example, certain Latin texts are written with words separated using an interpunct, not spacing. An example of this can be seen in figure 2. By contrast to Latin, the interpunct is often used to indicate points of stress within the verbal complex in the orthography of Old Irish editions and learning material, but not necessarily at word boundaries. Latin text requires that tokens be separated at points where an interpunct is used, however, this may be inappropriate

for Old Irish where the interpunct serves a different purpose. Therefore, it was necessary to create discrete tokenisers for Latin and Old Irish, each of which treat the interpunct as appropriate for the language in question.

Word spacing has also been identified as problematic when tokenising historical languages. Many Latin texts were written in *scriptio continua*, without any punctuation or spacing separating words from each other (see again figure 2). Meanwhile Thurneysen notes that generally, in Old Irish manuscripts, "words which are grouped round a single chief stress and have a close syntactic connexion with each other are written as one" (1946, 24). In either case, it is difficult to create an automatic tokeniser which can accurately separate such compounded words without large quantities of training data (Doyle et al., 2019). The workbench, therefore, allows users to manually identify the exact boundaries between tokens in their texts by highlighting some quantity of text which they consider to be a single token. By this means it is even possible for users to create tokens which contain space characters, as may be required, for example, where a nasal has been separated from the following word in Old Irish (see figure 3).

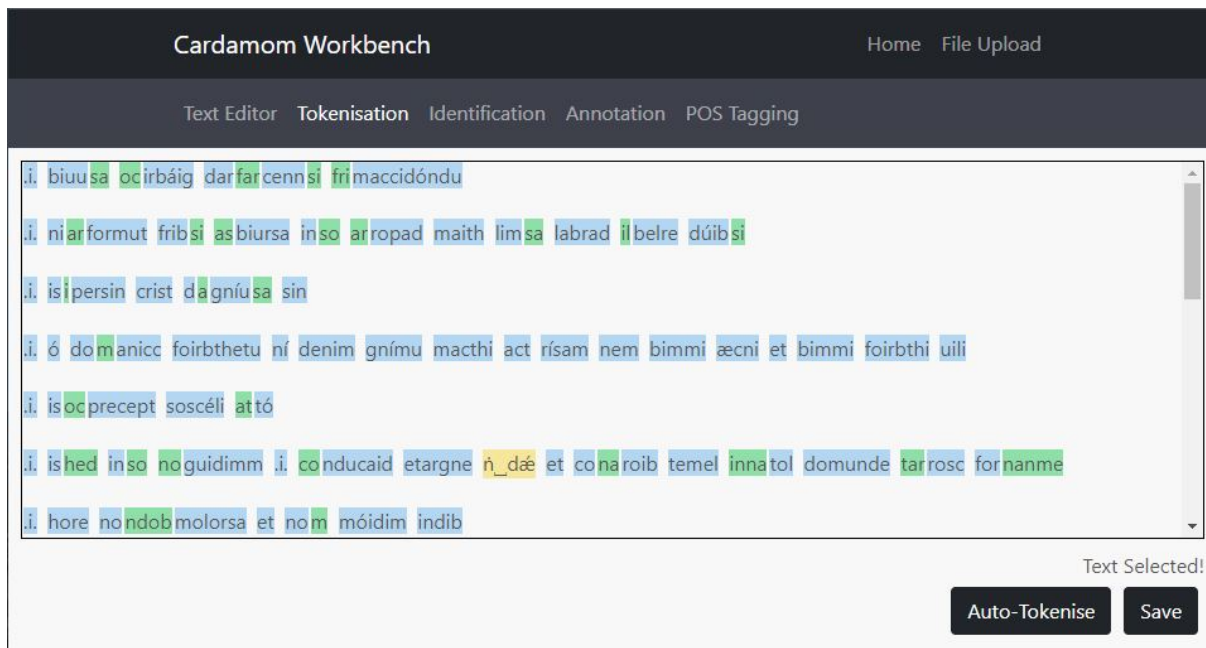


Figure 3: Cardamom Workbench, Old Irish Text: Space Character within the Selected Token (Yellow).

5.2.2 Language Identification, and Related Techniques

A considerable amount of Cardamom research has focused on the identification of various linguistic features and characteristics within a text. This includes, but is not limited to, identification of language and dialect (Goswami et al., 2020; Rani et al., 2022), authorship identification, and cognate detection. In the context of the workbench, these technologies may be of use to users working with texts which contain some degree of code switching. Identifying tokens which are not from the primary language of the text will allow for improved results in POS-tagging. These techniques may also be of interest to scholars of languages like Old Irish, for which "Contemporary divergences, such as would point to dialectal peculiarities, are very rare" (Thurneysen, 1946, 12).

5.2.3 POS-tagging

The Cardamom Workbench follows Universal Dependencies (UD) guidelines (Zeman, 2016) for tokenisation and POS-tagging. As such, the workbench utilises the same seventeen POS tags used in UD treebanks. This decision was made because UD has already established itself as a common standard, capable of facilitating the requirements of a wide range of languages. As such, it is reasonable to expect it will be suitable also for the various under-resourced and historical languages which are the target of the workbench. Moreover, adherence

to such a well supported standard as UD, means that extant validation tools can be utilised to ensure the quality of data created and annotated by users.

As has been mentioned above, users can POS tag their text both automatically and manually. Automatic POS-taggers were trained for various languages using lexical data primarily drawn from UD treebanks. These models can be improved both when UD repositories are updated, and when workbench users POS tag their own text. Tagged text is colour-coded in the GUI to enable users to quickly and intuitively assess POS-tagged tokens (see figure 4). A future iteration of the workbench is expected to expand this token-level tagging to include headword identification to support digital lexicography, and lexical feature identification in accordance with UD guidelines.

5.2.4 Other Annotations

Various other forms of annotation are possible aside from language and POS tagging of tokens. The Annotations tab allows users to apply annotation not only to tokens, but at sub-token and meta-token levels also. Users can highlight any quantity of text and add an annotation to it. This is useful, for example, in digital editions of historical language texts where, in the manuscript, text may have been lost due to damage, or abbreviated using a variety of symbols (Thurneysen, 1946, 25). Users may wish to indicate that they have supplied or restored text in such instances, and can do so easily by providing

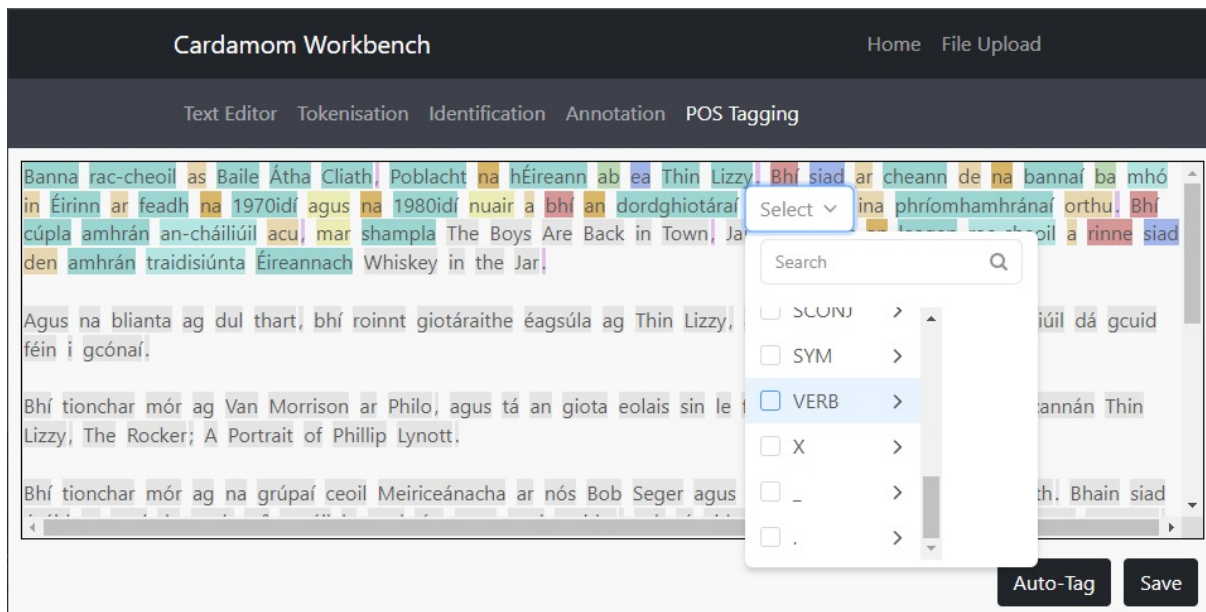


Figure 4: Cardamom Workbench, Modern Irish Text: POS-tagging with POS Tags Differentiated by Colour .

annotation in this manner. Here again, Cardamom technologies are available to help automate the process, for example, by suggesting the most likely annotation required based on the text selected by the user. In a future iteration of the workbench it is expected that users will have the option of exporting their text annotated with TEI markdown (TEI-Consortium, 1994), however, at launch the primary function of such annotations is to enhance resulting digital editions with metadata.

5.3 Value for Stakeholders and Future Work

The primary goals at launch are to ensure accessibility of current Cardamom technologies to users, and to provide a simple means of digitally publishing texts. Cardamom intends to provide free web hosting for users-submitted texts on servers owned and operated by the Insight Centre for Data Analytics, and permanent URLs will be provided for these once published. Once the period of funding has ceased for the Cardamom project itself, responsibility for continued support of the workbench, and hosting of both the application and digitally published texts, will be transferred to the Insight Centre for Data Analytics. This will ensure long-term accessibility of user-supplied content, which is beneficial both for users who will be appropriately credited with contributing the text, and for NLP researchers who will have access to more text data for under-resourced and historical languages. The quality of uploaded text and annotations can be tightly controlled using extant validation tools,

and manual oversight.

As has been mentioned throughout this paper, updates to the workbench’s functionality are expected as development continues after launch. Work is ongoing on a tool which utilises word embeddings to allow users to track orthographic and semantic changes in a lexeme over time, and to find words which are semantically or morphologically similar to an entered search term. It is envisioned that this functionality could be useful to historical linguists editing obscure manuscript passages, where one possible reading must be chosen over another. Generic tools such as concordancers are also intended to be implemented in future revisions, and extended functionality will be added for texts both as the workbench is developed, and in accordance with the level of annotation provided by users. For example, POS-tagging and headword annotation of tokens will enable linking to external lexical resources for a given language.

It is expected that once a sufficient interest has been demonstrated by users in the workbench, it will be possible to develop an expert peer-review and support network. This will further ensure the quality of submitted texts, allowing language experts to provide commentary and critique on a text before it is published. It will also be possible to credit reviewers when updates are made to published texts based on their recommendations. Such a network would also allow linguistic experts to advise on future development of the workbench to

support language-specific requirements, increasing its value to users going forward.

5.4 Related Tools

A number of extant tools may be compared to the workbench presented here, both as regards providing users with similar technologies, and simplifying interaction with annotated corpora. It is important to acknowledge these tools in order to appreciate the features and use cases which distinguish the Cardamom Workbench from them. The value proposition of the workbench, as well as its intended user base, are the primary distinguishing factors. As has been mentioned above, the intent of the workbench is to create value for two groups of researchers with distinct sets of requirements in order to improve their particular research prospects.

The historical focus of Cardamom research creates value in an area for which discrete solutions are required, and certain tools have already been made available in this area in an attempt to provide such solutions. *TEITOK* is an open source, web-based tools which enables users to create and distribute corpora (Janssen, 2016, 16). Users can align manuscript pages with transcribed text, and transcribe directly from manuscript images. Annotation is enabled using TEI, and users are given tools for visualising annotations such as dependency grammars and parse trees. As such, this tool is possibly the closest extant resource to the workbench in terms of its historical focus, and its corpus creation and annotation support. A few things set the two apart, however, the foremost of which is the technology stack provided by Cardamom. Workbench users benefit from these tools not as merely as static resources, but as dynamic ones. They can play a role in improving their performance by contributing more text and annotations. Thus, while the focus of *TEITOK* appears to be to facilitate corpus creation and annotation, the focus of the workbench is to provide users with tools which will empower them to process and annotate texts more efficiently, and to constantly improve the tools available to users.

Some extant resources provide users with technologies comparable to those of the Cardamom Workbench. The *IMS Open Corpus Workbench* (Evert, 2008) provides users with open source corpus query tools and is intended for use with large text corpora. On the one hand this is very useful for users who have access to large text corpora, though

it is an unrealistic scenario for under-resourced or historical languages. The aim of Cardamom research has been to close the resource gap by creating tools which can be both trained and used on relatively small text corpora. On the other hand, according to the *IMS Open Corpus Workbench's* website, "It is intentionally not very user friendly", requiring that users interact with it using secondary software which abstracts away from the technology stack. By contrast, the Cardamom workbench was designed from the beginning with user friendliness in mind, as its intended user base is specifically those who do not have the technical skill-set to use Cardamom technologies if it means downloading scripts from repositories like GitHub and running them using command line interface. *Persides* is an editing platform for Classics texts which allows large groups of users to partake in "allows for the participation of a large group of users in the process of editing, publishing, and analyzing ancient documents" (Almas and Beaulieu, 2013, 502). It is based on the principle that "a well-organized crowdsourcing effort can accomplish far more work than any lone scholar and the work ultimately produced benefits from the variety of perspectives included" (Almas and Beaulieu, 2016, 172). This contrasts with the work presented here in that the Cardamom workbench aims to empower individual scholars to annotate and publish their work with minimal effort or collaboration. Another web-based application, the *INCEpTION* annotation environment (Klie et al., 2018), provides users near free rein over how they annotate their corpora. While it provides predefined elements, like knowledge bases, layers and tag-sets, it also allows users to modify these, or to create their own annotations. While the Cardamom workbench allows users to provide their own annotations where desired, the streamlined annotation process is designed to ensure users' output meets a single common NLP standard as closely as possible for tasks like tokenisation and POS-tagging (Zeman, 2016). Moreover, the workbench provides users with a suite of NLP tools specifically designed to aid in such annotation for historical and under-resourced languages.

Possibly the most well known extant tool in this area is *Sketch Engine*, a web-based corpus management system which also provides users with text analysis functionalities. Some of the analysis tools provided by *Sketch Engine* overlap with those of the Cardamom Workbench, for example,

it allows POS-tagging for a wide range of supported languages. It also provides a "summary of a word's grammatical and collocational behaviour" (Kilgarriff et al., 2014, 9), however, to support such features *Sketch Engine* requires that tools like a tokeniser, lemmatiser, POS-tagger, and morphological parser must already exist for a given language (2014, 18). Being a commercial tool, it is not free to use, however, a feature-limited free counterpart, *NoSketch Engine*, does exist. While *Sketch Engine* provides very valuable technologies to lexicographers, translators, language learners, and institutes like universities, its primary focus seems to be on making extant tools more accessible rather than developing or improving language tools. Here again the Cardamom Workbench provides value to users. Both *Sketch Engine* and the Cardamom Workbench cater more to some languages, for which more language resources are readily available, than to other less resourced languages. Cardamom, however, provides users with the possibility of creating such resources, and harnessing them to improve built-in language tools as they use the workbench. The suite of technologies built into the Cardamom Workbench is also more extensive than that of *Sketch Engine*, and these are targeted towards the kinds of processing and annotation tasks which will allow users to create the most useful language resources using their supplied text.

6 Conclusion

This paper has presented the Cardamom Workbench, a tool which provides language experts with modern NLP tools which can be easily applied to their own texts. It also aims to provide users with a streamlined means of digitally publishing text content which may be of value to both traditional linguists and to NLP researchers, meanwhile allowing appropriate credit to be given to users who produce and annotate the digital text.

Acknowledgements

This publication has emanated from research supported by the Irish Research Council under grant IRCLA/2017/129 (CARDAMOM-Comparative Deep Models of Language for Minority and Historical Languages) and co-funded by Science Foundation Ireland (SFI) under grant SFI/12/RC/2289_P2 (Insight_2) and grant SFI/18/CRT/6223 (CRT-Centre for Research Training in Artificial Intel-

ligence). We would like to acknowledge the hard work carried out by all Cardamom members, past and present, on the workbench presented here. We would also like to acknowledge the use of language data from Universal Dependencies treebanks in training and testing certain Cardamom models.

References

- Dick Ahlstrom. 2014. [How the Irish Helped Weave the Web](#). *The Irish Times*.
- Bridget Almas and Marie-Claire Beaulieu. 2013. [Developing a New Integrated Editing Platform for Source Documents in Classics](#). *Literary and Linguistic Computing*, 28(4):493–503.
- Bridget Almas and Marie-Claire Beaulieu. 2016. *The Perseids Platform: Scholarship for All!*, pages 171–186. Ubiquity Press.
- Diego Alves, Božo Bekavac, Daniel Zeman, and Marko Tadić. 2023. [Analysis of Corpus-based Word-Order Typological Methods](#). In *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)*, pages 36–46, Washington, D.C. Association for Computational Linguistics.
- Joseph D. Becker. 1988. [Unicode 88](#). Standard, Unicode Consortium, Palo Alto.
- Emily Bender. 2019. [The #BenderRule: On Naming the Languages We Study and Why It Matters](#). *The Gradient*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language Models are Few-Shot Learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Roisin Burke. 2018. [Creator of Ireland's First Website Logs off after 34 Years](#). *EchoLIVE*.
- Bharathi Raja Chakravarthi, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John Philip McCrae. 2020. [A Sentiment Analysis Dataset for Code-Mixed Malayalam-English](#). In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 177–184, Marseille, France. European Language Resources association.

- Christopher Cieri, Mike Maxwell, Stephanie Strassel, and Jennifer Tracey. 2016. [Selection Criteria for Low Resource Language Programs](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4543–4549, Portorož, Slovenia.
- Oksana Dereza, Theodorus Fransen, and John P. McCrae. 2023a. Do not Trust the Experts: How the Lack of Standard Complicates NLP for Historical Irish. In *Proceedings of the 3d Workshop on Insights from Negative Results in NLP, EACL 2023*. In print.
- Oksana Dereza, Theodorus Fransen, and John P. McCrae. 2023b. Temporal Domain Adaptation for Historical Irish. In *Proceedings of the 10th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial), EACL 2023*. In print.
- Adrian Doyle, John P. McCrae, and Clodagh Downey. 2018. [Preservation of Original Orthography in the Construction of an Old Irish Corpus](#). In *Proceedings of the LREC 2018 Workshop: "CCURL2018 – Sustaining Knowledge Diversity in the Digital Age"*, pages 67–70, Miyazaki, Japan.
- Adrian Doyle, John P. McCrae, and Clodagh Downey. 2019. [A Character-Level LSTM Network Model for Tokenizing the Old Irish text of the Würzburg Glosses on the Pauline Epistles](#). In *Proceedings of the Celtic Language Technology Workshop*, pages 70–79, Dublin, Ireland. European Association for Machine Translation.
- Eoin English. 2018. [UCC-based Developer of Ireland's First Webpage Logs off](#). *The Irish Examiner*.
- Stephanie Evert. 2008. [The IMS Open Corpus Workbench \(CWB\)](#). Retrieved: February 02, 2023.
- Koustava Goswami, Sourav Dutta, and Haytham Assem. 2021a. [Mufin: Enriching Semantic Understanding of Sentence Embedding using Dual Tune Framework](#). In *2021 IEEE International Conference on Big Data (Big Data)*, pages 2034–2039.
- Koustava Goswami, Sourav Dutta, Haytham Assem, Theodorus Fransen, and John P. McCrae. 2021b. [Cross-lingual Sentence Embedding using Multi-Task Learning](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9099–9113, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Koustava Goswami, Rajdeep Sarkar, Bharathi Raja Chakravarthi, Theodorus Fransen, and John P. McCrae. 2020. [Unsupervised Deep Language and Dialect Identification for Short Texts](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1606–1617, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Michael A. Hedderich, Lukas Lange, Heike Adel, Jan-nik Strötgen, and Dietrich Klakow. 2021. [A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics.
- Maarten Janssen. 2016. [TEITOK: Text-Faithful Annotated Corpora](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4037–4043, Portorož, Slovenia. European Language Resources Association (ELRA).
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. [The State and Fate of Linguistic Diversity and Inclusion in the NLP World](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.
- Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The Sketch Engine: Ten Years On. *Lexicography*, 1:7–36.
- Jan-Christoph Klie, Michael Bugert, Beto Boullosa, and Iryn de Castilho, Richard Eckart and Gurevych. 2018. [The INCEPtion Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation](#). In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics. Event Title: The 27th International Conference on Computational Linguistics (COLING 2018).
- Teresa Lynn. 2012. [Medieval Irish and Computational Linguistics](#). *Australian Celtic Journal*, 10:13–27.
- John P. McCrae and Theodorus Fransen. 2019. [Cardamom: Comparative Deep Models for Minority and Historical Languages](#). In *Proceedings of the 1st International Conference on Language Technologies for All*, pages 276–279, Paris, France. European Language Resources Association (ELRA).
- John P. McCrae, Atul Kumar Ojha, Bharathi Raja Chakravarthi, Ian Kelly, Patricia Buffini, Grace Tang, Eric Paquin, and Manuel Locria. 2021. [Enriching a terminology for under-resourced languages using knowledge graphs](#). In *Proceedings of The Seventh Biennial Conference on Electronic Lexicography, eLex 2021*, pages 560–571.
- Barbara McGillivray, Thierry Poibeau, and Ruiz F. Pablo. 2020. [Digital Humanities and Natural Language Processing: "Je t'aime... Moi non plus"](#). *Digital Humanities Quarterly*, 14(2).
- Joakim Nivre. 2015. Towards a Universal Grammar for Natural Language Processing. In *Computational Linguistics and Intelligent Text Processing*, pages 3–16, Cham. Springer International Publishing.

- Spandan Patil, Lokshana Chavan, Janhvi Mukane, Deepali Vora, and Vidya Chitre. 2022. [State-of-the-Art Approach to e-Learning with Cutting Edge NLP Transformers: Implementing Text Summarization, Question and Distractor Generation, Question Answering](#). *International Journal of Advanced Computer Science and Applications*, 13(1).
- Priya Rani, John P. McCrae, and Theodorus Franssen. 2022. [MHE: Code-Mixed Corpora for Similar Language Identification](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3425–3433, Marseille, France. European Language Resources Association.
- Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. [Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism](#). *CoRR*, abs/1909.08053.
- David Stifter, Nina Cnockaert-Guillou, Beatrix Färber, Deborah Hayden, Máire Ní Mhaonaigh, Joanna Tucker, and Christopher Guy Yocum. 2021. [Developing a Digital Framework for the Medieval Gaelic World; Project Report](#). Technical report, Developing a Digital Framework for the Medieval Gaelic World.
- TEI-Consortium. 1994. [Text Encoding Initiative](#). Retrieved: February 24, 2023.
- Rudolf Thurneysen. 1946. *A Grammar of Old Irish*, 2 edition. The Dublin Institute for Advanced Studies, Dublin.
- Marta Villegas, Nuria Bel, Carlos Gonzalo, Amparo Moreno, and Nuria Simelio. 2012. [Using Language Resources in Humanities research](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3284–3288, Istanbul, Turkey. European Language Resources Association (ELRA).
- Nora White. 2012. [Ogham in 3D](#). Retrieved: February 24, 2023.
- Dan Zeman. 2016. [UD Guidelines V2](#). Retrieved: February 24, 2023.
- Donnchadh Ó Corráin, Hiram Morgan, Beatrix Färber, Gregory Toner, Benjamin Hazard, Emer Purcell, Caoimhín Ó Dónaill, Hilary Lavelle, Seán Ua Súilleabháin, Julianne Nyhan, and Emma McCarthy. 1997. [CELT: Corpus of Electronic Texts](#). Retrieved: February 24, 2023.

Sentiment and Natural Language Inference

Sentiment Inference and Gender Classification for Gender Profiling

Manfred Klenner

Department of Computational Linguistics

University of Zurich, Switzerland

klenner@cl.uzh.ch

Abstract

In this paper we describe the further development of an existing rule-based system for sentiment inference. We have created new resources, trained models for the novel language-specific task of gender classification of nouns and applied it to German gender-tailored profiling in newspaper texts. We discovered an imbalance wrt. gender denoting nouns and the role they take as sources or targets of verbs denoting positive or negative relationships. Our goal was to get empirical access to the perception of gender, their roles and their reciprocal relations as portrayed in the news. Our empirical findings are based on statistical hypothesis testing.

1 Introduction

The identification of gender denoting expressions in texts might serve various purposes. For instance, it could be used to identify bias or other forms of imbalance like gender stereotypes as portrayed by the media. We focus on the detection of polar relations (in favor of, against) and polar roles (e.g. positive or negative actor) that gender referring expressions occupy in three Swiss newspaper texts. Given the sentence *Merkel cheats the people*, we are entitled to infer that the writer claims that Merkel acts against her nation and that she should be regarded as a villain. We have compared the distribution of male and female denoting expressions in such contexts on the basis of 380 polar verbs that express a positive (in favor) or negative (against) relation between the actor (the source) and the theme, patient or recipient (the target). We use the term *sentiment inference* for this task, because the identification of relations or roles is not in every case just a simple lexicon lookup. Subordination and negation has to be taken into account. Take the sentence *The land criticizes that Europe (not) supports the Ukraine*: from the unnegated version we can infer that - among others - the mentioned land is against the EU and the Ukraine. The inference

pattern here is: if some actor A (land) is against something (support) that is good for another actor B (Ukraine), than A is against B, at least in a situation specific way. The negated version with *not* gives rise to the opposite inference, that A is in favor of B. In a couple of papers e.g. (Klenner and Amsler, 2016), (Klenner et al., 2017a), (Klenner et al., 2017b), (Klenner, 2018) and (Göhring et al., 2021) we have described the resources and principles behind sentiment inference¹.

In this paper, we focus on (the usage of) a new system component that allows us to do gender tailored analysis, namely our gender aware animacy classifier. Moreover, we not only are carrying out an intrinsic evaluation but also an extrinsic end-to-end evaluation. The goal was to find out whether these two components - the rule-based inference system and the gender classifier are suitable means for gender profiling. *Gender Profiling* strives to identify the contexts male and female denoting expressions² occupy according to e.g. the media and whether the distribution is uniform are imbalanced³. A finding contributing to the female profile could be, for instance, that female nouns are significantly more often the targets of particular verbs than male denoting nouns.

First, we describe our rule-based approach to sentiment inference, then we introduce our new gender classifier and then we discuss the empirical results of applying these two components to newspaper texts from 2004 to 2022. We also try to find out whether the gender profiles have changed, i.e. whether there is a difference between 2004-2014 and 2018-2022.

¹See <https://pub.cl.uzh.ch/demo/stancer/index.py> for an online demo.

²Certainly, we do not claim that gender is a binary category; but gender-denoting nouns without explicit indications (e.g. ‘*’) do have a binary reference that we cannot overcome.

³We avoid the stronger notion of *bias*, since we cannot determine whether the incidents reported by the news are facts or stem from a biased world view.

2 Sentiment Inference

Sentiment (or opinion) implicature (Deng and Wiebe, 2014) aims to predict positive or negative attitudes of opinion holders towards other persons, groups, etc. or towards inanimate entities (targets). We would like to adopt a broader view and call the resulting task *sentiment inference*⁴. If we read that someone has honored, punished or even hurt someone else, then, strictly speaking, we do not know whether there is an *attitude* of the initiator towards the target: we only know that some action was carried out that affected the target in a positive or negative way⁵. Sentiment inference as we put it is the prediction of positive and negative relations holding between a source (an opinion holder or not) and a target.

The central resource of our model is a verb lexicon comprising about 1,000 different verbs. Verbs might have more than a single reading, so in principle, disambiguation was needed. However, there is no verb disambiguator available for German and we do not have the resources to train one. Fortunately, it turned out that disambiguation partly can be done on the basis of dependency parsing, selectional restrictions and animacy detection (see Klenner and Göhring (2022)).

At first glance, animacy detection seems to be related to semantic role labeling. The semantic role *actor* most naturally would be animate. However, we have shown that existing semantic role labeler for German are not reliable in this respect (Klenner and Göhring, 2022). One problem of the task is metonymy, where e.g. a capital city stands for a government (e.g. in *Wien criticizes Brussels*).

Before we have an example, let us first discuss the kind of information a verb in our lexicon carries. Take *to cheat*. As most of the polar verbs, it has two polar roles, a source and a target. It also expresses a directed relation (here: *against*) that holds between the two. Here, the source is (acting) against the target. Moreover, the source of *cheat* might be regarded as negative actor, a villain, the target as the victim (given that the sentence is factual, i.e. not negated or in modal mode).

Table 1 and table 2 illustrate the kind of specifications in our lexicon. Table 1 defines the first frame of *sorgen für* (Eng. care for). The upper part

⁴The notion has been used in the past, see e.g. (Choi et al., 2016), who defined it as *directed opinion*.

⁵In the sentence *The government destroyed all our hopes*, the government is a negative source, but not an opinion holder.

of the table are restrictions that must be fulfilled in order to instantiate the polar frame (below the line). The (dependency) parse must comprise exactly a subject (subj) and prepositional phrase (pp) with the preposition *für* and the subject and the noun of the pp must be animate (+a).

dep. label	subj	pp-obj
lex. restr.	-	prep=für
sel. restr.	+a	+a
polar role	source	target
polar rel.	in favour	-
polar effect	+actor	+effect

Table 1: Frame I of *sorgen für* (Eng. care for). Dependency label (dep. label), lexical restriction (lex. restr.) and selectional restriction (sel. restr.) as well as the polar profile are shown.

If this is given, the filler of the subject is regarded as the source of an in-favor relation towards the target which is the noun of the pp. The source is claimed to be a positive actor (+actor) and the target to receive a positive effect (+effect).

dep. label	subj	pp-obj
lex. restr.		prep=für
sel. restr.	+a	-a
polar restr.		+pos
polar role	source	target
polar rel.	in favour	
polar effect	+actor	

Table 2: Frame II of *sorgen für* (Eng. care for)

Table 2 specifies frame II of the same verb. It is also an example where animacy is a disambiguating factor. The subcategorization frame II is the same (incl. the preposition) as frame I, but the filler of the pp noun is inanimate (-a). A German example sentence would be: *Sie sorgte für gute Stimmung*. The English translation is: she provided a good atmosphere. A different verb is used in English. Please note that frame II has an additional polar restriction, namely that the filler of the pp noun should be positive (+pos). We have implemented a phrase-level polarity composition on the basis of a polarity lexicon⁶ (see Clematide and Klenner (2010)) and composition rules (see Moilanen and Pulman (2007) for the principles of sentiment composition). Here *good atmosphere* is recognized as

⁶German Polarity Lexicon: download from the IGGSA website under <https://sites.google.com/site/iggsahome/downloads>

a positive phrase. Only if the pp is positive, the actor is a positive actor, if it is negative (frame 3, not shown) like in *bad atmosphere* the actor also is negative. Given a neutral actor like in (*Sie sorgt für Papier*, Eng. She ensures that there is enough paper), no polar relation or polar role at all should be set.

The selectional restrictions are not gender-specific. But the selectional restriction animate (+a) is fulfilled if either a male or female denoting noun is found as a filler.

For the present study, we used those 368 out of the 1,000 verbs that passed a particular frequency threshold (discussed in section 5). We further divided these 380 verbs into 3 subclasses: verbs denoting physical events (119 cases) like *to hit*, verbs denoting emotional events (101 cases) like *to enjoy* and verbs denoting communicative acts (160 cases) like *to blame*. This subdivision allowed us to focus on differences on a more fine-grained level. A couple of verbs cannot be assigned a definite category, e.g. *to hurt* could happen as a physical or an emotional incident. Such verbs are kept in both classes⁷. Table 3 shows some examples.

verb DE	verb EN	p	e	c
töten	kill	+	-	-
zerstören	destroy	+	-	-
quälen	torture	+	-	-
sorgen	care	-	+	-
verabscheuen	detest	-	+	-
ärgern	annoy	-	+	-
beschuldigen	blame	-	-	+
beschimpfen	insult	-	-	+
anprangern	accuse	-	-	+

Table 3: Verbs for 3 subclasses: p (physical), e (emotional), c (communicative)

Although the division into 3 subclasses is a step towards a more fine-grained analysis, there are commonalities across classes in terms of the strength of a verb. In psychology, but also in the Natural Language Processing (NLP) community, words have been characterized not only in terms of polarity (positive, negative) but also in terms of arousal and dominance (see Mohammad (2018)). Arousal quantifies the intensity of emotion provoked by a stimulus, and dominance the degree of control exerted by a stimulus. For instance, *tired* has low arousal and low dominance, *angry* has high arousal

⁷So the sum of the verbs from all classes is more than 368.

and medium dominance and *vanquish and defeat* has high arousal and high dominance. We used the VAD resource of Mohammad (2018)⁸ to assign scores for arousal and dominance to our verbs. 80 out of 380 were not found in this resource. We used fastText (Joulin et al., 2017) embeddings to find scores for these out of vocabulary (oov) verbs. We took the most similar verb of an oov verb and transferred its score to the oov verb. Most of the time, synonyms were found, but sometimes also antonyms. Thus, we manually inspected the pairings and approved the transfer or corrected it, if needed (choosing the best fitting similar word). The higher the arousal and dominance values of negative verbs, the clearer is the source of such a verb regarded as a villain and the target as a victim. This as well might reveal some gender-specific differences.

The model architecture up to the point where we started to create a version of the system for the task of gender profiling consisted of a lexicon of verbs, specifying their polar properties and selectional restrictions, a dependency parser and an animacy classifier. The gender classifier is new, also the classification of verbs as belonging to one of three verb classes and the arousal and dominance assignment to these verbs.

3 Grammatical Gender Classification

The grammatical gender of an animacy denoting expression in German can be either male or female⁹. Detecting male or female reference, i.e. reference to men or women, thus boils down to identify the grammatical gender of animacy denoting expressions. Other gender identifies only recently have been included by using the gender star etc. However, in our texts they are not being used. The most indicative part of a gender denoting expression, e.g. a noun phrase is the nominal head. If we had a complete list of gender denoting nouns, grammatical gender classification might be regarded as a simple lexicon look-up. However, such a list would be huge and could not be claimed to be complete, since e.g. new professions might come into existence. We have a list of 30,000 profession denoting nouns, 13,000 of which are female forms. Some of them are rather specific and probably will never be used in newspaper texts. Rather than searching

⁸The VAD resource is available under <https://saifmohammad.com/WebPages/nrc-vad.html>

⁹There are only very few cases where a neutral noun can refer to an animate (human) referent, i.e. *Mädchen*, Eng. girl.

through such a over-specific, but inherently incomplete list each time a noun has to be classified, a learned model for gender classification might be more reasonable since it also has some generative capacity, i.e. is able to classify nouns never seen before. Such a model should learn the footprint of gender denoting nouns as apposed to non-animacy denoting nouns. Word embeddings seem to be the perfect basis for such classifiers, since they capture relatedness. Still we cannot expect that pretrained word embeddings already provide the three needed (class) clusters: male, female, inanimate. But a machine learning approach might be able to properly weight embedding dimension in order carve out the class-specific profiles.

In [Klenner and Göhring \(2022\)](#) we have introduced German animacy classification. On the basis of 13,000 German nouns that were manually classified as denoting either animate or inanimate entities¹⁰ we trained a logistic regression classifier using fastText embeddings (see our paper for the various experiments and a full discussion). The overall accuracy was 96.67%.

In order to create a gold standard for grammatical gender classification, we manually selected those nouns that could be used to refer to women or men. Examples of female denoting nouns are *Schwester*, *Gastgeberin*, *Schauspielerin* (Eng. sister, hostess, actress, respectively).

It turned out that the class frequencies were imbalanced, more male than female denoting nouns. In German, by adding the suffix *in* to the end of male denoting noun (most of the time) a female denoting noun can be created, e.g. *Helfer* → *Helferin* (Eng. helper). If such a derived wordform was found in a corpus at least twice, it was added to the female list.

These lists of (fe)male denoting nouns were further augmented by exploiting a list of first names. Again, we only kept firstnames which also were found in a corpus and were above a threshold (here: 10 occurrences). Table 4 shows the final distribution (frequency counts) of male, female and inanimate denoting nouns¹¹. Our gold standard comprises more than 18,000 nouns.

We then had slightly more female than male nouns. However, since female nouns are in our text corpus - as we had found out - less frequent

	inanimate	female	male
nouns	5826	5637	5002
first names	-	966	966
Σ	5826	6603	6200

Table 4: Frequency counts of the three classes

than male nouns, we intentionally kept the resulting (little) bias.

The accuracy of the classifier on a random 75/25 train/test split is 96.0%, see table 5 for precision, recall and f-measure of that split. The mean accuracy of a ten-fold cross validation was 95.20%. Since the train set and the test set are exclusive, the good performance of the classifier indicates that word embeddings for this kind of nouns seem to be a proper basis for learning.

	inanimate	female	male
precision	96.0%	96.9%	94.9%
recall	95.7%	97.6%	94.5%
f1	95.8%	97.1%	94.7%

Table 5: Performance of the three-way, gender-aware animacy classification model

Not all German female denoting nouns possess the *in* ending. In our list of female denoting nouns, 50 have endings other than *in* (e.g. *Frisöse*, Eng. hairdresser). On the other hand, a word with an *in* ending is not a reliable indicator of a female noun. In a corpus of 25 million nouns, we found 67,823 words (tokens) ending with *in*. For 36,247 cases of these *in*-words our classifier predicted *female*. The remaining 31,576 *in*-nouns correspond to 4,035 types. We manually classified 1,000 and found only 5 female denoting words. Classifying *in*-words immediately as female denoting nouns would produce quite some errors. This is not what our fastText-based classifier does, although it uses sub-word splitting.

The performance of our classifier with respect to the non-*in* female denoting nouns cannot reliably be evaluated at the moment. It is future work to train models able to deal with such rare cases.

4 Corpus, Corpus Split and Gender Reference in German

Gender profiling in our study is restricted to the monitoring of polar roles and polar relations male and female denoting nouns occupy in newspaper texts. Different profiles then can be identified on

¹⁰Download at: <https://zenodo.org/record/7630043#.Y-aCU9LMJH4>

¹¹The list of male first names was reduced to the size of the female first names.

the basis of different distributions. Especially, uneven distributions are of interest, since they can be interpreted as gender specific. The basic assumption behind our approach is that the overall prior distribution of each gender should also more or less be reflected in the frequency of the polar roles they take and in the polar relations they enter in. We, thus, were interesting in constellations where the genders are involved less or more often than their prior (gender) probability suggests. We interpret these cases as polar imbalance that reveals the gender-specific perception these newspapers cast.

We have data for different periods of the same three Swiss newspapers (2004-2022). Only the last period from 2018 to 2022 was sampled by us for this study, the former data are provided by colleagues. The data points of the 2004-2014 data come without a timestamp, and only the plain sentences are available, not the texts. No coreference resolution was possible, thus. This reduces the number of hits, but should not skew the underlying distribution too much: there is no reason to believe that female denoting nouns are more or less often pronominalized than male ones. Only this would distort the prior gender probabilities we have found on the basis of gender denoting nouns. Since in German, inanimate objects also might have male or female grammatical gender (e.g. *Brücke* is female, Eng. bridge), counting male and female pronouns cannot provide any additional information about the gender distribution. Also, the plural use of *sie* (Eng. she) in German might refer to male, female or neutral animate or inanimate referents. Again, corpus statistics would not help. We thus only looked at cases where the gender classifier triggered and we omit pronoun fillers.

In German, the male word form of e.g. a profession for a long time was used generically to refer to either gender. This was true for singular and for plural. For instance, *Lehrer* (Eng. teacher) was used as a singular and a plural form to refer to all genders. However, for over 20 years now distinct word forms have been used in newspapers. Singular male *Lehrer* and female *Lehrerin*, plural *Lehrer* is now reserved for male reference, while *Lehrerinnen* is used for female reference. In recent years in the course of the discussions of a non-binary gender inclusive language usage, apart from special characters like the gender star (*') like in *Lehrer*innen* or the colon (':') like in *Lehrer:innen* the nominalized participle present of verbs is meant to refer to

all genders¹². For instance, the participle present of the verb *lehren* (Eng. teach) is *lehrend* (Eng. teaching), the nominalized plural form *Lehrende* (Eng. roughly: *teachings* to represent teachers) is used as an all-inclusive reference. This ongoing language change does not affect our current study. Special characters are not used in the three newspapers, they consequently used male and female forms and avoid the participle present¹³.

Our experiments are carried out over the whole corpus but partly also period-wise. In the period-wise mode we also tried to find out whether there is some change in the perception of gender. The most recent period, 2018 to 2022, was compared with the oldest one, from 2004 to 2014. Period 2015-2017 was viewed as a transition period.

We dependency parsed all sentences, extracted predicate argument structure from the parse trees (incl. passive voice normalization), applied the gender classifier to all nouns and run the sentiment inference system. We further analyzed those verb instantiations where the source was classified as male or female. The target was allowed to be animate or inanimate.

5 Empirical Setup

The maximum likelihood estimation (MLE) of the probability of the female gender wrt. whole corpus is 0.183 (2,671,140 out of 14,577,122 gender nouns). The assumption, the null hypothesis H_0 , was that the overall prior gender probability should also be reflected in the distribution of the sources and targets of the polar verbs. For instance, in 18.3% of all instantiations of e.g. the verb *beschuldigen* (Eng. denounce) the source should be a female denoting noun. If this expectation is significantly violated a gender-specific imbalance is found that is, the null hypothesis H_0 is rejected.

As an operationalization of this research question we relied on hypothesis testing on the basis of the binomial distribution. Male and female denoting nouns are binomially distributed per verb frame

¹²This, however, is only possible if a verb form is available for the noun which is not the case for e.g. *Professor*; *Professorin* (Eng. professor).

¹³The participle present nominalization - according to German grammar books - should be used to indicate that some person is involved only temporarily (or even only at the moment) in the task denoted by the participle. *Singende* (singing people) are different from *Sänger* (singer), they only currently are singing. The new usage is not conform with this view, however if it gains acceptance, the grammar books had to be rewritten.

role. For instance the source role of *betrügen* (Eng. cheat) requires an animate filler which either could be denoted by a male or female noun. If a gender occupies a particular verb position significantly less or more often than the prior probability suggests, than an imbalance is found. Henceforth, we call under represented (less often) genders *scarce* and over represented (more often) genders *abundant*. For instance, if female nouns are significantly less often sources of a verb, we say that the verb is *female scarce* for that role. We omit the reference to the role name if it is clear from the context.

We give a schematic example of the statistical procedure: if a transitive (active voice) verb has $n = 2000$ instantiations (and thus 2000 sources) and $s = 100$ sources are female, then we determine the cumulative probability of up to 100 cases given 2000 trials with $p = 0.183$ as $\sum_{i=0}^{100} \text{binom}(i, 2000, 0.183)$. If this value is below $\alpha = 0.025$, then we reject H_0 and adopt H_1 , i.e. we can conclude that female nouns occur significantly less often as sources than male nouns, the verb is, thus, female scarce. It might be the case (but not necessarily) that male nouns occur significantly more often as sources of the same verb. To check this, the probability of having 1900 or more occurrences of male sources given that $p=0.817$ is determined ($1 - \sum_{i=0}^{1900} \text{binom}(i, 2000, p = 0.817)$).

We only kept verbs where a normal distribution could be assumed. This is given if $np \geq 5$. Resolved for n we have $n \geq 5/0.183 \geq 27.3$. Overall 380 verbs out of 1,000 verbs are above this threshold.

Most of the verbs are negative verbs. This is not only due to the imbalance in our verb lexicon (70% negative verbs), but also presumably due to the fact that news more often are negative than positive. In our discussion we thus focus on negative verbs and only refer briefly to positive cases in the last subsection of section 6.

6 Empirical Study

We first identified the gender-specific distribution of source and target roles given the set of polar verbs: for which gender which verbs (verb roles) are scarce and for which abundant. A particular verb role might be scarce for one gender and abundant for the other one (and vice versa)¹⁴. In these

¹⁴Please note that if female is scarce for some verb, male must not necessarily be abundant (and vice versa): the cumulative probabilities (even of complementary priors) do not necessarily distribute the mass of 1.

cases the imbalance is complementary. We call these verbs *gender prompted*. Table 6 shows an example of the constellation *gender prompted*.

	female	male
source	scarce	abundant

Table 6: Example of *gender prompted*: source of verb *ermorden* (Engl. to kill)

We not only looked at the distribution of a single role, but also at the combination of roles, the possible source-target pairings: female-female, female-male, male-male and male-female. If for a particular verb the source role is abundant for one gender and at the same time the target role is abundant for the other one, the verb reveals a gender opposition (because the verb expresses a negative relationship). We call these verbs *gender settled*. These cases represent the strongest gender-specific claim we can make. Table 7 gives a example of the constellation *gender settled*.

	female	male
source	-	abundant
target	abundant	-

Table 7: Example of *gender settled*: verb *bedrängen* (Eng. to harass)

6.1 Source Role

In this setting, we determined the gender specific occupation of the source role. 72 out of the 380 verbs (19%) are either scarce or abundant for some gender, the rest of the verbs shows no significant gender-specific instantiation pattern.

Out of the 72 verbs, 8 verbs are male scarce, 61 male abundant; 51 verbs are female scarce and 11 female abundant (72=61+11). The intersection of male abundant and female scarce (and vice versa) gives us those verbs that we called gender prompted, i.e. the role in question (here: source) is preoccupied by one gender and rarely ever filled by the other one. All 51 female scarce verbs are male abundant. Also all 8 male scarce verbs are female abundant. Thus, 59 of the 72 verbs are gender prompted verbs, that is about 15% of the 380 verbs.

In order to find out whether these scarce or abundant verbs might show a verb class specific gender distribution, we assigned each verb its verb class and determined for each gender and prompt type

(scarce, abundant) a distribution (see Table 8).

	physical	emotional	communicative
↓ ♀	47.06	17.65	35.29
↓ ♂	12.50	12.50	75.00
↑ ♀	9.09	18.18	72.73
↑ ♂	45.90	16.39	37.71

Table 8: Verb class specific distribution of gender (female ♀ or male ♂) scarce (↓) and abundant (↑): source

Each row shows the gender-specific verb class distribution of a type, e.g. female scarce (♀ ↓): 47.06 physical, 17.65% emotional and 35.29% communicative verbs. To give an example of a gender prompted constellation: the communicative verbs where male are scarce (75%) and female abundant (72.73%) are gender prompted, they are often verbs of accusation (8 out of 11, in italics): *accuse, betray, blame, denounce, discriminate, dismiss, incriminate, hate, avenge, reject, sue*.

We can see symmetrical pattern: female scarce is mainly in class physical (47.06%), which is the major group of male abundant (45.9%). On the other hand: male scarce (72.73%) is in the class communicative, which is the major group of female abundance (75%). Thus, male nouns are more often sources of physical violence, while female nouns are more often sources of verbal oppositions. We call it *opposition* instead of *violence*, since the class *communicative* is more heterogeneous than the class *physical*. A communicative negative verb might be one that hurts the patient verbally like *insult*, but also one that might be regarded a defense like *reproach* or *accuse*.

Male nouns are abundant sources of verbs like: *abuse, assault, attack, beat, coerce, complain, condemn, deny, despise, destroy, distort, harass, harm, hurt, insult, kill, murder, rage, rape, slaughter, terrorize*.

The source role of the gender prompted verbs in some cases can be further qualified with the strong notion of a villain. For instance, the source of *slaughter* is a highly negative actor, a villain. On the other hand, the actor of *reproach* cannot be further classified on a polar dimension. The most negative verbs are those that refer to physical (to kill), emotional (to hate) or verbal (to excoriate) violence. These verbs are modeled in our lexicon as having a negative actor. For each gender we determined the percentage of negative actorship. For male abundant, 43% out of the 51 verbs are of that

type, male denoting nouns can be regarded as negative actors in these cases. For female abundance this is just about the half, 23%.

6.2 Arousal and Dominance

As discussed, words (verbs) also carry arousal and reveal dominance. Can we also find gender-specific differences for these two parameters? For verbs with female and male actors: Is the gender specific arousal (dominance) associated with the prompted verbs in line with the prior probability?

What does arousal mean in the context of a polar verb? A high arousal of a negative verb indicates that the source is regarded as a rather negative actor (a villain) and the target as someone highly negatively affected (a victim). Dominance means that the target is in a clear subordinate position.

The overall prior probability for female was 0.183. Now that we are looking for the gender-specific arousal mass for source (actor) roles, we rather should use the MLE estimation of the gender-specific probability of being the source (and later the target), not the overall prior. For female nouns the probability of filling the source role is 0.164 (78,643 female sources out of 478,165 sources). The arousal (dominance) mass for female should thus be 16.4% of the total arousal (dominance) mass.

The gender-specific arousal (dominance) mass is the product of the arousal (dominance) value of a verb (with a particular gender as source) multiplied by the frequency of that verb. The total mass is the sum of both gender masses.

The total arousal mass of male and female verb tokens is 13,677 (rounded). Female arousal level should correspond to 16.4% of this mass, which is 2,246, but only 260 (1.9%) was found. The same is true for dominance, the overall mass is 20,836 but only 416 actually has been seen for female (2% instead of 16.4%). We can interpret this in the following way: compared to female, male (negative polar) actions are dominating and are much more negative emotion evoking. The only reason for the imbalanced mass distribution can be the magnitude of the arousal (dominance) level per verb. Male denoting nouns must occur (more often) as sources of verbs with high arousal (dominance) scores than female denoting nouns.

If we look at the arousal and dominance levels for the target (i.e. patient) role, we find that this time female nouns are much more affected than

their prior probability predicts. The MLE estimation of the female prior of the target role is 0.177 (32,264 female targets out of 182,530 targets). The arousal mass of verbs with female/male as targets is 2632. Female nouns cover 40.3% of it instead of 17.7%. The negative load for female targets is drastically higher than for male targets. Note however that in this setting the actor might be male or female (see section 6.4 for the gender-paired view). For dominance we get 41%.

6.3 Target Role

In this setting, we determined the gender specific occupation of the target role. 43 out of the 380 verbs (11.3%) are either scarce or abundant for some gender, the rest of the verbs shows no significant gender-specific instantiation pattern. 34 out of 43 are gender prompted (i.e. scarce for one gender, abundant for the other one).

	physical	emotional	communicative
↓ ♀	21.43	7.14	71.43
↓ ♂	91.00	9.00	0.00
↑ ♀	95.00	5.00	0.00
↑ ♂	30.43	4.35	65.22

Table 9: Verb class specific distribution of gender (female ♀ or male ♂) scarce (↓) and abundant (↑): target

All 14 female scarce verbs are male abundant and all 20 male scarce are female abundant. 34 of the 43 verbs are, thus, gender prompted verbs, that is 8.9% of the 380 verbs.

If we look at the verb classes (Table 9), female are scarcely targets of negative communication (71.43%), while male are (65.22%). Male are scarcely targets of (particular) physical violence (91%), while female are (95%). Note that high scarce male and high abundant male wrt. to a verb class are not contradicting, because the gender-wise intersection of scarceness verbs and abundant verbs is empty: some physical verbs are scarce, some abundant.

Almost 24% (source: 15% + target: 8.9%) of the 380 verbs are gender prompted. For these verbs female and male denoting nouns are complementary (scarce, abundant) fillers of the source or target role. This indicates a significant gender imbalance.

6.4 Verbs of Gender Opposition

Now that we have for each gender the information for which verb role it is abundant, we can find

cross gender cases of opposition, namely the constellation which we have called settled: verbs for which male abundant holds for one role and female abundant for the other one (and vice versa).

We have found 11 verbs with male sources and female targets that show gender opposition : *harass, molest, murder, shoot, abuse, coerce, terrorize, kill murder, rape, injure, assault* . All verbs are expressing physical violence.

For the inverse setting (with female source and male targets) three verbs are found: *denounce, incriminate, accuse* . All verbs of the class *communicative* .

From a very condensed point of view we might say that male denoting nouns cover villain roles (female being the victim), while female denoting nouns cover accuser roles (male being the accused).

We could also look into the gender internal pairings. Only rare cases were found. In the pairing male-male the following verbs are settled: *arrest, convict* . For female-female only *discriminate* was found. There are more cases of cross-gender that gender internal opposition.

7 Gender Profile Change

So far, we have discussed gender profiles on the basis of all data from the whole period. An interesting question might be whether this has changed over the years or whether it is a constant pattern in newspaper texts. We have compared the period OLD (2004-2014) with the period NEW (2018-2022). Period 2015-2017 was left out as a potential transmission period.

First of all, the prior probabilities of gender have changed. In period OLD the probability of a female denoting noun is 0.169, in period NEW 0.196. We carried out our experiments with these period-specific probabilities.

	physical	emotional	communicative
↓ ♀	50 (54.8)	16.7 (9.7)	33.3 (35.5)
↓ ♂	20 (11.1)	0 (0)	80 (88.9)
↓ ♀	20 (11.1)	0 (0)	80 (88.9)
↓ ♂	46.9 (50)	16.3 (13.6)	34.7 (36.4)

Table 10: The distribution of verb class instantiations for the source role: format 2004-2014 (2018-2022)

Table 10 shows the results for the source role for period OLD and NEW (with NEW in brackets). Slight tendencies can be noticed. Male are even more abundant in the class physical (50% instead

of 46.9%) and female less (11.1% instead of 20%). Also there is an increase in female abundance for communication verbs (from 80% to 88.9%) while the increase for this class for male is less high.

	physical	emotional	communicative
↓ ♀	10 (0)	20 (0)	70 (100)
↓ ♂	100 (100)	0 (0)	0 (0)
↑ ♀	95 (93.3)	0 (0)	5 (6.7)
↑ ♂	22.7 (45.5)	9.09 (0)	68.2 (54.5)

Table 11: The distribution of verb class instantiations for the target role: format 2004-2014 (2018-2022)

Table 11 shows the target role development. The most striking change is in verb class *physical* for male. Whereas in the period OLD 22.7% were male abundant, in NEW we have 45.4%. At the same time male is less abundant in the class *communicative* (a drop from 68.23% to 54.5%).

As we can see from the two tables, the profiles have slightly changed. What is surprising is the fact that the number of female denoting nouns has increased only by 2.7% (from 16.9% to 19.6%). We would have guessed a higher increase, given that gender awareness seemed to have raised in recent years.

8 The Positive Dimension

As mentioned previously, negative verbs are much more frequent than verbs expressing a positive relationship. We have focused, thus, on the against relation in this study. However to complete the picture we might have a brief look into positive relations and the gender specific patterns in this section. We start with the source role. Female abundant verbs are *honor, celebrate, rejoice, win, help, love, like, fall in love, forgive, appreciate*. All female abundant verbs are gender prompted (are at the same time male scarce). Male abundant verbs are *accept, liberate, insist, affirm, respect, care, concede, reveal*. Also all male abundant are gender prompted.

	physical	emotional	communicative
↓ ♀	12.50	50.00	37.50
↓ ♂	27.27	63.64	9.09
↑ ♀	27.27	63.64	9.09
↑ ♂	11.11	44.44	44.44

Table 12: The distribution of verb class instantiations for the source role of positive verbs

Table 12 shows the verb class distribution. It is interesting to see that emotion verbs are much more prominent for positive verbs than for negative ones. Communicative verbs are least abundant for female (9.09%) which is quite the opposite to negative verbs (where it was 72.73%). Physical verbs are less important for the positive relationships.

The statistics for the target role case are too meager to be of any significance. There are 5 verbs that are gender prompted, namely *honor, encourage, love, care, fall in love*. They are female abundant and male scarce. Male is scarcely patient of these verbs while female are abundantly often. Statistics for positive gender cross abundance cannot be found in our data set.

9 Related Work

Bias detection and debiasing are important research topics (see [Stanczak and Augenstein \(2021\)](#) for a survey). Researchers use e.g. pointwise mutual information (PMI) to measure the association of words with gender ([Stanczak et al., 2021](#)). We are rather interested in statistically supported claims about gender-specific instantiation patterns of verbs.

In an approach more closely related to ours, [Sun and Peng \(2021\)](#) observe a gender-specific tendency to combine personal and professional events in the Wikipedia pages of celebrities, an asymmetric association where e.g. women’s personal events appear more often in the career section than for men. They also establish higher efficiency when extracting events (verb denotations) over analyzing raw text for detecting this gender bias. To this aim, they use the odds ratio (OR), calibrate over synthetic sentences to estimate real occurrence frequencies, and select the events with the largest gender differences.

We are not aware of other animacy detection approaches for German. Also there is no gender classifier available apart from ours. In [Klenner et al. \(2023\)](#), the initial version of our gender classifier applied to gender-tailored role labeling was introduced.

Gender classification in English is primarily restricted to predicting the gender of text author(s) (e.g. bloggers, see [Mukherjee and Liu \(2010\)](#)). Other researchers analyzed the ACL anthology to find gender specific research topics ([Vogel and Jurafsky, 2012](#)). However this is restricted to the recognition of the gender of person names. [Campa](#)

et al. (2019) aim to identify whether the subject of an article is female or male based on (the content of) the headlines. A gold standard of headlines was created and used where male and female reference could be found. Among others, a CNN approach reached an accuracy of 86.7%. In contrast, we do not identify the gender of the subject of the whole text, but of source and target roles of verbs.

Gender profiling is also a task in the area of computational forensic linguistics (Sousa-Silva, 2018), see e.g. the shared task on Bots and Gender Profiling 2019¹⁵. The task is to determine whether a tweet is from a human or a bot and if human which gender. Again, the gender of the author is profiled, not as in our case the gender of text referents.

We are not aware of any sentiment inference approach to German others than ours. For English, a couple of approaches exist. A rule-based approach to sentiment inference is Neviarouskaya et al. (2009). Each verb instantiation is described from an *internal* and an *external* perspective. For example, “to admire a mafia leader” is classified as affective positive (the subject’s attitude towards the direct object) given the internal perspective while it is (as a whole) a negative judgment, externally (here the concepts introduced by the Appraisal Theory are used, cf. Martin and White (2005)).

Rashkin et al. (2016) introduce connotation frames to represent various types of connotations using typed relations. They consider the writer’s perspective, the entity’s perspective, effects, values as well as mental states. For each predicate, they infer a connotation frame composed of 9 relationship aspects. In contrast to our setting (real sentences), their experiments are based on crowd sourcing with artificial, rather simple sentences (just subject/object, no subclauses).

Choi and Wiebe (2014) address methods for creating a sense-level lexicon for opinion inference. They consider expressed opinions towards events that have positive or negative effects on entities. As words have mixtures of senses among the three classes (+/-effect and Null), they develop a sense-level rather than word-level lexicon. The resulting resource is based on WordNet senses, annotated with one of the aforementioned classes. In contrast, our annotations consider not only effects on entities but also relations between entities as well as actors.

A more recent approach is described in Park

¹⁵See <https://pan.webis.de/clef19/pan19-web/author-profiling.html>

et al. (2021). The authors call the underlying task *direct sentiment extraction to question answering (DSE2QA)* which essentially is what others have called sentiment implicature (cf. Deng et al. (2014)). On the basis of a manually labeled corpus on the 2016 U.S. presidential election and on COVID-19, a method is developed that is utilizing BERT-like pretrained transformers. Questions (Does X has negative sentiment towards Y) on whether a particular relationship exists or not are used, answers are aggregated to make a final guess. This approach actually anticipates recent developments in the context of GPT-like models like ChatGPT. The authors of Zhang et al. (2023) show that ChatGPT outperforms existing approaches in the area of stance detection. Moreover, it is also able to explain its answer. The authors claim that this is a crucial new property of such models. We have carried out a couple of initial experiments with ChatGPT as well. A sentence and prompt like *Mister Tiber refuses to help his sick neighbor. Is he in favor or against her?* is answered with *Mister Tiber’s refusal to help his sick neighbor suggests that he is against her*, after removal of *sick* the chatbot now finds the prompt *difficult to determine*. This hesitant reaction was typical in our experiments. To find the right prompt is the task to solve in such contexts. As soon as the concrete training procedure behind it has been published, stance-tailored versions of ChatGPT might finally prove superior to other approaches. The chatbot is also able to do gender identification. The following question-prompt pair was correctly resolved: *Die ZDF-Moderatorin log die Verantwortliche des Aufsichtsrats an. Wer ist weiblich?*¹⁶ (Eng. *The ZDF presenter lied to the person in charge of the supervisory board. Who is female?*)¹⁷. The correct answer is *ZDF-Moderatorin, Verantwortliche*. Since the idea of science is not to just develop prompting skills, we have to wait until we have access to the exact methodological details of such models.

10 Conclusion and Outlook

In this paper, we focused on a gender-tailored analysis of newspaper texts. We searched for the gender profiles in terms of the gender-specific roles newspapers convey. We strived to fix those events (denoted by verbs) that are gender prompted, i.e.

¹⁶It is not the grammatical (female) gender, ChatGPT referred to - we checked this.

¹⁷When we tried the same question one day later, ChatGPT failed to give an answer.

descriptions where male or female denoting nouns are occurring significantly less or more often than expected. An even stronger, gender opposition indicating case are gender settled verbs, where the source role is abundantly filled by nouns of one gender and the target role by nouns of another one.

The profiles that we have found clearly cast male nouns as filling negative actor roles while female nouns are as targets negatively affected. Moreover female nouns as source role fillers are accusers and male the accused. The primary goal of this work is not a particular statistical screening but the development of a methodology which allows to validate (confirm or reject) claims that otherwise must be regarded as mere long-shot guesses. Our approach may also be used in other genres (e.g. fiction instead of news) in which a particular imbalance (e.g. men committing physical violence) may not (claim to) reflect reality, but rather some potential bias in the data that must be checked.

From a technical perspective, we introduced the first gender-specific classifier (as far as we are aware of). We combined it with a rule-based sentiment inference system for gender profiling. Our empirical study was carried out in the established statistical setting of hypothesis testing.

In future work, we like would to apply our approach to new data where coreference resolution is possible in order to increase the statistical basis of our claims. Also, other expressions like e.g. noun phrases with polar adjectives modifying gender denoting nouns could supplement our verb-specific view. The overall goal is an ever more fine-grained apparatus for gender profiling. At some point, we also will focus on gender inclusive reference and how to combine this with our current approach.

Acknowledgements

This work was supported by the Swiss National Foundation (SNF) under the project number 105215_179302 from 2018 to 2022. I would like to thank Anne Göhring for her collaboration in the project. Thanks also to Alison Jong-Ju Kim and Dylan Massey for their valuable support.

Discussion of Limitations

Our method detects gender imbalance by using an existing rule-based system and a grammatical gender classifier. Neither performs perfectly, and we do not claim that our sampling methods produce representative data drawn from the whole popula-

tion. Rather, we work with a subset that can be identified by our tools. Generalizing from the subset to the population is not our intention; our approach is a attempt to carry out gender-tailored sentiment analysis. We do not claim to find biases in the data, we instead speak of imbalance. Whether the cause of imbalance is bias would require an additional qualitative analysis of the results.

References

- Stephanie Campa, Maggie Davis, and Daniela Gonzalez. 2019. [Deep & machine learning approaches to analyzing gender representations in journalism](https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1194/reports/custom/15787612.pdf). In <https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1194/reports/custom/15787612.pdf>.
- Eunsol Choi, Hannah Rashkin, Luke Zettlemoyer, and Yejin Choi. 2016. [Document-level sentiment inference with social, faction, and discourse context](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 333–343, Berlin, Germany. Association for Computational Linguistics.
- Yoonjung Choi and Janyce Wiebe. 2014. [+/-effectwordnet: Sense-level lexicon acquisition for opinion inference](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar*, pages 1181–1191.
- Simon Clematide and Manfred Klenner. 2010. [Evaluation and extension of a polarity lexicon for german](#). In *Proceedings of the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA)*, pages 7–13, Lisbon, Portugal.
- Lingjia Deng and Janyce Wiebe. 2014. [Sentiment propagation via implicature constraints](#). In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 377–385, Gothenburg, Sweden. Association for Computational Linguistics.
- Lingjia Deng, Janyce Wiebe, and Yoonjung Choi. 2014. [Joint inference and disambiguation of implicit sentiments via implicature constraints](#). In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 79–88, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.
- Anne Göhring, Manfred Klenner, and Sophia Conrad. 2021. [Deinstance: Creating and evaluating a german corpus for fine-grained inferred stance detection](#). In *17th Conference on Natural Language Processing (KONVENS 2021)*, pages 213–217. ACL Anthology.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. [Bag of tricks for efficient text classification](#). In *Proceedings of the 15th Conference of the European Chapter of the Association*

- for *Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.
- Manfred Klenner. 2018. **Offensive language without offensive words (olwow)**. In *KONVENS 2018 (The Conference on Natural Language Processing. GermEval Task 2018 – Shared Task on the Identification of Offensive Language)*. GSCL.
- Manfred Klenner and Michael Amsler. 2016. **SenTiframes: a resource for verb-centered german sentiment inference**. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1–4, Paris, France. European Language Resources Association (ELRA).
- Manfred Klenner, Simon Clematide, and Don Tuggener. 2017a. **Verb-mediated composition of attitude relations comprising reader and writer perspective**. In *Computational Linguistics and Intelligent Text Processing - 18th International Conference, CICLing 2017, Budapest, Hungary, April 17-23, 2017, Revised Selected Papers, Part II*, volume 10762 of *Lecture Notes in Computer Science*, pages 141–155. Springer.
- Manfred Klenner, Anne Goehring, Alison Yong-Ju Kim, and Dylan Massey. 2023. **Gender-tailored semantic role profiling for german**. In *12th International Conference on Computational Semantics (IWCS)*, Nancy, France.
- Manfred Klenner and Anne Göhring. 2022. **Semantic role labeling for sentiment inference: A case study**. In *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, pages 144–149, Potsdam, Germany. KONVENS 2022 Organizers.
- Manfred Klenner and Anne Göhring. 2022. **Animacy denoting german nouns: Annotation and classification**. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 1360–1364, Marseille, France. European Language Resources Association (ELRA).
- Manfred Klenner, Don Tuggener, and Simon Clematide. 2017b. **Stance detection in Facebook posts of a German right-wing party**. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 31–40, Valencia, Spain. Association for Computational Linguistics.
- J. R. Martin and P. R. R. White. 2005. *Appraisal in English*. Palgrave, London.
- Saif Mohammad. 2018. **Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words**. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.
- Karo Moilanen and Stephen Pulman. 2007. **Sentiment composition**. In *Proceedings of the Recent Advances in Natural Language Processing International Conference (RANLP-2007)*, pages 378–382, Borovets, Bulgaria.
- Arjun Mukherjee and Bing Liu. 2010. **Improving gender classification of blog authors**. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 207–217, Cambridge, MA. Association for Computational Linguistics.
- Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. 2009. **Semantically distinct verb classes involved in sentiment analysis**. In *Proceedings of the IADIS International Conference Applied Computing 2009, 19-21 November, Rome, Italy, 2 Volumes*, pages 27–35. IADIS Press.
- Kunwoo Park, Zhufeng Pan, and Jungseock Joo. 2021. **Who blames or endorses whom? entity-to-entity directed sentiment extraction in news text**. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4091–4102, Online. Association for Computational Linguistics.
- Hannah Rashkin, Sameer Singh, and Yejin Choi. 2016. **Connotation frames: A data-driven investigation**. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 311–321, Berlin, Germany. Association for Computational Linguistics.
- Rui Sousa-Silva. 2018. **Computational forensic linguistics: An overview of computational applications in forensic contexts**. *Language and Law*, 5(2):119–143.
- Karolina Stanczak and Isabelle Augenstein. 2021. **A survey on gender bias in natural language processing**. *CoRR*, abs/2112.14168.
- Karolina Stanczak, Sagnik Ray Choudhury, Tiago Pimentel, Ryan Cotterell, and Isabelle Augenstein. 2021. **Quantifying gender bias towards politicians in cross-lingual language models**. *CoRR*, abs/2104.07505.
- Jiao Sun and Nanyun Peng. 2021. **Men are elected, women are married: Events gender bias on Wikipedia**. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 350–360, Online. Association for Computational Linguistics.
- Adam Vogel and Dan Jurafsky. 2012. **He said, she said: Gender in the ACL Anthology**. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 33–41, Jeju Island, Korea. Association for Computational Linguistics.
- Bowen Zhang, Daijun Ding, and Liwen Jing. 2023. **How would stance detection techniques evolve after the launch of ChatGPT?** *CoRR*, abs/1207.0016.

Multimodal Offensive Meme Classification with Natural Language Inference

Shardul Suryawanshi¹ and Mihael Arcan¹ and Suzanne Little² and Paul Buitelaar¹

^{1,2} Insight SFI Research Centre for Data Analytics

¹ Data Science Institute, University of Galway, Ireland

² Dublin City University, Ireland

{shardul.suryawanshi, mihael.arcan, suzanne.little,
paul.buitelaar}@insight-centre.org

Abstract

Multimodal offensive meme classification is a challenging classification task, where a multimodal meme needs to be classified as offensive or not offensive based on the provided text and image. A well-known approach to solving this problem is to fuse the text and image features captured either by the text and image encoder or by a transformer architecture to form a multimodal meme representation. In our work, we argue that the image features captured by the image encoder are unable to capture the abstract representation like language. Hence, we propose to transform the multimodal offensive meme classification task into an unimodal offensive text classification task for which we leverage the Natural Language Inference (NLI) task. Firstly, we carefully generate image captions using an off-the-shelf image captioner and automatically transcribed the meme as if it was explained to a visually impaired individual. Later, these meme transcriptions and labels (image-text-label) have been transformed into NLI format (premise-hypothesis-label). To evaluate our approach, we run benchmark analysis on Memotion, Hateful memes and MultiOFF datasets (in their NLI format) using four baselines finetuned on Emotion Analysis, Sentiment Analysis, Offensive tweet Classification, and NLI task. We achieve state-of-the-art (SOTA) results for the MultiOFF dataset and close to SOTA results for Memotion while achieving competent evaluation scores on the Hateful Memes dataset.

1 Introduction

Memes in the social media context are means of expressing emotions and ideas (Du et al., 2020). They easily propagate across various cultures due to their ability to mutate and spread (Dawkins, 2016). Hence, memes have become an integral part of online communication. But sadly, they have become the means of spreading hatefulness and offensiveness towards an individual or a group based

on but not limited to their ethnicity, sexual orientation, and religion (Suryawanshi et al., 2020). A multimodal or Image-with-text (IWT) (Du et al., 2020) offensive meme contains an image embedded with the text with either an image or text or both being offensive. Hence, it is necessary to consider both the image and text modality for the multimodal offensive meme classification.

The multimodal offensive meme classification task (Suryawanshi et al., 2020; Sharma et al., 2020a; Kiela et al., 2020) is a classification task where one needs to classify if the meme is offensive based on the image and text modalities associated with the meme. The nature of the task is multimodal since both the image and text modalities are required for the classification. The research community has been actively organizing shared tasks (Sharma et al., 2020a; Suryawanshi and Chakravarthi, 2021) and competitions (Kiela et al., 2020) to solve this challenging task.

Previous research in this area proposed novel approaches that combined both the image and text modalities using deep learning techniques, most of which leverage VL pre-training, which involves a large corpus of image and text. However, VL pre-trained models are susceptible to domain shifts when finetuned on a small multimodal offensive memes dataset (Singh et al., 2020); Additionally, the quality of global multimodal representations learnt during the VL pre-training might degrade after finetuning on out-of-domain datasets (Singh et al., 2020).

Language is more abstract than image. It condenses information better than the image. For example, when we refer a word “cat”, we could imagine cat from cartoons shows such as “Tom and Jerry”, “Garfield” to a real world cat. The word in itself condenses all the information. A well documented human knowledge is in text which could be learnt from language models. On the other hand, if we consider the Selena Gomez meme from Figure

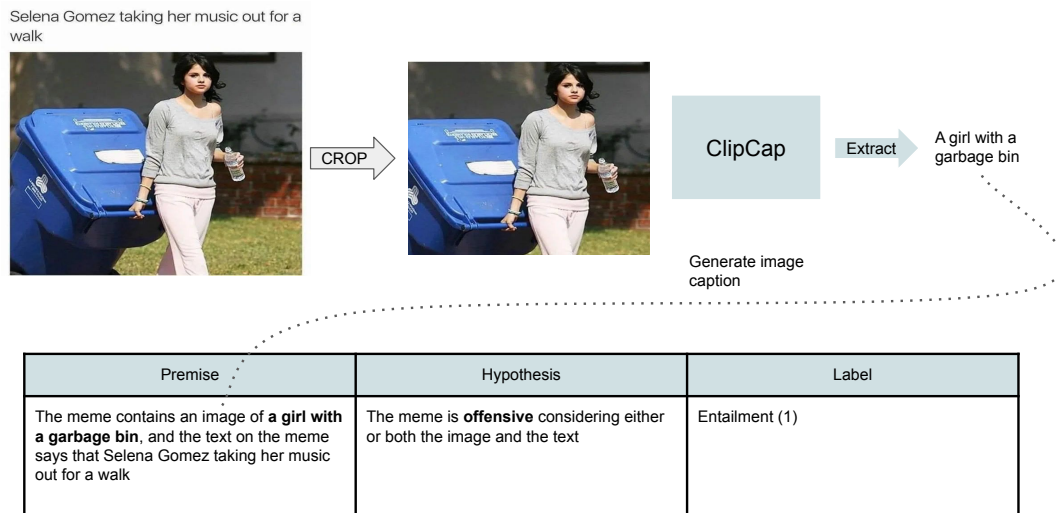


Figure 1: Step by step method of NLIfication of the multimodal data: First the image is cropped to get rid of the text. Later, the image caption is generated using ClipCap which is incorporated in the meme transcription along with the text associated with the meme. The old label (OFF or NOT) translated into a hypothesis^a. Lastly, the new label is assigned as 1 (entailment) if the old label was OFF else 0 is assigned.

^aHypothesis is identical for each data sample

1 whereby her music is called out as rubbish, the meme is still offensive even after Selena Gomez is replaced by any other musician. In such case, a VL pre-trained will tune its parameter to accommodate multiple variations. But, by transforming the task to unimodal will reduce such variations. Because the meme captions generated for each such meme would be “a girl/boy/person with a garbage bin”. Moreover, research by (Prajwal et al., 2019) shows that memes could be made accessible to a visually impaired individual through meme transcriptions. Hence, we propose to transform the multimodal offensive meme classification task into an unimodal offensive text classification task. Hence, we hypothesize that the transforming images into text could aid in our task.

In our research, we design a systematic framework that utilises NLI task to transform the multimodal task into the unimodal one. Firstly, we transform the data from image-text-label format into premise-hypothesis-label format. Later, the newly transformed data is used for finetuning three RoBERTa models previously finetuned on Emotion Analysis, Sentiment Analysis, Offensive tweet classification and NLI task respectively. We performed three ablations for each model to gain a better understanding of each model’s behaviour. Furthermore, we lay out detailed quantitative and qualitative error analysis of the task

2 Related Work

The research community has been actively facilitating supervised datasets (Sharma et al., 2020a; Kiela et al., 2020; Suryawanshi et al., 2020) to contribute towards solving the multimodal offensive meme classification task. However, unlike their text counterparts (Zampieri et al., 2019, 2020; Risch et al., 2021) these supervised multimodal datasets are smaller. In our research, we are utilizing three popular datasets: Memotion, Hateful memes and MultiOFF datasets.

Initially, researchers opted for a sequence to sequence (Seq2Seq) architecture for capturing text and image features with two encoders and later on fusing them to classify if the meme is offensive. But due to the efficiency of transformers over Seq2Seq, and the advent of VL pre-training, the research has been shifted towards transformer-based architectures such as LXMERT (Tan and Bansal, 2019), Visualbert (Li et al., 2019), VILBERT (Lu et al., 2019), UNITER (Chen et al., 2020). A Seq2Seq approach proposed by (Sharma et al., 2020b) fuses image features derived from InceptionNet and text features derived from the GloVe embedding. The feature fusion proposed in their research uses Bi-LSTM initialized with image features as hidden and cell state and calculated attention over the text features. They were able to score first rank with a macro-average F-score of 0.52907 on the Memotion shared task in subtask B: Humour Classification. A winning solution (AUROC: 0.8449, Accu-

racy: 0.7320) for the Hateful memes challenge by (Zhu, 2020) proposes an ensemble model that combines VL-BERT, UNITER-ITM, VILLA-ITM and ERNIE-Vil. Moreover, the authors extracted the entity, gender and race of the individuals from the meme by using face extraction with Mask-RCN. (Zhong et al., 2022) proposed injecting an external knowledge base in the form of entity recognition from the meme text to enhance the semantic representation of the meme. However, they relied on the raw image features captured via VGG. Their approach established new SOTA results (precision: 0.670, recall: 0.671, f-score: 0.671) for the MultiOFF dataset. In summary, all of these top-scoring approaches are multimodal. However, we propose an unimodal approach where we use the image caption of a meme (meme caption) as a text feature as a replacement for the raw image features. We are comparing our results with these current SOTAs and baselines in Section 4.

We take inspiration from (Prajwal et al., 2019), they suggest that memes could be transcribed to the visually impaired individual using carefully generated facial image captions. We argue that one might lose crucial information from the meme by just concentrating on facial image captions. Hence, we crop the meme to get rid of the unnecessary meme text, we consider the cropped meme as whole over just faces while generating meme captions. (Yin et al., 2019) proposes a framework that leverages the NLI task for zero-shot text classification. We closely follow this approach in our work but unlike their research, we use our framework to finetune the text classifier rather than zero-shot classification. Moreover, we just use one hypothesis for each data sample rather than generating true and false hypothesis for each sample.

All the Multimodal SOTA’s are complex and computationally heavy due to millions of trainable parameters. Moreover, they ensemble multiple VL models which is less practical since such models are complex to deploy in the real world. Hence to make the solution more simpler and practical, we propose to transform the multimodal offensive meme classification problem into an unimodal offensive text classification problem by leveraging the NLI task.

3 Data Pre-processing

As shown in the Figure 1, first we crop the image to avoid the text embedded in the meme. The meme is

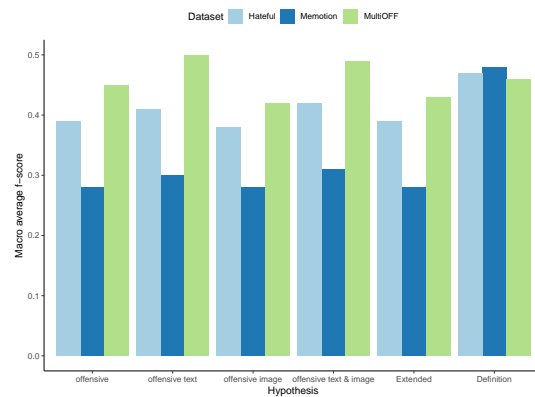


Figure 2: Clustered bar graph of macro f-score of RoBERTa in zero-shot setting for each Memotion, Hateful memes and MultiOFF dataset. Y-axis represent macro f-score while x-axis shows the hypothesis.

cropped on both the length (l) and breadth (b) by $\frac{1}{4}$ margin based on the manual inspection of random sample drawn from each dataset. Later, we generate the image caption using the ClipCap (Mokady et al., 2021) image captioner. We incorporate the generated captions inside the meme transcription along with the text associated with the meme which is used as a premise. Finally, the offensive label is converted into a natural sentence and a new label i.e. entailment label is assigned to either 1 or 0 if the meme is offensive or not offensive respectively.

3.1 Meme Transcription

CLIP by (Radford et al., 2021) gives competent results for Hateful Memes dataset in the zero-shot setting. Hence, we opted for ClipCap image captioner based on the CLIP image encoder with GPT2 prefix decoder (pre-trained on the MS-COCO dataset (Chen et al., 2015)) to generate meme captions at inference time. We transcribed memes by combining these meme captions with the meme text (the text embedded on the meme provided along with each dataset). The template used to automatically transcribe the meme is “The meme contains an image of meme caption, and the text on the meme says that meme text”. For example, the meme in Figure 1 is transcribed as “The meme contains an image of a girl with a trash can, and the text on the meme says that Selena Gomez is taking her music out for a walk”. In this example, the text “a girl with a trash can” is a meme caption, and the text “Selena Gomez is taking her music out for a walk” is a meme text.

3.2 NLI-fication

Figure 1 shows the overview of transforming data from image-text-label to premise-hypothesis-label

Dataset	Train size	Val size	Test size	Epochs	Learning rate	Batch size	Weight decay	Grad acc
Memotion	5,940	660	1,878	10	5.0e-7	8	0.001	4
Hateful	7,650	850	2,000	10	5.0e-5	8	0.001	16
MultiOFF	445	149	149	10	1.0e-5	8	0.001	8

Table 1: On the left side of the vertical line: Data statistics in terms size of training, validation and test for each Memotion, Hateful and MultiOFF dataset. On the right side of the vertical line: Hyper-parameter settings for all the three RoBERTa for Memotion, Hateful memes and MultiOFF dataset in terms of number of epochs, learning rate, batch size, weight decay and gradient accumulation steps .

format. We refer to this procedure as NLI-fication in the rest of the article. The meme transcription acts as a premise in the context of the Natural Language Inference (NLI) task, while the label i.e. “OFF” has been transformed into the hypothesis. We experimented with the hypothesis on three levels: primary, extended, and definition. Primary level does not use the natural sentence to explain the label, rather it just uses the label i.e. “offensive” as a hypothesis. The motivation behind the NLI-fication of the data comes from the fact that the model would get a better understanding of the label once it has been translated into a hypothesis in natural sentence. We tried different versions of the primary hypothesis by adding more words—offensive text, offensive meme, offensive image and text, offensive image or text—to the primary hypothesis. The extended hypothesis is just the natural sentence that describes the label “The meme is offensive considering either or both the image and the text”. At the definition level, we add the definition for the hate or offensive content provided by (Kiela et al., 2020) i.e. “A direct or indirect attack on people based on characteristics, including ethnicity, race, nationality, immigration status, religion, caste, sex, gender identity, sexual, orientation, and disability or disease. We define attack as violent or dehumanising (comparing people to non-human things, e.g. animals) speech, statements of inferiority, and calls for exclusion or segregation. Mocking hate crime is also considered hate speech.” We report the zero shot results on each dataset in Figure 2, the figure shows the averaged macro f-score of RoBERTa finetuned on NLI data (SNLI and MNLI) across all three (Memotion, Hateful memes, Multioff) datasets. We chose RoBERTa because it achieved SOTA results on NLI task despite being a small compared to bigger language models such as T5, BART, BigBird. The figure shows that the model yielded the highest mean macro f-score at “Definition level”. This points out that the hypothesis with more offence related keywords worked the best. Because, the

rest hypothesis (other than the definition) just has the word “offensive” as offence-related keyword, while definition has more keywords such as attack, dehumanising, mocking, hate, crime. The fact that “Definition level” works best in the zero shot setting emphasises that the model has a knowledge of the offensive keywords which could be improved upon further finetuning. We maintained identical “Definition level” hypothesis for each data sample across all three datasets. This does not only simplify our approach but also removes manual overhead of hypothesis tuning based on each sample. Hence, making our approach more generalizable to new multimodal offensive datasets.

4 Experimental Settings

4.1 Baselines

The baselines are based on the use of the finetuned dataset. Emotion and Sentiment of the text acts as a auxiliary information to offensiveness of the text (Mnassri et al., 2023). Hence, emotions and the sentiment of the text play an important role in identifying the offensiveness of the text. Moreover, the model finetuned on the offensive tweets could prove as a strong baseline due to the inter-training on closely related offensive tweet classification dataset (Choshen et al., 2022). We use RoBERTa fine-tuned on the Emotion, Sentiment and Offensive tweet classification data (Barbieri et al., 2020) as baselines. Specifically, we chose “twitter-roberta-base-sentiment¹”, “twitter-roberta-base-emotion²”, and “roberta-base-offensive³” respectively for Emotion, Sentiment, and Offensive RoBERTa baselines. These models are finetuned on the short text i.e. tweets, which is similar to the text captions embedded in the memes. These models are loaded with pop culture knowledge since they are finetuned on 54M tweets before

¹<https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment>

²<https://huggingface.co/cardiffnlp/twitter-roberta-base-emotion>

³<https://huggingface.co/cardiffnlp/roberta-base-offensive>

finetuning on the Emotion, Sentiment Analysis and Offensive datasets respectively. Hence, we believe that it is easier for these models to adapt to the domain of our task. We deliberately use RoBERTa for each of baseline to maintain the consistency and comparability of the results across the baselines and our approach. Table 1 shows the data statistics and hyperparameters used for each each dataset across all experiments. Moreover, we compare our results with current SOTAs along with mentioned baselines. We already cover their details in Section 2, we collectively call them Multimodal SOTA irrespective of the dataset.

4.2 Significance test

We performed a 5X2 significance test (Dietterich, 1998) for each model pair for the Memotion dataset to show that they are significantly different from each other. The significance test is primarily five-fold cross-validation performed two times. The macro-averaged f-score is recorded for each fold, resulting in 10 macro-averaged f-score which are used later to calculate the p-value and t-statistics for each pair of models: Inference Vs Emotion RoBERTa, Inference Vs Sentiment RoBERTa, Inference Vs Offensive RoBERTa, Emotion Vs Sentiment RoBERTa, Emotion Vs Offensive RoBERTa, and Offensive Vs Sentiment. Here, the null hypothesis is these pairs do not differ significantly from each other.

4.3 Our Approach

Based on the text classification framework proposed by Yin et al. (2019), we finetuned RoBERTa (on binary NLI dataset with 28k samples labelled as "entailment" and "not entailment") which was previously finetuned on the NLI datasets such as Stanford Natural Language Inference (SNLI) (Bowman et al., 2015) and Multi-Genre Natural Language Inference (MNLI) (Williams et al., 2018). We chose RoBERTa based on its state-of-the-art performance on the NLI task. Moreover, it was pre-trained not only on masked language modelling (MLM) but also on the sentence prediction objective which we thought would be helpful in our task since we are dealing with two different texts i.e. premise and hypothesis. Since our classification task is binary (labels: OFF or NOT), we decided to finetune RoBERTa on a binary NLI dataset. For this purpose, we sampled 28,000 examples from the combined SNLI and MNLI dataset and converted their labels into *Entailment* (1) or *Not entailment* (0)

Models	p-value	t-statistics
Inference Vs Emotion	1.15e-05	17.39
Inference Vs Sentiment	6.00e-06	19.85
Inference Vs Offensive	1.07e-06	28.07
Emotion Vs Sentiment	1.52e-06	26.16
Emotion Vs Offensive	2.70e-09	93.11
Offensive Vs Sentiment	5.47e-08	50.99

Table 2: 5 X 2 significance test results for Inference, Emotion and Offensive RoBERTa (against each other) with respect to Memotion dataset.

by encoding the *neutral* and *contradiction* label of the original dataset into Not entailment. Amongst 28,000 examples, 4,000 each were sampled randomly for the validation set and test set. The model was trained for five epochs and the best model with the least validation loss was saved during the training. We used this saved model later on to finetune the multimodal offensive meme datasets.

4.4 Ablations

We performed three ablations on each of the experiments. The first ablation uses just the meme caption (meme-captions-only), and the second ablation uses just the text from the meme (meme-text-only). In these ablations, we intend to study the impact of each text individually on the performance of the experiments evaluated with precision, recall and f-score. In the third ablation (no-NLI-fication), we removed the NLI-fication of the data from the data pre-processing pipeline and used just the premise as a text by removing the hypothesis altogether. In the last ablation, we intend to study the effect of NLI-fication on each Emotion, Sentiment, Offensive, and Inference RoBERTa model.

5 Quantitative error analysis

We refer to scores reported in (Mokady et al., 2021) for quantitative error analysis of the ClipCap image captioner whereby it is evaluated with 32.15 using Bleu@4, 27.1 using METEOR, 108.35 using CIDEr, and 20.12 using SPICE evaluation scores. These scores are close to that of other image captioning models such as BUTD, VLP, and OSCAR.

Table 2 shows results from 5 X 2 significance test. It could be seen in the table that all the p-values are less than 0.05. Hence, we do not have enough confidence to accept the null hypothesis: all the pairs of the models are not significantly different from each other. Hence, we reject the null hypothesis. Emotion Vs Offensive RoBERTa

Dataset	Models	Class	Precision	Recall	F-score
Memotion	Multimodal SOTA	macro	-	-	52.90
	Emotion	OFF	63.21	87.87	73.53
		NOT	43.20	15.28	22.57
		macro	53.20*	51.57	48.05
	Sentiment	OFF	62.44	84.63	71.86
		NOT	38.14	15.70	22.24
		macro	50.29	50.16	47.05
	Offensive	OFF	62.71	65.50	64.08
		NOT	38.32	35.50	36.86
		macro	50.52	50.50	50.47
	Inference (our approach)	OFF	64.54	56.28	60.13
		NOT	40.26	48.80	44.12
macro		52.40	52.54	52.12↓	
Hateful	Multimodal SOTA	Accuracy/AUC-ROC	73.20	0.8449	-
	Emotion	HATE	51.22	39.33	44.49
		NOT	68.05	77.52	72.48
		macro	59.63	58.43	58.49
	Sentiment	HATE	59.00	39.33	47.20
		NOT	69.67	83.60	76.00
		macro	64.33	61.47	61.60
	Offensive	OFF	55.87	52.67	54.22
		NOT	72.54	75.04	73.77
		macro	64.21	63.85	64.00
	Inference (our approach)	HATE	61.93	40.13	48.71
		NOT	70.34	85.20	77.06
macro		66.14	62.67	62.88	
		Accuracy/AUC-ROC	68.23↓	0.3662↓	-
MultiOFF	Multimodal SOTA	macro	67.00	67.00	67.00
	Emotion	OFF	54.29	65.52	59.37
		NOT	74.68	64.84	69.41
		macro	64.48	65.18	64.39
	Sentiment	OFF	54.67	70.69	61.65
		NOT	77.03	62.64	69.09
		macro	65.85	66.66	65.37
	Offensive	OFF	45.00	93.10	60.67
		NOT	86.21	27.47	41.67
		macro	65.60	60.29	51.17
	Inference (our approach)	OFF	59.09	67.24	62.90
		NOT	77.11	70.33	73.56
macro		68.10↑	68.79↑	68.23↑	

Table 3: The quantitative results of experiments: The report presents the detailed evaluation results in terms of class-wise and macro-averaged precision, recall and F1 score. Accuracy and AUC-ROC* denotes Accuracy and AUC-ROC scores for Hateful Memes dataset for comparing our approach with SOTA. The \uparrow and \downarrow in Inference section indicates if our approach surpassed the current SOTA or not.

shows the largest t-statistics which means they are more significantly different than any other pair.

Table 3 shows the detailed classification report –with class-wise and macro averaged precision (p), recall (r) and f-score (f)– of each Emotion, Sentiment, Offensive and Inference RoBERTa on Memotion, Hateful memes and MultiOFF datasets. The highlighted bold cased score shows the macro averaged p, r, f score for each RoBERTa model, and with * denoting the highest macro-averaged score. In this table we are evaluating our approach in two ways. Firstly, we evaluate against the baselines (Emotion, Sentiment, and Inference RoBERTa) whereby we compare macro-averaged p, r, f scores. Hence, these scores highlighted in bold for better

readability. Secondly, we evaluate against Multimodal SOTAs whereby we either use \uparrow or \downarrow to specify if our approach has surpassed the SOTAs or not respectively.

For the first part of the evaluation, it could be seen clearly that the macro averaged evaluation score is increased in the inference RoBERTa over Sentiment and Emotion RoBERTa across all datasets except for the fact that macro averaged precision of Emotion RoBERTa (53.20%) is greater than that of the Inference RoBERTa (52.40%). However, the difference between the macro-average recall of the two models was significant (4.08%). This difference shows that Inference RoBERTa shows more balanced class-wise

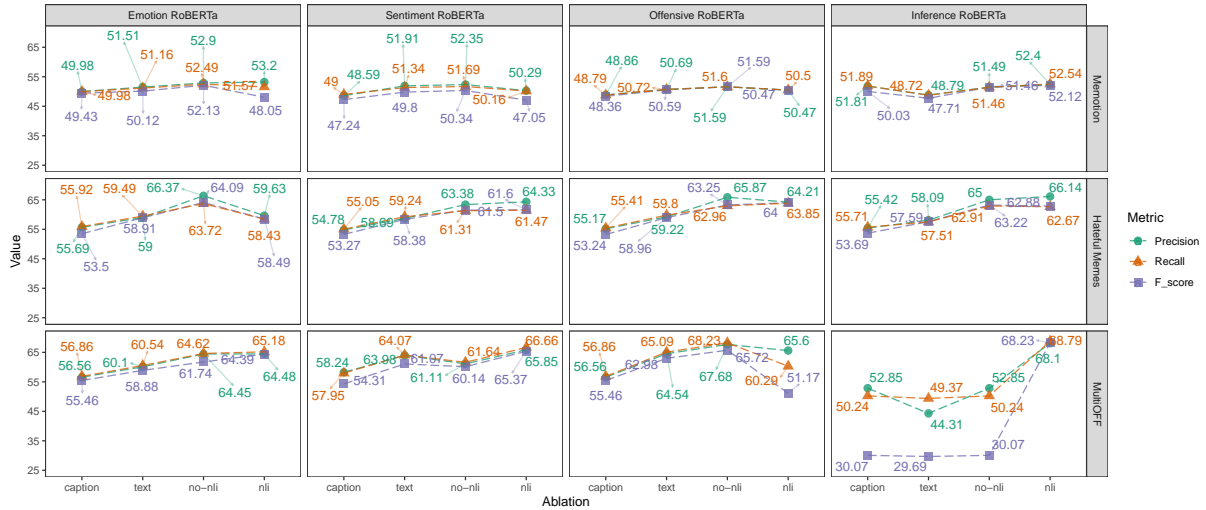


Figure 3: Line plot for the macro-averaged p, r, f-score for each ablation—meme-caption-only (caption), meme-text-only (text), no-NLI-fication (no-nli)— along with original experiment (nli).

precision and recall with less false positive and false negative count than that of Emotion RoBERTa. The increment in the evaluation scores of Inference RoBERTa compared to the other two could be credited to the NLI finetuning performed on binary NLI dataset (28k samples labelled as “entailment” and “not entailment”). Offensive RoBERTa shows better macro averaged recall and f-score compared to that of Inference RoBERTa for Hateful Memes dataset. This shows that Offensive RoBERTa while retains more offensive memes compared to the Inference RoBERTa at the expense of lesser macro averaged precision. It is interesting to observe that Offensive RoBERTa performs better than Emotion and Sentiment RoBERTa in the case of Memotion and Hateful memes dataset, but it fails to beat the two in case of MultiOFF dataset. This shows that even after offensive tweet classification being closely related to our downstream task, less training data leads to poor generalisation.

For the second part of the evaluation, in the case of the Memotion dataset, we achieve close to Multimodal SOTA performance (our macro-averaged f-score: 52.12%, SOTA: 52.90%). Since the official evaluation metric is accuracy and AUC-ROC for the Hateful memes challenge, we calculated both the metric for the best performing RoBERTa. Inference RoBERTa with the highest macro-averaged p, r, f score showed an accuracy of 68.30% and AUC-ROC of 0.3662 which is less than that of the Multimodal SOTA with an accuracy of 73.20% and AUC-ROC of 0.8449. The difference in the performance could be attributed to the reduced complexity of our approach since the winning solution used a complex ensemble technique that leveraged complex pre-trained models such as VL-BERT, UNITER-

ITM and VILLA-ITM. If this complexity is taken into account then the difference in the accuracy (4.9%) is not more. However, the less AUC-ROC of Inference RoBERTa shows that the model is more susceptible to threshold change when compared with the SOTA. In case of the MultiOFF dataset, we beat the Multimodal SOTA (Zhong et al., 2022) (p: 67.10%, r: 67.00%, f: 67.10%). Since the MultiOFF dataset consists of only 743 examples (# train: 445, # validation: 149, # test: 149), our approach shows robust performance in terms of new SOTA results even with fewer training samples.

Figure 3 shows the detailed evaluation report – across all of the three datasets and three models– in terms of macro-averaged p, r, f score on the three ablations (meme-caption-only, meme-text-only, no-NLI-fication). One common trend amongst graphs of Emotion, Sentiment and Offensive RoBERTa showed the least macro-averaged p, r, f score in meme-caption-only ablation compared to the rest. Moreover, it could be seen that the p, r, f scores for meme-text-only ablations are better than that of the meme-caption-only ablations for the three models. This shows that the meme text plays a more vital role than the meme caption at identifying offensive memes in the case Emotion, Sentiment and Offensive RoBERTa across all three datasets. However, the Inference RoBERTa trained on the Memotion and MultiOFF dataset show contradictory trend where meme-caption-only ablation shows better evaluation score than that of meme-text-only. But the same model in meme-text-only ablation shows improvement in evaluation scores over the meme-caption-only ablation in case of Hateful Memes dataset. This indicated that meme captions are more reliable features for Inference

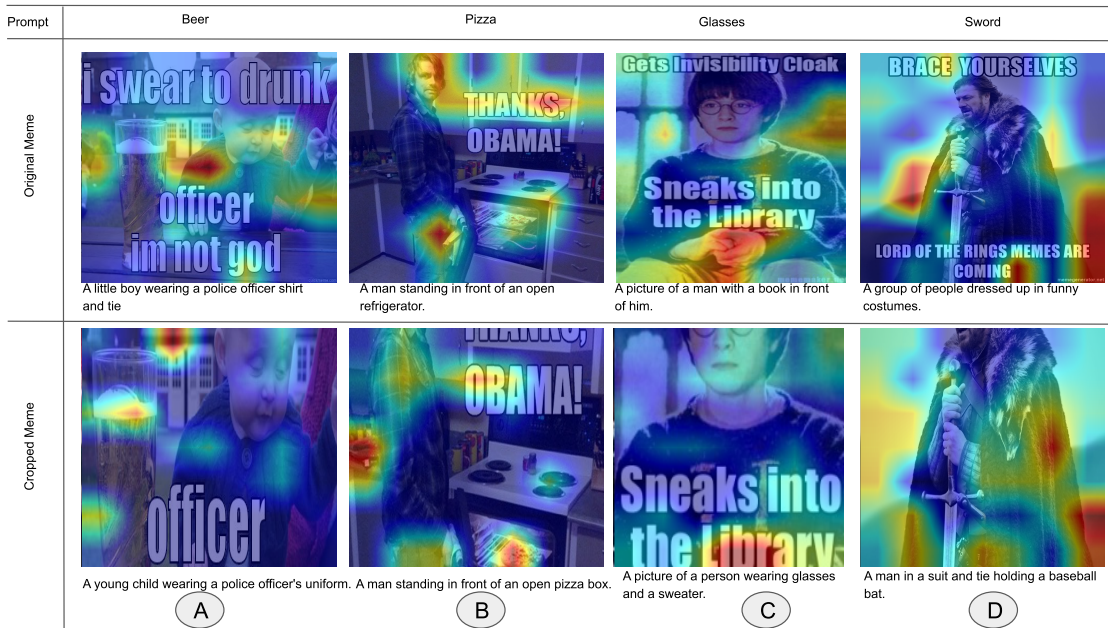


Figure 4: The first row represents original memes, and the second row represents cropped memes. The captions of the meme are mentioned below each image.

RoBERTa than meme texts when it comes to offensive dataset such as Memotion and MultiOFF. Moreover, the contradictory trend shows the difference in the working of Inference RoBERTa against the other two corroborated by the significance test results mentioned in the Table 2. The inference RoBERTa shows the peak performance in NLI settings than any other ablation. This illustrates that Inference RoBERTa is getting a bump in the performance due to the finetuning of the binary NLI data. Similarly, Emotion RoBERTa –in the case of MultiOFF dataset– Sentiment RoBERTa–in the case of the Hateful memes dataset and MultiOFF dataset– showed improvement in the performance in NLI settings compared to no-NLI-fiction ablation. However, Emotion RoBERTa showed a decline in the performance in NLI settings in the case of the Memotion and Hateful Memes dataset while Sentiment RoBERTa showed similar results in the case of the Memotion dataset. Irrespective of these differences in the performance of each model in Hateful memes and Memotion datasets, we could see that all three models show improvement in macro-averaged p, r, f score in NLI settings for MultiOFF dataset. On the one hand, Emotion RoBERTa in no-NLI-fiction ablation beats the Inference RoBERTa in the NLI setting in the case of the Memotion and Hateful Memes dataset. On the other hand, Offensive RoBERTa showed highest

evaluation scores across all the three datasets in no-NLI-fiction settings. This highlights the competency of the RoBERTa finetuned on the emotion and offensive tweet classification data. This shows that the emotion and offensive tweet classification could also be leveraged for offensive meme classification. It could be seen that the performance of the inference RoBERTa in NLI settings shows the highest evaluation scores which beat the current SOTA (p:67%, r:67%, f:67%). Moreover, our Offensive RoBERTa in no-NLI-fiction settings comes close to SOTA (p:67.08%, r:68.23%, f:65.72%). This shows the robustness of our approach in low sample settings. The significant marginal difference between no-NLI-fiction ablation and the original NLI shows that the Inference RoBERTa can perform better while leveraging the NLI knowledge gained after finetuning on the NLI dataset specially for small dataset.

6 Qualitative error analysis

In this section, we analyse the inference for the examples from the test samples of each Memotion, Hateful Memes and MultiOFF dataset. Firstly, we would like to highlight some of the examples from the ClipCap prefix image captioner in Figure 4. The first row in the Figure shows the prompt used to generate the heatmap. In the context of computer vision, heatmaps are used to identify the regions in

Image				
Meme captions	A man in a suit and tie standing next to another man in a suit and tie	A close up of a person wearing a suit and tie	a close up of a small animal near a field of grass	An old picture of a woman holding an umbrella
Dataset	Memotion	Memotion	MultiOFF	Hateful Memes
Inference RoBERTa	OFF	NOT	NOT	NOT
Emotion RoBERTa	OFF	OFF	NOT	HATE
Sentiment RoBERTa	OFF	OFF	NOT	NOT
Gold Label	OFF	OFF	NOT	HATE

Figure 5: Inference result on examples from Memotion, MultiOFF and Hateful memes for each Inference, Emotion and Sentiment RoBERTa. Please note that the meme captions mentioned here are generated after cropping the meme.

an image that are likely to contain objects of interest. Higher intensity or warmer colors (e.g., red or yellow) in the heatmap indicate higher confidence or probability of the object being present at those locations, while lower intensity or cooler colors (e.g., blue or green) indicate lower confidence. We used gScoreCAM (Chen et al., 2022) to generate heatmap from CLIP. These maps are generated top 1000K channels out of total 3072K channels. The centre of the map is dark red and turns to the lighter shade outwards. This indicates the confidence of the heatmap which is higher at the centre and lowers in the outward direction. The second row in the figure represents original memes along with their caption stated below the meme. Similarly, the last row represents cropped memes with their captions stated below the meme. We chose prompt based on wrong noun predicted in either original or cropped meme caption.

The first example circled (A) shows that the image captioner correctly identified the young child in the meme in both the original and cropped meme. However, it falsely identified the word `officer` from the meme text being printed on the child’s top. Furthermore, the important object here to be detected was `beer`. We prompted CLIP with the word `beer` on both original and cropped meme. It can be seen that no heatmap is present near the

original, but the cropped meme shows two such heatmaps on the object `beer`. This shows that the CLIP is more confident at selecting the required object in the cropped version. The third example circled (C) shows that the image captioner correctly identified meme captions after cropping unlike the meme caption for the original meme which emphasized the word `library`. Moreover, heatmaps generated for the prompt `Glasses` is present on the glasses on the cropped meme, but nowhere seen near the glasses in original meme. The example circled (B) shows improvement in the quality of the meme caption after cropping the meme as it could be seen that the object falsely recognized as `refrigerator` has been replaced by the correct one i.e. `pizza box`. Furthermore, if we prompt both memes the word `pizza`, the original meme shows bigger heatmap concentrated around meme text `THANKS`, and a small less confident heatmap on the `pizza`. However, heatmap on the cropped meme is concentrated on the object `pizza` as well as `pizza-like` object on the left side of the meme. All the examples (A), (B) and (C) shows that the cropped meme not only generated better captions but also helps CLIP to capture useful image feature. In this case, the meme text is acting like an adversary while capturing useful image feature. However, in example circled (D), the image cap-

tioner falsely recognized `sword` as a `baseball bat` instead after cropping the meme. If word `sword` is prompted to CLIP, the original as well as cropped meme fails to capture the useful image features as shown in the heatmap. This could be attributed to the fact that word `baseball bat` has been observed more in MS-COCO dataset, hence the captioner is biased towards such words. All examples show the sensitivity of the image captioner towards the meme text. Hence, although our approach works optimally with the current state of the image captioner, meme caption quality could be improved by completely removing the meme text from the meme.

Figure 5 shows the detailed report on the inference results for each Inference, Emotion and Sentiment RoBERTa on examples from the test set of Memotion, MultiOFF and Hateful memes datasets. The first column in the table illustrates an example from the Memotion dataset. This example is offensive as it intends to demean Obama⁴. All the models were able to correctly identify the given example as OFF. On the other hand, the example from the second column which is labelled as OFF has been incorrectly classified as NOT by Inference RoBERTa. But if we take a look at the meme from the example, it does not mean to harm or attack anyone. Hence, it could be labelled as NOT. This shows the noisiness of the dataset which could be attributed to the annotation process. The example from the third column belongs to the MultiOFF dataset. Here, all the models were able to correctly classify the given meme into the NOT category. The example in the fourth column showed interesting results since the meme has been incorrectly captioned which led to the failure of Inference and Sentiment RoBERTa while Emotion RoBERTa succeeded. This difference in the performance of the models shows a difference in their pattern recognition ability which has already been proven by the 5X2 significance test shown in Table 2.

7 Conclusion

All the experiments and their ablation suggest that the transforming multimodal offensive meme classification into unimodal offensive text classification problem not only simplifies the approach but also achieves SOTA results. It could also be seen that the NLI-fication of the multimodal data could improve the evaluation metric, especially in the

case of a smaller dataset (MultiOFF). Emotion RoBERTa outperformed its counterpart after removing the NLI-fication of the data while Sentiment RoBERTa fell short by a minute margin. This shows that the models finetuned on the Emotion and Sentiment Analysis task could prove useful in the offensive meme classification task. Overall, it is a viable option to translate the multimodal offensive meme classification into a unimodal (text) classification problem to get competent evaluation scores. Moreover, NLI-fication is not only simple but also effective at training on smaller out of domain dataset.

Limitations

In the qualitative error analysis, we observed the sensitivity of the CLIP prefix image captioner towards the meme text. This approach may generate an out of context meme caption which later could harm the performance of the model. Moreover, Figure 4 (D) shows inferior image captions upon cropping. Hence, fixed cropping $\frac{1}{4}$ margin along length and breadth could lead to information loss which results in incorrect captions. To tackle this issue in future, we plan to use an in-house image captioner model which will ignore the noise generated from the meme text without cropping it. To better understand the image captioning errors, we plan to train our model on a small subset of manually human-generated image caption.

Ethics Statement

The definition of "offensive" content is highly subjective and can vary across different cultures and communities. Hence, the same content that is deemed for certain group or community might not be offensive to others. Therefore, marginalised groups may be disproportionately affected by the model's decisions.

Acknowledgements

This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2 for the Insight SFI Research Centre for Data Analytics, co-funded by the European Regional Development Fund.

⁴44th President of the United States

References

- Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. [TweetEval: Unified benchmark and comparative evaluation for tweet classification](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.
- Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. [A large annotated corpus for learning natural language inference](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.
- Peijie Chen, Qi Li, Saad Biaz, Trung Bui, and Anh Nguyen. 2022. [gscorecam: What objects is clip looking at?](#) In *Proceedings of the Asian Conference on Computer Vision*, pages 1959–1975.
- Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. 2015. [Microsoft coco captions: Data collection and evaluation server](#).
- Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. [Uniter: Universal image-text representation learning](#). In *ECCV*.
- Leshem Choshen, Elad Venezian, Shachar Don-Yehia, Noam Slonim, and Yoav Katz. 2022. [Where to start? analyzing the potential value of intermediate models](#). *arXiv preprint arXiv:2211.00107*.
- Richard Dawkins. 2016. *The selfish gene*. Oxford university press.
- Thomas G Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923.
- Yuhao Du, Muhammad Aamir Masood, and Kenneth Joseph. 2020. [Understanding visual memes: An empirical analysis of text superimposed on memes shared on twitter](#). In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 153–164.
- Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. [The hateful memes challenge: Detecting hate speech in multimodal memes](#).
- Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. [Visualbert: A simple and performant baseline for vision and language](#). *arXiv preprint arXiv:1908.03557*.
- Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. [Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks](#). *arXiv preprint arXiv:1908.02265*.
- Khoulood Mnassri, Praboda Rajapaksha, Reza Farahbakhsh, and Noel Crespi. 2023. [Hate speech and offensive language detection using an emotion-aware shared encoder](#). *arXiv preprint arXiv:2302.08777*.
- Ron Mokady, Amir Hertz, and Amit H. Bermano. 2021. [Clipcap: Clip prefix for image captioning](#).
- K R Prajwal, C V Jawahar, and Ponnurangam Kumaraguru. 2019. [Towards increased accessibility of meme images with the help of rich face emotion captions](#). In *Proceedings of the 27th ACM International Conference on Multimedia, MM '19*, page 202–210, New York, NY, USA. Association for Computing Machinery.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision](#). *CoRR*, abs/2103.00020.
- Julian Risch, Anke Stoll, Lena Wilms, and Michael Wiegand. 2021. [Overview of the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments](#). In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 1–12, Duesseldorf, Germany. Association for Computational Linguistics.
- Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020a. [SemEval-2020 task 8: Memotion analysis- the visiolingual metaphor!](#) In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 759–773, Barcelona (online). International Committee for Computational Linguistics.
- Mayukh Sharma, Ilanthenral Kandasamy, and W.b. Vasantha. 2020b. [Membusters at SemEval-2020 task 8: Feature fusion model for sentiment analysis on memes using transfer learning](#). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1163–1171, Barcelona (online). International Committee for Computational Linguistics.
- Amanpreet Singh, Vedanuj Goswami, and Devi Parikh. 2020. [Are we pretraining it right? digging deeper into visio-linguistic pretraining](#). *arXiv preprint arXiv:2004.08744*.
- Shardul Suryawanshi and Bharathi Raja Chakravarthi. 2021. [Findings of the shared task on troll meme classification in Tamil](#). In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 126–132, Kyiv. Association for Computational Linguistics.
- Shardul Suryawanshi, Bharathi Raja Chakravarthi, Michael Arcan, and Paul Buitelaar. 2020. [Multimodal meme dataset \(MultiOFF\) for identifying offensive content in image and text](#). In *Proceedings of the*

Second Workshop on Trolling, Aggression and Cyberbullying, pages 32–41, Marseille, France. European Language Resources Association (ELRA).

- Hao Tan and Mohit Bansal. 2019. **LXMERT: Learning cross-modality encoder representations from transformers**. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.
- Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. **A broad-coverage challenge corpus for sentence understanding through inference**. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.
- Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. **Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach**. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923, Hong Kong, China. Association for Computational Linguistics.
- Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86.
- Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447.
- Qi Zhong, Qian Wang, and Ji Liu. 2022. Combining knowledge and multi-modal fusion for meme classification. In *MultiMedia Modeling*, pages 599–611, Cham. Springer International Publishing.
- Ron Zhu. 2020. **Enhance multimodal transformer with external label and in-domain pretrain: Hateful meme challenge winning solution**. *CoRR*, abs/2012.08290.

**Pre-Trained
Language Models and
Knowledge Probing**

MEAN: Metaphoric Erroneous ANalogies dataset for PTLMs metaphor knowledge probing

Lucia Pitarch

Universidad de Zaragoza
lpitarch@unizar

Jorge Bernad

Universidad de Zaragoza
jbernad@unizar.es

Jorge Gracia

Universidad de Zaragoza
jogracia@unizar.es

Abstract

Despite significant progress obtained in Natural Language Processing tasks thanks to Pre-Trained Language Models (PTLMs), figurative knowledge remains a challenging issue. This research sets a milestone towards understanding how PTLMs learn metaphoric knowledge by providing a novel hand-crafted dataset, with metaphoric analogy pairs where per correct analogy pair, other three erroneous ones are added controlling for the semantic domain and the semantic attribute. After using our dataset to fine-tune SoTa PTLMs for the multiclass classification task we saw that they were able to choose the correct term to fit the metaphor analogy around the 80% of the times. Moreover, thanks to the added erroneous examples on the dataset we could study what kind of semantic mistakes was the model making.

1 Introduction

Metaphors are not only very common devices but also key elements in language. They both help us express ourselves and shape the way we think by using a concept to reference and delimit another (Lakoff and Johnson, 1980).

For instance, let’s look at the following example extracted from The Guardian:

The intriguing echo of Eliza in thinking about ChatGPT is that people regard it as **magical** even though they know how it works – as a “**stochastic parrot**” (in the words of Timnit Gebru, a well-known researcher) or as **a machine for “hi-tech plagiarism”** (Noam Chomsky). (Naughton, 2023)

In the same paragraph three views about ChatGPT¹ are compared: either it is conceived as a magical device, as a ‘stochastic parrot’ meaning

it only repeats statistical patterns, or as a plagiarism tool. The metaphors and narratives we use to talk about Artificial Intelligence tools such as GPT have a huge impact on the sentiment we have towards them, being an already flagged concern at the European Parliament (Boucher, 2021).

Despite the pervasiveness and impact of metaphors in language and culture, processing them remains challenging for Natural Language Processing. Approaches taken towards them have shifted from pattern and statistical-based discovery since Shutova et al. (Shutova, 2015), towards Language Model exploitation for their discovery and interpretation (Ge et al., 2022). While the second approach is providing more efficient models and accurate results, in comparison to pattern-based methods it lacks interpretability. Moreover, it has been stated that PTLMs lack figurative knowledge (Liu et al., 2022) and have trouble processing it (Czinczoll et al., 2022). Though uncovering the kind of knowledge PTLMs encode has been a major concern since their origins (Petroni et al., 2019), attention to the figurative knowledge they keep has just gained attention in the last year. And if interpretability is a major concern in the Artificial Intelligence community (Bender et al., 2021) it should be even more relevant when treating metaphors, as they are especially sensitive devices that can be used to change the way we perceive the world (Semino et al., 2017).

At the moment, questions such as the following ones are being researched:

1. Do PTLMs encode figurative knowledge? (Liu et al., 2022; Aghazadeh et al., 2022)
2. Do PTLMs have figurative analogical reasoning? (Czinczoll et al., 2022; Chen et al., 2022)
3. What kind of figurative knowledge is the most challenging one? (Liu et al., 2022)

¹Open AI’s generative large language model

Our work follows the goal of understanding how PTLMs process figurative language, particularly the one dealing with metaphors, it adds a new research question to the ones already addressed in the literature, namely: ‘How do PTLMs acquire figurative knowledge?’, and contributes towards it in the following ways:

1. We provide MEAN, a novel manually curated dataset² with selected metaphoric analogies from MetaNet (Dodge et al., 2015) enriched with erroneous examples. Its main aim is to uncover what aspects of the metaphor PTLMs learn.
2. We test our dataset on the metaphoric analogy completion task and provide novel baselines for it.
3. We obtain promising results in the metaphor analogy task, suggesting PTLMs after fine-tuning can acquire semantic inference abilities for metaphor interpretation tasks.

2 Related Work

Probing language models to understand what linguistic and common ground knowledge they encode has been a major research line since 2019 with the arrival of Pre-Trained Language Models with transformer architecture (PTLMs) (Devlin et al., 2019). Simultaneously, computational metaphor processing has also benefited from such PTLMs and regained attention, leading to huge advances in metaphor identification, interpretation, and generation tasks (Ge et al., 2022; Rai and Chakraverty, 2020). Yet, just very recently, in 2022, these two interests are being aligned (PTLMs probing and computational metaphor processing), resulting in works such as (Liu et al., 2022; Chen et al., 2022; Czinczoll et al., 2022; Aghazadeh et al., 2022), where researchers try to uncover the figurative knowledge encoded in PTLMs.

When conducting probing tests in metaphor detection tasks, Aghzadeh et al. (2022), came to the conclusion that PTLMs do encode figurative knowledge, particularly in their middle layers, yet other authors (Liu et al., 2022) when experimenting with probing in metaphor generation and interpretation tasks highlight the inability of PTLMs to capture figurative language. The mentioned works probe

PTLMs in fill in the mask tasks. This kind of setting has as limitation that several words can correctly fill in the gap in the sentence, and if just one or two options are given as gold standard the possibilities of not having a match between the predicted token and the gold one are high. The solutions they apply to minimize this effect are using Mean Reciprocal Ranking metrics and (Chen et al., 2022; Czinczoll et al., 2022) also search if the synonyms of the predicted tokens match their gold standard. Additionally, the fill-in-the-mask setting, has trouble dealing with multi-words, as only one token is selected to fill in the mask, yet metaphoric expressions are usually multi-words. Thus, the experimental setting we choose is more similar, though still different to the one proposed by Liu et al. (2022) who instead of conducting a fill-in-the-mask task, perform classification experiments. Particularly they provide as the first part of the sentence a verbalized metaphor and as the second part of the sentence the verbalized explanation of the metaphor. Given the metaphor and two possible explanations, the model has to select the best fit between both. In their experiment, they claim that even if in zero-shot environment figurative language understanding is extremely challenging for PTLMs, they can in fact learn it after some fine-tuning. Moreover, by annotating the kind of background knowledge needed to understand the inputted metaphors, they observe object and commonsense metaphors were easier to interpret while sarcastic metaphors were the most difficult ones. The later research is the most similar to our own one, as it focuses on probing the knowledge of figurative language in PTLMs through a metaphor interpretation task, while they focus on paraphrasing we focus on metaphoric inference by the completion of metaphoric analogies. Moreover, we explore where the semantic challenge relies (either on the semantic domain or attribute) by manually selecting the errors.

3 MEAN Dataset

If we understand metaphor as a linguistic device used to express something in terms of another thing (Lakoff and Johnson, 1980), this means two conceptual domains are involved, the source domain is the one that the speaker is using in the text and the target domain is the implicit one, trying to be expressed.³ Source domain is expressed in the

²Our code and dataset are openly available at <https://github.com/sid-unizar/MEAN.git>

³In metaphor literature conceptual domains are understood as the background knowledge needed to understand

text by particular lexical entries which make reference to different elements involved in the source domain. These elements have their corresponding elements in the target domain, which is implicitly referenced through the explicit expression of the source domain elements. Such process of drawing correspondences between the source and target domain in a metaphor through the expression of the individual elements involved is called *metaphor mapping* (Kövecses, 2016). A natural way of representing such correspondences and inputting them to PTLMs is via analogical reasoning as in (Czinczoll et al., 2022). That is, we can rewrite the metaphor mapping as "source domain is to target domain what source element is to target element".

For instance in this quote from an article in Nature: ‘Although OpenAI has tried to put guard rails on what the chatbot will do, users are already finding ways around them.’⁴ The metaphor being expressed there would be: ‘Artificial Intelligence is a moving vehicle’, the source domain would be ‘moving vehicle’ and the target domain ‘Artificial Intelligence’, the lexical entries being used metaphorically in the text (or in other words, the source element) are ‘putting guard rails around’ and ‘them’ in ‘users are already finding ways around them’ the metaphoric mapping from this lexical entry to its correspondent one in the Artificial Intelligence domain would be ‘firewall’ or ‘security measures’ to avoid things such as bias or missusage of the tool.

Our dataset consists of analogy pairs where the first part of the analogy contains the metaphor source and target domains and the second part consists of the individual lexical entries that could serve as instances in the text of the metaphor. Both the source and target domains and the first set of lexical entries proposed in the dataset are a subsample extracted from MetaNet (Dodge et al., 2015). MetaNet is a repository of metaphors and frames containing almost 700 conceptual metaphors, design to aid the computational exploration of corpora. From them we just selected the ones which had assigned one or more metaphor mappings between the different frame entities and which had the pattern ‘A are B’. We extend MetaNet data by adding curated erroneous endings to the analogy. The three erroneous target elements per analogy were manually selected following linguistic crite-

ria to control what the model is learning and to which semantic aspect of the metaphor it is paying attention to. If the criteria for a target element to properly fit the analogy is that it has to share the semantic domain with the target domain and the semantic attribute with the source element, then erroneous examples are when one of these criteria fails. We consider as semantic domain the general category to which the target domain and target element belong. Semantic attribute is the specific role that an individual element within that domain might play; for instance the semantic domain of ‘hospital’ would be ‘healthcare’ and the role it plays inside the healthcare domain would be ‘location’. In our dataset, an element is added per analogy for each of the three erroneous possibilities found when these criteria are not met. Namely:

1. the target element fits the same semantic domain as the target domain of the metaphor, but has a different attribute than the proposed source element (shortened as sDdA in Tables 1 and 4);
2. the target element shares the same attribute as the source element, but does not share the semantic domain with the target domain (shortened as dDsA in Tables 1 and 4);
3. or it has both different semantic domains and attributes from the needed ones (shortened as dDdA in Tables 1 and 4).

The resulting dataset contains 166 analogies (composed of a source domain, a target domain, a source element, a four target element candidates withing which just one is correct) made for 71 different metaphors (composed by a source and target domain pair) and 100 different source and target metaphor domains. At the moment the dataset exists just for English. A sample of our dataset can be found in Table 1.

4 Experiments

In this section we describe the different choices taken for fine tuning the model and testing our approach.

4.1 Multiple choice task

We fine-tune and test BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) models both in their large and base versions for multiple choice

a text (Clausner and Croft, 1999).

⁴In <https://tinyurl.com/NatureAnon2023>

Source Domain	Target Domain	Source Element	Target Element (Gold)	Target Element (Erroneous)		
				sDdA	dDsA	dDdA
anger	fire	anger level	fire intensity	wood	flood magnitude	unicorn
taxation	punishment	taxer	punisher	prison	vampire	amusement

Table 1: Sample of the dataset.

classification using hugging’s face library⁵. The task consists of providing the model with a beginning and four possible endings of a sentence, among which just one is correct. The first part of the sentence is the verbalized first pair of the analogy with the source and target domains of the metaphor. The second part of the sentence contains the individual source and target elements of the metaphor, where the last element (target element) varies to cover the four possible choices of our dataset. In Table 2 the different templates to verbalize the analogies are summarized. We experiment with three different verbalization which range from minimal templates with just punctuation to larger templates with more complex phrasings, following previous literature on prompting (Schick and Schütze, 2022).

Start template	End template	id
' W1 ' : ' W2 '	' W3 ' : ' W4 '	T1
' W1 ' is to ' W2 '	what ' W3 ' is to ' W4 ' .	T2
If ' W1 ' is like ' W2 ' ,	then ' W3 ' is like ' W4 ' .	T3

Table 2: Templates and identifiers used along the paper to identify them. In order to create an input sequence for a language model, the start and end templates are joined with the sep token, and, in the case of BERT models, the tokens of the start and end templates have a different token type.

This kind of task in comparison with fill-in-the-mask settings, benefits from being able to deal with whole sequences of tokens, facilitating dealing with multiword expressions. Moreover, as the answer is selected from a closed set of items we can better control the model output and what it is learning by biasing each of the possible answers with a particular linguistic restriction (in our case different domain and attribute selection).

As our dataset is very small, the provided results for the PTLMs consist of the mean accuracy of a 10-fold cross-validation and a 95% confidence interval for the mean accuracy calculated by bootstrapping (Efron, 1979).

⁵Original code, setup and documentation from hugging face at: https://huggingface.co/docs/transformers/tasks/multiple_choice

Fine tuning setting. To fine-tune the models, we used the following hyperparameters: batch size of 8, Adam optimizer with weight decay of 0.01 and learning rate of 2e-5, no warm-up, and training during 5 epochs.

4.2 Baselines

To compare whether fine-tuning with the metaphors provided in our dataset improved the model’s output we compare the results obtained to the static 300-dimensional embeddings from three different models: GloVe (Pennington et al., 2014), word2vec (Mikolov et al., 2013), and fastText (Bojanowski et al., 2017). All models were retrieved via Gensim (Rehurek and Sojka, 2011). To avoid Out of Vocabulary words the following strategy, similar to the one in (Speer et al., 2017), was followed: for a word, remove the last character until the word is found in the model. To deal with multiword expressions, the mean of the word embeddings were calculated.

Since the problem is posed as an analogy task, the cosine similarity is used to discover the best target element from a set of predefined ones following (Mikolov et al., 2013). That is, given a source and target domain word embeddings, s_d and t_d , a source element s_e , and a set of target elements $T = \{t_{e_1}, \dots, t_{e_k}\}$, solve the following equation:

$$\operatorname{argmax}_{t_e \in T} \{\cos(s_e + t_d - s_d, t_e)\}$$

4.3 Error analysis

Additionally to analyse with which semantic feature the model is having more trouble (attribute or domain distinction) when choosing the correct analogy we report percentages of the different error types made by the model.

5 Results and discussion

Table 3 shows the accuracy of RoBERTa and BERT models for each of the provided templates and compares them to GloVe, word2vec, and fastText baselines. A huge improvement can be observed when finetuning the model and shifting from static to contextual embeddings. The high results obtained

point to the ability of PTLMs to learn metaphorical analogy inference, coincidentally with the conclusions obtained by (Liu et al., 2022).

	Acc.	CI
Baselines		
GloVe	32.5	-
word2vec	33.7	-
fastText	45.8	-
BERT large		
T1	85.5	(81.5, 89.7)
T2	69.3	(52.5, 83.6)
T3	87.3	(83.7, 91.0)
RoBERTa large		
T1	84.9	(72.3, 93.3)
T2	86.7	(83.2, 90.8)
T3	74.7	(65.6, 83.8)
BERT base		
T1	84.3	(76.5, 91.1)
T2	78.9	(68.5, 87.2)
T3	84.9	(78.6, 90.8)
RoBERTa base		
T1	75.3	(61.7, 85.1)
T2	80.1	(74.4, 86.2)
T3	88.0	(81.7, 93.5)

Table 3: Results for baselines and fine-tuned PTLMs. The reported accuracy for PTLMs is the mean of a 10-fold cross-validation. For these latter cases, it is also reported a 95% confidence interval (CI) calculated by bootstrapping.

In Table 4 the percentages per error type in the classification are shown. On all models and templates, the most errors were made by predicting a target element that shared the same domain as the source element but had a different attribute than the target domain. This could point to a lesser knowledge of PTLMs regarding semantic roles. Further research should be done on this line. In future work, we will experiment with injecting this kind of linguistic knowledge into PTLMs models for metaphor interpretation tasks.

6 Conclusions and Future Work

By experimenting with our novel dataset with selected erroneous answers: MEAN, we conclude PTLMs can learn, through fine tuning, metaphoric analogical reasoning, improving the baselines stated by static embeddings. We also observed most errors were made by confusing the needed

	sDdA	dDsA	dDdA
Model			
BERT base	76.2	20.0	3.8
BERT large	58.8	33.8	7.5
RoBERTa base	84.5	12.7	2.8
RoBERTa large	56.0	29.9	14.2
Templates			
T1	64.0	27.2	8.8
T2	72.3	23.5	4.2
T3	63.6	24.8	11.6
Total (all models and templates)			
	66.6	25.2	8.2

Table 4: Percentage of errors per error type, calculated for each model, template and totals.

attribute of the word to meet the metaphor analogy restrictions and thus we propose the injection of such linguistic features as a possible research line for future work. Additionally, in further iterations of this research line, we would like to expand our dataset with more analogies and to other languages such as Spanish.

Acknowledgements

Supported by the Spanish project PID2020-113903RB-I00 (AEI/FEDER, UE), by DGA/FEDER, by the *Agencia Estatal de Investigación* of the Spanish Ministry of Economy and Competitiveness and the European Social Fund through the “Ramón y Cajal” program (RYC2019-028112-I), and by the EU research and innovation program HORIZON Europe through the “4D PICTURE” project under grant agreement 101057332.

References

- Ehsan Aghazadeh, Mohsen Fayyaz, and Yadollah Yaghoobzadeh. 2022. Metaphors in pre-trained language models: Probing and generalization across datasets and languages. In *Annual Meeting of the Association for Computational Linguistics*.
- Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

- Philip Boucher. 2021. [What if we chose new metaphors for artificial intelligence?](#)
- Weijie Chen, Yongzhu Chang, Rongsheng Zhang, Jia-shu Pu, Guandan Chen, Le Zhang, Yadong Xi, Yijiang Chen, and Chang Su. 2022. Probing simile knowledge from pre-trained language models. *arXiv preprint arXiv:2204.12807*.
- Timothy C. Clausner and W. Bruce Croft. 1999. [Domains and image schemas*](#). *Cognitive Linguistics*, 10(1):1–31.
- Tamara Czinczoll, Helen Yannakoudakis, Pushkar Mishra, and Ekaterina Shutova. 2022. Scientific and creative analogies in pretrained language models. *arXiv preprint arXiv:2211.15268*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Ellen K Dodge, Jisup Hong, and Elise Stickles. 2015. Metanet: Deep semantic automatic metaphor analysis. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 40–49.
- B. Efron. 1979. [Bootstrap Methods: Another Look at the Jackknife](#). *The Annals of Statistics*, 7(1):1 – 26.
- Mengshi Ge, Rui Mao, and Erik Cambria. 2022. [A survey on computational metaphor processing techniques: From identification, interpretation, generation to application](#). Preprint.
- Zoltán Kövecses. 2016. Conceptual metaphor theory. In *The Routledge handbook of metaphor and language*, pages 31–45. Routledge.
- George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. The University of Chicago Press.
- Emmy Liu, Chen Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the ability of language models to interpret figurative language. *arXiv preprint arXiv:2204.12632*.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. [Efficient estimation of word representations in vector space](#). In *1st International Conference on Learning Representations, ICLR*.
- John Naughton. 2023. [Chatgpt isn’t a great leap forward, it’s an expensive deal with the devil](#).
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. [GloVe: Global vectors for word representation](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.
- Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.
- Sunny Rai and Shampa Chakraverty. 2020. A survey on computational metaphor processing. *ACM Computing Surveys (CSUR)*, 53(2):1–37.
- Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).
- Timo Schick and Hinrich Schütze. 2022. [True few-shot learning with prompts—a real-world perspective](#). *Transactions of the Association for Computational Linguistics*, 10:716–731.
- Elena Semino, Zsófia Demjén, Andrew Hardie, Sheila Payne, and Paul Rayson. 2017. *Metaphor, cancer and the end of life: A corpus-based study*. Routledge.
- Ekaterina Shutova. 2015. Design and evaluation of metaphor processing systems. *Computational Linguistics*, 41:579–623.
- Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI’17*, page 4444–4451. AAAI Press.

**Corpora
and
Annotation**

An Empirical Analysis of Task Relations in the Multi-Task Annotation of an Arabizi Corpus

Elisa Gugliotta and Marco Dinarelli*

Univ. Grenoble Alpes, CNRS, Grenoble INP*, LIG, 38000 Grenoble, France;

* Institute of Engineering Univ. Grenoble Alpes; Groupe *getalp*

elisa.gugliotta@univ-grenoble-alpes.fr; marco.dinarelli@univ-grenoble-alpes.fr

Abstract

In this study, we deal with the design of computational-linguistic resources and strategies for the analysis of under-resourced languages. In particular, we present empirical analyses aiming at identifying the best path to semi-automatically annotate a dialectal Arabic corpus via a neural multi-task architecture. Such an architecture is used to automatically generate several levels of linguistic annotation which can be evaluated by comparison with the gold annotation. Changing the order in which annotations are produced can have an impact on the quantitative results. Through multiple sets of experiments we show how to get the best performances with this methodology.

1 Introduction

In this paper we present an empirical investigation of the relations between different levels of linguistic annotation of a dialectal Arabic corpus. In fact, linguistic annotations, such as Part-of-Speech (POS) tagging or lemmatisation, are an important prerequisite for many NLP applications and in particular, for those concerning under-resourced languages such as Arabic Dialects (ADs) (Elhadi and Alfared, 2022). The development of NLP resources and systems for under-resourced languages requires awareness of their functioning in order to study them from a computational perspective. This type of awareness derives from the analytical study of the language in question. However, while high-resourced languages present many detailed linguistic studies, often under-resourced languages usually lack comprehensive, in-depth and up-to-date descriptions of their morphological and syntactic systems. Moreover, they are often characterised by graphic variations and the lack of a standard orthography. In many cases, the spelling is not standardised and reflects geolinguistic

variations (Bernhard et al., 2021).¹ This is also the case of the ADs, for which building resources such as linguistic annotated corpora, is a necessary stage to study and process them automatically. This is the reason why in the last couple of years there have been many projects focused on the creation of resources for the ADs.² A popular methodology to avoid the creation of AD corpora from scratch is the adaptation of resources, for example built for Modern Standard Arabic (MSA), in order to process ADs Harrat et al. (2018); El Mekki et al. (2021); Qwaider et al. (2019). However, MSA is used to perform language tasks completely different from those performed by using ADs. With this regards, Hary (1996) defines *multiglossia* as the linguistic situation in which different varieties coexist side-by-side in a language community, and where each variety is employed in different circumstances and has different functions. Therefore, in order to process ADs, the ideal solution should be to build dialect-centered resources from scratch, instead of adapting MSA resources, even though it involves a considerable effort. However, considering the enormous amount of work required to build resources from scratch, a possible strategy is adapting other existing AD tools to the AD under investigation, especially if the dialects belong to the same geographical areas (e.g. Tunisian and Algerian belong to the same area, namely the *Maghreb*). This is because ADs share much more with each other than with MSA.³ In fact, a number of features and variations within ADs seem to transcend regional boundaries and effectively escape the most traditionally accredited typology, which classifies the ADs into six major dialectal areas, from East (*Mashreq*) to West (*Maghreb*). A possible explanation resides into the

* This article was prepared jointly by the two authors and is based on Gugliotta's post-doctoral research work supervised by Dinarelli. However, for the requirements of the Italian Academy, Gugliotta must be considered responsible for sections 1, 2.2, 3, 5.1 and 6, while sections 2.1, 4 and 5.2 must be attributed to Dinarelli.

¹ Common phenomena are variations in pronunciation, as well as morphological variations, where inflected or derived forms vary according to location, or lexical variations. Furthermore, the absence of standard spellings leads to interpersonal variation.

² See Ahmed et al. (2022) for a review on free Arabic corpora.

³ For a study of the degree of similarity and dissimilarity between MSA and ADs, and among ADs, see Kwaik et al. (2018).

huge amount of migration, inter-dialectal contacts and many waves of diffusion which have brought specific linguistic features across the Arabic-speaking world (Benkato, 2019; Magidow, 2021; Benkato, 2020).

The creation of annotated corpora from scratch can be speed up by semi-automatic annotation using machine learning tools (Gugliotta and Dinarelli, 2020). In the case of multiple levels of annotation like in this work, a further benefit in using machine learning techniques can be obtained by exploiting Multi-Task (MT) learning, and in particular with neural models. MT neural learning approaches factorize information among learned tasks, improving results on all of them compared to individual tasks taken separately. Whether MT is performed in a parallel or cascaded fashion, it allows for sharing the representation of information of different tasks at intermediate layers (Caruana, 1997). MT has been proven to be particularly beneficial for ambiguous data, considering its ability to reduce sparsity, and helping to process complex patterns which involve multiple features. This is the case, for example, of POS-tagging (Rush et al., 2012; Søggaard and Goldberg, 2016; Alonso and Plank, 2016; Bingel and Søggaard, 2017; Hashimoto et al., 2016), which is particularly relevant to the morphological richness of Arabic, (as addressed by Inoue et al. (2017)) or dialectal Arabic (Zalmout and Habash, 2019).

For all these reasons and with the goal of basing our work particularly on AD, we found useful to exploit two resources recently created for the processing of Tunisian Arabic (Gugliotta and Dinarelli, 2022). The first resource is a MT neural architecture (see Section 2.1), built to help in annotating on multiple levels a Tunisian Arabizi Corpus. The second resource is the corpus itself (see Section 2.2). Concerning Arabizi, we must emphasize the spontaneous nature of this Roman orthography, which originated in digital environments where informal exchanges take place. Spontaneity plays a main role in the degree of encoding freedom left to native users, and this has an impact on the performance of MT systems. Other elements that play an influential part in MT learning systems include the design of the architecture itself and the order in which tasks are addressed. Beyond few exceptions, much of the existing work on MT learning systems focuses on learning one target task and one, or more, accurately selected auxiliary tasks (Changpinyo et al., 2018). There are various studies on multi-task learning, but it is not clear when this may be beneficial for all the tasks planned for the sys-

tem, or when it may instead produce a phenomenon known as negative transfer, that also depends on the interrelations among the tasks (Ruder, 2017).⁴ One of the keys to investigate this issue concerns the degree to which tasks are interrelated. A logical hypothesis is that morphological tasks may help syntactic tasks. With regard to the mentioned previous work on multi-task annotation, summarized in Gugliotta and Dinarelli (2022), the goal was to produce accurate annotations while facilitating manual checking work. Therefore, five levels of annotation were produced in a cascaded chain, via a MT learning system without delving, from a computational-linguistic point of view, into the degree of task interrelation. In this work, through exploiting these tools, we aim at finding possible task relations, and possibly improve previous results on each task by investigating such issue.

In order to explore this topic comprehensively, first of all, in Section 2, we will describe the architecture and the data on which we are relying for our study. Secondly, in Section 3, we will present the main related works. In Section 4, we will present the adopted methodology to address this issue. In Section 5, we will outline the experiments performed, drawing attention to some emerging trends. In the same section, we will discuss our results from a global point of view. Finally in Section 6 we will conclude the article.

2 MT Architecture and Data Structure

Like deep learning in general, multi-task learning is inspired by human learning. To learn new tasks, humans often transfer knowledge gained from prior related tasks. The possibility that certain cognitive structures may be prerequisites or have a positive or negative influence on the acquisition of new knowledge has been discussed by many researchers in the fields of didactics, pedagogy, cognitive linguistics, and psycholinguistics (Piaget, 2003; Vygotsky and Cole, 1978; Bransford and Johnson, 1972; Kole and Healy, 2007; Gick and Holyoak, 1980). However, the views of scholars are still too heterogeneous to explain the mechanisms and processes operating during human acts of comprehension and acquisition. Still it is well established that appropriate prior knowledge must be activated in order to be used effectively in the acquisition process. In a similar manner, Ruder (2017) motivates MT learning from the perspective of machine learning, viewing it as a form of inductive transfer. Indeed, the author explains that inductive

⁴See Section 3 for an outline of the existing work on MT learning systems and tasks interrelations.

transfer can help to improve a model by introducing an inductive bias, leading the model to prefer some assumptions over others. The inductive bias can be introduced by auxiliary tasks. Auxiliary tasks in MT learning can serve as conditions or suggestions for the main task. At the same time, related tasks can reinforce each other to form coherent predictions through shared representations. This strategy often leads to solutions that generalize better. However, according to Ruder (2017), our understanding of the degree of relationship or similarity between tasks is still limited, and we need to study them more in depth to better understand the generalization capabilities of MT learning by better fruiting their potential. Thus, one of the prerequisites of MT learning is the correlation between different tasks and data (Zhang et al., 2022).

2.1 The MT Architecture

The MT neural architecture employed in this work is an *encoder-decoder* system designed originally for the Tunisian Arabish Corpus (TArC) annotation. The MT system is able to instantiate as many decoders as the number of levels of linguistic annotations employed in the data, the different decoders operate in a cascade fashion, and it has been recently released (Gugliotta and Dinarelli, 2022). The MT system is designed to train LSTM or Transformer models. For our experiments we employed the LSTM model. As pointed out in (Gugliotta and Dinarelli, 2022), Transformers are in general preferred and very accurate for several NLP problems, especially when dealing with very large amount of data. However, they present limitations when modelling tasks with structured outputs (Weiss et al., 2018; Hahn, 2020). Since in our experiments outputs are always, at least partially structured, we employed mainly LSTM models. (Gugliotta and Dinarelli, 2022) shows indeed a significant performance gap between LSTM and Transformer models in experiments involving the TArC corpus, the same data we use in this work (please see the next section, for data description). Whatever the used model, the linguistic information that can be output by the MT system are: Code-Switching classification, normalization into CODA* (Habash et al., 2018), tokenization, POS-tagging and lemmatisation.

Concerning the classification of code-switching, it is provided at word level, in order to filter the Arabizi text from the foreign words, which are indeed classified as *foreign*. Table 1 presents the classification (**Class.** in the table header), the CODA* transliteration (**CODA***), the tokenization (**Token.**),

the POS-tagging (**POS**) and the lemmatisation (**Lemma**) of the following Arabizi sentence of TArC.

- (1) *Inchalah cycle ejjay wala eli ba3dou,*
/nšālla cycle əž-zāy walla əlli baʔd-u/,
 ‘God willing next time, or the time after that’.

Arabizi	Class.	CODA*	Token.	POS	Lemma
Inchalah	Az.	ان شاء الله	ان شاء الله	INTERJ	ان شاء الله
cycle	Fr.	Fr.	Fr.	Fr.	Fr.
ejjay	Az.	الجاي	الهابي	DET+ADJ	جاي
wala	Az.	ولا	ولا	CONJ	ولا
eli	Az.	اللي	اللي	REL_PRON	اللي
ba3dou	Az.	بعده	بعده	ADV+	بعد
				PRON_3MS	

Table 1: Example of the annotation levels. "Az." means "Arabizi", "Fr." means "foreign".

Each of these annotation level is processed by a dedicated decoder. As for the Arabizi input, it is converted into context-aware hidden representations by the MT system’s encoder. Each decoder is equipped with a number of attention mechanisms corresponding to the number of preceding modules (including the encoder). Hence, each decoder receives as input the hidden state of the encoder together with the hidden state of each previous decoder. Each decoder generates also its predicted output, which is used to learn the corresponding task by computing a loss function comparing the predicted output to the expected output. The entire architecture is learned end-to-end by calculating a global loss through the sum of each individual loss (Gugliotta and Dinarelli, 2022).

2.2 The Data

The data we used for the study presented in this paper are the TArC corpus (Gugliotta and Dinarelli, 2022) and the MADAR corpus (Bouamor et al., 2018). The first one contains 4,797 sentences produced by Tunisian users in digital contexts such as blogs, forums and social networks. These sentences are encoded in Arabizi, the Latin encoding employed for online written conversations. The MADAR corpus, on the other hand, is a parallel corpus of several Arabic dialects, including Tunisian (both from Tunis and Sfax cities). In our previous work, we exploited 2,000 sentences of the MADAR corpus, by proving their usefulness for the MT system learning (Gugliotta and Dinarelli, 2022). Also for experiments in this work we decided to use both corpora. In particular the MADAR data are concatenated to the TArC training data to create a single, bigger training set.

3 Related Work

Intuitively determining the degree of similarity between tasks is still a common practice especially in the design stages of MT architectures, when one does not yet have data on which to rely otherwise (Worsham and Kalita, 2020). In general, until a few years ago, methods for identifying task relationships focused on expert intuition. However, recent research increasingly takes into account the fact that neural networks do not need to operate on the same principles as human learning. More and more scholars, such as Alonso and Plank (2016), are arguing that the selection of MT learning tasks should be guided by the properties of the data, not by the intuition of what a human performer might consider easy. In fact, they conduct a number of studies showing that the best auxiliary tasks are neither too easy to predict nor too difficult to learn. In particular, for the mentioned study, they use a state-of-the-art architecture based on biLSTM models and evaluate its behavior on a motivated set of main and auxiliary tasks. The performance of the MT system is evaluated both by experimenting with different combinations of main and auxiliary tasks and by applying a frequency-based auxiliary task to a set of languages, processing tasks and evaluating its contribution. LSTM networks were also analyzed by Reimers and Gurevych (2017) for a wide variety of sequence tagging tasks, in order to find LSTM network architectures that can perform robustly on different tasks. Five classical NLP tasks were chosen as benchmark tasks: POS tagging, Chunking, Named Entity Recognition (NER), Entity Recognition and Event Detection. Guo et al. (2018) addressed multitask and curriculum learning to improve training of subsets of multiple tasks, starting with smaller and simpler tasks first. Zamir et al. (2018) computed an affinity matrix between tasks based on whether the solution of one task can be read easily enough by the representation trained for another task. Their approach, being fully computational and representation-based, avoids imposing prior (possibly incorrect) assumptions about the relationships between tasks. In addition, Standley et al. (2019), using the Taskonomy dataset (Zamir et al., 2018), found that, unlike affinities between transfer tasks, affinities between multiple tasks depend strongly on a number of factors such as dataset size and network capacity. A similar work to ours was presented by Bingel and Søgaard (2017), who conducted a study on ten traditional NLP tasks (including POS tagging), comparing the performance of MT and Single-Task (ST) learning, where hyperparameters of

ST architectures are reused in the MT configuration. Changpinyo et al. (2018) conducted extensive empirical studies on eleven sequence labeling tasks. They obtained interesting pairwise relationships that reveal which tasks are beneficial or detrimental to each other. Such information correlated with MT learning outcomes using more than two tasks. They also studied the selection of only advantageous tasks for joint training, showing that this approach, in general, improves MT learning performance, and highlighting thus the need to identify tasks to be learned jointly. Similar experiments, but specific to the domain of question answering, were performed by Vu et al. (2020) who conducted an in-depth study of the relationships between various tasks (question answering and sequence tagging) and proposed a task-embedding framework to predict these relationships. Sun et al. (2020) sought to enable adaptive sharing by learning which levels are used by each task through model training. More recently, Aribandi et al. (2021) proposed a massive collection of various supervised NLP tasks in different domains and task families in order to study the effect of multi-task pre-training on the largest scale to date and analyze the transfer of co-training between common task families. The researchers addressed the issue of inter-language transfer from high-resource languages to low-resource languages. They presented a model capable of automatically selecting the language from which to transfer a given task, based on inter-lingual criteria. Fifty et al. (2021) proposed a procedure for selecting subtasks based on task gradients.

4 The Adopted Methodology

During the annotation process of the TArC (Gugliotta and Dinarelli, 2022), a specific order of linguistic annotation production has been set out. Starting from the Arabizi as input, this specific order was: classification (to filter the code-switching elements), transliteration into CODA*, tokenization, POS-tagging and lemmatisation. This order of annotation was chosen based on principles of both linguistic reasoning and empirical observation of MT system performances. Starting from the premise that providing too much information to an algorithm can slow it down and lead to inaccurate results, it is important to think carefully about what information is most relevant to a specific goal. The ultimate goals of Gugliotta and Dinarelli (2022) were **1.** to produce precise annotation levels and at the same time **2.** to ease the work of manually checking and correcting the annotations predicted by the architecture. Therefore, in Gugliotta and Dinarelli (2022),

the chosen order of tasks was oriented toward simplifying both the tasks involved in the semi-automatic annotation (the automatic classification and the manual correction). In fact, it was considered useful to find a good compromise between proceeding in hierarchical order, from the simplest to the most complex annotation (observing the performance of the MT system in annotation), while respecting the relationships between the various levels of annotation based on linguistic reasoning. Concerning the choice of processing easy tasks first, it is possible to define what is the easiest task among others for a model by observing its learning progress or the result precision in case of classification tasks (Guo et al., 2018). For example, as we noticed by observing experimental results in Gugliotta and Dinarelli (2022), the task of transliteration from Arabizi into CODA*, resulted to be the most difficult task for the architecture. In our opinion, this difficulty comes from the ambiguity of Arabizi, being a spontaneous orthographic system. On the other hand, it results more complicated to establish what task can be the most difficult for a human annotator, because this depends on his specific previous experiences, which are hard to evaluate and are in any case unlikely to match exactly the goal of the annotation at hand. For manual checking of data, for example, annotators will make use of their prior skills and the annotation guidelines, and they will apply this knowledge to the new task, gradually becoming faster and more effective. In fact, we can consider them as learners. As a result, if we apply the same logic as the one used in language acquisition theories, the ease of a task is closely related to the concept of support, in terms of knowledge, that is made available to perform the task.⁵ This is to say that, for example during a manual correction phase, an annotator may find easier to correct various levels simultaneously, instead of correcting them one-by-one. Two possible reasons are (A.) the same error may have been transferred between different annotation levels, so it is easier to correct the various levels together. (B.) The presence of the other levels can help the annotator to better understand the error. The annotator will not only dispose of the text semantics, but also of the other levels of annotation (morpho-syntactic in the case of Gugliotta and Dinarelli (2022)). Therefore, generalising the prob-

⁵Concerning human language acquisition knowledge there are several theories, like for example the one called Zone of Proximal Development (ZPD) (Vygotsky and Cole, 1978). The ZPD represents the interval between what a learner is able to do unsupported and what he can achieve with support. Support may come from someone else with wider knowledge or skills (namely the teacher).

lem, we might conclude that the "simple-to-complex" order can work as well for deep learning systems as for human learners (including annotators). However, as already mentioned in Section 3, we must consider that what is possibly an auxiliary task for an annotator does not help a MT learning system in the same way. The experiments in the following section are aimed at investigating this concern.

4.1 Experimental Procedure

We organized different groups of experiments with the aim of identifying the best order of tasks to be performed by the architecture, and this in order to maximize the results on each of them. The first two groups of experiments are a mixture of ST (Single-Task) and MT (Multi-Task) strategies, organized into an iterative procedure. The procedure starts with using two annotation levels, one as input and the other as output task, where all possible combinations of two levels are tested to find the best order, results are shown in the tables 2 and 3. The order is thus chosen based on the best performing one. Performances on all tasks are measured with Accuracy (see Gugliotta and Dinarelli (2022)). Table 4 instead presents the grouping of particular intermediate experiments, in order to answer specific task relation questions. The iterative procedure continues using the annotation level detected as the easiest to predict, measured with empirical results, as the input to the system, and all the remaining annotation levels as output, both one at a time with ST experiments and with specific combinations of two or more annotation levels in MT experiments. This allows to select again the easiest task based on the empirical results. Results are given in the tables 5 and 6. We take care of using as much as possible Arabizi or CODA* as input to the system since these are the formats in which data may be naturally found, and needing to be transliterated, into CODA* for Arabizi and into Arabizi for CODA* (Gugliotta and Dinarelli, 2022), in addition to being annotated with the other levels of the TArC corpus (Gugliotta and Dinarelli, 2020) used also in this work for our analyses. Considering the spontaneous nature of Arabizi and the small amount of our data, having Arabizi text as input exposes to the risk of transferring errors obtained on the first task to the rest of the MT chain, hiding possibly the task-relation potential. For this reason, we performed two sets of experiments, one with Arabizi as input and one with Arabic script as input, the latter follows a conventional orthography (CODA*) and thus allows possibly to overcome the error transfer problem

implied by the use of Arabizi as input. The other experiments are based on MT learning. In fact, we want to compare the results obtained with the ST strategy with the same experiments performed in a MT setting. For these sets of experiments, we test different MT chains, that present different task orders, to observe which one is giving the best results, again testing both Arabizi (tables 7 and 10) and CODA* (tables 8 and 9) as input.

5 Results and Discussion

In this section we present the results of all our experiments. In Section 5.1 we present the preliminary experiments (mix of ST and MT strategies), while in the section 5.2 we present the results of the MT experiments. The experiments described in the Section 5.1, refer to a procedure centred on the observation of the best results of ST experiments, which then contribute to the definition of a precise task order in MT experiments. Therefore within this section, these MT experiments, which respect the order deduced from the ST results, will also be described. In order to provide a comprehensive description of the results and highlight the correlation between them, we will also globally discuss the results at the end of the paper (Section 6).

5.1 Preliminary Experiments

Table 2 shows our results on the test sets of TARc in the ST (Single-Task) experiment setting, using Arabizi and CODA* as input to the model.⁶ We defined these experiments as the *Starting ST experiments*, considering them as the first stage to define a task order for the MT architecture. When the input was the Arabizi text we also performed the classification task (*class.* in the table header), in order to filter the code-switched tokens not to process. In the column *Arabizi input* we thus report also the classification accuracy for each experiment, in brackets. Experiments are performed using both Arabizi and CODA* as input since the system can be used in some cases to transliterate Arabizi data into CODA* encoding, like for the TARc corpus, in some cases for transliterating CODA* encoded data into Arabizi, like for the MADAR corpus (Gugliotta and Dinarelli, 2022), in addition to the other annotation levels when these are available to train the model for doing so.

The ST tasks performed for these experiments are the tokenization of the input, the Part-of-Speech tagging, the lemmatisation and the transliteration of Arabizi into CODA* (for the experiments having Arabizi

Tasks	Arabizi input (class.)	CODA* input
Token.	80.0% (93.0%)	95.4%
POS	73.8% (92.5%)	54.5%
Lemma	75.5% (92.8%)	89.5%
Translit.	79.0% (92.8%)	67.2%

Table 2: Starting ST Experiments

as input), or of CODA* into Arabizi (in case of the experiments having CODA* as input). These tasks are reported in the table, in the column **Tasks**, with the respective entries: **Token.**, **POS**, **Lemma** and **Translit.**. Some results are in bold because they represent the best among the experiments reported within the table. As we can observe, both in the case of Arabizi and CODA* as input, the *easiest* task seems to be the tokenization, on which the system respectively achieved the accuracy of 80% and 95.4%. The former result is not surprising observing that, when using Arabizi as input, the transliteration task obtains one point less (79%) than the tokenization task (80%), these seem two very correlated annotation levels given the result on the tokenization task when using CODA* as input (95.4%). In fact, the tokenization implies the transliteration of the token, being both encoded in CODA* (as shown in Table 1). It is also interesting to observe the result on the classification task (93%) performed together with the tokenization, using Arabizi as input. Even if the difference is small, this is the best classification result. Thus, it seems that the classification benefits from the information of the tokenization task. It is also worth to highlight that both the tokenization and the lemmatisation performed from a CODA* input, obtain relatively high results, respectively 95.4% and 89.5%. While results on the POS (54.5%) and the transliteration into Arabizi (67.2%), using CODA* as input, are the lowest results, also compared to results obtained using Arabizi as input. Tokenisation and lemmatisation involve simpler processes than POS-tagging (identification of both the morphological class and the features of the token). In addition, we should consider that the CODA* conventional orthography is also employed to encode the tokenization and the lemmatisation levels. Indeed, these tasks result in *easy* operations for the model having as input the text in CODA*. This is not the case of the transliteration, where the system must convert the Arabic-encoded input into Latin-encoded information. In fact, it is surprising that the transliteration into CODA* is still obtaining a good result (79%) starting from an Arabizi input. This can be due to the fact that, as previously

⁶Please see Gugliotta et al. (2020); Gugliotta and Dinarelli (2022) for further details on the data and the architecture.

mentioned, the Arabizi encoding is a spontaneous, ambiguous script, while CODA* is a normalized encoding. Consequently, transliterating an ambiguous script into its normalization (i.e. many variations into one encoding) results to be an *easier* task in comparison to the opposite operation (CODA* into Arabizi, i.e., one encoding into one of the many encoding possibilities).

Once assessed that the tokenization task is the easiest using both Arabizi and CODA* as input, we continued the iterative procedure by using the detected easiest annotation level as input, and the other remaining annotation levels as output, both one at a time and all together in a MT learning setting. More precisely, we first performed ST experiments using the tokenization as input to the model, and alternatively POS and lemmas as output. These results are shown in the first two lines of Table 3, and they show that the easiest task between POS tagging and lemmatization, when using tokenization as input, is the lemmatization.

Input	Tasks	Accuracy
Token.	POS	86.2%
Token.	Lemma	92.4%
Token.	Lemma - POS	92.8% - 87.6%
Token.	POS - Lemma	87.3% - 92.6%

Table 3: Intermediate Experiments

By comparing the results of these two experiments, we can confirm our previous consideration about the fact that lemmatisation, in comparison to POS-tagging, is in general a simpler process to be performed starting from the token. The information that most helps the lemmatisation of a token is its morphological class. This information is contained in the POS, and more precisely in what we can define as the *main* part of the POS (namely only the morphological class, such as "verb", "noun", "adjective" etc., without its features, such as gender and number). The prediction of the *main* POS is a much easier task than the prediction of a POS with all the morphological features. In fact, the lemmatisation task obtains 92.4% of accuracy, 6.2 points more than the results on the POS tagging (86.2%). In the same table we also report two additional experiments that compare the combination of the two tasks (POS and lemmatisation) in the two possible orders, thus in a MT (Multi-Task) setting. We can observe that the combination achieving the best results is the first one (namely **Lemma - POS**), where the model obtained 92.8% and 87.6% of accuracy on the two tasks, respectively. While the margin of improvement is small with respect to the other possible order (POS - Lemma), this confirms that the

lemmatization is the easiest task using tokenization as input. Moreover it is interesting to see that in the two MT experiments results are always better than those obtained with ST experiments. This means that the two tasks help each other, which is what we expect in a MT learning setting. Given these results, we considered useful to explore the question further by means of additional experiments, shown in Table 4.

Input	Tasks	Accuracy
CODA*	Lemma - POS	89.2% - 84.2%
CODA*	POS - Lemma	85.9% - 90.5%
CODA*	Token. - POS	95.3% - 85.2%
CODA*	POS - Token.	85.6% - 95.2%

Table 4: Additional Experiments for Tasks Relations

These experiments present the grouping of particular intermediate annotation levels, using CODA* as input to the system. The aim of the experiments was to discover what task, between lemmatisation and tokenization, helped more the POS task, and in which order. For this reason we needed the tokenization not to be the input for the model, and among Arabizi and CODA* we preferred to have the input in CODA* in order to avoid introducing a bias in these experiments due to the errors depending on the Arabizi ambiguity. If we had to guess which task helps POS prediction more, we would have chosen tokenization rather than lemmatisation. The former is in fact a morphological task, as much as POS, while the latter is primarily a lexical (but also morphological) task. However, by observing the results, we can confirm what already observed in Table 3, namely that it is the lemmatisation the task helping more the POS tagging. In fact, the experiment showing the best results on POS is the second one, where the POS is followed by the lemmatisation. This result on the specific order (POS-Lemma) seems to be inconsistent with what has just been stated by commenting on Table 3, where slightly better results were obtained by keeping the Lemma-POS order. However, what makes the difference between the experiments in the tables 3 and 4 is the input. That is, when the input is the tokenized text, the Lemma-POS and POS-Lemma order obtain similar results (Table 3), whereas when the input is in CODA* (Table 4) there is a considerable difference in the two possible orders between POS and Lemma (POS improves of 1.7 accuracy points, Lemma improves of 1.3 points with the POS-Lemma order). Instead, we have non-significant differences by inverting the order between *Token.* and POS. Thus, it seems that the system has more difficulties in extracting the

Exp. ID	Accuracies on tasks			
	Token.	Lemma	POS	Arabizi
I	95.4%	-	-	-
II	95.3%	89.8%	-	-
III	96%	90.7%	86.2%	-
IV	94.4%	88.9%	84.5%	67.8%

Table 5: Chain based on ST experiments - CODA* input

Exp. ID	Accuracies on tasks				
	Class.	Token.	Lemma	POS	CODA*
I	86.2%	-	-	-	-
II	93%	80%	-	-	-
III	95%	80%	78.2%	-	-
IV	94.1%	78.9%	77.5%	77.8%	-
V	94.2%	78.9%	77.3%	78.6%	79.5%

Table 6: Chain based on ST experiments - Arabizi input

lemma from the CODA*, without the intermediate step of POS tagging, which instead obtains better results (85.9%) directly on the CODA*, than on the lemma (84.2%), also helping to improve the results on the lemmatisation, which rises by 1.3 points (90.5% vs 89.2%), if placed after the POS level. This is also evident if we compare these results with those obtained in the ST experiment (CODA* - Lemmatization: 89.5%) in Table 2. The results on lemmatisation improve (by 1 point) when it follows the POS task (90.5%), thus, the two tasks (POS and lemmatisation) help each other. Once these considerations have been made, we can present the results in Table 5 and Table 6, that present the experiments aiming at identifying the final MT learning chain based on ST experiments, having CODA* (Table 5), or Arabizi (Table 6) as input. The progressive Roman numerals in the ‘**Exp. ID**’ columns of these tables indicate the sequential order in which the experiments were performed. These numerals will also be used to refer to the experiments while discussing the results. Concerning these experiments, both in the case of an input in Arabic characters (CODA*) and in the case of an input in spontaneous Latin orthography (Arabizi), it emerges a tendency for improved results due to the presence of auxiliary tasks. With regards to Table 5, we can observe that thanks to the presence of the tokenization task, the lemmatisation improves of 0.3 points at the experiment II (second line of Table 5), in comparison with the lemmatisation experiment as a ST in Table 2. Observing the experiment III (*Exp.* from now on for short) reported in Table 5, we can notice that thanks to the presence of the POS task, the tokenization task improves of 0.7 points with respect to the ST experiment on tokenization, reported in Table 2.

Also the lemmatisation task obtains better results, improving by 0.9 points, thanks to the presence of the POS task, at the Exp. III, in comparison with the Exp. II in Table 5. Finally the transliteration task from CODA* into Arabizi improves by 0.6 points, thanks to the previous tasks (at the Exp. IV in Table 5, in comparison to the transliteration as an ST experiment in Table 2). However, by adding the transliteration to the chain of tasks, the model is subject to much more difficulty, as can be noticed at the Exp. IV of Table 5, where all the previous tasks undergo the negative transfer effect, due to the presence of the transliteration into Arabizi.⁷ From Table 6 we can draw very similar observations. From the Exp. II, we can observe an improvement of 6.8 points of the classification task, in comparison with the Exp. I, thanks to the tokenization task. On the next step (Exp. III), classification continues to improve (by 2 points) thanks to the lemmatisation task, which also improves by 2.7 points (thanks to the tokenization) in comparison with the ST experiment on lemmatisation in Table 2. Finally, at the Exp. V, we can observe how, thanks to the normalization of Arabizi into CODA*, POS-tagging improves of almost one point (0.8), in comparison with the previous step (Exp. IV) in Table 6. Also the transliteration task obtains better results, 0.5 points in comparison with the ST transliteration reported in Table 2, thanks to the previous tasks. By observing Table 6, *the most difficult task* for the model seems to be the POS tagging. In fact, at the Exp. IV, while the POS task improves by an impressive 4.8 points (in comparison with the ST experiment in Table 2), all the previous tasks lose about one point, compared with the results of the previous step (Exp. III).

5.2 Multi-Task Experiments

In our multi-task system, as previously stated, variables come into play, such as the factorization of the information shared among the decoders, the presence of attention mechanisms, etc. For this reason, we decided to compare the results obtained from ST experiments with those of the MT experiments. Therefore, in the following tables we can observe different combinations of tasks performed sequentially by the MT architecture. The goal is to check whether or not the ST task-chain matches with the MT task-chain that gives better results than other combinations or than the combinations that would seem logical from a linguistic point of view (e.g.: Arabizi - Classification - CODA* - Lemmatization - Tokenization - POS). Each

⁷This phenomenon has been mentioned in the section 1.

line of the following tables represents an experiment with all the tasks in a specific order. The order of a task is specified in brackets as a footnote of the corresponding accuracy result. When such note is not present, the order of the task is the one corresponding to the column of the table. For instance in the Exp. I in Table 7, the task order is *Class. - CODA* - Token. - POS - Lemma*, where the order of *Class.* and *CODA** is the one given by the corresponding column in the table since their accuracy has no footnote; while for *Token.*, *POS* and *Lemma* the order is given by the index in footnote to their accuracy. This notation allows to give several task orders in the same table keeping the same table headers. We also keep the same experiment identifier naming with roman cardinals as in the previous tables, e.g. Exp. I mentioned above.

Table 7 presents the MT experiments with the Arabizi text as input. For the experiments reported in this table, the first tasks are always the classification and the transliteration into CODA*. Concerning the last two line of the table (lines VII and VIII), they summarize the results of two experiments, where the model receives the Arabizi input and processes the tasks of lemmatisation and transliteration into CODA* as a second and third task, respectively.

Exp. ID	Accuracies on tasks				
	Class.	CODA*	Lemma	Token.	POS
I	97.3	82.6	82.3(5)	82.3(3)	71.4(4)
II	99	84.2	82.8(4)	83.5 (3)	83.1 (5)
III	92.9	78.5	54.2(4)	75.9(5)	78(3)
IV	94.3	78.3	76.4(5)	77.9(4)	78.1(3)
V	97.9	84.3	83.6 (3)	82.3(4)	82.3(5)
VI	98.8	83.5	82.4(3)	82.3(5)	82.3(4)
VII	98.5	83.7(3)	83(2)	83(4)	82.6(5)
VIII	93.2	77.8(3)	78.6(2)	76(5)	78.2(4)

Table 7: Chain based on MT experiments - Arabizi input

At the end of the section 5.1, by discussing the preliminary experiments, we stated that POS-tagging is the most difficult task, together with the transliteration into Arabizi. In particular, we have deduced this by looking at Table 6. In fact, we remind that for these experiments we imposed a task order based on ST experiments described in the section 4.1. We also recall that, in Table 6 (experiments concerning the MT-chain based on ST experiments) the highest result obtained on POS tagging was 78.6%.

Concerning the Multi-Task (MT) experiments and looking at Table 7, we can see that the highest result on POS is 83.1%. We can also note that on all tasks, except for lemmatisation, better results are achieved with the Exp. II, where POS is the last task processed by the

MT architecture. Thus, it seems that POS prediction is benefiting of all the previous task information. The POS results in the Exp. II (83.1%) are improved of 4.5 points in comparison with the best result of Table 6 (78.6%). At the Exp. II, it is also interesting to observe how the lemmatisation task, processed between tokenization and POS, contributes to the improvement of both tokenization and POS, though it loses almost one point (0.8) compared to its highest result, obtained when lemmatisation is in the third position (see the Exp. V). In fact, at the Exp. V in Table 7, we can see that lemmatisation improves by 0.8 points if it follows the transliteration task and if it is followed by the tokenization task. The difficulty introduced by the POS task is evident from the tables 6, 7 and 3. In the latter one we also observed the encouraging results obtained on the lemmatisation task, using tokenization as input.

We also performed the experiments reported in Table 8, in order to identify the best task sequence for predicting Arabizi strings from CODA* strings. Considering that the input for these experiments is already filtered by the *foreign* tokens, we did not perform the classification task. Except for the transliteration into Arabizi, which is always the last task, the order of the tasks for each experiment are shown again through footnotes with a number in brackets.

Exp. ID	Accuracies on tasks			
	Lemma	Token.	POS	Arabizi
I	88.9(3)	94.8(1)	84.1(2)	68
II	88.9(2)	94.4(1)	84.5(3)	67.8
III	89.7 (2)	95.1 (3)	85.1 (1)	68.5
IV	89.4(3)	94.7(2)	84.6(1)	68.4
V	89.7 (1)	95 (2)	84.7(3)	68.2
VI	89.1(1)	95.2 (3)	85 (2)	68.4

Table 8: Chain based on MT experiments - CODA* input

Even in Table 8 we can observe that MT experiments produced better results if compared to those of the task sequence established with the ST logic in Table 5. In fact, we defined the transliteration into Arabizi as the most complex task starting from an input in CODA*. In Table 5 the result obtained on transliteration was 67.8%, while in Table 8 we can see how in several experiments we obtained better results, and in general on all tasks. The chain established through the sequential logic of ST experiments, shown again in Table 8 as Exp. II, actually appears to be the worst combination for both tokenization and transliteration. We note, on the other hand, that the best over all tasks is the one that, in the Exp. III, sees POS in the first position of the task chain. Again, like

in Table 7, POS is separated from the rest of the tasks by the intermediate presence of the lemmatisation task, and followed by tokenization. It is very interesting to observe that in Exp. III POS gets as much as one point more than in the Exp. I of the same table, where it was the second task, after the tokenization task. We remind that according to the linguistic logic, the tokenization being a morphological task, it should support the morpho-syntactic tasks.

Finally, we performed experiments with different task combinations, considering the possibility that annotations, such as lemmas or POS-tags, are introducing negative a bias for the task of CODA* transliteration into Arabizi encoding, and that the classification can instead help in it. These are reported in Table 9. Concerning the experiments reported in the last two lines of the table (lines VII and VIII), these treated the lemmatisation as a second task, after the classification (which is always the first task) and before the task of transliteration into Arabizi. In fact, the latter is always the second task performed during the previous experiments reported in the same table (experiments 1-6).

Exp. ID	Accuracies on tasks				
	Class.	Lemma	Token.	POS	Arabizi
I	97.2	88.8(5)	94.5(3)	83.6(4)	68.8 (2)
II	98.1	89.3(4)	95.3(3)	83.4(5)	68.3(2)
III	98.1	89.1(4)	95.2(5)	83.4(3)	68.5(2)
IV	97.4	88.6(5)	94.7(4)	83.3(3)	68.4(2)
V	97.8	88.9(3)	95.2(4)	84.3(5)	68.7(2)
VI	97.5	89.2(3)	94.4(5)	83.4(4)	68.3(2)
VII	97.5	89.3(2)	95(4)	83.6(5)	68.7(3)
VIII	98.3	89.2(2)	95.4 (5)	84.8 (4)	68.6(3)

Table 9: Other MT experiments to predict Arabizi

The goal of experiments reported in Table 10, instead, is to predict the CODA* transliteration from the Arabizi input. Thus, the transliteration into CODA* is always the last task, while the classification is always the first task.

Exp. ID	Accuracies on tasks				
	Class.	Lemma	Token.	POS	CODA*
I	94.1	76.3(4)	77.9(2)	77.9(3)	78.1
II	94.2	77.3(3)	78.9 (2)	78.6(4)	79.5
III	94	77.2(3)	78.2(4)	78.5(2)	78.2
IV	93.8	76.3(4)	78.1(3)	78.1(2)	78
V	94	77.2(2)	78.4(3)	78.5(4)	78.5
VI	94.2	77.3(2)	78.7(4)	78.8(3)	78.7

Table 10: Other MT experiments to predict CODA*

In these last two tables, 9 and 10, we have reported, for the sake of completeness, experiments with

additional combinations of tasks. Both seem to confirm the concept with which we would like to conclude our analysis. Namely, specific task ordering in a MT learning setting, in the case of a robust model provided with attention mechanisms, matters up to a certain point. In fact, looking at the last two tables, where we aimed at improving transliteration into Arabizi (Table 9) and CODA* (Table 10), we can notice first that the tasks exhibit roughly always the same accuracy values in all experiments. As a second observation, two different strategies are adopted. In Table 9 the transliteration task in Arabizi is always in the second position (except for experiments VII and VIII), while in Table 10 transliteration in CODA* is always the last task. By comparing the results of the strategy in Table 9 with those obtained on the Arabizi transliteration task in Table 8 (where Arabizi is always the last task), we can say that the strategy of tackling Arabizi as the second task yields better results, although the difference is small. We can draw the same conclusion by looking at the results on the transliteration task into CODA*, comparing the results in Table 7 to those in Table 10. In the former, transliteration is always addressed as the second task (except in the experiments VII and VIII), and doing so yields better results than those reported in Table 10, where the transliteration task is always the last one.

6 Conclusions

In this work, we presented empirical analyses in order to pinpoint the best approach for semi-automatic annotation of a dialectal Arabic corpus through a multi-task neural architecture. The experiments performed highlight a number of factors that may play a role in the outcome of good data annotation. Among the ones discussed are the interrelations between the tasks processed by the architecture, the difficulty the architecture faces in performing the tasks and the impact that determining specific orders of data annotation may have on the results, especially if to infer the relationship between tasks, we rely *only* on linguistic intuitions. By observing the experiments performed by this study, it clearly emerges the existence of relations between tasks, and these are especially evident when observing ST experiments. In fact, it turned out that morphological information does not necessarily support morphological tasks (Table 4), whereas it supports, for example, lemmatisation. At the same time, lemmatisation appears to play a key role in supporting the POS task, which difficulty is evident from the tables 6, 7 and 3. In the latter one we also observed

the encouraging results obtained on the lemmatisation task, using tokenization as input. The optimal choice therefore is to isolate the POS task, leaving it as the last task to be processed and preceding it by all simple tasks such as tokenization or lemmatisation. The latter is probably more effective, as intermediate task between tokenization and POS, in that it consists in fewer operations to be performed by the model, which is then able to generalize better on lemmatisation, especially once the tokenization is performed as a previous task (see Table 3). In other words, the lemmatisation task, positioned between tokenization and POS, can provide a cushioning effect to the negative transfer introduced by the POS task (see for example the POS negative transfer effects on the tokenization at the Ex. V in Table 7). We also remind that, in section 5.1, by observing Table 4, we noted that: (1.) The best results on the POS, having the input in CODA, are obtained at the experiment where the POS is side-by-side with the lemmatisation instead of the tokenization. (2.) The accuracy on lemmatisation improves (by 1 point) in comparison with the ST accuracy (Table 2). This seems to mean that the reason why the lemmatisation level succeeds in "absorbing" the negative transfer of POS-tagging on the rest of the MT system, lies in two reasons. The **first** is that lemmatisation, basically, is an easy task (especially if based on CODA* transliteration, as shown in the tables 2 and 5), and the **second** is that the operations to perform POS-tagging are essentially a prerequisite to those implemented to solve the lemmatisation task. In fact, although POS-tagging is a complex task, it does not affect the lemmatisation results (as it does instead with the other tasks), actually POS improves the lemmatisation by disambiguating the string. In short, the two tasks are strongly related. However, imposing specific orders on tasks, according to such relations in ST learning logic has been shown to be an uncertain strategy in comparison to the MT strategy. Regarding the latter, we believe that what really has an influence on the results in terms of improvement of individual tasks is not so much the relation between tasks, but the inherent difficulty of tasks. In fact, there seems to be a tendency for general improvement in results on the various tasks if the tasks that require greater architectural capacity are tackled at the initial positions in the chain of tasks.

References

Arfan Ahmed et al. 2022. Free and accessible Arabic corpora: A scoping review. *Computer Methods and Programs in Biomedicine Update*, page 100049.

- Héctor Martínez Alonso and Barbara Plank. 2016. When is multitask learning effective? Semantic sequence prediction under varying data conditions. *arXiv preprint arXiv:1612.02251*.
- Vamsi Aribandi, Yi Tay, Tal Schuster, Jinfeng Rao, Huaixiu Steven Zheng, Sanket Vaibhav Mehta, Honglei Zhuang, Vinh Q Tran, Dara Bahri, Jianmo Ni, et al. 2021. Ext5: Towards extreme multi-task scaling for transfer learning. *arXiv preprint arXiv:2111.10952*.
- Adam Benkato. 2019. From medieval tribes to modern dialects: On the afterlives of colonial knowledge in Arabic dialectology. *Philological Encounters*, 4(1-2):2–25.
- Adam Benkato. 2020. Maghrebi Arabic. *Arabic and contact-induced change*, 1:197.
- Delphine Bernhard et al. 2021. Collecting and annotating corpora for three under-resourced languages of france: Methodological issues. *Language Documentation & Conservation*, 15:316–357.
- Joachim Bingel and Anders Søgaard. 2017. Identifying beneficial task relations for multi-task learning in deep neural networks. *arXiv preprint arXiv:1702.08303*.
- Houda Bouamor et al. 2018. The madar Arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on LREC*.
- John D Bransford and Marcia K Johnson. 1972. Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of verbal learning and verbal behavior*, 11(6):717–726.
- Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.
- Soravit Changpinyo, Hexiang Hu, and Fei Sha. 2018. Multi-task learning for sequence tagging: An empirical study. *arXiv preprint arXiv:1808.04151*.
- Abdellah El Mekki et al. 2021. Domain adaptation for Arabic cross-domain and cross-dialect sentiment analysis from contextualized word embedding. In *Proceedings of the 2021 Conference of the North American Chapter of the ACL: Human Language Technologies*, pages 2824–2837.
- Mohamed Taybe Elhadi and Ramadan Alsayed Alfared. 2022. Adopting Arabic taggers to annotate a libyan dialect text with a pre-tagging processing and term substitutions. *ISTJ*.
- Chris Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn. 2021. Efficiently identifying task groupings for multi-task learning. *Advances in Neural Information Processing Systems*, 34:27503–27516.
- Mary L Gick and Keith J Holyoak. 1980. Analogical problem solving. *Cognitive psychology*, 12(3):306–355.
- Elisa Gugliotta and Marco Dinarelli. 2020. TArC: Incrementally and Semi-Automatically Collecting a Tunisian Arabish Corpus. In *Proceedings of the Twelfth LREC*, pages 6279–6286.

- Elisa Gugliotta and Marco Dinarelli. 2022. TArC: Tunisian Arabish Corpus first complete release. In *Proceedings of the Thirteenth International Conference on LREC*.
- Elisa Gugliotta, Marco Dinarelli, and Olivier Kraif. 2020. Multi-task sequence prediction for Tunisian Arabizi multi-level annotation. *arXiv preprint arXiv:2011.05152*.
- Michelle Guo, Albert Haque, De-An Huang, Serena Yeung, and Li Fei-Fei. 2018. Dynamic task prioritization for multitask learning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 270–287.
- Nizar Habash et al. 2018. Unified guidelines and resources for Arabic dialect orthography. In *Proceedings of the Eleventh International Conference on LREC*.
- Michael Hahn. 2020. [Theoretical limitations of self-attention in neural sequence models](#). *Transactions of the ACL*, 8:156–171.
- Salima Harrat, Karima Meftouh, and Kamel Smaïli. 2018. Maghrebi Arabic dialect processing: an overview. *Journal of International Science and General Applications*, 1.
- Benjamin Hary. 1996. The Importance of the Language Continuum in Arabic Multiglossia. *Understanding Arabic: essays in contemporary Arabic linguistics in honor of El-Said Badawi*, pages 69–90.
- Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka, and Richard Socher. 2016. A joint many-task model: Growing a neural network for multiple nlp tasks. *arXiv preprint arXiv:1611.01587*.
- Go Inoue, Hiroyuki Shindo, and Yuji Matsumoto. 2017. Joint prediction of morphosyntactic categories for fine-grained Arabic part-of-speech tagging exploiting tag dictionary information. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 421–431.
- James A Kole and Alice F Healy. 2007. Using prior knowledge to minimize interference when learning large amounts of information. *Memory & Cognition*, 35(1):124–137.
- Kathrein Abu Kwaik, Motaz Saad, Stergios Chatzikyriakidis, and Simon Dobnika. 2018. A lexical distance study of Arabic dialects. *Procedia computer science*, 142:2–13.
- Alexander Magidow. 2021. The old and the new: Considerations in Arabic historical dialectology. *Languages*, 6(4):163.
- Jean Piaget. 2003. *The psychology of intelligence*. Routledge.
- Chatrine Qwaider, Stergios Chatzikyriakidis, and Simon Dobnik. 2019. Can modern standard Arabic approaches be used for Arabic dialects? sentiment analysis as a case study. In *Proceedings of the 3rd Workshop on Arabic Corpus Linguistics*, pages 40–50.
- Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of lstm-networks for sequence tagging. *arXiv preprint arXiv:1707.09861*.
- Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.
- Alexander M Rush, Roi Reichart, Michael Collins, and Amir Globerson. 2012. Improved parsing and pos tagging using inter-sentence consistency constraints. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 1434–1444.
- Anders Søgaard and Yoav Goldberg. 2016. Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of the 54th Annual Meeting of the ACL (Volume 2: Short Papers)*, pages 231–235.
- T Standley, AR Zamir, D Chen, L Guibas, J Malik, and S Savarese. 2019. Which tasks should be learned together in multi-task learning? *arxiv e-prints. arXiv preprint arXiv:1905.07553*.
- Ximeng Sun, Rameswar Panda, Rogerio Feris, and Kate Saenko. 2020. Adashare: Learning what to share for efficient deep multi-task learning. *Advances in Neural Information Processing Systems*, 33:8728–8740.
- Tu Vu et al. 2020. Exploring and predicting transferability across nlp tasks. *arXiv preprint arXiv:2005.00770*.
- Lev Semenovich Vygotsky and Michael Cole. 1978. *Mind in society: Development of higher psychological processes*. Harvard university press.
- Gail Weiss, Yoav Goldberg, and Eran Yahav. 2018. [On the practical computational power of finite precision RNNs for language recognition](#). In *Proceedings of the 56th Annual Meeting of the ACL (Vol 2: Short Papers)*, pages 740–745, Melbourne, Australia. ACL.
- Joseph Worsham and Jugal Kalita. 2020. Multi-task learning for natural language processing in the 2020s: where are we going? *Pattern Recognition Letters*, 136:120–126.
- Nasser Zalmout and Nizar Habash. 2019. Adversarial multitask learning for joint multi-feature and multi-dialect morphological modeling. *arXiv preprint arXiv:1910.12702*.
- Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. 2018. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3712–3722.
- Zhihan Zhang, Wenhao Yu, Mengxia Yu, Zhichun Guo, and Meng Jiang. 2022. A survey of multi-task learning in natural language processing: Regarding task relatedness and training methods. *arXiv preprint arXiv:2204.03508*.

Crowdsourcing OLiA Annotation Models the Indirect Way

Christian Chiarcos

Applied Computational Linguistics
University of Augsburg, Germany

christian.chiarcos@philhist.uni-augsburg.de

Abstract

The paper describes a technology to complement established documentation workflows in two linguistic community projects with the possibility to automatically create OLiA Annotation Models, i.e., formal, ontological representations of their annotation schemas. For this purpose, we provide a domain-specific extractor that consumes MediaWiki wikitext, extracts sections headers and tables and produces an OWL2/DL ontology as a result. This ontology can be further processed with standard technology as established in the context of the Linguistic Linked Open Data (LLOD) community. The main contribution we provide effectively eliminates the entry barrier into LLOD technology and OLiA for two potential user communities, and that this setup can be trivially adopted to any comparable community project – as long as it uses Wiki technology and Wiki lists for documenting tags and abbreviations.

1 Background and Motivation

The Ontologies of Linguistic Annotation (Chiarcos, 2008; Chiarcos and Sukhareva, 2015, OLiA) serve as a central hub for linguistic annotation terminology on the web of data, and they constitute a formative element of the Linguistic Linked Open Data (LLOD) cloud in that they provide machine-readable semantics for linguistic annotations. These ontologies define reference concepts and relations that can be used to annotate linguistic data in a standardized way, making it easier to share and compare data across different languages and domains.

Applications of OLiA include the mapping of tags from one annotation schema to their closest counterparts in another schema (Chiarcos and Ionov, 2021), to perform cross-corpora queries across different corpora (Chiarcos and Gã-tze, 2007), to aggregate information across heterogeneous tagsets in ensemble combination architectures (Chiarcos, 2010) or in multi-source annota-

tion projection (Sukhareva and Chiarcos, 2016). Being based on RDF technologies, all of this can be achieved on-the-fly by identifying the shortest paths between different OLiA ontologies by means of a W3C-standardized query language (SPARQL). As schemas differ in their granularity, this mapping is not free of information loss, but its dynamic aspects sets OLiA apart from other attempts to establish interoperability between different annotation schemas such as EAGLES (Calzolari and Monachini, 1996) or the Universal Dependencies (De Marneffe et al., 2021), in that it does not require a transformation of the original annotations, but instead, leaves the original annotations untouched, and only complements them with a more interoperable interpretation.

For more than 100 languages, OLiA covers different aspects of linguistic annotation, including Part of Speech (PoS) annotation, syntax, and inflectional morphologies. Aspects of discourse semantics (discourse structure, discourse relations, information structure, anaphora, coreference, named entities) are subject to a separate discourse extension (Chiarcos, 2014). Despite its potential benefits in interoperability and interpretability, it can be complicated for the developer of a corpus or an NLP tool to produce a certain type of annotations to provide an OLiA Annotation Model, because this requires a set of technical skills that neither most linguists nor most web developers, nor most NLP specialists, possess.

This paper aims to address the challenge to create annotation models. For the integration of a language resource into the OLiA ecosystem, this normally represents the first step to take, but a relatively hard one for, say, a linguist working on an annotated corpus, or a developer not intrinsically familiar with RDF technology. Our proposed solution is to integrate ontology development into established documentation workflows, so that users are creating an ontology along with their regular

work without even noticing it.

2 The Ontologies of Linguistic Annotation

The OLiA ontologies define a set of reference categories for linguistic annotations. On the one hand, this pertains to linguistic concepts as used in tagsets, annotation schemes and lexical resources (OLiA Reference Model),¹ on the other hand, OLiA provides formalizations of entire annotation schemas or (families of) language resources (OLiA Annotation Models).²

Annotation Model concepts are linked to OLiA Reference Model concepts by means of `rdfs:subClassOf`/`rdfs:subPropertyOf` relationships, exploiting the full band-width of OWL2/DL semantics (i.e., class intersection \sqcap , union \sqcup and complement \neg operators). Every annotation model resides in a separate, stand-alone ontology, and for every annotation model, there is at least one linking model in which the mapping to OLiA Reference Model concepts is provided.³ This declarative, machine-readable mapping helps to disentangle definition and interpretation, and, moreover, it facilitates debugging, future revisions and portability across different platforms. Also, it is a feature that sets OLiA apart from other, past and present, standardization efforts such as EAGLES (Calzolari and Monachini, 1996), ISOcat (Kemps-Snijders et al., 2008) or the Universal Dependencies (De Marneffe et al., 2021) – all of these employ(ed) opaque scripts to produce standard tags which can only be debugged and consulted *in code* – if publicly available at all.

In a similar way, the OLiA Reference Model is also linked with other, community-maintained reference terminologies such as ISOcat (Kemps-Snijders et al., 2008) or the General Ontology of Linguistic Description (Farrar and Langendoen, 2010), and the OLiA Reference Model partially builds on these, but further domain-, theory- or language-specific reference terminologies are likewise integrated with OLiA (Chiarcos et al., 2020a). This includes, for example, UniMorph (McCarthy et al., 2020, specific to inflection morphology), Lex-

Info (McCrae et al., 2017, specific to linguistic terminology for lexical resources in OntoLex-Lemon), or the BLL Thesaurus (Chiarcos et al., 2016, linguistic metadata for a linguistic bibliography).

In the context of LLOD, OLiA serves mostly as an additional layer of interoperable annotations over language resources such as corpora (Bosque-Gil et al., 2018), but also, it is a central component of the NLP Interchange Format, and thus, of web services that dynamically cater linguistic annotations (Hellmann et al., 2013). Yet, OLiA provides potential users and contributors with a certain entry bias, as it is based on RDF technologies as its technical backbone. This paper aims to address one of the aspects of the challenge, the creation of annotation models.

We provide three components designed for bootstrapping OLiA Annotation Models from conventional annotation documentation: (1) a configurable tool to convert MediaWiki source files into OWL ontologies, (2) a novel Annotation Model for morphological analyzers from Apertium, and (3) an Annotation Model for linguistic glosses from Wikipedia. Our converter is a relatively small, but generic piece of code. It can be configured for different constellations, and it requires the source data to provide Wiki tables with one row corresponding to one individual in the end. It is optimized for the extraction tasks at hand, but it is sufficiently that, for any data that comes in a similar form, it can be either directly employed or easily adapted.

3 An Annotation Model for Apertium

Apertium⁴ is an open-source machine translation (MT) system, developed by a large community of volunteers and enthusiasts. Apertium focuses on symbolic, rule-based approaches on machine translation, which are particularly fruitful for closely related language pairs with insufficient resources to train a neural or statistical MT system on. Indeed, rule-based generation requires textbook expertise and bilingual word lists for its development, but not necessarily parallel corpora.

The Apertium ecosystem comprises

1. a machine translation engine,
2. tools to manage the necessary linguistic data for a given language pair, and

¹Namespace prefix `olia:`, reference URL <http://purl.org/olia/olia.owl#>.

²As an example, the Penn Treebank schema, namespace prefix `penn:`, resides under <http://purl.org/olia/penn.owl#>.

³For the Penn Treebank tagset, the linking model resides under <http://purl.org/olia/penn-link.rdf>.

⁴<https://www.apertium.org>

3. language resources (morphological analyzers, dictionaries) for 51 languages and 53 language pairs considered stable (plus 135 languages and 249 language pairs with experimental support and at different degrees of maturity).⁵

3.1 Apertium Morphosyntactic Annotations

Apertium implements symbolic, transfer-based machine translation, where source language input is first morphologically and syntactically analyzed, then, the lemmas are word-wise translated into the target language, where restructuring rules and surface generation takes place. As such, it provides or wraps a large ensemble of morphological generators and analyzers, often based on finite state transducers (FST).

Apertium tags and morphosyntactic features are not standardized across languages, but they share some common conventions.⁶ To some extent, these are in a continuous state of flux, as new language pairs are coming in (and bring in new terminology), while the community presses for more consistency across them. These update processes are relatively slow, as new languages are coming in at a moderate rate, so, any annotation model built from this documentation is likely to remain valid for the coming years, but still needs to be regularly updated. As there is no overall versioning applied across all Apertium language pairs, the documentation and any OLiA Annotation Model derived from it reflects the status at a particular state in time, and requires a timestamp as metadata to make this explicit.

Here, we focus on morphological analyzers within Apertium, and, normally, these represent the first component to be provided for any particular language – and, in fact, for some language pairs, machine translation is or can be implemented using only the FST technology that is also underlying the morphological analysis. This is somewhat different from earlier approaches on connecting Apertium with LLOD technology, as this was solely focusing on the dictionaries also contained in Apertium (Gracia et al., 2018; Chiarcos et al., 2020b; Gracia et al., 2020), and the most recent version of this data includes a manually verified mapping from ab-

brevisions/tags to the LexInfo 3.0 ontology,⁷ and thus, indirectly, to OLiA. However, this is necessarily incomplete, as the dictionaries account for open-class lexemes and selected parts of speech only, but not for morphological processes, function words and their morphosyntactic features – all of as these are handled via hand-crafted grammar rules in Apertium, but not by the Apertium dictionaries.

As opposed to this, we aim to provide a more exhaustive mapping that also allows the future development of RDF-based web services as wrappers around Apertium *analyzers*, the LLOD publication of Apertium-compliant corpora, or the linking of such corpora with Apertium-based and other OntoLex dictionaries. It is to be noted, however, that we rely exclusively on the available documentation and provide a fully automated conversion only. If there are omissions or errors in the documentation, or if any particular tool does not adhere to the overall recommendations, these aspects will not be covered by our annotation model.

3.2 Conversion to RDF

Apertium symbol definitions are provided in a wiki page⁶ with tables for different kinds of annotations, separated by headlines (see Fig. 4 in the appendix). For converting Apertium data, we operate with wiki text (MediaWiki source code). This is because in established Apertium workflows, the list of symbols is designed to be scrapeable, it provides additional information in its comments, and explicit guidelines for systematicising tables, headline formatting and the marking of tags.

We aim for a generic tool, so we do not *depend* on these conventions (also cf. Fig. 4 as an illustration for the degree of variation observed on the page), but we *respect* them. Our conversion operates as follows:

1. We retrieve the original wikitext using the flag `?action=raw` (cf. Fig. 1).
2. We create the class `:Symbol` as a top-level class, using a user-provided base URI as namespace.
3. For every headline under which (directly or indirectly) at least one table is found, we create a class from the label enclosed in `<!-- ... -->`, if this is not available, we operate with the section title, instead. The class

⁵https://wiki.apertium.org/wiki/List_of_language_pairs

⁶https://wiki.apertium.org/wiki/List_of_symbols

⁷lexinfo.net/

```

==Part-of-speech Categories== <!-- POS -->
{|class=wikitable
! Symbol          !! Gloss                !! Notes                !! Universal POS
|-
| <code>n</code>    || Noun                  || ''see 'np' for proper noun''           || NOUN
|-
| <code>vblex</code> || Standard ("lexical") verb || ''see also: vbser, vbhaver, vbmod, vaux, vbdo'' || VERB
|-

```

Figure 1: Apertium list of symbols (wikitext, excerpt).

name is normalized by enforcing CamelCase, removal of whitespaces, and URL encoding. Also, if the class name happens to have been previously created during the conversion, we produce a unique name by attaching a numerical suffix. The original section header is given as an `rdfs:label`.

4. Based on the hierarchy of headlines, every class is assigned a super-class generated from its header, resp., `:Symbol` for top-level section headers.

Output generated so far from the snippet given above is:

```

:POS rdfs:subClassOf :Symbol;
  rdfs:label
    "Part-of-speech Categories"@en .

```

5. For every witable, we determine the column labels from its header how, splitting at `!!`. Column labels are normalized by camelCase conversion, lower-casing of the first word and whitespace removal. These will become RDF properties when processing the following rows. If a witable does not provide a header row, we re-use the last established header row. If no header has been established before, the table is skipped with a warning.

For the table in Fig. 1, the normalized column labels are `symbol`, `gloss`, `notes`, and `universalPOS`.

6. For every row within a table, we split its columns at `||` and align them with the column labels provided in the header.

For the first row in the snippet above, this yields (shown here as a JSON dictionary):

```

{"symbol" : "<code>n</code>",
 "gloss"  : "Noun",
 "notes"  : "'see 'np' ...'",
 "universalPOS" : "NOUN" }

```

7. For every row, determine its identifier by following a sequence of user-provided column names (by default `symbol`, `symbols`, `tag`, `xmlTag`, `xmlAttributeValue`, as needed for the Apertium page): for the first of these column labels found in the current table, we retrieve the cell value as label. We remove XML markup from this label, normalize whitespaces and punctuation to `_` and apply lowercasing and URI encoding to obtain (the local name for) the URI. If the resulting symbol is not unique, we attach a numerical suffix. The row URI is assigned the class derived from its section header as an `rdf:type`:

```

:n a :POS .

```

8. For every column in the current row, we create a triple where the property (derived from the normalized column labels) provides the cell content (stripped of markup and white-space normalized) as a string value:

```

:n :symbol "n" ;
  :gloss "Noun" ;
  :notes "'see 'np' ...'";
  :universalPOS "NOUN" .

```

This conversion is applicable to any wikitext page that provides wiki tables with explicit headers. Section headers are optional. It is required, though, that a user provides the base URI and a (normalized) column label that determines how to identify the columns from which row URIs are to be created.

3.3 Introducing Standard Vocabularies

An additional parameter that a user can provide is a mapping from normalized column labels to RDF vocabularies, provided as a JSON dictionary. The defaults account for converting the Apertium page:

```

{
  ":symbol" : "olias:hasTag",
  ":symbols": "olias:hasTag",
  ":tag"    : "olias:hasTag",

```

```

    ":xmlAttributeValue": "olias:hasTag",
    ":xmlTag": "olias:hasTag",
    ":gloss": "rdfs:label",
    ":notes": "rdfs:comment",
    ":means": "rdfs:comment",
    ":description": "rdfs:comment"
  }
}

```

Properties not listed here are preserved. In the Apertium data, this applies to `:appearsInAttributeNotes`, `:appearsInXMLTagsNotesExamples`, `:universalFeature`, `:universalFeatures`, and `:universalPOS`. With these replacements, we arrive at the following representations of the first row in our data set:

```

:POS rdfs:subClassOf :Symbol;
    rdfs:label
      "Part-of-speech Categories"@en .

:n a :POS .
:n olias:hasTag "n" ;
  rdfs:label "Noun" ;
  rdfs:description
    "'see 'np' for proper noun'";
  :universalPOS "NOUN" .

```

What remains to do to qualify this as an OLiA annotation model is to declare this file an ontology and to provide elementary metadata:

```

<.../apertium.owl> a owl:Ontology ;
  rdfs:comment
    "OLiA Annotation Model for
    Apertium ..." ;
  rdfs:isDefinedBy
    <https://wiki.apertium.org/wiki/
      List_of_symbols> ;
  owl:versionInfo "2023-03-07 12:06:48" .

```

The object URI of `rdfs:isDefinedBy` is extrapolated from the base URI – unless explicitly specified by the user. As OLiA Annotation Models are traditionally provided as RDF/XML, the resulting Turtle file is converted with off-the-shelf tools. The resulting OWL file can be loaded and processed with off-the-shelf Semantic Web tools, e.g., with the ontology browser Protégé, cf. Fig. 2.

4 Wikipedia Glossing Abbreviations

Wikipedia⁸ is the prime example for a collaboratively constructed, community-maintained resource, and it is acknowledged as that since more than two decades. Unsurprisingly, it also found some popularity among people interested in or professionally working with language, and as such, it serves as a knowledge hub for linguistically relevant topics, and often the first place to look for orientation.

⁸<https://www.wikipedia.org/>

One such application is that Wikipedia seems to be used by students and linguistic practitioners as a central point to collect and to document glosses used as abbreviations in linguistic literature, in particular in the context of interlinear glossed text (IGT, cf. Appendix Fig. 5).⁹ IGT is a format consisting of multiple lines where the first line usually represents a source language string, the following lines provide linguistic analyses, e.g., a transliteration, linguistic glosses, morphological segmentation, morpheme glosses, etc. Typically, the last line comprises a translation into the description language.

This formalism is widely used for educational purposes, for language documentation and in linguistic typology, and it has also been converted to a Linked Data representation and produced a native RDF vocabulary specifically for this purpose, Ligt (Chiarcos and Ionov, 2019; Nordhoff, 2020; Ionov, 2021). Ligt, however, only captures the *structure* of IGT formats, for the semantics of the tags used in that context, it relies on OLiA – which provides a small number of IGT-relevant annotation models, e.g., the UniMorph schema (Chiarcos et al., 2020a) and the glossing guidelines of Dipper et al. (2007), which incorporated the Leipzig Glossing Rules (Committee of Editors of Linguistics Journals, 2008/2015) and extended them to syntax and information structure.

A second usage in the context of Wikipedia itself is that it provides templates for producing interlinear glossed text as part of Wikipedia pages, and these abbreviations are recommended for use. At a future point in time, they may actually be automatically linked to the current website if mentioned in the template, but at the moment, the automated linking operates on a shorter, and older excerpt of these abbreviations. Both the Wikipedia templates and their surface rendering are illustrated in the appendix (Fig. 6). As of March 1, 2023, the English Wikipedia contains 7,639 instances of the interlinear template on 651 pages,¹⁰ plus an unknown number of applications of derived templates (e.g., `fs_interlinear` or language- or script-specific templates).

The Wikipedia gloss labels are not directly tied to any particular data, but their usage in combina-

⁹https://en.wikipedia.org/wiki/List_of_glossing_abbreviations

¹⁰<https://bambots.bruceymyers.com/TemplateParam.php?wiki=enwiki&template=Interlinear>

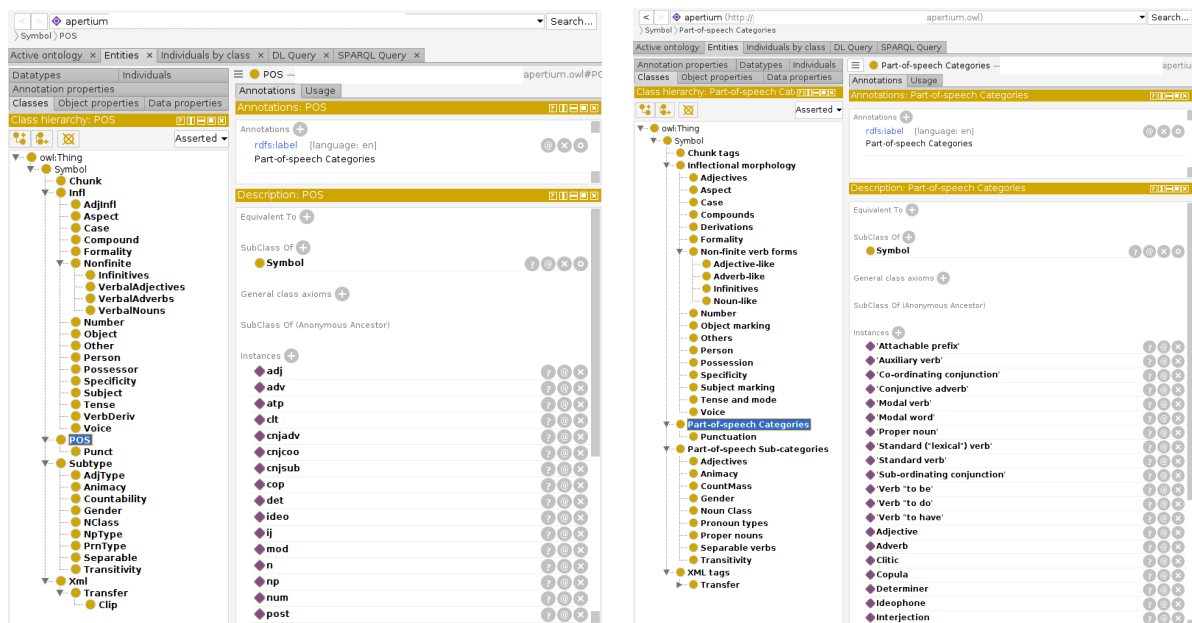


Figure 2: Apertium Annotation Model, visualized with Protégé, configured to display of URIs (left), resp. labels (right).

tion with Wikipedia templates for interlinear glossing is recommended. Furthermore, they are frequently consulted (and extended) by practitioners in the field, in particular by students, so that they attain a certain near-normative function. In the context of efforts to mine scientific papers for machine-readable versions of interlinear glossed text comprised in them (Lewis and Xia, 2010; Nordhoff and Krämer, 2022), it becomes increasingly relevant also to provide machine-readable semantics for the abbreviations, especially if such data is to become the basis for further linguistic research or language technological solutions for low-resource languages as repeatedly proposed over the years (Bender et al., 2014; Zhou et al., 2019).

It is to be noted, however, that glosses and concepts used in the literature reside in an $n:m$ relationship, so that the same abbreviation is used for one purpose by a particular researcher, but for another by another person. As an example, the abbreviation AC is defined as “motion across (as opposed to up/down-hill, -river)”, as “animacy classifier”, or as “accusative case”. This is why “conventional glosses” have been singled out, and except for a small number of exceptions, these provide a 1:1 mapping. For the specific case of AC, this is not considered a conventional gloss at all (because of its ambiguity), and for the functions mentioned before, only accusative case (with the tag ACC) receives that status.

The application of our converter to Wikipedia was straight-forward. The extraction was performed via the Wikipedia API, but the resulting wikitext followed the same conventions (albeit much less constrained than in Apertium). Beyond that, user parameters (base URI, source URI, column labels and their mapping to properties) were adjusted: The row URI is taken from the column with the (normalized) labels `conventionalGloss` (for grammatical abbreviations, punctuations and numbers), and `2LetterGloss` (for kinship terms). As not every row provides a conventional gloss, we also added the column `variants` to the list of URI-defining columns: By ordering preferences, this is used for URI generation only of neither `conventionalGloss` nor `2LetterGloss` are found.

Our conversion of abbreviation variants is lossless in the sense that these are preserved, but only as attribute values, we do not create a distinct tag with its specific `alias:hasTag` for each of them. This was done in order to properly distinguish preferred (readings of) glosses from dispreferred (glosses or readings). The original objective of distinguishing conventional and variant glosses in Wikipedia seems to be that the same gloss was used for different, unrelated meanings (1:m mappings), while at the same time, the same meaning could be expressed by a variety of tags. The current distinc-

tion has been introduced to enable a 1:1 mapping (even though this is not fully achieved).

The resulting ontology is analogous in structure and vocabulary to the Apertium ontology. A difference is that the concept hierarchy of Wikipedia is much shallower, grouping all morphosyntactic features and categories together under the umbrella of `:GrammaticalAbbreviations`.

5 Automatically Supported Linking

To facilitate the creation of OLiA Linking Models, we provide a command-line tool that takes three main parameters, one reference model (that provides concepts that represent superclasses in the linking), one annotation model (that provides concepts and individuals that are assigned superclasses in the linking) and one linking model (specifying the file into which the resulting mapping is to be written).¹¹ By default, the linking procedure only creates `rdfs:subClassOf` links between concepts and `rdfs:subPropertyOf` links between properties, but with the flag `-indiv`, it also creates `rdf:type` links between annotation model instances and reference model classes.

The comparison is performed in several steps. If one step produces no linking candidates, it resorts to the next. For a given annotation model concept (or individual), check all reference model concepts in the following way:

1. Convert local names of the URIs from camel case to lower-cased whitespace segmentation. If both strings match, the reference model URI is a linking candidate.
2. Convert local names and RDF/SKOS labels to lower-cased whitespace segmentation. If two strings match, the reference model URI is a linking candidate.
3. Convert local names and RDF/SKOS labels to lower-cased whitespace segmentation and retrieve the set of words used for describing for both URIs. If there is an overlap between both sets of words, the reference model URI is a linking candidate.

The linking tool is interactive, and for every annotation model word for which at least two candidates are found, it presents these to the user as an ordered

¹¹This tool is not specific to OLiA, so we use lower case spelling. Indeed, any pair of ontologies can be linked in that manner.

list. The user can manually select one of the candidates by entering its number, optionally add a comment or state that no linking candidate is applicable. If there is one linking candidate, it is automatically linked (and marked by an `rdfs:comment` in the Linking Model), if there are none, this is marked by an `rdfs:comment`.

This way of linking is restricted, as it is incomplete and heuristic, but it is also *very fast*. In most cases, processing an Annotation Model concept requires 2-3 key strokes: the number of the selected reference model concept (or 0 for no match) and `<ENTER>`. Yet, manual refinement is highly recommended, and automated comments are generated to guide the way.

We can bootstrap a baseline linking with the OLiA Reference Model from the existing LexInfo linking for Apertium dictionary – but this accounts only for parts of speech, not for grammatical features. In total, 197 Apertium Wiki tags can be linked in this way. Overall, the Apertium ontology comprises 37 classes (headlines) and 301 instances (tags). In addition to this, the automated procedure produced 26 `rdfs:subClassOf` and 22 `rdf:type` links against the OLiA Reference Model, and 15 `rdfs:subClassOf` links against the OLiA Top Model. The limited coverage of linking for instances is partially due to the degree of underspecification they are presented in the table. In parts, however, it is also due to gaps in OLiA. As such, OLiA does currently not support Bantu nominal classes (that alone accounts for 3% of the gaps) and other features specific to certain languages or language families. While language-specific features are generally beyond scope for OLiA, we strongly suggest to extend it with features relevant to entire language families.

6 Manual Linking for Wikipedia Glossing Abbreviations

For Wikipedia glosses, we found that only 82 (16%) were previously covered by the Linking Models for UniMorph (68 in total) or the Dipper et al. (2007) model (42 in total, 28 in both). This linking exploits that the same set of conventional tags were inherited from the literature into these models, but with the automatically supported linking, this number could only be increased by 9 `rdf:type` links. On the one hand, this indicates a certain level of underspecification and idiosyncrasy in both resources, as clearly evident from the brevity of definitions

in Wikipedia, for example; in parts, this is due to gaps in OLiA (for example, it doesn't currently account for kinship terms as there do not seem to exist any corpora that contain or tools that produce such annotations, kinship terms alone represent 7.5% of conventional Wikipedia glosses). On the other hand, this discrepancy may also indicate a fundamental difference between Wikipedia glossing abbreviations (resp., the scholarly tradition from which these emerge) and OLiA (developed with a focus on linguistically annotated corpora, not text book examples).

In order to explore this further, we resort to manual linking of Wikipedia glossing abbreviations, and we expect that this process may lead to a number of suggestions regarding extensions or restructuring of the OLiA Reference Model as a side-product of the process: The annotation model developed so far represents a solid basis from which a concept hierarchy can be manually crafted in an ontology editor. Unfortunately, the current data is represented in a relatively shallow way, as a limitation for Wikipedia glosses is that (except for the basic distinction between punctuation and numbers, grammatical abbreviations and kinship terms), they are relatively unstructured: Abbreviations are provided as an alphabetically organized list, without being grounded in an overarching taxonomy. The task is thus to pick instances (representing conventional or variant glosses) from an unstructured list and to put them into the OLiA categories they belong to, ideally using drag-and-drop mechanisms.

Protégé is a seminal OWL editor and it allows both to manually create a concept hierarchy and provides an interface for quickly re-classifying individuals by means of drag and drop.¹² To this end, we created a novel ontology and imported both the generated Wikipedia ontology and the OLiA top-level ontology and manually classified the Wikipedia glosses according to their type. The top-level ontology defines the root concepts of OLiA, i.e., types of units (e.g., `oliat:Word`) and features (e.g., `oliat:MorphosyntacticFeature`, `oliat:GenderFeature`, etc.). Although this coarse-grained classification does not yet establish a proper linking between Wikipedia glosses and the OLiA Reference Model, it allows for a rough classification that can be the basis for subsequent re-

finements, or serve to evaluate future linking methods. Figure 3 illustrates the manual reclassification procedure.

At the moment, this process of re-classification is still ongoing. Preliminary findings indicate that many Wiktionary glosses are ambiguous or underspecified in that they really act like *abbreviations* for terms, not like *tags* for linguistic annotation. And the same term may occur in different contexts. As such, the conventional tag REP stands for 'repetitive', but the meaning is further explained as either 'repetitive aspect' (otherwise referred to as iterative aspect), 'repeated word in repetition' (echo word) or 'repetitive numeral' (numeral formed by reduplication of a basic numeral).¹³ A linking to existing OLiA Reference Model concepts is possible, and using OWL2/DL semantics, the ambiguity can be expressed in OLiA:

```
wiki:screp ∈
  olia:IterativeAspect ⊔
  olia:EchoWord ⊔
  (olia:Reduplication ⊔ olia:Numeral)
```

Such a complicated linking cannot be established with the automated linking procedure described below, nor with manual the drag-and-drop method, both of which only support direct type assignments. The necessary anonymous classes representing intersections or unions have to be constructed manually, and this is also supported by Protégé. Moreover, this example also illustrates to some extent *why* the linking is failing at times: The OLiA terms 'iterative aspect', 'echo word', and 'reduplication' have no counterpart in the Wikipedia description.

7 Summary and Discussion

This paper described the automated creation of OLiA Annotation Models for different community projects, based on the conversion of wikitext and its layout conventions for section headings and tables. The converter and the associated linking tool are published under open source as part of the OLiA GitHub repository.¹⁴ Both tools are relatively

¹³According to Turner (1967, p.285), the Chontal phrase *núli núli* 'completely' is a repetitive numeral based on *núli* 'one'.

¹⁴<https://github.com/acoli-repo/olia/tree/master/tools>. Also, the ontologies are provided there, currently under <https://github.com/acoli-repo/olia/tree/master/owl/experimental/meta>. Later on, they are expected to migrate to the stable release (<https://github.com/acoli-repo/olia/tree/master/owl/stable>

¹²This functionality is available from the "Individuals by type" view, not enabled by default, but available via Window|Views|Individual views (Protégé 5.5.0, Desktop).

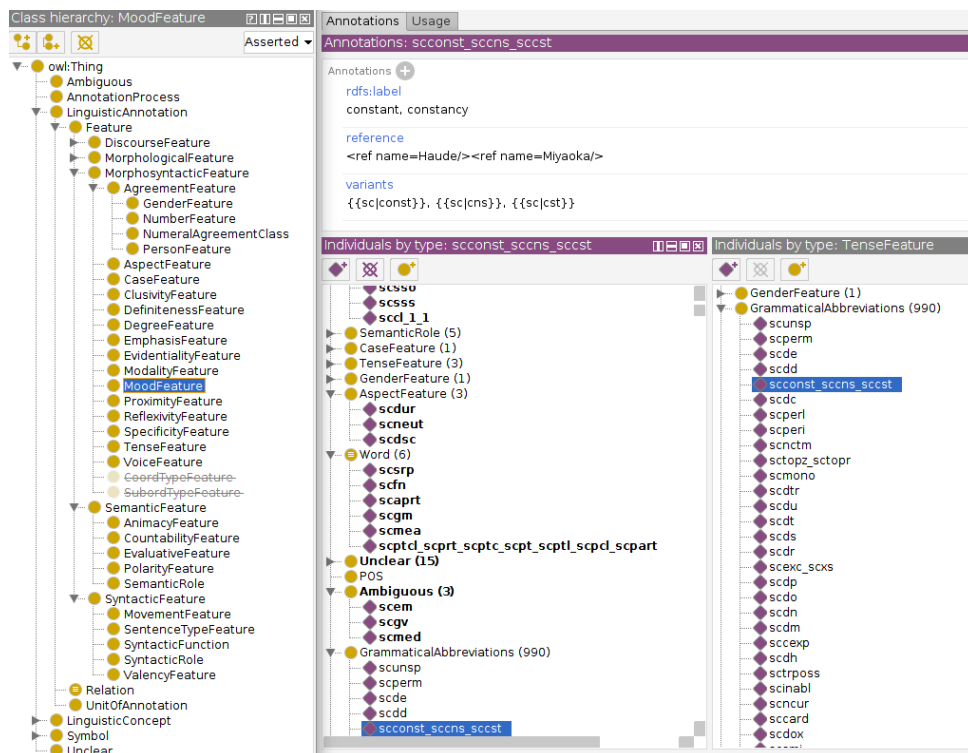


Figure 3: Drag-and-drop classification of Wikipedia abbreviations against OLiA top-level concepts: Between both "Individuals per type" tabs, RDF individuals can be moved by drag and drop. The Annotation view tab above shows the annotations of the current individual. On the left, you see (and can edit) the concept hierarchy.

simple command-line tools with a high level of genericity. The converter is applicable to any MediWiki content with tables, the linker is applicable to any pair of ontologies.

Conversion from HTML and other web formats is a standard task and has been conducted countless times. For example, DBpedia,¹⁵ DBnary¹⁶ and UniMorph¹⁷ are all based on extraction templates applied over Wikipedia, resp. Wiktionary – although for different types of data. DBpedia and DBnary are also routinely updated in this manner, whereas UniMorph data is frozen and conversion scripts do not seem to be publicly available. Our approach differs in that we do not extract a dataset (ABox), but an ontology (TBox), and that it operates on a much more fine-grained scale. This allows, for example, to expose the result of the build process directly to the user, again.

In particular, the build process can be extended to produce either a graphical representation of the resulting ontology or to apply an interactive browser to the result, so that users can dynamically explore, browse and search their annotation mod-

els with off-the-shelf tooling. The integration of existing documentation with such visualizations remains, however, a subject of future efforts, as different possibilities exist for this purpose, and the preferences within the communities need to be taken into consideration. The classical approach to ontology visualization is to convert RDF to the Dot language and to generate a static image with GraphViz.¹⁸ Similarly, SVG and SVG renderers can be used for the same end.¹⁹ The downside of this approach is that the image is to be manually uploaded to or updated in the respective wiki. Alternatively, it is possible to directly link interactive visualization tools such as WebVOWL, along with the URL that contains the ontology to be visual-

¹⁸This has been the basis for a number of classical ontology/RDF visualizers integrated in the Protégé ontology viewer. At the present day, the conversion to Dot can also be performed by a web service, e.g., <https://www.easyrdf.org/converter?out=dot&raw=1&uri=>, followed by the ontology URL. For generating an actual image, different layout schemes can be employed, and we recommend using a local installation of GraphViz, because this is more easily scriptable than online services such as WebGraphViz (<http://www.webgraphviz.com/>).

¹⁹As provided, for example, by yWorks: <https://www.yworks.com/use-case/visualizing-an-ontology>.

¹⁵<https://www.dbpedia.org/>

¹⁶<http://kaiko.getalp.org/about-dbnary/>

¹⁷<https://unimorph.github.io/>

ized.²⁰

What is interesting about the approach is that it allows to fully automatically create formal ontologies (OLiA Annotation Models) on the basis of established community workflows. We could *build* on established Apertium conventions for their list of symbols, and we could *build* on the current practices in the maintenance and development of the Wikipedia glossing abbreviations. (And, as both as community-maintained, if these conventions would ever be broken by another contributor, and this is noted by our tools, we can fix those issues directly.) At no point did we have to *enforce* new requirements to enable the creation of an OLiA Annotation Model, and neither did we ask Apertium or Wikipedia contributors to operate with a cumbersome tool for handling RDF and linked data. In other words, the entry barrier for OLiA and LLOD technology has been almost eliminated for these groups of users. This also sets it apart from solutions such as VocBench (Stellato et al., 2020) or OpenRefine (Miller and Vielfaure, 2022), which already require their users to have an innate interest in Linked Data or Semantic Web technologies, so that they are actively operating towards this goal with the intent to create a mapping into a machine-readable format. This is not required here, as, instead, the converter is already provided. Moreover, we are concerned with crowd-sourced, community-maintained data, which has a certain quality of being in a continuous update and revision process. So, extraction needs to be repeated relatively frequently – but OpenRefine and VocBench are not designed for repeated conversion, as these are highly interactive tools.

The creation of Linking Models, then, requires a higher level of technical expertise, of course, but this does not have to be provided by an Apertium or Wikipedia contributor, instead, it can come from the LLOD community. And if more technically oriented community members see scientific or technological value in that kind of data *for their own purposes*, this is likely to happen.

It should be noted that the approach to create ontologies as a side-product of established community conventions for maintaining and creating their

²⁰At the time of writing, the recommended URL for that purpose would be <http://vowl.visualdataweb.org/webvowl-old/webvowl-old.html#iri=>, followed by the ontology URL. However, as the `-old` link indicates, the system is currently in transition to a novel backend, so that link might change.

documentation, is not the first of its kind either. We conducted an earlier, unpublished experiment that infused RDFa attributes into Jekyll templates, so that HTML pages generated from Markdown (as used by the Universal Dependency community to document their annotation schemas) would already contain a machine-readable representation of these schemas. The technology worked very well, and a prototype over an older version of UD guidelines with RDFa markup is still online,²¹ and using an RDFa reader on the published HTML pages, a full-fledged ontology could be derived on the fly and queried with SPARQL. From the perspective of a UD contributor, nothing changed, and the process was taking advantage of established conventions originally intended to streamline the layout, especially the usage of explicit variables for certain aspects, and the section structure of the Markdown document. A downside here was that the build process was relatively unstable, and it turned out to take too long for efficiently debugging and maintaining this setup (several minutes, but sometimes more), so that eventually, this experimental prototype was discontinued, and with a change of layout and Markdown conventions with the transition from version 1.0 to 2.0 of the Universal Dependencies, they have not been updated.

With our converter, we do not rely on such a complicated setup. Instead, we provide a simple script for building Annotation Models, and using a cron job, they can be repeatedly called to provide up-to-date RDF data for Annotation Models and visualizations. If deployed on a web server, these can be produced by a third party, independently from the infrastructure of the particular community involved.

Our tools and annotations have been integrated into the OLiA GitHub repository,²² so they will remain accessible to the community as long as OLiA remains a relevant resource. Moreover, they will be subject to any long-term sustainability solution developed for OLiA in the future.

²¹See <http://fginter.github.io/docs/>. Note the small RDF logos that trigger the RDFa parsing process. However, these URLs contain a GET request at a public web service for RDFa parsing, after more than a decade of successful operation, was shut down mid-last year, so that these links yield a status page, not RDF data in Turtle, anymore. Alternative web services are available, but the links in this prototype have not been updated, yet.

²²<https://github.com/acoli-repo/olia/>

Acknowledgements

We would like to thank three anonymous reviewers for comments and feedback, which have been integrated into this paper, the authors of the Wikipedia glossing page and the Apertium documentation, upon whose work we build here, and the developers of the earlier Apertium-Lexinfo mapping, most notably Julia Bosque-Gil and Max Ionov.

References

- Emily M. Bender, Joshua Crowgey, Michael Wayne Goodman, and Fei Xia. 2014. [Learning grammar specifications from IGT: A case study of chintang](#). In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 43–53, Baltimore, Maryland, USA. Association for Computational Linguistics.
- Julia Bosque-Gil, Jorge Gracia, Elena Montiel-Ponsoda, and Asunción Gómez-Pérez. 2018. Models to represent linguistic linked data. *Natural Language Engineering*, 24(6):811–859.
- Nicoletta Calzolari and Monica Monachini. 1996. EAGLES Proposal for Morphosyntactic Standards: in view of a ready-to-use package. In G. Perissinotto, editor, *Research in Humanities Computing*, volume 5, pages 48–64. Oxford University Press, Oxford, UK.
- Christian Chiarcos. 2008. An ontology of linguistic annotations. *LDV Forum*, 23(1):1–16.
- Christian Chiarcos. 2010. Towards robust multi-tool tagging. An OWL/DL-based approach. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 659–670.
- Christian Chiarcos. 2014. Towards interoperable discourse annotation. discourse features in the ontologies of linguistic annotation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 4569–4577.
- Christian Chiarcos, Christian Fäth, and Frank Abromeit. 2020a. Annotation interoperability for the post-ISOCat era. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5668–5677.
- Christian Chiarcos, Christian Fäth, and Maxim Ionov. 2020b. The acoli dictionary graph. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3281–3290.
- Christian Chiarcos, Christian Fäth, and Maria Sukhareva. 2016. [Developing and using the ontologies of linguistic annotation \(2006-2016\)](#). In *Proceedings of the 5th Workshop on Linked Data in Linguistics: Managing, Building and Using Linked Language Resources (LDL-2016)*, pages 63–72, Portorož, Slovenia.
- Christian Chiarcos and Michael GÄ-tze. 2007. A linguistic database with ontology-sensitive corpus querying. In *Datenstrukturen für linguistische Ressourcen und ihre Anwendungen. Frühjahrstagung der Gesellschaft für Linguistische Datenverarbeitung (GLDV 2007)*, Tübingen, Germany.
- Christian Chiarcos and Maxim Ionov. 2019. Ligt: An llod-native vocabulary for representing interlinear glossed text as rdf. In *2nd Conference on Language, Data and Knowledge (LDK 2019)*. Schloss Dagstuhl-Leibniz-Zentrum Für Informatik.
- Christian Chiarcos and Maxim Ionov. 2021. Linking discourse marker inventories. In *3rd Conference on Language, Data and Knowledge (LDK 2021)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.
- Christian Chiarcos and Maria Sukhareva. 2015. OLiA - Ontologies of Linguistic Annotation. *Semantic Web Journal*, 518:379–386.
- Committee of Editors of Linguistics Journals. 2008/2015. Leipzig glossing rules, conventions for interlinear morpheme-by-morpheme glosses. Technical report, University of Leipzig, Germany.
- Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.
- Stefanie Dipper, Michael Götze, and Stavros Skopeteas. 2007. Information structure in cross-linguistic corpora: Annotation guidelines for phonology, morphology, syntax, semantics, and information structure. *Interdisciplinary Studies on Information Structure (ISIS), Working papers of the SFB 632*, 7.
- Scott Farrar and D Terence Langendoen. 2010. An owl-dl implementation of gold. *Linguistic Modeling of Information and Markup Languages*, pages 45–66.
- Jorge Gracia, Christian Fäth, Matthias Hartung, Max Ionov, Julia Bosque-Gil, Susana Veríssimo, Christian Chiarcos, and Matthias Orlikowski. 2020. Leveraging linguistic linked data for cross-lingual model transfer in the pharmaceutical domain. In *The Semantic Web–ISWC 2020: 19th International Semantic Web Conference, Athens, Greece, November 2–6, 2020, Proceedings, Part II 19*, pages 499–514. Springer.
- Jorge Gracia, Marta Villegas, Asuncion Gomez-Perez, and Nuria Bel. 2018. The Apertium bilingual dictionaries on the web of data. *Semantic Web*, 9(2):231–240.
- Sebastian Hellmann, Jens Lehmann, Sören Auer, and Martin Brümmer. 2013. [Integrating nlp using linked data](#). In *12th International Semantic Web Conference, 21-25 October 2013, Sydney, Australia*.
- Maxim Ionov. 2021. Apics-ligt: Towards semantic enrichment of interlinear glossed text. In *3rd Conference on Language, Data and Knowledge (LDK 2021)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

- Marc Kemps-Snijders, Menzo Windhouwer, Peter Wittenburg, and Sue Ellen Wright. 2008. ISOcat: Corraling data categories in the wild. In *Proceedings of the 2008 International Conference on Language Resource and Evaluation (LREC)*.
- William D Lewis and Fei Xia. 2010. Developing odin: A multilingual repository of annotated language data for hundreds of the world’s languages. *Literary and Linguistic Computing*, 25(3):303–319.
- Arya D McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J Mielke, Garrett Nicolai, Miikka Silfverberg, et al. 2020. Unimorph 3.0: Universal morphology. In *Proceedings of The 12th language resources and evaluation conference*, pages 3922–3931. European Language Resources Association.
- John P McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The ontolex-lemon model: development and applications. In *Proceedings of eLex 2017 conference*, pages 19–21.
- Meg Miller and Natalie Vielfaure. 2022. Openrefine: An approachable open tool to clean research data. *Bulletin-Association of Canadian Map Libraries and Archives (ACMLA)*, 170.
- Sebastian Nordhoff. 2020. Modelling and annotating interlinear glossed text from 280 different endangered languages as linked data with ligt. In *Proceedings of the 14th Linguistic Annotation Workshop*, pages 93–104.
- Sebastian Nordhoff and Thomas Krämer. 2022. Imtvault: Extracting and enriching low-resource language interlinear glossed text from grammatical descriptions and typological survey articles. In *Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference*, pages 17–25.
- Armando Stellato, Manuel Fiorelli, Andrea Turbati, Tiziano Lorenzetti, Willem Van Gemert, Denis Dechandon, Christine Laaboudi-Spoiden, Anikó Gerencsér, Anne Waniart, Eugeniu Costetchi, et al. 2020. Vocbench 3: A collaborative semantic web editor for ontologies, thesauri and lexicons. *Semantic Web*, 11(5):855–881.
- Maria Sukhareva and Christian Chiarcos. 2016. Combining ontologies and neural networks for analyzing historical language varieties. A case study in Middle Low German. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC-2016)*, pages 1471–1480.
- Paul R Turner. 1967. Highland chontal phrase syntagmemes. *International Journal of American Linguistics*, 33(4):282–286.
- Zhong Zhou, Lori S. Levin, David R. Mortensen, and Alexander H. Waibel. 2019. Using interlinear glosses as pivot in low-resource multilingual machine translation. *arXiv: Computation and Language*.

Appendix: Illustrative Sample Data

Others [\[edit\]](#)

Symbol	Gloss	Notes
abbr	Abbreviation (e.g. <i>etc.</i> , <i>Mr.</i>)	Acronyms are also included (see acr)
date	Dates, years...	
email	Electronic Mail	Shorten form of Electronic Mail
file	Filenames	
mon	Money	
percent	Percentage	e.g. 25%, 0.9%
time	Time	
url	Web address	
web	Links and Emails	
year	Years	
maj	Large script in which every letter is the same height	
min	small script in which every letter is the same height	

Compounds [\[edit\]](#)

Symbol	Gloss	Notes	Universal feature
cmp	Compound Noun		

Chunk tags [\[edit\]](#)

Tag	Description
<SN>	Noun phrase / noun group (<i>sintagma nominal</i>)
<SA>	Adjective phrase / adjective group
<SV>	Verb phrase / verb group (<i>sintagma verbal</i>)

XML tags [\[edit\]](#)

Note: All XML tags are explained in depth in the PDF [documentation](#), see also the [dix.dtd](#) and [dix.rng](#) files in the GitHub repository.

XML tag	Means	Appears in XML tags / notes / examples
<dictionary>	Mono- or bilingual dictionary	Toplevel tag for all dictionaries
<alphabet>	Set of characters in the language	In <dictionary>
<sdefs>	Symbol definitions	In <dictionary>

Figure 4: Apertium list of symbols (excerpt).

VOL		volitive mood; volitional (cf. AVOL avolitional)	[112][117]
	VP	verbal particle	[19]
V_r	VR, v.r.	verb, reflexive (e.g. as a covert category)	[129]
	VSM	verb-stem marker	[67][23]
V_t	VT, v.t.	verb, transitive (e.g. as a covert category)	[129][15]
	WH.EX	exclamatory <i>wh</i> - clause ('what a ...!')	[citation needed]
	WH	interrogative pronoun (<i>wh</i> -word), <i>wh</i> - agreement	[56][16]
WHQ	WH.Q	<i>wh</i> - question	[16][131][20]
WIT		witnessed evidential (cf. EXP)	[38][16]
	WP, WPST	witnessed past	[80][99]
X	?	(unidentified morpheme)	[32][31]
	YNQ, PQ, P.INT, PI	yes–no question, polar question/interrogative (e.g. PC vs CQ)	[131][16][19][1]
	-Z	-(al)izer (e.g. ADJZ adjectivizer, NZ nominalizer, TRZ transitivity marker, VBZ verbalizer)	
ZO		zoic gender (animals)	[132]

Kinship [edit]

It is common to abbreviate grammatical morphemes but to translate lexical morphemes. However, kin relations commonly have no precise translation, and in such cases they are often glossed with anthropological abbreviations. Most of these are transparently derived from English; an exception is 'Z' for 'sister'. (In anthropological texts written in other languages, abbreviations from that language will typically be used, though sometimes the single-letter abbreviations of the basic terms listed below are seen.) A set of basic abbreviations is provided for nuclear kin terms (father, mother, brother, sister, husband, wife, son, daughter); additional terms may be used by some authors, but because the concept of e.g. 'aunt' or 'cousin' may be overly general or may differ between communities, sequences of basic terms are often used for greater precision. There are two competing sets of conventions, of one-letter and two-letter abbreviations.^{[133][134][47][24]}

1-Letter Gloss	2-Letter Gloss	Meaning	Equivalent sequence of nuclear relations
A	Au	aunt	= MZ or FZ / MoSi or FaSi
B	Br	brother	[basic term]
C	Ch	child	= S or D / So or Da
	Cu	cousin	= MZD, MZS, MBD, MBS, FZD, FZS, FBD, FBS = MoSiDa, MoSiSo, MoBrDa, MoBrSo, FaSiDa, FaSiSo, FaBrDa, FaBrSo
D	Da	daughter	[basic term]
e, E	o, el	elder/older	(e.g. eB, eZ) ^[54]

Figure 5: Wikipedia list of glossing abbreviations (excerpt).

```

{{interlinear|lang=jig|spacing = 3| box = yes
|Nyama-baji imimikin-bili-rni-rni ardalakbi-wurru-ju
|DEM-PL old.woman-ANIM.DU-F-ERG hot-3PL-do
|'The two old women feel hot.'}}

```

Nyama-baji *imimikin-bili-rni-rni* *ardalakbi-wurru-ju*
DEM-PL old.woman-ANIM.DU-F-ERG hot-3PL-do
'The two old women feel hot.'

Figure 6: Wikipedia template Interlinear and its rendering, example from <https://en.wikipedia.org/wiki/Template:Interlinear>.

Towards ELTeC-LLOD: European Literary Text Collection Linguistic Linked Open Data

Ranka Stanković

University of Belgrade, Serbia
ranka.stankovic@rgf.bg.ac.rs

Miloš Utvić

University of Belgrade, Serbia
misko@matf.bg.ac.rs

Christian Chiarcos

University of Augsburg, Germany
christian.chiarcos@uni-a.de

Olivera Kitanović

University of Belgrade, Serbia
olivera.kitanovic@rgf.bg.ac.rs

Abstract

This paper describes a case study on the generation of Linked Data text corpora using the NLP Interchange Format (NIF). The ELTEC corpus subset, which consists of 900 novels from the period 1840-1920 for 9 European languages, served as the basis for this research. The annotated version of the novels, in the so-called TEI level-2 format, was transformed into NIF, an RDF/OWL-based format that aims to achieve interoperability between NLP tools, language resources, and annotations. In this paper, we present our approach for transformation, and the implemented pipeline, and offer the code and results for similar use cases.

1 Introduction

Linguistic data science is a specialized area within the broader field of data science. It concentrates on the structured analysis and investigation of extensive data sets, employing various techniques and methodologies to extract valuable insights.¹ A crucial aspect of this field is the development of use cases that facilitate the integration of different language data types into a standardized ecosystem. This process utilizes tools and open standards established by the W3C to enable intelligent access, integration, and distribution of language data that caters to various user requirements. (Bosque-Gil et al., 2021)

Here, we illustrate the application of this approach to a subset of the ELTEC corpus (Burnard et al., 2021; Schöch et al., 2021; Stanković et al., 2022), which consists of 900 novels from the period 1840-1920 for 9 European languages. While working on the development of the ELTeC text collection, which includes numerous novels in many under-resourced languages, the concept of transforming the collection into linked data and adding it

to the Linguistic Linked Open Data (LLOD) cloud was conceived. This would have the advantage of enhancing the exposure of under-resourced language data by linking it with other language resources already present in the LLOD cloud, thereby increasing its visibility.

The ELTeC core collection² has 12 corpora of 100 novels comparable in their internal structure. The ELTeC plus corpora take the total number of available full-text novels to 338 and ELTEC extension 547, with the ELTeC extensions, more than 2000 full-text novels are included in ELTeC. This research is focused on transformation and publishing a set of novels from ELTEC text collection from period 1840-1920 as open linked data according to best practice and guidelines fostered by CA18209 - European network for Web-centred linguistic data science (NexusLinguarum)³.

The ELTeC novels format was developed within the COST Action CA16204 Distant Reading for European Literary History (D-Reading) (Burnard et al., 2021) in the so-called XML/TEI level-2⁴. Given the current lack of comparable corpus data in the LLOD cloud, they represent a particularly valuable resource for LLOD, as this technology allows not only interlinking different language versions, but potentially, also integrates dictionaries of the respective languages, prosopographical networks, geographical information, and other knowledge bases. The contribution is especially important since several low-resourced languages have ELTeC sub-collections with 100 novels. An overview of part of the ELTeC collection that was used in this case study will be presented in Section 1.3.

This paper will present a data model in Section 2.2 and approach for transformation from XML/TEI (Text Encoding Initiative) into NIF⁵

¹This is the definition adopted by the Cost Action CA18209, *Nexus Linguarum - European network for Web-centred linguistic data science* (2019-2023), <https://nexuslinguarum.eu/> (Declerck et al., 2020)

²<https://www.distant-reading.net/eltec/>

³<https://nexuslinguarum.eu/>

⁴<https://distantreading.github.io/Schema/eltec-2.html>

⁵<http://bpmlod.github.io/report/nif-corpus/index.html> (Un-

(NLP Interchange Format) in Section 2.3. The description of the results of transformation in the form of RDF graphs will be discussed in Section 3.1 and the examples of SPARQL query in Section 3.2. Discussion with open issues, dilemmas, difficulties, and constraints in research will be given in Section 4, followed by current results and plans for further activities in Section 5.

1.1 Motivation

In our research, results of literary scholars and the digital humanities community developed within the Cost Action D-Reading, are brought together with technologies for web-centered linguistic data science semantic networks developed in the Cost Action NexusLinguarum, fostering interdisciplinary research in these two areas. In the digital humanities community, the XML-based standards of the Text Encoding Initiative (TEI)⁶ represent the prototypical approach to publishing electronic text and data. Yet, they have been criticized for not establishing a sufficient degree of interoperability, and their synchronization with formal semantics and web standards such as RDF and OWL have been repeatedly suggested since the 2000s (Bański, 2010; Ciotti and Tomasi, 2016). With the development of the Linguistic Linked Open Data (LLOD) cloud (Chiarcos et al., 2012; Pareja-Lora et al., 2019; Cimiano et al., 2020b), interest in formalizing this bridge has been intensified, albeit, so far, with a focus on lexical data (Bellandi, 2023).

ELTeC as a carefully selected and balanced text collection for each language, when available in LLOD could become a playground for various types of research in different scientific disciplines. The main contribution is a complex project which includes the preparation and publishing of 900 novels in LLOD. The developed procedure could be used for other ELTEC sub-collections and other XML/TEI corpora, and thus serve as a point of orientation for future publication workflows of multilingual corpus data on the web. This activity is directly related to the activities of Nexus Linguarum Working Group 1 ‘Linked-Data based language resources’ that include creation, interlinking, enrichment, and evolution of the linguistic resources, especially in the context of a designated task of the action regarding the Development of the LLOD cloud for under-resourced languages and domains.

official Draft)

⁶<https://tei-c.org/guidelines/>

Motivation for this research was found in several previous successful use cases of transformation and publication using the NLP Interchange Format (NIF) (Hellmann et al., 2013), a community standard for representing the linguistic annotations of textual data in RDF, as produced by conventional NLP tools available at the time. It has been primarily designed for NLP web services but is also applicable for linguistically annotated corpora if their annotations do not exceed a certain level of complexity. Its primary goal has been to provide interoperable web services connecting NLP services, data, and applications and to build modular, flexible workflows on that basis (Hellmann et al., 2012; Cimiano et al., 2020c). NIF supports the annotation of named entities, part-of-speech tags, dependency parses, sentiment analysis, and other types of linguistic information. By its use of string URIs, NIF also supports multilingual text resources, enabling the representation of text in multiple languages and the alignment of annotations and translations across languages by means of RDF properties.

1.2 Related Research

Examples of electronically edited *text* in TEI and linked data complementing include the recent application of the Web Annotation standard to annotate TEI editions (del Rio Riande and Vitale, 2020). While such standoff annotation with JSON-LD is appropriate for *completed* editions, digital editions that are being worked on at the time of Linked Data annotation require a representation in *inline* XML, as demonstrated, by the experimental edition of a Middle French medical treatise (Tittel et al., 2018), as well as the Diachronic Spanish Sonnet Corpus: TEI and linked open data encoding, data distribution, and metrical findings (Ruiz Fabo et al., 2021). Aside from JSON-LD standoff and XML inline annotation with RDF, a third line of research on electronically edited text as Linked Data includes the full conversion of individual texts, structured corpora, and annotations. This is what is being pursued here. Normally, this line of research is conducted on data that follows conventions in the NLP and corpus linguistics communities rather than the DH communities, and here, tabular formats or, more recently, JSON have fully superseded the XML formats of the early 2000s. Cimiano et al. (2020a) presented prototypical applications of Linguistic Linked Data in Digital Humanities technologies and LOD resources in Digital Human-

ities as well as frequently used vocabularies. We see a special contribution to our work in discussing how to establish bridges between Linked Data technologies developed for NLP and TEI data produced and consumed in digital humanities.

Hellmann et al. (2010) and Brümmer (2015) described early experiments on the application of NIF to corpus data, and Brümmer et al. (2016) introduces the DBpedia Abstract Corpus - an open, large-scale corpus of annotated Wikipedia texts in six languages. The corpus contains over 11 million texts and more than 97 million entity links. The paper discusses the characteristics of the Wikipedia texts, the process of creating the corpus, its format, and interesting use cases, such as training and evaluating Named Entity Linking. NIF (Hellmann et al., 2013) was used as the corpus format to provide DBpedia compatibility using Linked Data as well as NLP tool interoperability. NIF is featured as a format for corpus data in the Best Practice Recommendations of the W3C Community Group Best Practices for Multilingual Linked Open Data (BP-MLOD).⁷ As an illustration of the capacities of NIF, FrameNet (FN), an extensive lexical database for the English language has been published into RDF Linked Open Data (LOD) format, along with a vast corpus of text that has been annotated using FN. Alexiev and Casamayor (2016) examined the FN-LOD representation, compares it with NIF, and proposes an approach for the integration of FN into NIF that does not require any custom classes or properties.

Another widely used standard for linguistic annotations in RDF is Web Annotation (Sanderson et al., 2013) (formerly known as Open Annotation), published as a W3C standard (recommendation) in 2017⁸. Unlike NIF, however, it does not provide specific data structures for linguistic annotation, but only formalizes markables (‘annotation targets’) and information they are annotated with (‘annotation bodies’) in a reified annotation property. As Web Annotation does not provide specifically linguistic annotation, we focus on NIF-based vocabularies, here.

Yet another RDF-based corpus format is POWLA (Chiarcos, 2012), a reconstruction of the

Linguistic Annotation Framework (Ide and Suderman, 2014, LAF, ISO 24612:2012) in OWL2/DL. As a proprietary standard, however, LAF seems not to be used much in the field, so the current role of POWLA seems to be primarily that of a companion vocabulary that serves to augment shallow data models such as NIF or Web Annotation data with generic data structures for linguistic annotation (Cimiano et al., 2020d). We are not aware of any corpus or annotation projects using POWLA independently of either NIF or Web Annotation since de Araujo et al. (2017), and for the rather shallow annotations of the ELTeC data, core NIF data structures are sufficient so we decided to focus on NIF.

Other RDF-based corpus formalisms we are aware of are either limited to a specific technology or software, e.g., the NewsReader Annotation Format (Fokkens et al., 2014, NAF-RDF), or the LAPPS Interchange Format (Ide et al., 2016, LIF), or they are focusing on a particular user community and their specific needs, e.g., the compatibility with tabular (‘CoNLL’) formats as used in NLP (Chiarcos and Glaser, 2020, CoNLL-RDF) or on the representation of interlinear glossed text (IGT) as used in language documentation, language teaching, and linguistic typology (Ionov, 2021, Ligt). CoNLL-RDF is based on a reduced core vocabulary taken from NIF, but it introduces its own URI schema, based on the counting of tokens and sentences. Unlike NIF, CoNLL-RDF URIs thus do not directly refer to a document, but only to a unit of annotation. Furthermore, CoNLL-RDF is more specialized in the annotation of syntax and semantics, whose treatment in NIF requires NIF extensions, whereas here, we focus on matters well covered by NIF, morphosyntactic annotation and named entities. Nevertheless, a future direction of our research is to compare NIF and CoNLL-RDF editions of our data with respect to verbosity and scalability issues.

In a recent overview of these and related vocabularies, Cimiano et al. (2020b) described the principles for annotating text data using RDF-compliant formalism, that are providing the basis for making annotated corporate and text collections accessible from the LLOD ecosystem. Because web documents may change, to preserve interpretability, it is recommended to include the full text of the annotated document in the RDF data.

Based on our literature overview and the char-

⁷However, these have not progressed beyond the level of a draft, available under <http://bpmlod.github.io/report/nif-corpus/index.html>, cf. <https://www.w3.org/community/bpmlod/>.

⁸<https://www.w3.org/TR/annotation-model/>

acteristics of our data, we decided to follow the BPMLOD draft recommendation and apply NIF 2.0 to our data. In the light of the alternatives, this offers a number of advantages:

- NIF is widely used (about as much as Web Annotation or CoNLL-RDF, but much more than tool- or community-specific RDF vocabularies or than generic formats such as LAF/POWLA).
- NIF provides explicit, native data structures for linguistic annotation (unlike Web Annotation).
- For the current annotations of the ELTeC corpus (morphosyntax, named entities), the native NIF vocabulary is sufficient. Additional data structures that could also account for morphological segmentation (as in *Ligt*), dependency syntax, and semantic role labeling (as in CoNLL-RDF) or generic linguistic annotations (as in POWLA) are not required.
- NIF is designed for standoff annotation, i.e., it uses string URIs to point to documents provided in their native formats on the web. Web Annotation is similar in this regard, but both are different from designated data models for linguistically annotated corpora whose basic unit of analysis is not the (primary text in the) document, but units of annotations imposed over these (e.g., CoNLL-RDF, *Ligt*). As an example, NIF URIs directly resolve against an offset in the annotated document, whereas CoNLL-RDF URIs are generated from sentence ID and token number, i.e., they require pre-processed documents.

For this reason, we eventually went with the NIF vocabulary for data modeling. It is to be noted though, that NIF has a number of potential downsides, including a high degree of verbosity (in comparison to tool- or domain-specific formats as well as to tabular formats as currently used in NLP – but probably less than or comparable to traditional XML-based formats such as LAF), so that one of the research questions we aim to contribute to is the discussion of scalability issues for such kind of data. Also, we would like to contribute to an effort of comparing and harmonizing data models for linguistic annotations on the web that has been initiated in 2020 in the context of the W3C Community Group Linked Data for Language Technology

(LD4LT).⁹ To the best of our knowledge, progress in this working group is slow. On the one hand, this can be attributed to external factors such as the involvement of many contributors in the development of a lexical companion vocabulary for corpus data, *OntoLex-FrAC* (Chiarcos et al., 2022a), which is in the process of finalization and which is expected to provide important stimuli for the discussion of annotations in LD4LT. On the other hand – and probably, more importantly –, the LLOD cloud diagram¹⁰ currently suffers from a lack of corpus data, to begin with, so only limited data is available that can serve as a basis for comparison and benchmarking to evaluate or demonstrate the potential of LLOD technologies for corpus data. With the data set produced as a result of our efforts, such a dataset becomes available for the first time. As this is a relatively large-scale, annotated parallel corpus, it allows to both explore the potential of RDF technology for cross-lingual linking, as well as for the linking of corpora with annotations or, prospectively, lexical resources – for which the application of LLOD technologies is by far more established, and for which tremendous amounts of data are available (Gracia et al., 2018).

The field of literature and the Semantic Web encompasses various research areas and applications where semantic technologies are applied to enhance the understanding, analysis, and organization of literary works. While the intersection of literature and the Semantic Web is relatively new, several notable works have explored this interdisciplinary domain. These works represent a fraction of the research carried out at the intersection of literature and the Semantic Web. The field continues to evolve, and ongoing studies explore novel ways to leverage semantic technologies for improved understanding, analysis, and accessibility of literary works.

The specific research questions that can be explored when transforming TEI literary corpus into a linked NIF corpus: RQ1) What are the challenges and potential improvements for named entities to be recognized and linked to external resources in the NIF corpus? RQ2) How annotations, such as part-of-speech tags and lemma, should be represented for the literary works in the linked NIF corpus? RQ3) How effectively does the linking of enti-

⁹https://www.w3.org/community/ld4lt/wiki/LD4LT_Annotaton_Workshop_Zaragoza_2021.

¹⁰<http://linguistic-lod.org/>.

ties in the NIF corpus contribute to the enrichment and integration of the literary works with other linked data sources, such as DBpedia, Wikidata, or other semantic web datasets?

1.3 ELTeC collection

ELTeC is a multilingual collection of roughly comparable corpora each containing 100 novels from a given national (or rather: language-based) literary tradition (Schöch et al., 2021). The multiple encoding levels are defined in the ELTeC scheme: at level zero, only the bare minimum of markup defined above is permitted, while at level 1 a slightly richer (though still minimalist) encoding is defined. At level 2, additional tags are introduced to support linguistic processing of various kinds, as discussed further below. (Burnard et al., 2021).

The current version comprises 10 languages: German (deu), English (eng), French (fra), Hungarian (hun), Polish (pol), Portuguese (por), Romanian (rom), Slovenian (slv), Spanish (spa), Serbian (srp), with level-2 annotations for 100 novels per language. Further in the paper ISO 639-2:1998 Codes for the representation of names of languages — Part 2: Alpha-3 code¹¹ will be used.

The obligatory annotations for ELTeC TEI level-2 are POS tags and lemma, but some of them have also NER (named entity recognition) layer and some of them have detailed grammatical descriptions for tokens. All annotated novels are publicly available and published as XML/TEI files under CC-BY license. Input data collection with novels in XML/TEI level-2 is available in the following repositories: <https://github.com/COST-ELTeC/ELTeC-Ing/tree/master/level2> where "Ing" is substituted with 3-letter code for language.

All language sub-collections are annotated with Universal Dependencies POS tag set and lemmatized. All, except French, have sentence boundaries marked with <s> XML element. NER tag sets do not have the same number of categories for different languages: most frequently used are PERS (person), ORG (organization), and LOC (location), but few also have DEMO (demonym, name of kinds of people: national, regional, political e.g. Frenchwoman, German, Parisians), ROLE (names of the profession, but also titles, nobility, office, military), WORK (titles of books, songs, plays, newspaper, paintings, sculptures, and other creations), EVENT (important events e.g. Christmas, Victory Day).

Some text collections (srp, slv, por) have unique IDs for paragraphs, sentences, and tokens, while others are without identifiers.

Metadata from 700 novels, named WikiELTeC is available in Wikidata. WikiELTeC was semi-automatically populated from TeiHeader using OpenRefine, QuickStatents, and custom-made procedures (Ikonić Nešić et al., 2022). Each item for a novel is connected with an appropriate item that is an instance of electronic edition (Q59466853), first edition (Q10898227), print edition (Q59466300), and digital edition (Q1224889) using property (P747) (has edition or translation), and every item of edition must be connected with a corresponding item for a novel with inverse property (P629) (edition or translation of). The list of all properties used for novels in Wikidata is documented in WikiProject_ELTeC¹².

2 Methods

2.1 Standards for linguistic annotation

There are two prominent RDF standards for linguistic annotation: NLP Interchange Format (NIF) and Web Annotation. Both standards use URIs (or IRIs) for addressing corpora, which coincides with the use of URIs in other formats such as TEI and XML standoff formats. However, these standards are relatively technical and not particularly user-friendly, and there is a need for clearer documentation that provides guidelines (GL's) and best practices (BP's) for implementation. Apart from NIF standards, two resources were used: 'Best Practices for Multilingual Linked Open Data' (BPMLOD) W3C community group, and the output of the LIDER project¹³

NIF is a community standard developed in a series of research projects at the AKSW Leipzig, Germany, and still maintained by that group. A typical UR/IRI consists of two main components, a base name that serves to locate the document, and an optional fragment identifier. For numerous media types and different file formats, different fragment identifiers have been defined, often as best practices (BPs; also referred to as Requests for Comments, RFCs) of the Internet Engineering Task Force (IETF).

Khan et al. (2022a) report that this is one area where there is a real necessity for documentation that provides clear GL's and BP's. The presented research could be a showcase for the use of NIF

¹¹<https://www.iso.org/standard/4767.html>

¹²https://www.wikidata.org/wiki/Wikidata:WikiProject_ELTeC

¹³<https://lider-project.eu>

and the transformation of TEI-compliant corpora to NIF. This paper contributes to this effort by providing a case study on NIF as an RDF-based format for describing strings in the novel, relying on the classes and properties that are formally defined within the NIF Core Ontology 2.0¹⁴. The reason not to use the latest version 2.1 of NIF Ontology is the lack of full documentation, but some features introduced in 2.1 version will be discussed.

2.2 ELTEC-NIF data model

An overview of the linguistic annotation of corpora by NLP tools in a way that integrates Semantic Web standards and technologies is given in (Khan et al., 2022b), focusing on NIF and Web annotation. For this case study we selected the NLP Interchange Format (NIF), designed to facilitate the integration of NLP tools in knowledge extraction pipelines, as part of the building of a Semantic Web toolchain and a technology stack for language technology on the web. NIF provides support for part-of-speech tagging, lemmatization, and entity annotation, enabling ELTeC level-2 layers transformation.

The first version of ELTeC novels excerpts in NIF format is produced using the INCEPTION tool (Klie et al., 2018). TTL files are available in JeRTeh (Serbian Society for Language Resources and Technologies) web portal¹⁵. Several changes were introduced, mostly related to named entities and metadata linking. Selected metadata from WikiELTeC (Ikonić Nešić et al., 2022) is linked with novel content triples. Figure 1 presents an outline of the model for ELTeC-NIF.

For named entities, several ontologies were consulted. From OLIA¹⁶ were user equivalents: `olia:Person`, `olia:Space`, `olia:Organization`, `olia:Event`. To link with DBpedia, `dbo`¹⁷ namespace is introduced, and for Wikidata `wd`¹⁸. To link the type of recognized named entities are used following classes: `dbo:Person = wd:Q5`, `dbo:Place = wd:Q7884789`, `dbo:Organisation = wd:Q43229`, `dbo:Event = wd:Q1656682`, `dbo:Profession = wd:Q28640`, `DEMO = dbo:demonym = wd:Q217438`, `dbo:Work = wd:Q386724`. The recognized named entities

are not linked with Wikidata or DBpedia items, they are just marked and classified in one of seven predefined types.

The presented research connects the previous results from the fields of Digital Humanities (Burnard et al., 2021; Schöch et al., 2021; Ikonić Nešić et al., 2022; Krstev, 2021; Stanković et al., 2022) and Linked Data (Hellmann et al., 2012; Brümmer, 2015; Alexiev and Casamayor, 2016; Cimiano et al., 2020c) which are traditionally considered separate areas of research. TEI is a widely used standard for encoding and representing textual data, while Linked Data focuses on interlinking and integrating diverse datasets. By bridging these two areas, the paper contributes to the integration of TEI-encoded literary resources with the broader Linked Data ecosystem.

2.3 Transformation procedure

A collab notebook was prepared for the transformation of XML/TEI into NIF. For Wikidata management *mkwikidata*¹⁹ library was used for working with RDF *rdflib*. The code is available as a Python notebook in the GitHub repository TEI2NIF²⁰. Code comprises classes: *Novel*, *Sentence*, *Token*, *NamedEntity* for appropriate transformation and set of additional functions.

For each novel in selected language in the set: $Lngs = \{deu, eng, fra, hun, pol, por, rom, slv, spa, srp\}$ the graph is created. Main function *write_gnovel* instantiate Graph with the following namespaces: *itsrdf*, *nif*, *olia*, *dc*, *dct*, *ms*, *wd*, *wdt*, *dbo*, *eltec*. After the instantiation of *Novel*, initial triples for the novel are added.

The parsing through selected XML/TEI level-2 version of the novel comprises several parts for generating triples: 1) novels metadata 2) sentences 3) named entities, and 4) words/tokens.

3 Results

3.1 NIF Terse RDF Triple Language (ttl)

From ELTeC level-2 described in Section 1.3, 900 novels from 9 language sub-collections with 100 *ttl* files were published. The number of sentences is limited to 1000 per novel in this edition. For the Serbian additional option, the dataset was prepared without a sentence limit.

¹⁹<https://pypi.org/project/mkwikidata/>

²⁰<https://github.com/rankastankovic/TEI2NIF>

¹⁴<https://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core/nif-core.html>

¹⁵<http://lloj.jerteh.rs/ELTEC/srp/NIF-INCEPTION/>

¹⁶http://purl.org/olia/discourse/olia_discourse.owl

¹⁷<https://dbpedia.org/ontology/>

¹⁸<https://www.wikidata.org/wiki/>

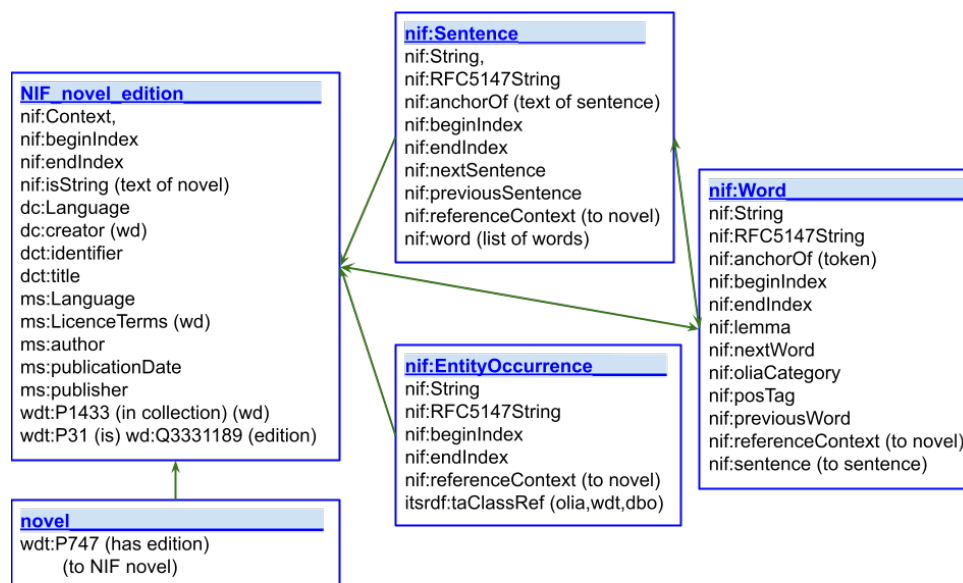


Figure 1: Data model for novel in ELTeC corpus .

Uncompressed files are accessible at: <http://l1od.jerteh.rs/ELTEC/lng/NIF/>, where $lng \in Lngs$, with Creative Commons Attribution 4.0 International license. Zipped files are available also: <http://l1od.jerteh.rs/ELTEC/lng/NIF-lng-1000.zip>, where $lng \in Lngs$ and they will be available on European Language Grid portal and other language repositories.

The core classes `nif:String` is used for the novel's content itself, described by `nif:beginIndex` and `nif:endIndex`. Dublin Core vocabulary is used for predicates related to the language, author, identifier, and title. The author is linked with Wikidata items for example Emili Bronthe is represented by (`wd:Q80137`). The novel "Wuthering Heights" (`wd:Q202975`) with the property *has edition or translation* (`wdt:P747`) is linked to digital version of the novel (`eltec:ENG18471.txt`) used as a source for NIF version. Further on, novel `eltec:ENG18471.txt` is linked by property *is published in* (`wdt:P1433`) with "engELTeC: English Literary Text Collection (ELTeC)" (`wd:Q111271624`). META-SHARE ontology²¹ is used to describe language, licence terms, author, publisher, and publication year:

```
wd:Q202975 wdt:P747 eltec:ENG18471.txt.
eltec:ENG18471.txt a nif:Context,
    nif:String, nif:RFC5147String;
    nif:beginIndex "0";
```

²¹<http://w3id.org/meta-share/meta-share/2.0.0>

```
nif:endIndex "98583";
nif:isString "Wuthering Heights A novel
, By Ellis Bell , ... and Mr. Hindley
will have to proceed to extremities
, see if he wont .";
dc:Language "en";
dc:creator wd:Q80137 ;
dct:identifier "ENG18471";
dct:title "Wuthering Heights :
ELTeC edition"^^xsd:string ;
ms:Language "en"^^xsd:string ;
ms:LicenceTerms wd:Q20007257 ;
ms:author "Bronte, Emily (1818-1848)";
ms:publisher "COST Action \"Distant
Reading for European Literary
History\" (CA16204)";
ms:publicationDate "1847";
wdt:P1433 wd:Q111271624;
wdt:P31 wd:Q3331189.
```

For illustration, a short sentence "This is certainly , a beautiful country !" from "Wuthering Heights" Emily Brontë (1847) is presented and illustrative parts will be discussed. Substring of the `nif:Context` can be: a single word, sentence, or named entity that is linked to the relevant Context resource via `nif:referenceContext`. Beginning and end indices refer to the string content (sentence) represented by the context. The previous and next sentences are references as well as a list of words.

```
<http://l1od.jerteh.rs/ELTEC/eng/NIF/
ENG18471.txt#char=285,327>
a nif:String, nif:RFC5147String,
nif:Context ;
nif:anchorOf "This is certainly ,
a beautiful country ! " ;
nif:beginIndex "285" ;
nif:endIndex "327" ;
```



```
nif:nextSentence
<http://llod.jerteh.rs/ELTEC/eng/NIF/
ENG18471.txt#char=328,450>;
nif:previousSentence
<http://llod.jerteh.rs/ELTEC/eng/NIF/
ENG18471.txt#char=169,284>;
nif:referenceContext eltec:ENG18471.txt ;
nif:word <http://llod.jerteh.rs/ELTEC/
eng/NIF/ENG18471.txt#char=285,289>,
<http://llod.jerteh.rs/ELTEC/eng/NIF/
ENG18471.txt#char=290,292>,
...
<http://llod.jerteh.rs/ELTEC/eng/NIF/
ENG18471.txt#char=325,326> .
```

Following listing presents triplets for tokens (words). The property `nif:anchorOf` is used to explicate the annotated string. Apart from indices, `nif:lemma` and `nif:posTag` are included, `nif:previousWord` and `nif:nextWord`, `nif:sentence` and `nif:referenceContext`.

```
<http://llod.jerteh.rs/ELTEC/eng/NIF/
ENG18471.txt#char=307,316> a
nif:String, nif:RFC5147String, nif:Word;
nif:anchorOf "beautiful" ;
nif:beginIndex "307" ;
nif:endIndex "316" ;
nif:lemma "beautiful" ;
nif:nextWord
<http://llod.jerteh.rs/ELTEC/eng/NIF/
ENG18471.txt#char=317,324>;
nif:posTag "ADJ" ;
nif:oliaCategory olia:Adjective ;
nif:previousWord
<http://llod.jerteh.rs/ELTEC/eng/NIF/
ENG18471.txt#char=305,306>;
nif:referenceContext eltec:ENG18471.txt;
nif:sentence
<http://llod.jerteh.rs/ELTEC/eng/NIF/
ENG18471.txt#char=285,327>.
```

Since the English corpus has not NER layer annotated, example is taken from the Portuguese corpus. One can see that `itsrdf:taClassRef` is used to link to the appropriate type on NER, in this case for person: `olia:Person`, `wdt:Q5`, `dbo:Person`:

```
<http://llod.jerteh.rs/ELTEC/por/NIF/
POR0100.txt#char=78337,78365> a
nif:Phrase, nif:String,
nif:RFC5147String ; nif:anchorOf
"D. Diogo Furtado de Mendonça";
nif:beginIndex "78337";
nif:endIndex "78365";
nif:referenceContext eltec:POR0100.txt;
itsrdf:taClassRef
olia:Person, wdt:Q5, dbo:Person .
```

Total size of the repository for all nine languages is 12.87 GB, which includes 900 txt files, 900 ttl and 900 zip files. Table 1 gives an overview per language. The calculation in Fuseki database is calculated for the Serbian corpus. The database has

Language	zip (MB)	txt+ttl (GB)
deu	118	1.6
eng	106	1.42
hun	103	1.4
pol	90	1.12
por	96	1.23
rom	78	1.01
slv	100	1.32
spa	124	1.64
srp	95	1.22

Table 1: Size of corpus file repositories.

20.7 GB (17 times more than the files in the repository). There are 21,416,099 triples, 99012 sentences, 1,731,440 words, 32625 persons (`wd:Q5`), 5937 places (`wd:Q7884789`) etc.

3.2 SPARQL Endpoint

Apache Jena Fuseki²² is used for uploading and testing Serbian ELTeC corpus (Krstev, 2021; Stanković et al., 2022) transformed to NIF, as a SPARQL server web application at JeRTeh site²³. Fuseki provides the SPARQL 1.1 protocols for query and update as well as the SPARQL Graph Store protocol. It is integrated with TDB (component of Jena for RDF storage) to provide a robust, transactional persistent storage layer, and incorporates Jena text query.

Six most frequent nouns in a novels of writer Jakov Ignjatović (`wd:Q570913`): `kuća` (house) 275, `otac` (father) 208, `dan` (day) 144, `mati` (mother) 140, `godina` (year) 127, `ruka` (hand) 123 can be found with following SPARQL query:

```
SELECT ?lemma (COUNT(?lemma) AS ?count)
WHERE {
  ?subject nif:lemma ?lemma ;
    nif:posTag "NOUN"^^xsd:string;
    nif:referenceContext ?novelid.
  # Jakov Ignjatović
  ?novelid dc:creator wd:Q570913.
}
GROUP BY ?lemma
ORDER BY desc(?count)
```

List of recognised named entities linked with entity types in Wikidata can be retrieved with following query :

```
SELECT ?subject ?nentity ?etype
WHERE {
  ?subject nif:anchorOf ?nentity ;
    itsrdf:taClassRef ?etype.
  FILTER (isURI(?etype) &&
    contains(str(?etype), ("wiki") ) )
}
```

²²<https://jena.apache.org/documentation/fuseki2/>

²³<http://fuseki.jerteh.rs/#/dataset/SrpELTeC/query>

The total numbers of recognised named entities grouped by type from Wikidata: person (Q5) 32625, name for a geographical entity or location (Q7884789) 5937, role - profession (Q28640) 24287, demonyms - name for a resident of a locality (Q217438) 5387, organization (Q43229) 451, events (Q1656682) 267, individual intellectual or artistic work (Q386724) 129, are retrieved with following query:

```
SELECT ?etype (COUNT(?etype) AS ?count)
WHERE {
  ?subject nif:anchorOf ?nentity ;
    itsrdf:taClassRef ?etype.
  FILTER (isURI(?etype) &&
    contains(str(?etype), ("wiki") ) )
}
group by ?etype
```

4 Discussion

The primary issue at hand concerned which version of NIF to use - 2.0 or 2.1. Although version 2.1 offered some additional features that could have been advantageous for our case study, such as detecting information and subsequently linking entities, we opted for version 2.0. This was because, to the best of our knowledge, version 2.1 was only a release candidate and lacked comprehensive documentation. The service introduced two NIF substring resources that had the potential to be named entities. Each of these substring resources contained multiple pieces of annotation information:

- Indicating that a particular substring had been identified as a probable reference to a named entity. In NIF 2.1, this was achieved by assigning the `nif:EntityOccurrence` class to the substring resource.
- Providing potential references to Linked Data identifiers for the mentioned named entities, as well as classifying or referencing the entities into one or more categories. To reference these entities, we used the `itsrdf:taIdentRef` property from IT-SRDF.

The dilemma related to NER was also mapping of NER types to appropriate ontology and choosing the best-fitting ontology class. We already mentioned that tagsets for NER classes are not the same for all languages and each language used specific tools and models. The general suggestion was to use 7 classes, that are mapped in our approach but some were used less and some more. For example, the Polish corpus is annotated with a very detailed tagset

including *MISC*, *nam_adj_country*, *nam_fac_road*, *nam_fac_square*, *nam_liv_god*, *nam_liv_person*, *nam_loc_country_region*, *nam_loc_gpe_city*, *nam_loc_gpe_country*, *nam_loc_nam_org_nation*, *nam_org_organization*, *nam_pro_media_periodic*, *nam_pro_title*,...²⁴ In order to keep those detailed information, this is encoded as:

```
<...POL0004.txt#char=17646,17662>
  a nif:RFC5147String ;
  nif:anchorOf "Marya błogosławi"^^xsd:string ;
  nif:beginIndex "17646"^^xsd:nonNegativeInteger ;
  nif:endIndex "17662"^^xsd:nonNegativeInteger ;
  nif:referenceContext eltec:POL0004.txt ;
  itsrdf:taClassRef "<nam_liv_person>"^^xsd:string .
```

For syntactic quality we are using custom Python scripts and SPARQL queries, while RDFUnit tool (Kontokostas et al., 2016) is used as an RDF Unit-Testing suite for semantic quality to validate the RDF data against the NIF Ontology.

Named entities annotated with the proposed dataset with seven categories are properly linked, but some collections, like Polish, have different NER tagset, which should be handled in the next version. Ongoing efforts are being made to develop a solution based on NIF corpus for entity linking with Wikidata.

The interlinking of entities in the NIF corpus offers the potential for new discoveries and valuable insights into literary works, authors, historical figures, and cultural contexts. Moreover, the linked NIF corpus holds the promise of shedding light on language variation, including dialectal differences, historical language evolution, and specific geographic or temporal language usage. This, in turn, can reveal patterns of language change, borrowings, and semantic shifts within literary works. The findings presented in the corpus can facilitate comparative analysis of literary works, genres, and authors, uncovering shared linguistic features, stylistic trends, and thematic connections.

The ELTeC-NIF corpora benefit various users and stakeholders in NLP tasks. NIF's flexibility and interoperability make it valuable for sharing and utilizing NLP data across different domains. Researchers can analyze linguistic annotations and extract features, Tool Developers can use NIF corpora for training or testing, Linguists can study language phenomena, and Semantic Web Developers can integrate NLP data with linked sources for advanced analysis and knowledge discovery.

5 Conclusion and future directions

Future plans include several activities. We would like to generate a version of our corpus adhering to

²⁴<https://github.com/CLARIN-PL/Liner2>

the CoNLL-RDF vocabulary (Chiarcos and Fäth, 2017), a direct rendering of the CoNLL format in RDF, that mimicks CoNLL’s original TSV-style layout, and describe a novel extension of CoNLL-RDF, introducing a formal data model, formalized as an ontology. The transformation will rely on the ontology as a basis for linking RDF corpora with other Semantic Web resources. (Chiarcos et al., 2021) Since CoNLL-RDF is easy to read, easy to parse, close to conventional representations and facilitates LLOD integration by applying off-the-shelf Semantic Web technology to CoNLL corpora and annotations, we would like to compare it with NIF. As it doesn’t use string URIs directly, CoNLL-RDF is probably less suitable for philological corpora than NIF or Web Annotation – these can directly be used to provide standoff annotations over a digitally edited text on the web, regardless of its format. At the same time, however, it is less verbose than NIF, but limited to a minimal core vocabulary from NIF, so it is possible that it has advantages in speed and scalability. Yet, with the limited amount of data published in both formats currently available, this suspicion cannot be directly evaluated, and such an evaluation would be a prospective goal of our efforts.

Next steps will be integration into the Linguistic Linked Open Data (LLOD) Cloud²⁵, coordinated effort of the the Open Linguistics Working Group (OWLG), its members and collaborating initiatives. The LLOD cloud is visualized by means of a cloud diagram that displays all the resources with their relative sizes and their connections. (Cimiano et al., 2020b) Finally, due to the available resources, the current version has limited the number of sentences to 1000, but the final version will be produced from the whole novels. Moreover, set of additional novels in extended edition and some novels for languages that do not have level-2 but have level-1 could be playground for testing web services for POS-tagging, morphosyntactic annotation, and named entities recognition and linking.

We also hope that soon an appropriate SPARQL endpoint with with the adequate capacity will become available, so that this valuable resource can be used in linguistic community working with linked data. Publishing RDF data on the web in a sustainable way has previously been proven challenging, and again, we would like to evaluate different approaches and the adequacy of existing host-

ing solutions for larger-scale data such as linguistic corpora. Also, in the context of European infrastructure initiatives for NLP services, the role of linked data remains somewhat underexplored,²⁶ and we expect our upcoming experiences in developing such a solution – both on a technological and a political level – to be of particular value for future initiatives on corpus data in RDF.

Also, last, but not least, publishing data is only the very first step in the process. The development of tools that allow their users to benefit from the advantages promised by the application of Linked Data technology to language resources (findability, federation, interoperability, ease of information integration, queriability) will be decisive for the future of LLOD technology. For lexical data, some of these effects can already be seen, as tools for lexicographers to become available, both with respect to automated support for lexicography (Gracia et al., 2021) and with respect to end-user tools for creating and maintaining dictionaries (Fiorelli et al., 2020). Although initial applications have been proposed for annotation engineering (Chiarcos et al., 2022b) and corpus querying (Ionov et al., 2020), the general progress on corpus data may be hampered by the limited amount of data previously available, as well as by the diversity of vocabularies applied for their publication.

Acknowledgements

This paper is based upon work from COST Action NexusLinguarum – “European network for Webcentered linguistic data science” (CA18209), supported by COST www.cost.eu, through Virtual mobility grant and other activities.

References

- Vladimir Alexiev and Gerard Casamayor. 2016. Fn goes nif: integrating framenet in the nlp interchange format. In *Proceedings of the LDL 5th Workshop on Linked Data in Linguistics: Managing, Building and Using Linked Language Resources*, pages 1–10.
- Piotr Bański. 2010. Why tei stand-off annotation doesn’t quite work. In *Balisage: The markup conference*.

²⁶LOD technology is usually seen as a key to achieve sustainable management of scientific data, and it has thus been integrated into the technology stack of initiatives such as SSHOC (Dumouchel et al., 2020). The wide usage of RDF and Linked Data for language resources substantially pre-dates the FAIR principles (Farrar and Lewis, 2007; Chiarcos et al., 2011), but has gained a lot of traction in the course of this development, more in Khan et al. (2022b).

²⁵<http://linguistic-lod.org/llod-cloud>

- Andrea Bellandi. 2023. Building linked lexicography applications with lexo-server. *Digital Scholarship in the Humanities*.
- Julia Bosque-Gil, Verginica Barbu Mititelu, Hugo Gonçalo Oliveira, et al. 2021. **Balancing the digital presence of languages in and for technological development. A Policy Brief on the Inclusion of Data of Under-resourced Languages into the Linked Data Cloud.**
- Emily Brontë. 1847. *Wuthering Heights: A novel by Ellis Bell*. London: T. C. Newby, London.
- Martin Brümmer, Milan Dojchinovski, and Sebastian Hellmann. 2016. Dbpedia abstracts: a large-scale, open, multilingual nlp training corpus. In *Proceedings of the Tenth International Conference on LREC'16*, pages 3339–3343.
- Martin Brümmer. 2015. Expanding the nif ecosystem. corpus conversion, parsing and processing using the nlp interchange format 2.0.
- Lou Burnard, Christof Schöch, and Carolin Odebrecht. 2021. In search of comity: Tei for distant reading. *Journal of the Text Encoding Initiative*, (14).
- Christian Chiarcos. 2012. Powla: Modeling linguistic corpora in owl/dl. In *The Semantic Web: Research and Applications: 9th Extended Semantic Web Conference, ESWC 2012, Heraklion, Crete, Greece, 2012. Proceedings 9*, pages 225–239. Springer.
- Christian Chiarcos, Elena-Simona Apostol, Besim Kabashi, and Ciprian-Octavian Truică. 2022a. Modelling frequency, attestation, and corpus-based information with ontolox-frac. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4018–4027.
- Christian Chiarcos and Christian Fäth. 2017. Conll-rdf: Linked corpora done in an nlp-friendly way. In *Language, Data, and Knowledge: First International Conference, LDK 2017, Galway, Ireland, June 19-20, 2017, Proceedings 1*, pages 74–88. Springer.
- Christian Chiarcos, Christian Fäth, and Maxim Ionov. 2022b. Querying a dozen corpora and a thousand years with fintan. In *Proceedings of the 13th LREC*, pages 4011–4021.
- Christian Chiarcos and Luis Glaser. 2020. A tree extension for conll-rdf. In *Proceedings of the 12th LREC*, pages 7161–7169.
- Christian Chiarcos, Sebastian Hellmann, and Sebastian Nordhoff. 2011. Towards a linguistic linked open data cloud: The open linguistics working group. *Trait. Autom. des Langues*, 52(3):245–275.
- Christian Chiarcos, Maxim Ionov, Luis Glaser, and Christian Fäth. 2021. An ontology for conll-rdf: Formal data structures for tsv formats in language technology. In *3rd Conference on Language, Data and Knowledge (LDK 2021)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.
- Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann. 2012. *Linked Data in Linguistics*. Springer.
- Philipp Cimiano, Christian Chiarcos, John P. McCrae, and Jorge Gracia. 2020a. *Linguistic Linked Data in Digital Humanities*, pages 229–262. Springer International Publishing, Cham.
- Philipp Cimiano, Christian Chiarcos, John P. McCrae, and Jorge Gracia. 2020b. *Linguistic Linked Open Data Cloud*, pages 29–41. Springer International Publishing, Cham.
- Philipp Cimiano, Christian Chiarcos, John P. McCrae, and Jorge Gracia. 2020c. Linked data-based nlp workflows. *Linguistic Linked Data: Representation, Generation and Applications*, pages 197–211.
- Philipp Cimiano, Christian Chiarcos, John P. McCrae, and Jorge Gracia. 2020d. Modelling linguistic annotations. In *Linguistic Linked Data: Representation, Generation and Applications*, pages 89–122. Springer.
- Fabio Ciotti and Francesca Tomasi. 2016. Formal ontologies, linked data, and tei semantics. *Journal of the Text Encoding Initiative*, (9).
- Denis Andrei de Araujo, Sandro José Rigo, and Jorge Luis Victória Barbosa. 2017. Ontology-based information extraction for juridical events with case studies in brazilian legal realm. *Artificial Intelligence and Law*, 25:379–396.
- Thierry Declerck, Jorge Gracia, and John P. McCrae. 2020. Cost action “european network for web-centred linguistic data science”(nexuslinguarum). *Procesamiento del Lenguaje Natural*, 65:93–96.
- Gimena del Rio Riande and Valeria Vitale. 2020. Recogito-in-a-box: From annotation to digital edition. *Modern Languages Open*.
- Suzanne Dumouchel, Emilie Blotière, Laure Barbot, et al. 2020. Triple project: building a discovery platform to enhance collaboration. In *ITM Web of Conferences*, volume 33, page 03005. EDP Sciences.
- Scott Farrar and William D Lewis. 2007. The gold community of practice: An infrastructure for linguistic data on the web. *Language Resources and Evaluation*, 41:45–60.
- Manuel Fiorelli, Armando Stellato, Tiziano Lorenzetti, et al. 2020. Editing ontolox-lemon in vocbench 3. In *Proceedings of the 12th LREC*, pages 7194–7203.
- Antske Fokkens, Aitor Soroa, Zuhaitz Beloki, et al. 2014. Naf and gaf: Linking linguistic annotations. In *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, pages 9–16.
- Jorge Gracia, Ilan Kernerman, and Besim Kabashi. 2021. Results of the translation inference across dictionaries 2021 shared task. In *CEUR workshop proc.*, ART-2021-131934.

- Jorge Gracia, Marta Villegas, Asuncion Gomez-Perez, and Nuria Bel. 2018. The apertium bilingual dictionaries on the web of data. *Semantic Web*, 9(2):231–240.
- Sebastian Hellmann, Jens Lehmann, Sören Auer, and Martin Brümmer. 2013. Integrating nlp using linked data. In *The Semantic Web–ISWC 2013: 12th International Semantic Web Conference, Sydney, NSW, Australia, 2013, Proceedings, Part II 12*, pages 98–113. Springer.
- Sebastian Hellmann, Jens Lehmann, Sören Auer, and Marcus Nitzschke. 2012. Nif combinator: Combining nlp tool output. In *Knowledge Engineering and Knowledge Management: 18th International Conference, EKAW 2012, Galway City, Ireland, 2012. Proceedings 18*, pages 446–449. Springer.
- Sebastian Hellmann, Jörg Unbehauen, Christian Chiarcos, and Ngonga Ngomo. 2010. The tiger corpus navigator. In *Ninth International Workshop on Treebanks and Linguistic Theories*, volume 91.
- Nancy Ide and Keith Suderman. 2014. The linguistic annotation framework: a standard for annotation interchange and merging. *Language Resources and Evaluation*, 48:395–418.
- Nancy Ide, Keith Suderman, Marc Verhagen, and James Pustejovsky. 2016. The language application grid web service exchange vocabulary. In *Worldwide Language Service Infrastructure: Second International Workshop, WLSI 2015, Kyoto, Japan, 2015.*, pages 18–32. Springer.
- Milica Ikončić Nešić, Ranka Stanković, Christof Schöch, and Mihailo Skoric. 2022. [From ELTeC text collection metadata and named entities to linked-data \(and back\)](#). In *Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th LREC*, pages 7–16, Marseille, France. ELRA.
- Maxim Ionov. 2021. Apics-ligt: Towards semantic enrichment of interlinear glossed text. In *3rd Conference on Language, Data and Knowledge (LDK 2021)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.
- Maxim Ionov, Florian Stein, Sagar Sehgal, and Christian Chiarcos. 2020. cqp4rdf: Towards a suite for rdf-based corpus linguistics. In *The Semantic Web: ESWC 2020 Satellite Events: ESWC 2020 Satellite Events, Heraklion, Crete, Greece, 2020*, pages 115–121. Springer.
- Fahad Khan, Christian Chiarcos, Thierry Declerck, et al. 2022a. [A survey of guidelines and best practices for the generation, interlinking, publication, and validation of linguistic linked data](#). In *Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th LREC*, pages 69–77, Marseille, France. ELRA.
- Fahad Khan, Christian Chiarcos, Thierry Declerck, et al. 2022b. [When linguistics meets web technologies. recent advances in modelling linguistic linked data](#). *Semantic Web*, (vol. 13, no. 6):987–1050.
- Jan-Christoph Klie, Michael Bugert, Beto Boullosa, et al. 2018. [The inception platform: Machine-assisted and knowledge-oriented interactive annotation](#). In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico.
- Dimitris Kontokostas, Christian Mader, Christian Dirschl, et al. 2016. Semantically enhanced quality assurance in the jurion business use case. In *The Semantic Web. Latest Advances and New Domains: 13th International Conference, ESWC 2016, Heraklion, Crete, Greece, 2016, Proceedings 13*, pages 661–676. Springer.
- Cvetana Krstev. 2021. [The serbian part of the eltec collection through the magnifying glass of metadata](#). *Infotheca - Journal for Digital Humanities*, 21(2):26–42.
- Antonio Pareja-Lora, Barbara Lust, Maria Blume, and Christian Chiarcos. 2019. *Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences*. The MIT Press.
- Pablo Ruiz Fabo, Helena Bermúdez Sabel, Clara Martínez Cantón, and Elena González-Blanco. 2021. The diachronic spanish sonnet corpus: Tei and linked open data encoding, data distribution, and metrical findings. *Digital Scholarship in the Humanities*, 36(Supplement_1):i68–i80.
- Robert Sanderson, Paolo Ciccarese, and Herbert Van de Sompel. 2013. Designing the w3c open annotation data model. In *Proceedings of the 5th Annual ACM Web Science Conference*, pages 366–375.
- Christof Schöch, Roxana Patras, Tomaž Erjavec, and Diana Santos. 2021. [Creating the european literary text collection \(eltec\): Challenges and perspectives](#). *Modern Languages Open*.
- Ranka Stanković, Cvetana Krstev, , Duško Vitas, et al. 2022. [Distant reading in digital humanities: Case study on the serbian part of the eltec collection](#). In *Proceedings of the LREC*, pages 3337–3345, Marseille, France. ELRA.
- Sabine Tittel, Helena Bermúdez-Sabel, and Christian Chiarcos. 2018. Using RDFa to link text and dictionary data for medieval french. In *Proceedings of the 6th Workshop on Linked Data in Linguistics (LDL-2016): Towards Linguistic Data Science. ELRA, Paris, France, Miyazaki, Japan*.

**Human-Machine
Annotation and Question
Answering in Linked Data**

Human-Machine Collaborative Annotation: A Case Study with GPT-3

Ole Magnus Holter

University of Oslo, Norway

olemholt@ifi.uio.no

Basil Ell

University of Oslo, Norway

Bielefeld University, Germany

basile@ifi.uio.no

Abstract

Within industry, it is vital to adequately communicate the qualities and features of what is to be built, and requirements are important artefacts for this purpose. Having machine-readable requirements can enhance the level of control over the requirements, allowing more efficient requirement management and communication.

Training a semantic parser typically requires a dataset with thousands of examples. However, creating such a dataset for textual requirements poses significant challenges. In this study, we investigate to what extent a large language model can assist a human annotator in creating a gold corpus for semantic parsing of textual requirements.

The language model generates a semantic parse of a textual requirement that is then corrected by a human and then added to the gold standard. Instead of incrementally fine-tuning the language model on the growing gold standard, we investigate different strategies of including examples from the growing gold standard in the prompt for the language model.

We found that selecting the requirements most semantically similar to the target sentence and ordering them with the most similar requirement first yielded the best performance on all the metrics we used. The approach resulted in 41 % fewer edits compared to creating the parses from scratch, – thus, significantly less human effort is involved in the creation of the gold standard in collaborative annotation. Our findings indicate that having more requirements in the gold standard improves the accuracy of the initial parses.

1 Introduction

Requirements describe the qualities that a physical product or a service must provide. They are an important part of industry communication, and often parts of contracts. Thus, the requirements legally bind the contractor and the supplier, and

failing to comply with them can mean both legal and economic undesired consequences.

Having the requirements expressed in a computer-understandable format would be beneficial. Manual tasks, such as requirement retrieval and documentation could be automated. In addition, it can lay the foundation for automatic compliance checking of project descriptions with the requirements. Ideally, requirements should natively be formulated in a machine-readable format, i.e., when they are created. However, the reality is that the industry must work with a large number of existing requirements, most of them embedded in complex domain-specific documents written for subject-matter experts.

To address this challenge, semantic parsing offers a promising solution by transforming natural language text into a logical representation. To create a semantic parser, however, we need training data, and manually creating logical representations is a tedious and error-prone task. Moreover, the complexity of the documents and the language of these texts makes it difficult to use techniques such as crowd-sourcing. Since it requires a considerable amount of expert hours, it is an expensive undertaking. Automatic or semi-automatic methods that help us to create training data could result in substantial savings in both cost and labour.

Recent advances in large language models (LLMs) have resulted in generic models that can solve many NLP tasks without fine-tuning them on a task-specific corpus (Liu et al., 2019; Raffel et al., 2020). While the typical LLM benchmarks do not include semantic parsing, some works demonstrate that LLMs are capable of producing accurate semantic parses (Shin et al., 2021; Roy et al., 2022).

There has, however, been little focus on using LLMs for semantic parsing in complex domains such as industry standards or requirements. Furthermore, to the best of our knowledge, no work has addressed human-in-the-loop LLM-supported

semantic parsing or LLM-supported creation of semantic parsing gold standard datasets that can then be used to train semantic parsers.

While some attention has been given to sample selection and ordering for in-context learning (as a means of few-shot learning), most studies focus on common datasets where the approaches have full access to a gold standard. To the best of our knowledge, no study has investigated sample selection for in-context learning (few-shot learning) from an iteratively growing set of possible examples of industry requirements. In our scenario, the initial set of examples is empty and is populated via human-machine collaboration.

In this paper, we investigate the possibility to use GPT-3 to reduce the effort of creating a gold standard for semantic parsing of industry requirements to description logic. To conduct the study, we compile and annotate a dataset consisting of requirement sentences, all written in English, from various industry domains. The sentences are sampled from documents by Det Norske Veritas (DNV), a global risk management and classification corporation with a focus on standards and requirements.

We hypothesize that while a semantic parser, based on a large language model, may not consistently produce logically correct formalizations, the generated formalizations are often close to the desired form. Consequently, correcting them is easier for a human than creating logical formalizations from scratch. Our focus in this study is not to create a semantic parser for a particular application, but rather to demonstrate that this method can be used to quickly create high-quality training data.

Furthermore, we investigate how sample selection and ordering affect the performance on this specific task with technical, complex input texts and description logic as output and an iteratively increasing number of available examples. We then examine the decrease in human effort between manually creating logical representations vs. correcting LLM-generated logical representations.

The remainder of the paper is structured as follows. Section 2 gives an overview of related work. Section 3 describes the problem in more detail. In Section 4, we describe the method, while in Section 5 we present the results of the experiments. The discussion and the conclusion are found in Section 6 and Section 7, respectively. In Section 8 we describe the limitations of this study and sketch ideas for future work.

2 Related work

LLM prompting The transformer model, introduced by (Vaswani et al., 2017), was followed by (Devlin et al., 2019), who pretrained a bidirectional transformer model (BERT) on a large text corpus. The BERT model, together with its many variants, has been used to solve many different tasks in NLP. It has been shown that these models already contain a vast amount of knowledge (Petroni et al., 2019; Roberts et al., 2020). While fine-tuning to a specific task has been the preferred way of using such models (Raffel et al., 2020), prompting has more recently been suggested as an alternative approach (Petroni et al., 2019) and has been used for many tasks.

While many LLM prompts are manually created, several works have investigated the automatic generation or improvement of prompts. Haviv et al. (2021) propose to automatically rewrite queries to learn how to better query an LLM, while Jiang et al. (2020) propose to mine patterns from a corpus. Sample selection and ordering in a prompt can also have a large impact on performance. It is, however, hard to predict which order is better than another as this can change from task to task and from model to model (Lu et al., 2022). Liu et al. (2022) find that choosing examples semantically similar to the target task improves GPT-3’s in-context learning performance on various tasks over a random baseline. They also observed that the ordering of the n most similar examples affects performance, but that different ordering performed best for different datasets. The impact of the ordering, however, was comparably small. Chang et al. (2021) propose to use clustering and select one element from each cluster to ensure good coverage of examples. They demonstrate that this strategy outperforms random selection. For a more detailed overview of prompting methods, strategies, and applications, see (Liu et al., 2023).

Prompt-based semantic parsing Several recent works on prompt-based semantic parsing have used constrained language models (Shin et al., 2021; Yang et al., 2022b). The models are constrained so that they will answer with a syntactically correct natural language equivalent of a semantic parse, i.e., a canonical form, that can be converted to a logical formalism by means of a synchronous context-free grammar. BenchCLAMP (Roy et al., 2022) was proposed as a benchmark specifically to evaluate se-

semantic parsing methods with constrained language models. A different approach was suggested by Rongali et al. (2022). The approach learns a mapping from natural language to a canonical form by jointly training a seq2seq model using masked prediction, denoising, and supervised semantic parsing examples using very little data.

While the constrained language model’s output to a canonical format can be considered a form of paraphrasing, another way to use an LLM as part of a semantic parsing pipeline is to use an LLM to augment real datasets or to synthesize training data for semantic parsing by paraphrasing real examples or examples generated by a grammar (Yang et al., 2022a; Rongali et al., 2022).

As an extension to manually created prompts for semantic parsing, prompt tuning was proposed by Schucher et al. (2022). In their study, a trainable embedding is prepended at all layers of the language model, which is shown to outperform a fine-tuned T5 model (Raffel et al., 2020). In addition, the authors demonstrate that the performance gap between generating a logical representation directly and using a canonical form reduces as the size of the T5 model increases.

Regarding sample selection for semantic parsing, Shin et al. (2021) propose to use GPT-3 to select the n most relevant examples for a target sentence. They do not, however, show how it compares to other sample selection methods or consider the sample ordering.

Training data generation Wang et al. (2021) suggest that instead of using LLMs to directly produce a label (few-shot) to solve a classification task, one could use a couple of examples and a label as a prompt to generate “gold” data. They achieve better results when using the generated data to fine-tune T5 (Raffel et al., 2020) than using few-shot. Using generated gold data and real gold data in combination, they achieved state-of-the-art results on the SuperGLUE tasks (Wang et al., 2019).

In the construction of the Penn treebank, the authors use simple models to create initial syntactic parses which were then manually corrected (Marcus et al., 1993).

While in this paper we use an LLM for a particular case of semantic parsing, our study differs from prompt-based semantic parsing in that we do not intend to solve the task by prompting the LLM. It is also different from training data generation by prompting LLMs in that we do not use the LLM to

generate synthetic data. It is similar to the approach taken by (Marcus et al., 1993), but we are not using heuristics or models pretrained for a particular task, but rather a generic large language model.

3 Preliminaries

3.1 Modelling of requirements

Klüwer and DNV GL (2019) proposed a logical framework for representing requirements using OWL 2 and description logic (DL) where a requirement is satisfied if and only if for every x that is a member of the class \mathcal{S} and satisfies the condition \mathcal{C} (which may be empty), x also satisfies the demand \mathcal{D} . The framework is appropriate for requirements because DL primarily deals with concepts rather than individuals. For an introduction to description logic see (Krötzsch et al., 2012). If \mathcal{S} , \mathcal{C} , and \mathcal{D} are (possibly complex) ontological class expressions, the requirement can be expressed as:

$$\mathcal{S} \sqcap \mathcal{C} \sqsubseteq \mathcal{D} \quad (1)$$

This means that a thing that is an \mathcal{S} needs to also be a \mathcal{D} if it is \mathcal{C} . E.g., if something is a “steel pipe” \mathcal{S} and it is “exposed to salt water” \mathcal{C} , it must have “corrosion protection” \mathcal{D} .

Ontological class expressions are either atomic classes, or expressions combining classes with conjunction \sqcap , disjunction \sqcup , negation \neg , or quantifiers with a property and a class expression (e.g., $\exists r.C$). We use square brackets after datatype to designate OWL 2 data ranges. E.g., $\exists \text{hasSize.float}[\geq 50]$ means that the concept has a `hasSize` relation to a float $f \in [50, \infty)$. We use expressions of the type $\exists \text{hasDescription.string}["a \text{ description}"]$ for expressions that are descriptive in nature or are either unnecessarily detailed or not expressible in DL.

The following requirement texts are taken from the document RU-Ship Pt4 Ch7 Sec 3 (Arrangements).¹ The DL statements are modelled by us.

Requirement [2.2.1] (sentence 2): *[...] the tank surfaces and bulkheads shall be insulated.*

```
TankSurface  $\sqcup$  Bulkhead
 $\sqsubseteq \exists \text{hasFeature. Insulation}$ 
```

¹All documents are copyrighted ©DNV. DNV does not take responsibility for any consequences arising from the use of this content.

Requirement [2.2.2]: *Coamings for stairs, pipe openings, etc. shall be of ample height.*

```
Coaming  $\sqcap$   $\exists$ usedFor .
  (Stair  $\sqcup$  PipeOpening)
 $\sqsubseteq$   $\exists$ hasHeight .
  (PhysicalQuantity
 $\sqcap$   $\exists$ hasDescription .
  string["of ample height"])
```

Requirement [3.1.1]: *The main inlet and main outlet pipes for thermal-oil at the fired heater and at the heater heated by exhaust gases shall have stop-valves, arranged for local manual and remote controlled operation from an easily accessible location outside the heater room.*

```
(MainInletPipe  $\sqcup$  MainOutletPipe)
 $\sqcap$   $\exists$ usedFor . ThermalOil
 $\sqcap$   $\exists$ connectedTo . (FiredHeater
 $\sqcup$  ExhaustGasHeater)
 $\sqsubseteq$   $\exists$ hasPart . (StopValve
 $\sqcap$   $\exists$ arrangedFor . ManualOperation
 $\sqcap$   $\exists$ arrangedFor . (RemoteOperation
 $\sqcap$   $\exists$ hasDescription . string["from an
  easily accessible location
  outside the heater room"])
```

3.2 Semantic parsing of requirements

To automatically find a logical representation of a sentence, we can use a semantic parser. In general, a semantic parser realizes a function $f : I \rightarrow O$ where the domain I is typically a set of utterances in natural language, such as in the form of sentences over an alphabet ($I \subseteq \Sigma^*$), and the codomain O is the set of machine-readable representations that for some utterance express a subset of its meaning that is relevant for some task. The set of representations can be a language L generated via a grammar M , i.e., $L(M)$. For example, it can be the set of expressions in first-order logic over a predefined set of predicates P and class names C .

The functionality of the semantic parser will vary depending on the type of input, the logical formalism, and the needs of the particular application. Therefore, it is necessary to create a custom semantic parser for a new application and domain. In our case, the function $f : I \rightarrow O$ represents a mapping from a set I of textual requirements to the set of meanings expressed using description logic syntax as in Equation 1. One way to create a semantic parser is by fine-tuning a neural network pretrained on language generation, using models such as BART (Lewis et al., 2020) or T5 (Raffel et al., 2020). Training a neural network this way, however, requires a large annotated dataset, which can be very expensive to obtain.

3.3 A case study with GPT-3

Given the high cost of obtaining training data for semantic parsing in technical domains, we investigate the potential benefits of incorporating a large language model, specifically GPT-3, as part of human-computer collaboration, for constructing a gold standard dataset for semantic parsing of technical requirement sentences. Specifically, we want to find out the following: *i)* To what extent can a Hybrid Human-Machine collaborative annotation with GPT-3 reduce the effort needed for developing gold examples for semantic parsing as opposed to human annotation only? *ii)* Does using semantically similar requirements as examples improve effectiveness over random selection? *iii)* Will the ordering of the semantically similar requirement examples affect the effectiveness of the approach? *iv)* How does the number of examples influence the result? *v)* If we cluster the requirements and pick the most central requirement for each cluster, thus ensuring good coverage from the start, can that improve the performance *a)* over a random baseline, or *b)* over using the semantically most similar requirements?

4 Method

4.1 Corpus creation

To create the corpus, we obtained 2225 unlabelled requirement sentences from 23 PDF documents from DNV² that were accessible online³ (see Table 3). To extract the text from the documents and create a semi-structured XML version of the PDF, we used Apache PDF box⁴ and regular expressions. We limit our work to sentences containing the modal verb “shall,” as DNV considers “shall” to be an indicator of a requirement (Det Norske Veritas, Ed. July 2022).

An annotation guideline was created and subsequently followed by the first author of the paper to produce the reference gold standard (RGS) consisting of 136 requirement sentences with a corresponding description logic formula. The second author of the paper verified the annotations to ensure the quality.

We do not make use of a predefined set of predicates or class names. However, by providing examples we implicitly specify the set of predicates and

²All documents are copyrighted ©DNV

³From <https://rules.dnv.com/> 21.9.2022

⁴v2.0.1

the set of class names, so that a model could learn which class names and predicates are preferred.

4.2 Human-machine collaborative annotation

We propose a novel method for gold standard creation using a large language model together with a human expert. The approach involves iterating over a set of unlabelled requirement sentences (R) and generate a prompt p which consists of a brief task description (see Appendix A), n examples selected using a sample selection method (m), and the target sentence s . The examples are on the form:

Input: [requirement sentence]
Output: [logical representation]

We use a large language model, specifically GPT-3, to generate an initial semantic parse (r') for the target sentence. Subsequently, a human expert reviews and corrects the model's output to ensure accuracy and consistency (r''). The gold standard (G), which is initially empty, is extended with (s, r''). This iterative process continues until all examples are annotated, resulting in a complete gold standard. The process is outlined in Algorithm 1. If n exceeds the size of the set G ($n > size(G)$), we are unable to select n samples. In such cases, we utilize all the samples in G if G is non-empty, or non at all if G is empty.

Algorithm 1 Creating a gold standard

```

procedure CREATEGOLDSTANDARD( $R, m$ )
   $G \leftarrow \emptyset$ 
  for  $s \in R$  do
     $p \leftarrow \text{createPrompt}(s, m, G)$ 
     $r' \leftarrow \text{GPT}(p)$ 
     $r'' \leftarrow \text{humanImprovement}(s, r')$ 
     $G \leftarrow G \cup \{(s, r'')\}$ 
  end for
return  $G$ 
end procedure

```

The initial task description is part of all prompts. The samples, however, may be different for each target sentence. We use three general sample selection methods from the growing gold standard. The first general sample selection method (RandomN) is to randomly select n examples for each target sentence. To investigate how the number of examples in the prompt influences the quality of GPT-3's answer, we perform four experiments using this method, where n is 5, 10, 20, and 30, respectively. Since Random20, Clustering, and the

MostSimilar requirements have the same number of examples, the Random20 can also serve as a baseline for the other sample selection methods.

The second general sample selection method (MostSimilar) is to use the n requirement sentences that are most semantically similar to the target sentence. To embed the sentences, we use the RoBERTa-large model from the sentence transformer library (Reimers and Gurevych, 2019) in Huggingface⁵. For each sentence s' in G , we calculate the cosine similarity between s' and the target sentence s . The sentences are sorted with the most semantically similar sentences first before we select the $k = 20$ most similar sentences. To investigate the impact of the order of the examples, we perform three experiments using this method, MostSimilarRandom, where the order of the n examples is randomized. MostSimilarFirst where we keep the original order of the n most similar sentences, and MostSimilarLast, where we sort the n most similar requirements from the least to the most similar.

The third general sample selection method (Clustering) is to use a fixed set of diverse requirements that ensure good coverage of topics. We used the KMeans clustering implementation in scikit-learn⁶. From each cluster k , we choose the data point that is closest to the cluster centroid. This gives us 20 sentences that, used as part of the prompt, will ensure high coverage of different types of requirements. This method will allow us to see if aiming for good coverage of different examples is better than random selection (RandomN) or selecting the most semantically similar sentences (MostSimilar). In the Clustering sample selection method, we label the sentences from the 20 clusters first.

4.3 Metrics

We estimate the effort, denoted by δ , of a human annotator to correct the logical representation with three metrics.

String Edit Distance Levenshtein Distance measures string similarity by counting the shortest edit sequence to transform one string into another. To compare DL formulas, however, we need a distance metric that considers their structure, thus we use a string edit distance metric that operates on the level of DL terms, operators, and individual string

⁵<https://huggingface.co/sentence-transformers/all-roberta-large-v1>

⁶v1.0.2

tokens counting the minimum number of insertions, substitutions, deletions, and transpositions.

For example, the difference between `Boiler` $\sqsubseteq \exists \text{hasFeature. Insulation}$ and `Compressor` $\sqsubseteq \exists \text{hasFeature. Insulation}$ is 1.

If the string edit distance exceeds the costs of turning an empty string into the reference parse, we return the edit distance of turning an empty string into the reference parse. This is reasonable because a human would discard a parse that would take more effort to correct than to create it from scratch.

Graph Edit Distance The string edit distance does not take into account that some binary operators, like conjunction and disjunction, are associative. For instance, the string edit distance between $A \sqcap B$ and $B \sqcap A$ is 2, even though the formulas are logically equivalent. To address this issue, we also use graph edit distance between the two DL formulas. Graph edit distance computes the minimum number of edits required to transform one graph g' into a graph isomorphic to another graph g .

We parse the DL formula and transform it into a graph with terms on the nodes and the edges representing the relationships between the nodes. For the axiom (\sqsubseteq), we attach numeric labels to the edges because changing the order of the edges would change the meaning of the axiom. For the unary and binary operators (conjunction and disjunction where the order of the operands is not relevant), we do not add labels for the edges. An example of the graph structure is given in Figure 1.

We use the following operations to compute graph edit distance: node insertion, node deletion, node substitution, edge insertion, edge deletion, and edge substitution (in the case the edge has a label). The cost of each operation is set to 1. Like the string edit distance, if the graph edit distance exceeds the cost of turning a graph containing only one node with the \sqsubseteq symbol into the graph of the reference parse, we return the edit distance between the graph of the reference parse and the graph containing only one node with the \sqsubseteq symbol (used to align the graphs).

Computing the graph edit distance is an NP-hard problem (Zeng et al., 2009). Therefore we use a timeout of 20 seconds and return the best result. If no result was found within the timeout, we assume the distance is high, and use the maximum distance instead.

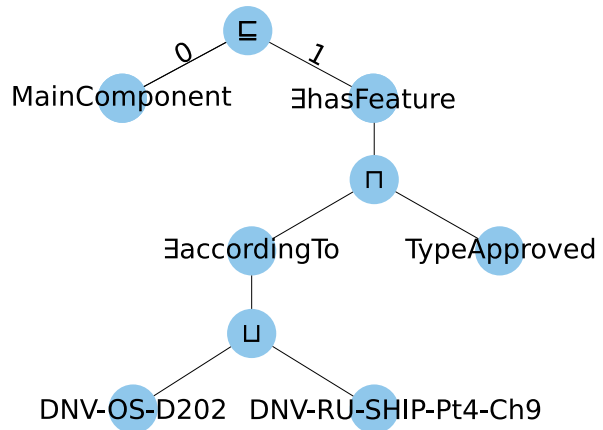


Figure 1: DL graph used for graph edit distance. Representing: `MainComponent` $\sqsubseteq \exists \text{hasFeature. (TypeApproved} \sqcap \exists \text{accordingTo. (DNV-OS-D202} \sqcup \text{DNV-RU-SHIP-Pt4-Ch9))}$.

Jaccard distance Furthermore, to say something about how similar the terms, operators, and tokens in the GPT-3 parse are to the reference gold standard, we use Jaccard similarity. Jaccard similarity is the fraction of items shared between two sets to the union of the items in the sets two sets, and Jaccard distance is the complement of Jaccard similarity. We split the parse proposed by GPT-3 and the reference parse into their individual DL tokens, operators, and string tokens, remove duplicates, and calculate the Jaccard distance between the two.

4.4 Experimental setup

Hyperparameters For all the experiments, we use the model `text-davinci-3`. The temperature was set to 0 to eliminate randomness. We request the model to only return the most probable parse. Max token was set to 256, and the newline character was used as a stop symbol.

Experiments To quantify the effort for a human annotator to create a logical representation from scratch, without receiving anything proposed by GPT-3, we use *i*) Empty, where instead of a proposal from GPT-3, we use an empty string. To investigate to what extent the prompt (p) affects the effectiveness of the approach, and to answer the questions stated in Section 3.3, we use the following methods for choosing which examples to include in p (the sample selection methods are described in detail in Section 4.2). *ii*) Random5, *iii*) Random10, *iv*) Random20, *v*) Random30, *vi*) MostSimilarRandom, *vii*) MostSimilarFirst, *viii*) MostSimilarLast. All the experiments that

involve random sampling or ordering were run five times, and we report average values and the standard deviation.

We follow Algorithm 1 for each of the sample selection methods described above and for each of the 136 requirement sentences. We start with an empty list as the gold standard G from which we choose examples. G is incrementally extended as we prompt GPT-3 with new sentences. For each of the experiments not using the Clustering method, the order of the target sentences was the same. In the experiment using the Clustering method, the 20 central concepts from the clusters were used first, and the rest of the target sentences were used in the same order.

Evaluation To estimate the difference in human effort with and without the proposals by GPT-3, we evaluate the prediction by GPT-3 against the reference gold standard (RGS) with the metrics introduced in Section 4.3. For each requirement sentence s and reference parse r , we compute the difference between r and the parse generated by GPT-3 (r') with string edit distance (δ^s), Jaccard distance (δ^j), and graph edit distance (δ^g). Δ^s , Δ^j , and Δ^g are the sum of the string edit distances, Jaccard distances, and graph edit distances, respectively. The evaluation procedure is outlined in Algorithm 2. Note that, since we are evaluating against RGS, we extend G with the reference parse r directly.

Algorithm 2 Evaluation

```

procedure EVALUATE( $RGS, m$ )
   $(\Delta^s, \Delta^j, \Delta^g) = (0, 0, 0)$ 
   $G \leftarrow \emptyset$ 
  for  $(s, r) \in RGS$  do
     $p \leftarrow \text{createPrompt}(s, m, G)$ 
     $r' \leftarrow \text{GPT}(p)$ 
     $\Delta^s \leftarrow \Delta^s + \delta^s(r, r')$ 
     $\Delta^j \leftarrow \Delta^j + \delta^j(r, r')$ 
     $\Delta^g \leftarrow \Delta^g + \delta^g(r, r')$ 
     $G \leftarrow G \cup (s, r)$ 
  end for
  return  $(\Delta^s, \Delta^j, \Delta^g)$ 
end procedure

```

5 Results

We sum all the string edit distances, graph edit distances, and Jaccard distance, and report the totals

and averages from the experiments described in Section 4.4 in Table 1.

GPT-3-assisted annotation The experiment using the empty string (Empty) gives a total string edit distance of 2,573 edits. The best-performing sample selection method uses 1,506 edits. This gives us a difference of 1067 edits. For graph edit distance, the numbers are 2,692 and 1,681, a reduction of 1011 edits. The average Jaccard distance decreases from 1 to 0.52.

The edit distance metrics depend on the size of the formula; short parses can have at most small edit distances, while long parses can have large edit distances. Therefore, to be able to observe a trend over time, we need to factor out the size of the formula. Consequently, we normalize the string edit distance by dividing the number of edits by the number of tokens in the correct parse. A normalized edit distance of 1 indicates that the entire formula needs to be changed. Although the metric shows much variation, we can observe a downward trend in string edit distance from the first to the last target sentence (see Figure 2). This trend is also visible for graph edit distance, as shown in Figure 3. Similarly, in Figure 4, we can see a comparable trend for Jaccard distance.

Sample selection methods We found that the MostSimilarFirst sample selection method obtained the shortest distance on all metrics. Specifically, it achieved a total string distance of 1,506, a total graph edit distance of 1,681, and an average Jaccard distance of 0.52. The MostSimilarLast method, however, was found to perform worse than the random ordering of the most similar examples on average.

All the experiments with the MostSimilar method yielded smaller string edit distances, graph edit distances, and Jaccard distances than the experiments with RandomN. The experiment with the Clustering method, however, obtained a better string edit distance than the experiment with the MostSimilarLast method, while MostSimilarLast performed better on the other metrics. The experiment with the Clustering method has a smaller string edit distance and graph edit distance than all the experiments with RandomN on average. The experiments with Random20 and Random30 performed better than the experiment with Clustering on Jaccard distance. The experiment with Random30 was better than all the other experiments

with RandomN on average, and the experiment with Random5 obtained the largest distance on all the metrics on average.

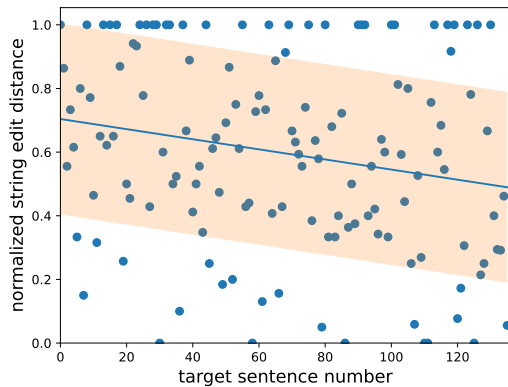


Figure 2: Normalized string edit distance from the first to the last target sentence using the MostSimilarFirst method.

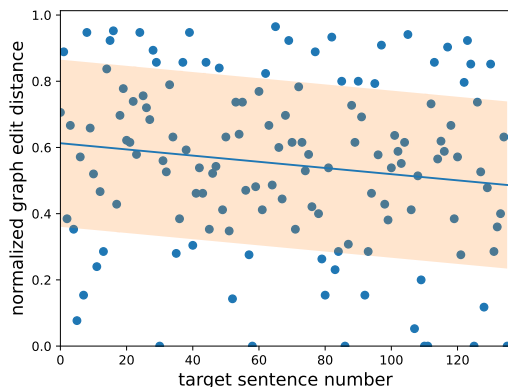


Figure 3: Normalized graph edit distance from the first to the last target sentence using the MostSimilarFirst method.

5.1 Examples of GPT-3 mistakes

First, we discuss the different types of errors we encounter. Often, multiple errors occur in one parse provided by GPT-3. Furthermore, we analyzed the frequencies of these errors on the same 20 requirements using three sample selection methods: Clustering, Random20, and MostSimilarLast. We randomly selected one of the experiments with Random20 for this analysis, and the error counts are presented in Table 2.

- i) **Wrong DL syntax** Although rare, this type of error typically affects the first one or two

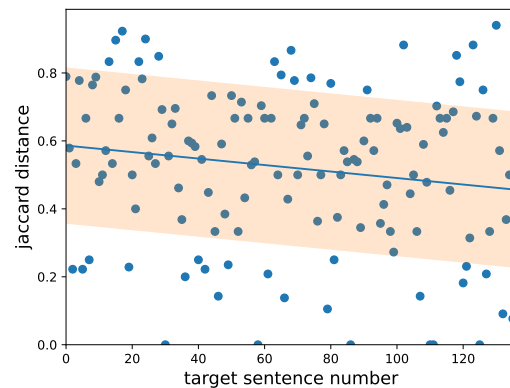


Figure 4: Jaccard distance from the first to the last target sentence using the MostSimilarFirst method.

sentences. Examples include the use of variables (e.g., $\exists c. \text{Component} \sqsubseteq \text{Type-Approved}(c)$) and multiple subclass axioms ($A \sqsubseteq B \sqsubseteq C$), neither of which is permitted in DL. We observed syntax mistakes both in Clustering and Random20.

- ii) **Different modelling choice** This type of error is not necessarily incorrect, but it affects the edit distance metrics. Different modelling choice is relatively frequent and takes many forms, such as using a concept as a property instead of a class or breaking down a requirement differently than we do, but in a plausible way. For example, we model accordance as $\exists \text{inAccordanceWith} \dots$, but we have observed instances where GPT-3 models it as $\exists \text{fitted} \dots$ ($\text{InAccordanceWith} \dots$).
- iii) **Element on the wrong side of axiom** Another common type of modelling mistake made by GPT-3 is to model a condition as a mandatory feature or create an axiom where the left side is not what the requirement is about. In these cases, the proposed axiom is often substantially different from the reference gold standard. For example, if a requirement says *There shall be a portable foam applicator in each boiler room*, modelling $\text{PortableFoamApplicator} \sqsubseteq \exists \text{hasLocation} \text{BoilerRoom}$ would be incorrect as it implies that all portable foam applicators must be located in boiler rooms.
- iv) **Too much or too little information as a string** Another type of mistake is including either too much or too little information

Method	String Distance		Graph Distance		Jaccard	
	Δ	μ	Δ	μ	Δ	μ
Empty	2,573	18.92	2,692	19.79	135.49	1.00
Clustering	1,640	12.06	1,821	13.39	80.18	0.59
Random5	1,809±47	13.30±0.34	1,909±21	14.04±0.15	84.45±1.74	0.62±0.01
Random10	1,744±42	12.82±0.31	1,903±29	13.99±0.21	80.94±1.86	0.60±0.01
Random20	1,724±17	12.68±0.13	1,843±1	13.55±0.01	79.77±1.00	0.59±0.01
Random30	1,683±27	12.38±0.20	1,848±29	13.59±0.22	77.89±1.39	0.57±0.01
MostSimilarRandom	1,569±29	11.54±0.21	1,759±25	12.93±0.18	72.64±1.47	0.53±0.01
MostSimilarFirst	1,506	11.07	1,681	12.36	70.96	0.52
MostSimilarLast	1,658	12.19	1,748	12.85	74.45	0.55

Table 1: The sum (Δ) and average (μ) values of edit distance, graph edit distance, and Jaccard distance for each of the experiments. For the RandomN experiments, we show the average of 5 runs with one standard deviation.

in a `hasDescription.String[]`. In some cases, GPT-3 may provide redundant information or use this construct for things that are easy to express in DL. In other cases, it may try to model something using DL that is not possible. We found this mistake to be most frequent in Random20, and least frequent in the experiment with the Clustering method.

- v) **Different terminology (plausible)** The use of different terminology is another common mistake, which can include synonyms, spelling differences, or using the plural instead of a single form, compared to the reference gold standard. For instance, `Fail-SafeFunctionality` instead of `FailSafeFunctionality`, `NewDesigns` instead of `NewDesign` are simple differences in spelling, and `Emergency` may be as good as `StateOfEmergency`. While similar terms could be interchangeable, they are all counted as equally different using our metrics. This type of error was found to be less frequent using the MostSimilarFirst method and the Random20 method and most frequent in the experiment with the Clustering method.
- vi) **Different terminology (not plausible)** Generating very long and complicated concepts or properties like `Within3MetersFromHazardousAreas` and `TwoIndependentAlternativesForPressurization` is another mistake GPT-3 makes. Although these names may be technically correct, they are unlikely to be found in a typical ontology. Instead of trying to break down complex concepts into more atomic ones, GPT-3 captures everything in a single concept or property.

This type of mistake was found to be most frequent in the experiment with Clustering, and least frequent with MostSimilarFirst.

- vii) **Confusing disjunction and conjunction** Another mistake GPT-3 makes is confusing conjunction and disjunction. This often occurs when using only one feature relation instead of multiple. For example, GPT-3 may model the requirement of having the two features A and B using a disjunction, as in $\exists r.(A \sqcup B)$. However, the correct representation should use a conjunction, as in $\exists r.A \sqcap \exists r.B$. This mistake was found to be most frequent in the experiment using the Clustering method.
- viii) **Missing or extra elements/clauses** Adding too much information or missing important details are also mistakes seen in GPT-3's parses. For instance, GPT-3 may add explanations and reasons behind a requirement, even though they are not needed in our framework. It may also miss some important details.

6 Discussion

GPT-3 assisted annotation The difference between creating the 136 parses from scratch and with the help of the best method using GPT-3 is 1067 edits, a reduction of the effort of about 41% in the number of string edits. For graph edit distance the reduction is 1011 edits, about 38%. This shows that the method is effectively reducing the human effort of creating the gold standard. Figures 2 and 3 indicate that the accuracy of the parses improves with more examples in the gold standard.

Considering Jaccard distance, we observe that, on average, there are differences between 52% of

Method	<i>i)</i> Wrong DL syntax	<i>ii)</i> Different modelling	<i>iii)</i> Wrong side of axiom	<i>iv)</i> Too much/little in string	<i>v)</i> Different terminology (plausible)	<i>vi)</i> Different terminology (not plausible)	<i>vii)</i> Confusing \sqcap and \sqcup	<i>viii)</i> More/less elements
MostSimilarFirst	0	6	3	7	8	2	2	7
Clustering	1	6	3	2	11	4	4	8
Random20	1	6	3	8	8	3	2	9

Table 2: Counts of different GPT-3 mistakes on the same 20 sentences. *i)* is wrong DL syntax, *ii)* different modelling choice, *iii)* element on the wrong side of axiom, *iv)* too much or too little information as a string, *v)* different terminology (plausible), *vi)* different terminology (not plausible), *vii)* Confusing disjunction and conjunction, *viii)* missing or extra clauses or elements .

the terms, symbols, and tokens. Hence, there is an overlap of 48 % between the terms in the predicted parses and the reference parses. This distance also decreases with more examples.

To create a correct formula from scratch, one needs more than just to write down the components, one has to identify good terms (the correct terms) to express this in a logical format and then structure it correctly. If we have many of the correct terms and parts of the structure, this is already helpful.

The evaluation metrics do not take into account the lexical and semantic similarity of DL terms. The metrics will, for example, regard a term as wrong if the term was written in plural form instead of in singular form. This is, however, easy to correct as opposed to identifying and using a new term. It may also be easier to substitute a semantically similar term with another if the annotator knows which is the correct one. Edit distance can also overestimate the human effort of deleting a series of tokens in a `∃hasDescription.String[]`-construct. If the model suggests making a long string literal which should not be included, it requires deleting multiple tokens, while a human can typically do this in one operation. If, however, both the proposed parse and the parse in the reference gold standard contain such a string literal, then the deletion of individual tokens would correspond to the actual effort.

Hence, we argue that string edit distance, graph edit distance, and Jaccard distance overestimate the human effort because to change a term into something completely different is more effortful than to change spelling or use a synonym. However, our metrics treat all changes as equally different. As seen in GPT-3 mistake *v)* in Section 5.1, many of

the mistakes with terms involve substituting plausible but incorrect terms.

Sample selection methods As expected, we find that selecting the examples that are most semantically similar to the target sentence is the most effective strategy which is confirmed by all the metrics. We also find that the ordering impacts performance, which is consistent with the results presented in (Liu et al., 2022). Specifically, we find that ordering the examples with the most semantically similar examples first achieved the best results. The Clustering method also yields better results than random sampling similar to what was found by (Chang et al., 2021). All sample selection methods, however, yield a reduction in the work needed to create the gold standard.

With the MostSimilarFirst method, the Jaccard distance was found to decline over time (see Figure 4). This trend can be attributed to the fact that as we accumulate more examples and consequently have access to more examples with similar topics and terms to the target sentence, the model will be increasingly exposed to sentences with similar terms and how these terms are represented in the DL parses. Our error counts support the observation that when creating the prompt using the MostSimilarFirst method it produces fewer terminology-related mistakes, indicating a better understanding of the DL vocabulary.

7 Conclusion

In our study, we propose a systematic approach to gold standard creation based on the concept of Human-Machine collaborative annotation. To evaluate the effectiveness of our approach, we con-

ducted a case study on a small corpus of industry requirements. Our results indicate that the best method reduced the annotation effort over manual annotation by about 41 % and 38 % using the string edit distance, and graph edit distance respectively. We argue that the actual reduction in effort is even greater, as the metrics we use overestimate the effort required to correct terms.

In our study, we find that selecting the semantically most similar requirements as examples and ordering them with the most similar example first was most effective. Additionally, we found that using 30 examples was better, on average, than 5, 10, and 20. It is worth noting, however, that the effectiveness of the model depends more on which examples it sees than the number of examples, demonstrated by the fact that both Clustering and MostSimilar resulted in fewer edits than all the experiments with RandomN even Random30, which use more examples.

8 Limitations and future work

Limitations The metrics we use to estimate the human effort to correct an initial parse, i.e., string edit distance, graph edit distance, and Jaccard distance, all assume that each operator, term, and token are equally difficult to change and thus overestimate the real effort as discussed in Section 6. The distance is measured between the parse proposed by the LLM and the parse in the reference gold standard. However, as there may exist multiple ways to represent one and the same requirement, it is possible that the proposed parse is equally valid as the reference parse, but simply on a different form. A human annotator could have accepted this parse (with or without modifications), however, our metrics are unable to capture such cases.

We were not able to measure how the approach affects the actual time it takes for a human to create the parses from scratch as opposed to correct the proposals by the LLM. This would have been a better measure than edit distance measures and Jaccard distance. To be able to estimate the actual time it takes for a human to create the parses, we would have needed to conduct all the experiments several times with multiple domain experts doing the corrections (to account for individual differences), something we did not have access to.

In addition, creating a consistent reference gold standard was challenging due to the many different topics and the lack of an ontology to ensure

consistent modelling of terms and constructs. The possibility of modelling the same requirements in different ways further complicated the process. Using a more narrow domain or having access to a concrete ontology and application could have facilitated the creation of the reference gold standard. In the future, however, we want to use our approach to create a gold standard for a real application.

Since this is a case study, we have focused on only one language model. However, it is important to notice that other models are likely to demonstrate different performances. Furthermore, we could have compared how a human subject performs compared to a language model on the task. It is possible that human performance also is sub-optimal.

Moreover, one may argue that a wrongly parsed requirement by GPT-3 may mislead the human annotator into creating a parse that is incorrect but looks plausible. It is, therefore, important to have annotators with both domain and modelling knowledge. To see if this is the case, one would have to have several groups of people annotate the same requirements with and without collaboration with GPT-3.

Future work It would be interesting to carry out similar studies with existing semantic parsing datasets and compare how the performance on this particular dataset differs from standard datasets. Working with several models and several datasets could provide insight into how effective this method is for gold standard creation for semantic parsing in general, and how the domain specificity affects the effectiveness in particular.

Another interesting direction for future work is to explore the possibility of including an existing vocabulary as part of the prompt. Since many of the mistakes come from using incorrect vocabulary or different concept breakdowns than the one proposed in the reference gold standard, a two-phase prompting approach, where one can make use of vocabulary from an existing ontology, could improve the performance of the method.

Finally, the correct understanding of a requirement often relies on factors such as domain knowledge, the surrounding context and the interplay with other requirements. Therefore, taking into account larger structures, such as paragraphs, sections or entire documents can provide essential information that could enhance parsing accuracy.

9 Acknowledgement

The research is funded by the SIRIUS centre⁷: Norwegian Research Council project number 237898. It is co-funded by partner companies, including DNV. We thank the reviewers for their valuable feedback.

References

- Ernie Chang, Xiaoyu Shen, Hui-Syuan Yeh, and Vera Demberg. 2021. [On training instance selection for few-shot neural text generation](#). In *ACL 21 / IJCNLP 21 (Volume 2: Short Papers)*, pages 8–13.
- Det Norske Veritas. Ed. July 2022. RULES FOR CLASSIFICATION: Ships. Technical report, DNV-RU-SHIP. ©DNV GL.
- Jacob Devlin, Ming-Wei Chang, et al. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *NAACL, Volume 1 (Long and Short Papers)*, pages 4171–4186.
- Adi Haviv, Jonathan Berant, and Amir Globerson. 2021. [BERTese: Learning to speak to BERT](#). In *EACL 2021: Main Volume*, pages 3618–3623, Online.
- Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. [How can we know what language models know?](#) *TACL*, 8:423–438.
- Johan W Klüwer and DNV GL. 2019. OWL Upper Ontology for Reified Requirements. <https://data.dnv.com/ontology/requirement-ontology/documentation/req-ont.pdf> accessed: 2023-01-13.
- Markus Krötzsch, Frantisek Simancik, and Ian Horrocks. 2012. A description logic primer. *arXiv:1201.4089*.
- Mike Lewis, Yinhan Liu, et al. 2020. [BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *ACL 2020*, pages 7871–7880.
- Jiachang Liu, Dinghan Shen, et al. 2022. [What makes good in-context examples for GPT-3?](#) In *DeeLIO 2022 Workshop*, pages 100–114.
- Pengfei Liu, Weizhe Yuan, et al. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.
- Yinhan Liu, Myle Ott, et al. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv:1907.11692*.
- Yao Lu, Max Bartolo, et al. 2022. [Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity](#). In *ACL 2022 (Volume 1: Long Papers)*, pages 8086–8098.
- Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. [Building a large annotated corpus of English: The Penn Treebank](#). *Computational Linguistics*, 19(2):313–330.
- Fabio Petroni, Tim Rocktäschel, et al. 2019. [Language models as knowledge bases?](#) In *EMNLP-IJCNLP 2019*, pages 2463–2473.
- Colin Raffel, Noam Shazeer, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21(1):5485–5551.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *EMNLP-IJCNLP 2019*, pages 3982–3992.
- Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. [How much knowledge can you pack into the parameters of a language model?](#) In *EMNLP 2020*, pages 5418–5426.
- Subendhu Rongali, Konstantine Arkoudas, Melanie Rubino, et al. 2022. Training naturalized semantic parsers with very little data. *arXiv:2204.14243*.
- Subhro Roy, Sam Thomson, et al. 2022. Benchclamp: A benchmark for evaluating language models on semantic parsing. *arXiv:2206.10668*.
- Nathan Schucher, Siva Reddy, and Harm de Vries. 2022. [The power of prompt tuning for low-resource semantic parsing](#). In *ACL 2022 (Volume 2: Short Papers)*, pages 148–156.
- Richard Shin, Christopher Lin, et al. 2021. [Constrained Language Models Yield Few-Shot Semantic Parsers](#). In *EMNLP 2021*, pages 7699–7715.
- Ashish Vaswani, Noam Shazeer, et al. 2017. Attention is all you need. *NIPS 2017*, 30.
- Alex Wang, Yada Pruksachatkun, et al. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *NeurIPS 2019*, 32.
- Zirui Wang, Adams Wei Yu, Orhan Firat, and Yuan Cao. 2021. Towards zero-label language learning. *arXiv:2109.09193*.
- Jingfeng Yang, Haoming Jiang, et al. 2022a. [SE-QZERO: Few-shot compositional semantic parsing with sequential prompts and zero-shot models](#). In *NAACL 2022*, pages 49–60.
- Kevin Yang, Olivia Deng, et al. 2022b. [Addressing resource and privacy constraints in semantic parsing through data augmentation](#). In *ACL 2022*, pages 3685–3695.
- Zhiping Zeng, Anthony KH Tung, et al. 2009. Comparing stars: On approximating graph edit distance. *Proceedings of the VLDB Endowment*, 2(1):25–36.

⁷<http://sirius-labs.no>

A Prompt

We used the following fixed prompt with GPT-3.

Below are some inputs and the outputs of a semantic parser of industry standards. It always transforms a sentence into its correct corresponding logical representation. The input is a requirement from an industry standard. The output is a logical representation in description logic (DL) format. The output represents classes, properties, individuals and restrictions. The symbols used in the DL syntax are: \exists , \sqsubseteq , \sqcap , \sqcup , and \neg . On the left-hand side of the \sqsubseteq is most often a physical object and possibly a condition on the object. The right-hand side of the \sqsubseteq is what is demanded of the object on the left side.

B Examples of modelling by GPT-3

DNV-RU-HSLC-Pt4 [1.2.2] (sentence 2) *Flexible pipes shall have suitable connections, be resistant to salt, water, oil and vibration, be visible, easily accessible and are not to penetrate watertight bulkheads.* GPT-3 proposes

```
FlexiblePipe  $\sqsubseteq$ 
 $\exists$ hasFeature.(SuitableConnection
   $\sqcap$  ResistanceToSalt
   $\sqcap$  ResistanceToWater
   $\sqcap$  ResistanceToOil
   $\sqcap$  ResistanceToVibration
   $\sqcap$  Visible
   $\sqcap$  EasilyAccessible)
 $\sqcap$   $\neg$  $\exists$ permits.(PenetrateWatertightBulkhead)
```

There are some problems with this. First, this formula states that the FlexiblePipe shall have one of these features, but the requirement states that it needs to have all the features. In addition, GPT-3 changes *resistant to water* to ResistanceToWater, while being a reasonable proposal it is different from the reference parse, thus it affects distance.

The reference parse is:

```
FlexiblePipe  $\sqsubseteq$ 
 $\exists$ hasFeature.SuitableConnection
 $\sqcap$   $\exists$ hasFeature.ResistantToSalt
 $\sqcap$   $\exists$ hasFeature.ResistantToWater
 $\sqcap$   $\exists$ hasFeature.ResistantToOil
 $\sqcap$   $\exists$ hasFeature.ResistantToVibration
 $\sqcap$   $\exists$ hasFeature.Visible
 $\sqcap$   $\exists$ hasFeature.EasilyAccessible
 $\sqcap$   $\neg$  $\exists$ permits.
  PenetrateWatertightBulkheads
```

This formula comes out with a string edit distance of 19 and a graph edit distance of 30, a Jaccard

distance of 0.5. The edit distance is high, but in reality, this formula helps the annotator much as the basic structure is there. One has to remove the parenthesis, and add \sqcap \exists hasFeature. before each of the features, change the spelling from Resistance to Resistant (and similar for the others).

DNV-RU-HSLC-Pt4 [1.5.1] (sentence 1) *Fuel with a flashpoint below 35°C shall not be used.* In this example, GPT-3 proposes the same as the reference parse:

```
Fuel  $\sqsubseteq$ 
 $\neg$  $\exists$ hasFlashpoint.(PhysicalQuantity
   $\sqcap$   $\exists$ hasValue.float[<35]
   $\sqcap$   $\exists$ hasUnit.string['C'])
```

DNV-OS-C103 [1.3.7] (sentence 1) *For new designs, and/or unproved design applications of designs where limited or no direct experience exists, relevant analyses and model testing, shall be performed in order to demonstrate that an acceptable level of safety is obtained.* GPT-3 proposes

```
NewDesigns  $\sqcup$  UnprovedDesignApplications
 $\sqsubseteq$   $\exists$ hasFeature.(RelevantAnalyses
   $\sqcup$  ModelTesting)
 $\sqcap$   $\exists$ permits.AcceptableLevelOfSafety
```

The reference parse is:

```
 $\exists$ hasFeature.(NewDesign
   $\sqcup$  (Design  $\sqcap$   $\exists$ hasFeature.
    (LimitedExperience  $\sqcup$  NoExperience)))
 $\sqsubseteq$   $\exists$ hasFeature.RelevantAnalysis
 $\sqcap$   $\exists$ hasFeature.ModelTesting
 $\sqcap$   $\exists$ permits.AcceptableLevelOfSafety
```

This solution gives a string edit distance of 14, a graph edit of 20, and a Jaccard distance of 0.5. Here we observe that GPT-3 has broken down the requirement differently from what the reference parse does. It puts the demand on the concept NewDesigns \sqcup UnprovedDesignApplications. We consider, however, that the requirement is not so much about the design, but the object that is being designed. On the right side of \sqsubseteq , it requires only one feature for something that is either a relevant analysis or a model testing, which is wrong. It should be two (different) features. The use of plural in NewDesigns, RelevantAnalyses is easy to correct, but affects the edit distances and Jaccard distance.

C Documents

Document code	Name
DNV-CG-0051	Non-destructive testing (January 2022)
DNV-CP-0231	Cyber security capabilities of systems and components (September 2021)
DNV-CP-0507	System and software engineering (September 2021)
DNV-OS-A101	Safety principles and arrangements (July 2019/August 2021)
DNV-OS-C101	Design of offshore steel structures, general - LRFD method (July 2019/August 2021)
DNV-OS-C102	Structural design of offshore ship-shaped and cylindrical units (July 2020/August 2021)
DNV-OS-C103	Structural design of column stabilised units - LRFD method (July 2020/August 2021)
DNV-OS-D101	Marine and machinery systems and equipment (July 2021)
DNV-OS-D201	Electrical installations (July 2022)
DNV-OS-D202	Automation, safety and telecommunication systems (July 2019/August 2021)
DNV-OS-D301	Fire protection (July 2019/August 2021)
DNV-OS-E301	Position mooring (July 2021)
DNV-OS-E402	Diving systems (July 2019/August 2021)
DNV-RU-HSLC-Pt3	High speed and light craft Part 3 Structures, equipment (August 2021)
DNV-RU-HSLC-Pt4	High speed and light craft Part 4 Systems and components (July 2022)
DNV-RU-NAVAL-Pt3	Naval vessels Part 3 Surface Ships (December 2015)
DNV-RU-NAVAL-Pt4	Naval vessels Part 4 Sub-surface ships (January 2018)
DNV-RU-NAV-Pt7	Naval vessels Part 7 Fleet in service (July 2022)
DNV-RU-OU-0101	Offshore drilling and support units
DNV-RU-OU-0104	Self-elevating units, including wind turbine installation units and liftboats (July 2022)
DNV-RU-SHIP-Pt4	Ships Part 4 Systems and components (July 2021)
DNV-SI-0166	Verification for compliance with Norwegian shelf regulations (January 2022)
DNV-ST-0111	Assessment of station keeping capability of dynamic positioning vessels (December 2021)

Table 3: The documents used in this study

All documents are copyrighted ©DNV. DNV does not take responsibility for any consequences arising from the use of this content.

LexExMachinaQA: A framework for the automatic induction of ontology lexica for Question Answering over Linked Data

Mohammad Fazleh Elahi

Bielefeld University, Germany
melahi@techfak.uni-bielefeld.de

Basil Ell

Bielefeld University, Germany
Oslo University, Norway
basile@ifi.uio.no

Philipp Cimiano

Bielefeld University, Germany
cimiano@techfak.uni-bielefeld.de

Abstract

An open issue for Semantic Question Answering Systems is bridging the so called *lexical gap*, referring to the fact that the vocabulary used by users in framing a question needs to be interpreted with respect to the logical vocabulary used in the data model of a given knowledge base or knowledge graph. Building on previous work to automatically induce ontology lexica from language corpora by using association rules to identify correspondences between lexical elements on the one hand and ontological vocabulary elements on the other, in this paper we propose LexExMachinaQA, a framework allowing us to evaluate the impact of automatically induced lexicalizations in terms of alleviating the lexical gap in QA systems. Our framework combines the LexExMachina approach (Ell et al., 2021) for lexicon induction with the QueGG system proposed by Benz et al. (Benz et al., 2020) that relies on grammars automatically generated from ontology lexica to parse questions into SPARQL. We show that automatically induced lexica yield a decent performance i.t.o. F_1 measure with respect to the QLAD-7 dataset, representing a 34% – 56% performance degradation with respect to a manually created lexicon. While these results show that the fully automatic creation of lexica for QA systems is not yet feasible, the method could certainly be used to bootstrap the creation of a lexicon in a semi-automatic manner, thus having the potential to significantly reduce the human effort involved.

1 Introduction

According to (Höffner et al., 2017), the benefit of Semantic Question Answering (SQA) systems from the perspective of end users is that they can access knowledge in knowledge bases or knowledge graphs i) without having to master a formal language such as SPARQL, and ii) without having knowledge about the (ontological) vocabularies used in the knowledge bases. One of the seven challenges identified by the authors for the development of SQA systems is handling the lexical gap, requiring to bridge between the way users refer to certain properties and the way they are modelled in a given knowledge base. Take the following examples involving a (relational) noun, a verb, and an adjective, respectively:

- ‘*Who is the husband of Julia Roberts?*’ In this case, ‘*husband*’ needs to be interpreted with respect to DBpedia as `dbo:spouse`¹ in order to map the question correctly to the following SPARQL query:

```
SELECT ?o WHERE {
  dbr:Julia_Roberts dbo:spouse ?o }
```

- ‘*Who stars in the Matrix?*’ In this case, ‘*stars in*’ refers to the property `dbo:actor`, so that the question can be mapped to the following SPARQL query:

```
SELECT ?o WHERE {
  dbr:The_Matrix dbo:actor ?o }
```

- ‘*How high is the Mulhacén?*’ In this case, ‘*high*’ needs to be interpreted in terms of the

¹In this paper we use compact URIs and use namespace prefixes that are defined as follows: `dbr:` <http://dbpedia.org/resource/>, `dbo:` <http://dbpedia.org/ontology/>, `dbp:` <http://dbpedia.org/property/>, `rdf:` <http://www.w3.org/1999/02/22-rdf-syntax-ns#>, `lemon:` <http://lemon-model.net/lemon#>, and `lexinfo:` <http://www.lexinfo.net/ontology/2.0/lexinfo#>.

DBpedia property `dbp:elevation` in order to map the question correctly to the following SPARQL query:

```
SELECT ?o WHERE {
  dbr:Mulhacén dbp:elevation ?o }
```

Existing QA systems have attempted to handle the lexical gap by using edit distances or similarity measures to recognize inflected forms of the same lemma and dealing with misspellings or spelling variants (Höffner et al., 2017). A frequently used lexical resource is WordNet (Miller, 1995) and has been used to recognize synonyms in QA systems (e.g., (Walter et al., 2012)). Some QA systems have also relied on pattern databases such as PATTY (Nakashole et al., 2012) to find constructions that verbalize a given relation or property. Word embeddings have also been used to discover related terms (Hakimov et al., 2017).

In this paper, building on our previous work (Benz et al., 2020; Elahi et al., 2021), we follow a different approach and induce a lexicon that is specific for a given knowledge base or vocabulary. We have shown that such lexica can be induced automatically to some extent using our LexExMachina approach (Ell et al., 2021) that builds on association rules to find correspondences between lexical elements and ontological vocabulary elements. However, it is unclear if this approach would help to effectively bridge the lexical gap prevailing in QA systems. In this paper, we thus leverage the LexExMachina approach to induce lexical knowledge relevant for QA, so that we call the approach *LexExMachinaQA*. In order to evaluate the impact of the automatically induced lexica, we build on the QA system proposed by Benz et al. (Benz et al., 2020) that relies on a lexicon-ontology model to automatically generate a grammar that allows to parse questions into SPARQL. While the approach in principle works for multiple languages, in this paper we restrict the evaluation to the English language as a proof-of-concept. Our evaluation is conducted with respect to QALD-7 as a benchmark. We contrast the results obtained with an automatically induced lexicon with the results of a lexicon created manually, comprising 806 lexical entries overall. We show that the automatically induced lexicon yields a decent performance of F_1 for the QA system proposed by Benz et al. (Benz et al., 2020) on the QALD-7 benchmark, corresponding to a performance degradation of between 34–56%

relative to the performance of a QA system based on the manually created lexicon. While this shows that it is still worth to invest into manual lexicon creation, the results are encouraging in the sense that the automatically induced lexicon could reduce significantly the human effort involved.

2 Method

In this section, first, we briefly describe our model-based approach to Question Answering (QueGG), detailed in previous work (Benz et al., 2020). QueGG makes use of an ontology lexicon to generate grammars from which questions in natural language are generated. Second, we describe how a lexicon can be created manually. Third, we briefly describe LexExMachina, our previous work on inducing correspondences between natural language and a knowledge base using association rule mining (Ell et al., 2021). Finally, we describe how we make use of the correspondences obtained via LexExMachina to automatically derive a lexicon that can then be used by QueGG.

2.1 Background: QueGG

QueGG (Benz et al., 2020), our previous work, is a model-based approach to QA in which a developer of the QA system provides a lexicon using the lemon-OntoLex model (Cimiano et al., 2016), specifying how the vocabulary elements are realized in natural language. The lemon-OntoLex model is an updated version of the lemon model (McCrae et al., 2011) and is the core representation used by the grammar generation in QueGG. The main benefit of the approach is that it is fully controllable in the sense that it can be predicted what the impact of extending the lexicon will have in terms of the questions covered by the system.

Our previous work on QueGG has shown that, leveraging on lemon lexica, question answering grammars can be automatically generated, and these can, in turn, be used to interpret questions and parse them into SPARQL queries. A QA web application developed in previous work (Elahi et al., 2021; Nolano et al., 2022) has further shown that such QA systems can scale to millions of questions and that the performance of the system is practically real-time from an end-user perspective.

The grammar generation from a lexical entry with a specific syntactic frame, detailed in Lex-Info (Cimiano et al., 2011), is controlled by a

generic template that describes how specific lexicalized grammar rules can be generated for a given lexical entry. The grammar generation supports the following syntactic frames:

- **NounPPFrame**: corresponding to a relational noun that requires a prepositional object such as ‘*spouse*’ (*of*), ‘*mayor*’ (*of*), ‘*capital*’ (*of*)
- **TransitiveFrame**: corresponding to transitive verbs such as (*to*) ‘*direct*’ and (*to*) ‘*marry*’.
- **InTransitivePPFrame**: corresponding to intransitive verbs subcategorizing a prepositional phrase such as ‘*star*’ (*in*), ‘*born*’ (*on*) or ‘*flow*’ (*through*)
- **AdjectivePredicateFrame**: covering intersective adjectives such as ‘*Spanish*’ and ‘*Afghan*’. This frame is used for both attributive and predicative use of the adjective.
- **AdjectiveSuperlativeFrame**: covering gradable adjectives such as ‘*high*’ and ‘*highest*’.

For the sake of self-containedness, we describe a lexical entry and the grammar rules for the transitive verb (*to*) ‘*direct*’. The lexicon entry is shown in Figure 1. The semantics of the lexical entry (*to*) ‘*direct*’ is expressed by the property `dbo:director`. The lemon entry also specifies that the subject of the property is realized by the direct object of the verb ‘*direct*’, while the object of the property is realized by the syntactic subject of the verb ‘*direct*’. The following grammar is generated automatically:

```
Rule 1:
S -> Who directs X? | Who directed X? |
     Which person directs X? | Which
     person directed X?
Rule 2:
S -> What is directed by X? | What was
     directed by X? | Which film is
     directed by X? | Which films are
     directed by X? | Which film was
     directed by X? | Which films were
     directed by X? | Give me all films
     directed by X?
Rule 3:
S -> How many films are directed by X? |
     How often did X direct?
Rule 4:
S -> film directed by X | films directed
     by X
```

```
1 :to_direct a lemon:LexicalEntry ;
2   lexinfo:partOfSpeech lexinfo:verb ;
3   lemon:canonicalForm :form_direct ;
4   lemon:otherForm :form_directs ;
5   lemon:otherForm :form_directed ;
6   lemon:synBehavior
7 :direct_frame_transitive ;
8   lemon:sense :direct_ontomap .
9 :form_direct a lemon:Form ;
10  lemon:writtenRep "direct"@en ;
11  lexinfo:verbFormMood lexinfo:infinitive .
12
13 :form_directs a lemon:Form ;
14  lemon:writtenRep "directs"@en ;
15  lexinfo:person lexinfo:thirdPerson .
16
17 :form_directed a lemon:Form ;
18  lemon:writtenRep "directed"@en ;
19  lexinfo:tense lexinfo:past .
20
21 :direct_frame_transitive a
22   lexinfo:TransitiveFrame ;
23   lexinfo:subject :direct_subj ;
24   lexinfo:directObject :direct_obj .
25
26 :direct_ontomap a lemon:OntoMap,
27   lemon:LexicalSense ;
28   lemon:ontoMapping :direct_ontomap ;
29   lemon:reference dbo:director ;
30   lemon:subjOfProp :direct_obj ;
31   lemon:objOfProp :direct_subj ;
32   lemon:condition :direct_condition .
33
34 :direct_condition a lemon:condition ;
35   lemon:propertyDomain dbo:Film ;
36   lemon:propertyRange dbo:Person .
```

Figure 1: Lemon entry for the transitive verb (*to*) ‘*direct*’.

2.2 Background: Manual Lexicon Creation

A necessary prerequisite for the grammar generation approach is the availability of a lemon lexicon that describes by which lexical entries the elements (classes, properties) of a particular dataset can be verbalized in a particular language. In particular, a lexicon is needed for each language to be supported by the QA system. We manually created a lexicon for English and DBpedia.² The manually created lexical entries,³ together with the automatically generated grammar, are available online.⁴ Table 1 shows the number of manually created lexical entries for QALD-7 training data for different frame types of LexInfo as well as the number of grammar rules automatically generated from these.

The creation of a single lexical entry took approximately 2–3 minutes. The total construction time for the lexicon comprising of 806 entities was approximately 30 hours.

²<https://downloads.dbpedia.org/2016-10/core-i18n/en/>

³<https://github.com/fazleh2010/multilingual-grammar-generator/tree/main/result/en/lexicalEntries>

⁴<https://github.com/fazleh2010/multilingual-grammar-generator/tree/main/result/en/grammar>

Frame Type	# Lexical Entries	# Grammar Rules
NounPP	722	1,444
Transitive	37	111
InTransitivePP	27	81
AdjPredicate	15	76
AdjSuperlative	5	15
Total	806	1,727

Table 1: An overview over the number of manually created lexical entries for QALD-7 training data for different frame types and the number of automatically generated grammar rules.

2.3 Background: LexExMachina

LexExMachina (Eli et al., 2021) is a methodology that induces correspondences between natural language and a knowledge base by mining class-specific association rules from a loosely-parallel text-data corpus (e.g., Wikipedia + DBpedia). These association rules can help to bridge from natural language to a knowledge base and from a knowledge base to natural language. In the context of question answering, we make use of those rules that bridge from natural language to a knowledge base.

For example, in the context of a question about a person where the question contains the adjective "Greek", the corresponding SPARQL query would contain a triple pattern such as `?x dbo:nationality dbr:Greece`, whereas in the context of a question about a settlement where the question contains the adjective "Greek", the corresponding SPARQL query would contain a triple pattern such as `?x dbo:country dbr:Greece`.

The association rule that specifies that if the term "Greek" occurs in a text about a politician, then this corresponds in DBpedia to the triple pattern with predicate `dbo:nationality` and object `dbr:Greece` is represented as follows:

$$\begin{aligned}
 & \text{dbo:Politician} \in c_e \wedge \text{"Greek"} \in l_e \Rightarrow \\
 & (e, \text{dbo:nationality}, \text{dbr:Greece}) \in G
 \end{aligned}$$

Here, c_e is the set of classes an entity e is instance of, l_e is a set of linguistic patterns (such as n-grams) that occur in the text that mentions the entity e , and G is the knowledge base that we bridge to (here: DBpedia). This rule is an example for the rule pattern $c_s, l_s \Rightarrow po$, one of the 20 types of association rules regarded by LexExMachina. In particular, the rule expresses

that for an entity e that is an instance of the class `dbo:Politician` where the linguistic pattern "Greek" occurs in the text that mentions or describes the entity e , within the knowledge graph G there is (or should be) a triple that expresses that the entity e is in relation `dbo:nationality` with the entity `dbo:Greece`.

The LexExMachina approach was previously applied to a subset of a loosely-parallel text-data corpus consisting of Wikipedia as a corpus and DBpedia as a knowledge graph, which resulted in 447, 888, 109 rules, published together with the original paper.

Association rules come with a set of measures. The general form of an association rule is $A \Rightarrow B$. For the types of rules that we regard in this paper, with $sup(A)$ we refer to the number of times the event described by the left hand side of an association rule occurred in the corpus (e.g., how often it occurred in the corpus that a text that mentioned or described a politician contained "Greek"). With $sup(B)$ we refer to the number of times the event described by the right hand side of an association rule occurred in the knowledge graph (e.g., how often it occurred in the knowledge graph that an entity is in relation `dbo:nationality` with the entity `dbo:Greece`). $sup(AB)$ refers to the number of times that both events occurred together (e.g., how often it occurred that a text that mentioned or described an entity of type politician contained "Greek" and this entity is in relation `dbo:nationality` with the entity `dbo:Greece` in the knowledge graph). The confidence of an association rule of the form $A \Rightarrow B$, denoted by $conf(A \Rightarrow B)$, is the estimated conditional probability $P(B|A)$ and is calculated as $sup(AB)/sup(A)$.

In practice, association rules with high confidence do not necessarily disclose truly interesting event relationships (Brin et al., 1997). Therefore, an *interestingness measure* quantifies the interestingness of an association rule. For example, the interestingness measure $Cosine(A \Rightarrow B)$ is defined as $\sqrt{P(A|B)P(B|A)}$. Note that $P(A|B)$ is equal to $conf(B \Rightarrow A)$, i.e., the confidence of the "reversed" rule.

2.4 Lexicon Generation based on LexExMachina

The starting point for our lexicon induction method is a knowledge graph. We retrieve all the prop-

erty URIs from the graph and mine class-specific association rules for each property, yielding lexicalizations for each property.

While LexExMachina defines 20 different types of class-specific association rules, in the context of LexExMachinaQA we rely only on two of those. In fact, we rely only on the two rules that predict a lexicalization for a subject of a given class and a property or for an object of a given class and a property. These rules are described in more detail in the following:

1. The rule pattern with the name $c_s, p \Rightarrow l_s$ has the following meaning: given a subject entity e that is an instance of the class c_s and given that e is in relation p to some term, then the relation can be expressed with the linguistic pattern l . The LexExMachina dataset contains 98, 317, 655 rules of this type.

$$\begin{aligned} & \text{dbo:FictionalCharacter} \in c_e \quad (1) \\ & \wedge \exists o : (e, \text{dbo:spouse}, o) \in G \\ & \Rightarrow \text{"husband of"} \in l_e \end{aligned}$$

2. The rule pattern with the name $c_o, p \Rightarrow l_o$ has the following meaning: given an object entity e that is an instance of the class c_o and given that some term is in relation p with e , then the relation can be expressed with the linguistic pattern l . The LexExMachina dataset contains 6, 499, 288 rules of this type.

$$\begin{aligned} & \text{dbo:Person} \in c_e \quad (2) \\ & \wedge \exists s : (s, \text{dbo:starring}, e) \in G \\ & \Rightarrow \text{"star in"} \in l_e \end{aligned}$$

The linguistic patterns found on the right-hand side of the above rules are n -grams found in the corresponding texts. In LexExMachina, n -grams with $1 \leq n \leq 4$ are considered.

Given an association rule, the creation of a lexical entry comprises the following steps:

1. We remove stop words (excluding prepositions) from the linguistic patterns on the right hand sides of the rules.
2. We use a part-of-speech tagger to tag the n -grams on the right-hand side of a rule. We rely on the Stanford tagger in particular.⁵

⁵<https://nlp.stanford.edu/software/tagger.shtml>

3. Relying on the part-of-speech sequence, patterns are classified into the syntactic frames discussed in Section 2.1. A noun followed by a preposition is classified as a NounPPFrame. A verb is either classified as a transitive verb (i.e., TransitiveFrame) or as an intransitive verb (i.e., InTransitivePPFrame), based on the English Wiktionary dictionary.⁶ Wiktionary also contains inflection forms of verbs, which are added to a lexical entry – see for example Figure 1 line 14 "directs" and line 18 "directed" in the entry for the transitive verb (*to*) 'direct'. An adjective is classified as an attributive adjective (i.e., AdjectivePredicativeFrame) or as a superlative adjective (i.e., AdjectiveSuperlativeFrame). We use Wiktionary for an adjective's classification and retrieve its inflection forms.

We describe how the actual lexical entries in RDF format are created by way of OTTR templates (Skjæveland et al., 2018). OTTR is a language for defining templates over RDF data. Thereby, consistency can be ensured and RDF graph instantiations are more human-readable than plain RDF data. Using OTTR enables us to separate the data about a lexical entry that we collect from LexExMachina and from Wiktionary from how we represent it. For example, in order to create the lemon entry for the relational noun 'husband' (*of*), shown in Figure 2, we need to have collected the canonical, singular and plural form of the noun, the preposition, the corresponding DBpedia property, and the property's domain and range. Then, when the OTTR template shown in the appendix in Figure 3 is instantiated using the OTTR template instantiation statement shown below, then RDF data similar⁷ to the data shown in Figure 2 is generated.

```
quegg:NounPPFrame (
  "husband"@en, "husband"@en,
  "husbands"@en, "of"@en,
  dbo:husband, dbo:Person,
  dbo:Person) .
```

⁶<http://en.wiktionary.org/>

⁷Instead of showing the actual RDF data as it is generated, which contains blank nodes such as `_:b0`, `_:b1` etc., for the purpose of readability we have replaced these with meaningful URIs.

Lexicon	# Entries NounPP*	# Entries Transitive*	# Entries InTransitivePP*	# Entries AdjPred*+AdjSuper*	# Entries Total
Rule Pattern $c_s, p \Rightarrow l_s$					
s-L1	280,219	27,703	26,072	34,175	368,169
s-L2	286,127	28,246	26,724	34,818	375,915
s-L3	572,254	56,492	53,448	69,636	751,830
s-L4	248,963	24,954	23,825	31,067	328,809
s-L5	497,926	49,908	47,650	62,134	657,618
Rule Pattern $c_o, p \Rightarrow l_o$					
o-L1	66,454	4,598	4,422	8,618	84,092
o-L2	42,416	4,701	3,908	7,203	58,228
o-L3	57,713	3,626	3,437	6,636	71,412
o-L4	43,654	2,644	2,597	4,739	53,634
o-L5	38,712	2,742	1,092	4,798	47,344

Table 2: The table shows the number of lexical entries per frame type generated with the two rule patterns for the best 5 lexicon configurations according to F -score. Here, AdjPred* refers to AdjectivePredicateFrame and AdjSuper* refers to AdjectiveSuperlativeFrame.

Lexicon	$sup(A)$	$sup(B)$	$sup(AB)$	$P(A B)$	$P(B A)$	$Cos.$	micro- P	Micro- R	Micro- F_1	Macro- P	Macro- R	Macro- F_1
Rule Pattern $c_s, p \Rightarrow l_s$												
s-L1	5	5	5	0.02	0.09	0.1	0.32	0.44	0.37	0.40	0.40	0.40
s-L2	5	5	5	0.02	0.02	0.1	0.32	0.44	0.37	0.40	0.40	0.40
s-L3	250	50	50	0.02	0.10	0.1	0.31	0.47	0.37	0.39	0.39	0.38
s-L4	250	5	5	0.02	0.10	0.1	0.30	0.46	0.36	0.38	0.39	0.39
s-L5	250	5	5	0.02	0.60	0.1	0.26	0.46	0.33	0.38	0.37	0.38
Rule Pattern $c_o, p \Rightarrow l_o$												
o-L1	5	5	5	0.09	0.02	0.1	0.15	0.36	0.21	0.22	0.23	0.27
o-L2	5	5	5	0.02	0.02	0.09	0.14	0.41	0.21	0.24	0.24	0.24
o-L3	5	5	5	0.1	0.02	0.09	0.14	0.37	0.21	0.23	0.23	0.23
o-L4	5	5	5	0.02	0.1	0.1	0.13	0.40	0.20	0.24	0.24	0.24
o-L5	5	5	5	0.02	0.1	0.09	0.13	0.40	0.20	0.23	0.23	0.23

Table 3: The table shows the configurations as well as micro-averaged and macro-averaged precision, recall, and F_1 scores for the 5 best lexicon configurations according to F -measure with respect to QALD-7 training data.

3 Evaluation

In this section we describe how we evaluate the manually created and the automatically generated ontology lexica and describe how we have optimized threshold values based on the parameters of LexExMachina rules to yield the best settings for LexExMachinaQA. We compare the results of the automatically generated lexica to the results obtained using the manually created lexicon as an upper baseline.

3.1 Lexicon Evaluation

We evaluate each lexicon using the QALD-7 benchmark (Usbeck et al., 2017). A QALD dataset consists of a set of tuples of the form (q, s) where q is a question in natural language and s is a corresponding SPARQL query that retrieves the answers to q from a knowledge graph (here: DBpedia).

An example (q, s) pair is the following: (*‘Who was the wife of U.S. president Lincoln?’*, SELECT ?o WHERE { dbr:Abraham_Lincoln dbo:spouse ?o }).

Given a lexicon, our approach generates grammars from which questions are generated – we call these QueGG questions. These questions have corresponding queries. Thus, we generate a set of (question, query) tuples.

We evaluate the QueGG answers for each QALD question using *Precision* (Eq. 3), *Recall* (Eq. 4) and *F-Measure* as defined by the QALD task (Usbeck et al., 2017).

Given a question-query pair (q, s) from QALD, we find the question-query pair (q', s') from QueGG such that the similarity between the questions q and q' is maximal. We use Jaccard similarity to measure the similarity between two questions:

$$(q', s') = \max_{(q', s') \in \text{QueGG}} JS(q, q')$$

The reason for using the Jaccard similarity measure is because it ignores word order and duplicate words, thus it emphasizes unique words shared by two questions. For example, for the QALD-7 question *‘When was the Titanic completed?’* we retrieve the QueGG question *‘When was RMS Ti-*

```

1 :husband_of a lemon:LexicalEntry ;
2   lexinfo:partOfSpeech lexinfo:noun ;
3   lemon:canonicalForm :husband_of_form ;
4   lemon:otherForm :husband_of_singular ;
5   lemon:otherForm :husband_of_plural ;
6   lemon:sense :husband_of_sense_1 ;
7   lemon:synBehavior :husband_of_nounpp .
8
9 :husband_of_form a lemon:Form ;
10  lemon:writtenRep "husband"@en .
11
12 :husband_of_singular a lemon:Form ;
13  lemon:writtenRep "husband"@en ;
14  lexinfo:number lexinfo:singular .
15
16 :husband_of_plural a lemon:Form ;
17  lemon:writtenRep "husbands"@en ;
18  lexinfo:number lexinfo:plural .
19
20 :husband_of_nounpp a lexinfo:NounPPFrame ;
21  lexinfo:copulativeArg :arg1 ;
22  lexinfo:prepositionalAdjunct :arg2 .
23
24 :husband_of_sense_1 a lemon:OntoMap,
25  lemon:LexicalSense ;
26  lemon:ontoMapping :husband_of_sense_1 ;
27  lemon:reference dbo:spouse ;
28  lemon:subjOfProp :arg2 ;
29  lemon:objOfProp :arg1 ;
30  lemon:condition :husband_of_sense_1_condition .
31
32 :husband_of_sense_1_condition a lemon:condition ;
33  lemon:propertyDomain dbo:Person ;
34  lemon:propertyRange dbo:Person .
35
36 :arg2 lemon:marker :husband_of_form_preposition .
37 ## Prepositions ##
38 :husband_of_form_preposition a
39  lemon:SynRoleMarker ;
40  lemon:canonicalForm
41  [ lemon:writtenRep "of"@en ] ;
42  lexinfo:partOfSpeech lexinfo:preposition .

```

Figure 2: Lemon entry for the relational noun ‘husband’ (of).

‘*Who is the daughter of the daughter of Jan Delay?*’ gets 100% similarity with the question ‘*Who is the daughter of Jan Delay?*’.

$$\text{precision}(q, s) := \frac{|\Omega_{s,G} \cap \Omega_{s',G}|}{|\Omega_{s',G}|} \quad (3)$$

$$\text{recall}(q, s) := \frac{|\Omega_{s,G} \cap \Omega_{s',G}|}{|\Omega_{s,G}|} \quad (4)$$

3.2 Parameter Optimization

The rules created by LexExMachina have a number of parameters (see Section 2.3). We make use of these parameters to specify which rules to use based on threshold values when creating a lexicon. We carry out grid search to find the best values (according to F_1 -measure) for these parameters on the QALD-7 training dataset.

The threshold parameters that we optimize and the grid intervals we explore are the following:

$$\begin{aligned} \text{sup}(A) &\in \{5, 50, 250\} \\ \text{sup}(B) &\in \{5, 50, 250\} \\ \text{sup}(AB) &\in \{5, 50, 250\} \\ P(B|A) &\in \{0.02, 0.09, 0.1, 0.6\} \\ P(A|B) &\in \{0.02, 0.09, 0.1, 0.6\} \\ \text{Cosine}(A \Rightarrow B) &\in \{0.02, 0.09, 0.1, 0.6\} \end{aligned}$$

In principle, this yields $3^3 \times 3^4 = 1728$ configurations to explore. However, there cannot be a rule where $\text{sup}(A)$ or $\text{sup}(B)$ is smaller than $\text{sup}(AB)$. For two configurations that only differ in, e.g., the $\text{sup}(A)$ threshold and both $\text{sup}(A)$ values are less or equal to $\text{sup}(AB)$, both configurations would yield the same lexicon. Thus, we exclude configurations where either $\text{sup}(A)$ or $\text{sup}(B)$ is set to a value lower than $\text{sup}(AB)$. Thereby, the number of configurations we explore in grid search is 896.

3.3 Results

Table 3 shows the parameters and scores for the 5 best lexicon configurations according to F_1 -measure. In general, we see that the variation of scores is low for the top 5 configurations within a pattern class. For example, the micro F_1 -measures for the rule pattern $c_s, p \Rightarrow l_s$ vary between 0.33 and 0.37. The micro F_1 -measures for rule pattern $c_o, p \Rightarrow l_o$ are generally lower, but show also smaller variation across configurations, ranging between 0.2 and 0.21.

Table 2 shows the number of lexical entries induced per frame type separately for the 5 best configurations for each rule in addition to the overall number of lexical entries. Over all configurations, a clear pattern emerges. First of all, it can be seen that the configurations for rule pattern $c_s, p \Rightarrow l_s$ are more productive, creating an order of magnitude more lexical entries compared to the pattern $c_o, p \Rightarrow l_o$. In terms of distribution of frame types, about 75% of the induced lexical entries are of type `NounPPFrame`, representing relational nouns. About 15% of the induced lexical entries are verb frames, with more or less an equal share of Transitive and IntransitivePP verb frames, and about 10% are adjective frames.

As can be seen in Table 4, in terms of micro F_1 measure the results using the automatically induced lexicon are 0.42 under the upper baseline using the manually created lexicon (micro F_1 of 0.79). This corresponds to a relative performance degradation of about 53%.

Training Data						
Lexicon	Micro- P	Micro- R	Micro- F_1	Macro- P	Macro- R	Macro- F_1
s-L1	0.32	0.44	0.37	0.40	0.40	0.40
o-L1	0.15	0.36	0.21	0.22	0.23	0.27
manual	0.84	0.75	0.79	0.61	0.62	0.61
Test Data						
Lexicon	Micro- P	Micro- R	Micro- F_1	Macro- P	Macro- R	Macro- F_1
s-L1	0.023	0.005	0.008	0.139	0.139	0.139
o-L1	0.015	0.004	0.007	0.093	0.093	0.093
manual	0.63	0.01	0.02	0.24	0.24	0.24

Table 4: Comparison of the evaluation results on the QALD-7 training data and test data for the best-performing lexicon automatically induced for $c_s, p \Rightarrow l_s$ rules and for the best-performing lexicon automatically induced for $c_o, p \Rightarrow l_o$ rules with results for the manually created lexicon.

Overall, these results clearly show that, while our method successfully induces many appropriate lexical entries, with the completely automatically generated lexicon the performance is far from the results obtained with a manually created lexicon.

System	Micro-P	Micro-R	Micro-F
WDAqua-core1	0.37	0.39	0.39
CNN-QA	–	–	0.29
$c_s, p \Rightarrow l_s$	0.32	0.44	0.37
$c_o, p \Rightarrow l_o$	0.15	0.36	0.21
manual	0.84	0.75	0.79

Table 5: Comparison of best result of LexExMachinaQA (i.e., $c_s, p \Rightarrow l_s$ and $c_o, p \Rightarrow l_o$) with the systems evaluated on QALD-7 dataset.

Table 5 shows the results of the evaluations using the best configurations for the rule patterns $c_s, p \Rightarrow l_s$ and $c_o, p \Rightarrow l_o$, for the rule-based systems WDAqua-core1 (Diefenbach et al., 2020), and for the machine learning-based approach CNN-QA (Sorokin and Gurevych, 2017). We compare the results of our approach to these two approaches as they have also been evaluated on QALD-7 training dataset. As can be seen from Table 5, the approach using the manually created lexicon outperforms state-of-the-art systems by a large margin (F_1 of 0.79 compared to 0.39 by the WDAqua-core1 system). This clearly shows the potential of our lexicon-based approach. Concerning the results using the automatically induced lexicon for rule pattern $c_s, p \Rightarrow l_s$, we see that our approach outperforms the CNN-QA approach (F_1 of 0.37 vs. 0.29) and has comparable performance to WDAqua-core1 (F_1 of 0.37 vs. 0.39). This is a remarkable result, showing that our approach can outperform state-of-the-art systems using a fully automatically generated lexicon. If a high quality lexicon is available, our approach outperforms SOTA systems by almost doubling performance.

3.4 Qualitative Analysis

In order to illustrate the working of our system, we analyze its behaviour in more detail by discussing 6 types of cases. Hereby, we rely on the best lexicon obtained from $c_s, p \Rightarrow l_s$ rules (i.e., s-L1). In particular, we sample 150 questions from the QALD-7 training set and classify them into six cases.⁸

Case 1 (Exact lexicalization): There are many cases in which the grammar generation based on an automatically induced lexicon generates exactly the same (question, query) pair as contained in the QALD-7 dataset. This is the case for 59 out of 150 (i.e., 39.33%) questions. An example here is the question ‘*In which year was Rachel Stevens born?*’

Case 2 (different variations but correct lexicalization): A second case is the one where our grammar generation based on the automatically induced lexicon generates a question that is semantically equivalent to a QALD-7 question, but that contains a synonym or variant of the lexical element in the ground truth question. In many cases, the generated question is grammatically correct and expresses the same meaning. According to our analysis, 12 out of 150 (i.e., 8%) questions are not identical but semantically equivalent. An example is the QALD-7 question ‘*When was the Titanic completed?*’ In this case, the most similar automatically generated question is ‘*When was RMS Titanic completed on?*’

Case 3 (different variations but incorrect lexicalization): For 9 out of 150 (i.e., 6%) questions, our approach generates a question that features an incorrect lexicalization of the relevant

⁸23% do not belong to any of these classes.

property. Consider the QALD-7 question ‘*What is the currency of the Czech Republic?*’ In this question, ‘*currency of*’ refers to the property `dbo:currency`. Our approach incorrectly induces that ‘*republic of*’ denotes the property `dbo:currency` and thus generates the question: ‘*What is the republic of Czech Republic?*’ which nevertheless retrieves the correct answer.

Case 4 (same lexicalization but different SPARQL query): There are cases where a question generated by an automatically induced lexicon is equivalent to a question in QALD-7, but the corresponding SPARQL queries differ. The question ‘*Who is the president of Eritrea?*’ is generated, but instead of relating ‘*president of*’ to `dbo:leader` as required to retrieve the correct answer in QALD-7, our lexicon induction approach relates ‘*president of*’ to `dbo:office`, thus generating the same question but with a different SPARQL query, thus retrieving a different answer. This is the case for 6 out of 150 (i.e., 4%) questions.

Case 5 (Ask query): 20 out of 150 (i.e., 13.33%) questions in QALD-7 are ASK queries. The grammar generation excludes ASK queries because many of these questions are those whose answer is No. In this case, the SPARQL query of the question generated by automatically induced lexicalization is different from QALD-7 ones.

Case 6 (complex query): QueGG allows handling questions that are realized by a simple query.⁹ QueGG has limited support for questions for which the corresponding query is complex, such as the following question-query pair:

Who is the mayor of the capital of French Polynesia?

```
SELECT ?uri WHERE { res:French_Polynesia dbo:capital ?x .
?x dbo:mayor ?uri . }
```

10 out of 150 (i.e., 6.6%) questions in QALD-7 are complex queries. The most similar question generated by the automatically induced grammar is ‘*What is the capital of French Polynesia?*’. In our case, none of these questions retrieves all answers

⁹A simple SPARQL query consists of a triple pattern with the predicate `rdf:type`, a triple pattern with the predicate `rdfs:label` and one more triple pattern.

as one or more lexicalization is not correct.

The qualitative evaluation thus shows that in some cases our approach generates correct questions with alternative but valid interpretations that do not match the QALD-7 gold standard. The evaluation thus underestimates the performance of our approach in some cases.

4 Related Work

The automatic acquisition of a lexicon from a corpus is not a new idea. For example, (Zernik, 1989) describes a method to automatically extract lexical entries, where an entry’s semantics is expressed via a semantic template, different configurations in which the syntactic arguments can be organized are recorded etc. Furthermore, *semi-automated semantic knowledge base construction and multilingual lexicon acquisition* was one of the foci of the Penman project, which started in 1978 (Hovy, 1993).

In the context of the task of Automatic Question Generation, one can distinguish between the generation of questions from natural language text, e.g., (Heilman and Smith, 2009; Curto et al., 2012; Zhang et al., 2021) and the generation of questions from a knowledge base, e.g., (Chaudhri et al., 2014; Bordes et al., 2015; Raynaud et al., 2018; Bi et al., 2020).

Question generation from text makes use of manually created rules or trained models that transform a sentence into a question.

Several works mine relation-specific patterns from corpora. The approach M-ATOLL by Walter et al. (Walter et al., 2014) mines textual patterns that denote binary relations between entities. The text corpus is dependency-parsed and natural language patterns are identified via a set of manually defined dependency graph patterns that are matched against the parsed text. The resulting patterns are represented in *lemon* format. In contrast to M-ATOLL, the LexExMachina approach does not rely on a pre-defined set of patterns, but mines the patterns inductively from data (that has not been dependency-parsed).

A good overview about Natural Language Generation (NLG) from RDF can be found in the context of the WebNLG challenge¹⁰ (Gardent et al., 2017). Approaches that tackle this challenge need to be able to carry out tasks such as sentence segmentation, lexicalization, aggregation, and surface real-

¹⁰<https://webnlg-challenge.loria.fr/>

isation. Several of these tasks could make use of an automatically generated lexicon as we generate from LexExMachina rules. Recent work by Moussallem et al. (Moussallem et al., 2020) presents an approach based on an encoder-decoder architecture that is capable of generating multilingual verbalizations. Explicit linguistic knowledge in the form of automatically generated lexica could probably be incorporated into their approach.

The (syntactic) frames we used represent only a small set of possible syntactic frames and overlap with frames defined in VerbNet (Kipper et al., 2008). Our frames are by nature mainly syntactically defined and differ from the more semantic frames defined in FrameNet (Baker et al., 1998).

5 Conclusions and Future Work

We have presented LexExMachinaQA, a framework that allows to evaluate the impact of automatically induced ontology lexica on Question Answering over Linked Data. The framework builds on the LexExMachina approach that mines class-specific association rules over a loosely coupled text and KG dataset. We show how the association rules can be transformed into lemon lexical entries and rely on the QueGG approach to automatically create a grammar from the induced lexicon that can be used to parse questions into SPARQL queries over the corresponding vocabulary. We have evaluated the impact of the automatically induced lexica with respect to the English part of the QALD-7 dataset in terms of F_1 -measure. While our method for lexicon induction yields many reasonable lexical entries that provide a baseline QA performance, our results show that it is not yet feasible to induce a lexicon that comes close to a manually created lexicon by fully automatic means. While not being able to fully replace a manually created lexicon, our method has clearly the potential to contribute to overcoming the lexical gap in Question Answering over Linked Data. In future work we will investigate if the proposed method works for other loosely-coupled datasets beyond Wikipedia/DBpedia and examine if the induced lexical knowledge can be used by QA approaches other than QueGG.

6 Acknowledgements

This research is part of the project eTaRDiS¹¹ (Exploration of Temporal and Spatial Data in Immersive Scenarios), which is funded by the Federal Ministry of Education and Research (BMBF). Basil Ell is partially funded by the SIRIUS centre: Norwegian Research Council project No 237898.

References

- Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics, Volume 1*, pages 86–90.
- Viktoria Benz, Philipp Cimiano, Mohammad Fazleh Elahi, and Basil Ell. 2020. Generating Grammars from Lemon Lexica for Questions Answering over Linked Data: a Preliminary Analysis. In *NLIWOD Workshop*, volume 2722 of *CEUR Workshop Proceedings*, pages 40–55.
- Sheng Bi, Xiya Cheng, Yuan-Fang Li, Yongzhen Wang, and Guilin Qi. 2020. Knowledge-enriched, Type-constrained and Grammar-guided Question Generation over Knowledge Bases. In *COLING 2020*, pages 2776–2786.
- Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015. Large-scale Simple Question Answering with Memory Networks. *CoRR*, abs/1506.02075.
- Sergey Brin, Rajeev Motwani, and Craig Silverstein. 1997. Beyond Market Baskets: Generalizing Association Rules to Correlations. In *ACM SIGMOD 1997*, pages 265–276.
- Vinay K Chaudhri, Peter E Clark, Adam Overholtzer, and Aaron Spaulding. 2014. Question generation from a knowledge base. In *EKAW 2014*, pages 54–65.
- Philipp Cimiano, Paul Buitelaar, John P. McCrae, and Michael Sintek. 2011. LexInfo: A declarative model for the lexicon-ontology interface. *JWS*, 9(1):29–51.
- Philipp Cimiano, John P. McCrae, and Paul Buitelaar. 2016. Lexicon Model for Ontologies: Community Report. In *W3C Community Group Final Report*.
- Sérgio Curto, Ana Cristina Mendes, and Luisa Coheur. 2012. Question Generation based on Lexico-Syntactic Patterns Learned from the Web. *Dialogue & Discourse*, 3(2):147–175.
- Dennis Diefenbach, Andreas Both, Kamal Singh, and Pierre Maret. 2020. Towards a question answering system over the semantic web. *Semantic Web*, 11(3):421–439.

¹¹<https://digital-history.uni-bielefeld.de/etardis/>

- Mohammad Fazleh Elahi, Basil Ell, Frank Grimm, and Philipp Cimiano. 2021. Question Answering on RDF Data based on Grammars Automatically Generated from Lemon Models. In *SEMANTICS 2021*.
- Basil Ell, Mohammad Fazleh Elahi, and Philipp Cimiano. 2021. Bridging the Gap Between Ontology and Lexicon via Class-Specific Association Rules Mined from a Loosely-Parallel Text-Data Corpus. In *LDK 2021*, pages 33:1–33:21.
- Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The WebNLG challenge: Generating text from RDF data. In *INLG*, pages 124–133.
- Sherzod Hakimov, Soufian Jebbara, and Philipp Cimiano. 2017. AMUSE: Multilingual Semantic Parsing for Question Answering over Linked Data. In *ISWC 2017*, page 329–346.
- Michael Heilman and Noah A. Smith. 2009. Question Generation via Overgenerating Transformations and Ranking. Technical report, Carnegie Mellon University.
- Konrad Höffner, Sebastian Walter, Edgard Marx, Ricardo Usbeck, Jens Lehmann, and Axel-Cyrille Ngonga Ngomo. 2017. Survey on Challenges of Question Answering in the Semantic Web. *Semantic Web*, 8(6):895–920.
- Eduard H Hovy. 1993. Natural Language Processing by the Penman Project at USC/ISI. Technical report, University of Southern California Marina del Rey Information Sciences Institute.
- Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2008. A large-scale classification of english verbs. *Language Resources and Evaluation*, 42:21–40.
- John P. McCrae, Dennis Spohr, and Philipp Cimiano. 2011. Linking Lexical Resources and Ontologies on the Semantic Web with Lemon. In *ESWC 2011*, volume 6643, pages 245–259.
- George A Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.
- Diego Moussallem, Dwaraknath Gnaneshwar, Thiago Castro Ferreira, and Axel-Cyrille Ngonga Ngomo. 2020. NABU – Multilingual Graph-Based Neural RDF Verbalizer. In *ISWC 2020*, pages 420–437.
- Ndapandula Nakashole, Gerhard Weikum, and Fabian Suchanek. 2012. PATTY: A taxonomy of relational patterns with semantic types. In *EMNLP 2012*, pages 1135–1145.
- Gennaro Nolano, Mohammad Fazleh Elahi, Maria Pia di Buono, Basil Ell, and Philipp Cimiano. 2022. An Italian Question Answering System based on grammars automatically generated from ontology lexica. In *CLiC-it 2022*.
- Tanguy Raynaud, Julien Subercaze, and Frédérique Laforest. 2018. Thematic Question Generation over Knowledge Bases. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 1–8.
- Martin G. Skjæveland, Daniel P. Lupp, Leif Harald Karlsen, and Henrik Forssell. 2018. Practical Ontology Pattern Instantiation, Discovery, and Maintenance with Reasonable Ontology Templates. In *ISWC 2018*, pages 477–494.
- Daniil Sorokin and Iryna Gurevych. 2017. End-to-end representation learning for question answering with weak supervision. In *Semantic Web Challenges*, pages 70–83, Cham. Springer International Publishing.
- Ricardo Usbeck, Axel-Cyrille Ngonga Ngomo, Bastian Haarmann, Anastasia Krithara, Michael Röder, and Giulio Napolitano. 2017. 7th Open Challenge on Question Answering over Linked Data (QALD-7). In *Semantic Web Evaluation Challenge*, pages 59–69.
- Sebastian Walter, Christina Unger, and Philipp Cimiano. 2014. M-ATOLL: A Framework for the Lexicalization of Ontologies in Multiple Languages. In *ISWC 2014*, pages 472–486.
- Sebastian Walter, Christina Unger, Philipp Cimiano, and Daniel Bär. 2012. Evaluation of a Layered Approach to Question Answering over Linked Data. In *ISWC 2012*, page 362–374.
- Uri Zernik. 1989. Lexicon Acquisition: Learning from Corpus by Capitalizing on Lexical Categories. In *IJCAI 1989*, pages 1556–1564.
- Ruqing Zhang, Jiafeng Guo, Lu Chen, Yixing Fan, and Xueqi Cheng. 2021. A review on question generation from natural language text. *ACM Transactions on Information Systems (TOIS)*, 40(1):1–43.

A OTTR template definition: NounPPFrame

```

1 @prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
2 @prefix ottr: <http://ns.ottr.xyz/0.4/> .
3 @prefix quegg: <http://example.org/quegg#> .
4 @prefix lemon: <http://example.org/lemon#> .
5 @prefix lexinfo: <http://example.org/lexinfo#> .
6
7 quegg:NounPPFrame[
8   ?main_URI, ?canonical, ?singular, ?plural, ?property, ?domain,
9   ?range, ?marker] :: {
10
11   ottr:Triple(?main_URI, rdf:type, lemon:LexicalEntry),
12   ottr:Triple(?main_URI, lexinfo:partOfSpeech, lexinfo:noun),
13   ottr:Triple(?main_URI, lemon:canonicalForm, _:form_1),
14   ottr:Triple(?main_URI, lemon:canonicalForm, _:form_2),
15   ottr:Triple(?main_URI, lemon:synBehavior, _:nounpp),
16   ottr:Triple(?main_URI, lemon:sense, _:sense_ontomap),
17
18   ottr:Triple(_:form_1, rdf:type, lemon:Form),
19   ottr:Triple(_:form_1, lemon:writtenRep, ?singular),
20
21   ottr:Triple(_:form_2, rdf:type, lemon:Form),
22   ottr:Triple(_:form_2, lemon:writtenRep, ?plural),
23
24
25   ottr:Triple(_:nounpp, rdf:type, lexinfo:NounPPFrame),
26   ottr:Triple(_:nounpp, lexinfo:copulativeArg, quegg:arg1),
27   ottr:Triple(_:nounpp, lexinfo:prepositionalAdjunct, quegg:arg1),
28
29
30   ottr:Triple(_:sense_ontomap, rdf:type, lemon:OntoMap),
31   ottr:Triple(_:sense_ontomap, rdf:type, lemon:LexicalSense),
32
33   ottr:Triple(_:sense_ontomap, lemon:ontoMapping, _:sense_ontomap),
34   ottr:Triple(_:sense_ontomap, lemon:ontoMapping, _:sense_ontomap),
35   ottr:Triple(_:sense_ontomap, lemon:reference, ?property),
36   ottr:Triple(_:sense_ontomap, lemon:subjOfProp, quegg:arg2),
37   ottr:Triple(_:sense_ontomap, lemon:objOfProp, quegg:arg1),
38   ottr:Triple(_:sense_ontomap, lemon:condition, _:condition),
39
40   ottr:Triple(_:condition, rdf:type, lemon:condition),
41   ottr:Triple(_:condition, lemon:propertyDomain, ?domain),
42   ottr:Triple(_:condition, lemon:propertyRange, ?range),
43
44   ottr:Triple(_:condition, lemon:propertyRange, ?range),
45
46   ottr:Triple(quegg:arg2, lemon:marker, ?marker),
47
48   ottr:Triple(quegg:of, rdf:type, lemon:SynRoleMarker),
49
50   ottr:Triple(quegg:of, lemon:canonicalForm, _:b1),
51   ottr:Triple(_:b1, lemon:writtenRep, ?marker),
52   ottr:Triple(_:b1, lexinfo:partOfSpeech, lexinfo:preposition)
53 } .

```

Figure 3: Definition of an OTTR template that can be used to create a lexical entry of type NounPPFrame.

Use Cases and Applications

Unifying Emotion Analysis Datasets using Valence Arousal Dominance (VAD)

Ryutaro Takanami

Lancaster University, UK
ryu.takanami212@gmail.com

Mo El-Haj

Lancaster University, UK
m.el-haj@lancaster.ac.uk

Abstract

This paper presents a novel approach to unifying various emotional datasets in Natural Language Processing (NLP) using the Valence Arousal Dominance (VAD) framework. Emotion analysis, which aims to deeply analyse emotions and understand user behaviour, is a complex research area that requires large, standard, and unified datasets. However, the lack of such datasets in NLP has been a challenge in advancing the field. Our approach maps diverse emotions from different datasets into four categories: joy, anger, fear, and sadness using the VAD framework. This process creates multidimensional emotional scores that are consistent across datasets, regardless of the number of emotions included. By unifying these datasets, we were able to train a BERT model on the combined data and improve the performance of emotion detection.

1 Introduction

Emotion detection is a crucial aspect of Natural Language Processing (NLP). There are two main approaches used in NLP for emotion detection: the categorical model and the dimensional model. The categorical model, based on the work of Ekman and Plutchik (Ekman, 1999; Plutchik, 1980), suggests that human emotions can be represented as basic emotions such as joy, sadness, and anger. On the other hand, the dimensional model, based on the work of Russell et al. (Russel, 1980), proposes that emotions can be captured as a point in a multidimensional space, with unconscious elements driving categorical feelings.

While the categorical model provides a straightforward approach to capturing emotions, it has some limitations. For example, it assumes that emotions are discrete categories, and fails to account for the possibility of ambiguity or mixed emotions. The dimensional model overcomes these limitations by representing emotions as points in a

multidimensional space, allowing for the possibility of mixed or ambiguous emotions.

Despite the advantages of the dimensional model, there are still challenges in emotion detection. One of the significant obstacles is the lack of standardised emotional datasets. The available datasets differ in terms of the number of emotions and the types of emotions annotated, making it challenging to train a single machine learning model. To tackle this issue, we propose a method of unifying annotations from different datasets using Valence Arousal Dominance (VAD) to convert labels into a unified VAD score that represents emotions in a 3-dimensional space. This approach provides a more comprehensive understanding of emotions and maximises the use of available datasets to train machine learning models.

In addition to unifying annotations, we address the issue of “weak emotions” by annotating such instances with a neutral VAD score. Sentences that contain conflicting emotions or those that do not exhibit a clear or strong emotional response are referred to as weak emotion sentences. Conventional annotation methods treat sentences with the same emotion equally, but VAD can detect and provide a more nuanced label by assigning a score range instead of a fixed annotation value.

This study has three main objectives:

1. To provide a flexible mapping model that can incorporate different types of emotions from different datasets and unify them into a polarity score of four emotions: joy, anger, fear, and sadness.
2. To improve the accuracy of emotion prediction compared to sentiment polarity detection.
3. To investigate whether the VAD scores can detect neutrality, or what we later refer to as ‘weak emotions’.

In conclusion, our approach to emotion detection provides a more nuanced understanding of emotions in text and helps to overcome some of the limitations of existing methods. By unifying annotations using VAD, we can train machine learning models with greater accuracy and provide more comprehensive insights into the emotions expressed in text.

2 Related Work

One of the earliest emotion detection approaches was the use of lexicons, pre-defined dictionaries of words and their associated emotional valence (Mohammad, 2018). This approach is simple and straightforward, but it is limited by the size and scope of the lexicon, as well as by the fact that words can have multiple meanings and connotations.

Another approach to emotion detection is the use of machine learning algorithms, which can learn to identify patterns in data and predict emotions expressed in text (Pang and Lee, 2004; El-Haj et al., 2016). However, machine learning algorithms require large amounts of labeled data to train effectively, and the lack of standardised emotion datasets has hindered progress in this field. To address this challenge, researchers have proposed unifying different emotion datasets to create a larger, more comprehensive dataset for training machine learning models (Mohammad, 2018; Abdul-Mageed and Ungar, 2017). By mapping varied emotions from different datasets into a common set of categories, these unified datasets can provide a more nuanced understanding of emotions in text, while also allowing for more accurate predictions of emotions.

Other approaches have been proposed to improve emotion detection in text, such as the use of lexicons, pre-defined dictionaries of words and their associated emotional valence (Mohammad, 2018). Another approach is the use of machine learning algorithms, which can learn to identify patterns in data and predict emotions expressed in text (Pang and Lee, 2004). However, machine learning algorithms require large amounts of labeled data to train effectively, and the lack of standardised emotion datasets has hindered progress in this field (Alwakid et al., 2022).

In recent years, there has been a growing interest in using the Valence Arousal Dominance (VAD) model as a way to detect and unify different emotion datasets (Kulkarni and Bhattacharyya,

2021; Luengo et al., 2010). The VAD model captures the affective quality of emotions and offers a more nuanced understanding of emotions in different contexts (Russel, 2003). By mapping different emotions to a common set of VAD scores, researchers can create a unified dataset that is more comprehensive and offers a more nuanced understanding of emotions in text. This approach has the potential to improve the accuracy of emotion detection algorithms and provide a more fine-tuned understanding of emotions expressed in text. To address this challenge, we propose unifying different emotion datasets using VAD, a multidimensional model of emotions that captures valence, arousal, and dominance. By mapping varied emotions from different datasets into four categories - joy, anger, fear, and sadness - we can create multidimensional emotional scores that work across different datasets, regardless of the number of emotions introduced in each. This approach enables us to train machine learning models on a unified dataset, which can improve emotion detection performance and provide more comprehensive insights into the emotions expressed in text.

3 Datasets

This research uses five different datasets mainly focusing on text written in English. Four of the studied datasets are annotated with coarse-grained categorical emotions, while the fifth has VAD labels.

3.1 Stance Sentiment Emotion Corpus (SSEC)

The Stance Sentiment Emotion Corpus (SSEC) is an annotation of the SemEval-2016 Task 4¹ Twitter stance. The corpus contains 4,870 tweets, each paired with eight emotional categories: Anger, Anticipation, Disgust, Fear, Joy, Sadness, Surprise, and Trust. Each tweet was annotated by three to six annotators who were undergraduate students of media computer science (Schuff et al., 2017). SSEC is a widely used dataset in the emotion detection field, and its focus on stance and emotions in tweets makes it particularly relevant to social media analysis.

3.2 SemEval-2018 Task 1 EC

SemEval-2018 Task 1 EC is a dataset of 3,259 English tweets paired with 11 categorized emotion

¹SemEval-2016 Task 4: Sentiment Analysis in Twitter: <http://alt.qcri.org/semeval2016/task4/>

labels: Anger, Anticipation, Disgust, Fear, Joy, Love, Optimism, Pessimism, Sadness, Surprise, and Trust (Mohammad et al., 2018). The dataset was created by having seven annotators label one or more emotions that represent the tweeter’s emotion from a sentence. This dataset is especially valuable for research that focuses on microblogging sites such as Twitter.

3.3 WASSA-2017 Shared Task on Emotion Intensity (WASSA)

WASSA-2017 is a dataset containing about 4,636 manually annotated tweets, categorized into four emotions: Anger, Fear, Joy, and Sadness (Mohammad and Bravo-Marquez, 2017). The authors gathered tweets containing emotional words representing each category. The emotional words were chosen using Roget’s Thesaurus (Chapman et al., 1977). The tweets were manually annotated using crowd-sourcing. WASSA-2017 is a useful dataset for emotion detection research because of its focus on emotion intensity.

3.4 SemEval-2017 Task 4 A (Polarity)

SemEval-2017 Task 4 A is a dataset from the Sentiment Analysis in Twitter challenge (Rosenthal et al., 2017). It contains 11,906 polarity-emotion annotated tweets, with polarity labels of "positive," "neutral," and "negative." Tweets that mentioned any internationally trending events on Twitter were chosen for data collection, and the tweets were annotated with 3-point scales (positive, neutral, and negative) (Rosenthal et al., 2017). This dataset is valuable for research that focuses on sentiment analysis and emotion detection.

3.5 EmoBank

EmoBank is a dataset containing 10,062 sentences paired with continuous VAD labels (Buechel and Hahn, 2017). It is the largest VAD-model text corpus to the best of our knowledge. The sentences were extracted from several online sources, such as blogs, essays, news headlines, and tweets. The dataset was annotated with 5-point scales (ranging from 1 to 5) by crowd workers (Buechel and Hahn, 2017). EmoBank is a valuable resource for emotion detection research because of its large size and its fine-grained VAD labels.

4 Pre-processing

In this section, we detail the pre-processing steps for the training set that will be used as input for our

BERT model.

The BERT model is trained to predict VAD values and to convert these values into categorical labels, based on the required emotion categories. For datasets, such as SemEval-2018 and SSEC (Section 3.1), which are annotated with multiple categorical emotions in a single sentence, we average the VAD values of each emotion to obtain the overall VAD value of that sentence before BERT model training. This is because the VAD value of a sentence should consist of only one score for the training of the machine learning BERT model. For instance, if a sentence is labeled with “joy”, “love”, and “trust”, the VAD scores for each will be something like: joy” = [980, 824, 794], “love” = [1000, 519, 673] and “trust” = [888, 547, 741]. The score of the sentence will then become a three-dimensional score of: Valence $V = (980+1000+888)/3 = 956$, Arousal $A = (824+519+547)/3 = 630$, and Dominance $D = (794+673+741)/3 = 736$.

In the SemEval-2018 and SSEC datasets, multiple labels can be assigned to a single sentence, but not if it is considered neutral. To account for this, we set the intermediate value in VAD space, 500, for sentences without any labels. This is because the range of each axis is a VAD score from 0 to 1000, and in this research, we choose 500 as the moderate strength of the emotion score, or what can be considered as no emotion but falls within the neutral score range, as we demonstrate later in Experiment 2 (Section 5.2).

For the EmoBank dataset (Section 3.5), the pre-existing VAD values range between 1 and 5 points, which is different from our VAD scale. In this work, we use a scale of 0 to 1000 for our VAD score annotations, as the NRC VAD lexicon (Mohammad, 2018) adopted the same scale. To transform the categorical labels in EmoBank to our scale of 0-1000 VAD scores, we use the following formula, where EmoBank-Score is the 1-5 Likert scale score given by the human annotators:

$$VADScore = (EmoBank - Score - 1) / 4 * 1000 \quad (1)$$

We also pre-process the text of the datasets. The majority of the sentences in the datasets are sourced from Twitter, so we pre-process the data by removing mentions and URLs, as they are considered unrelated to expressing emotions. On the other hand, hashtags are retained, as they can help capture cases where emotions are directly included in

the hashtag, such as “#love”.

5 Experimental Work

The experimental work is divided into two phases. In the first phase, we train a BERT machine learning model to predict categorical emotions from the unified representation of multiple datasets using the VAD model. In the second phase, we demonstrate how the model can be adapted to capture what we refer to as “weak emotions” which are neutral emotions found in sentiment datasets such as SemEval-2017 (Section 3.4).

5.1 Experiment 1: Predicting Categorical Emotions

This experiment addresses the first two objectives of the research as outlined in the Introduction (Section 1).

In this experiment, we create a combined prediction model from multiple differently annotated datasets and evaluate if the accuracy can be improved compared to training on individual datasets. The combined model was trained on the EmoBank, SemEval-2018, and SSEC datasets (denoted as “All”). Additionally, separate models were trained for each individual dataset (denoted as “Emo”, “Sem”, and “SSEC”, respectively), as shown in Table 1².

We use the WASSA dataset (Section 3.3) as the test set for this experiment, as each sentence in WASSA is annotated with a single categorical label (joy, anger, fear, or sadness), making it an appropriate dataset to evaluate our models. The results of the BERT model are expressed in terms of VAD scores and are labeled according to the WASSA categories for comparison. This is done by calculating the Euclidean Distance between the predicted VAD scores and the VAD scores of each of the four emotions as labeled in WASSA, and the emotion with the minimum distance becomes the predicted label for a given sentence.

5.2 Experiment 2: Detecting Weak Emotions

This experiment addresses the third objective of the research by investigating whether the VAD scores can detect neutrality (weak emotions).

For this experiment, we use the SemEval-2017 dataset as the testing set, as it has a polarity annotation of positive, neutral, and negative emotions.

²WASSA and SemEval-2017 datasets are used as testing sets and were therefore not included in the training process

The Valence dimension (“V” axis) in VAD is used to predict the polarity emotions. Valence is known to be the most stable dimension in VAD space, where individual perceptions are represented (Hoffman et al., 2012).

We use the VAD score prediction models trained in Experiment 1 to predict the polarity emotions by using SemEval-2017 as the test data. Before comparing the results to the true labels, the predictions are visualised in a scatter plot to show how the combination of multiple datasets increases the representation of emotions estimated by the BERT model (Section 6.2). After predicting the sentiment of a sentence in dimensional space, we convert the predicted V score into categorical emotion labels: positive, neutral, and negative.

Since the test data is annotated with categorical variables, we need to change the predicted V-values, represented by the V-dimension, to categorical values. To do this, we set polarity emotion thresholds for the V-dimension at 300 and 700. It seems reasonable to classify emotions less than 300 as negative, emotions between 300 and 700 as neutral, and emotions above 700 as positive, dividing the V-Score range of 0-1000 into three semi-equal ranges.

6 Results and Evaluation

6.1 Experiment 1

The results of the emotion prediction accuracy for the four emotions (joy, anger, fear, and sadness) tested using the WASSA dataset are shown in Table 1. The results demonstrate that training the BERT model on a combination of different emotion-based datasets (denoted as ‘All’) produces results that are equivalent to training using a single dataset. This suggests that mapping the differently annotated datasets is capable of producing comparable results, and the combination of different datasets did not result in a decrease in accuracy. In particular, when some of the models trained individually (denoted as SEEC) had lower accuracy, the combination of several datasets helped the BERT model learn better how to predict emotions.

	All	SEEC	Emo	Sem
Four emotions	0.44	0.25	0.41	0.45

Table 1: Emotion prediction accuracy.

The number of sentences per emotion is shown

in Table 2 and Figure 1. The imbalance in the data resulted in a bias in emotion prediction, which is expected since anger and joy are the most frequent classes. This can be seen in the results of the models by emotion, shown in Table 3. As a potential solution, future experiments could reduce the number of emotions and increase emotions that are close in the VAD space (e.g., fear and sadness).

	All	SEEC	Emo	Sem
anger	10555	1997	7734	824
joy	3966	1472	1091	1403
fear	2265	1324	270	671
sadness	1405	77	967	361

Table 2: Number of sentences by emotion.

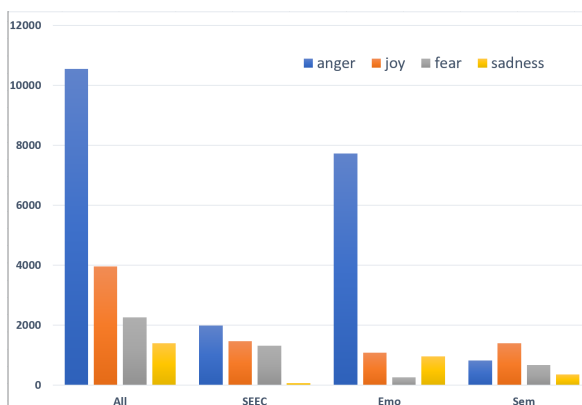


Figure 1: Number of sentences by emotion.

	All	SEEC	Emo	Sem
anger	0.75	0.10	0.96	0.68
joy	0.76	0.34	0.39	0.90
fear	0.12	0.34	0.01	0.15
sadness	0.01	0.34	0.01	0.01

Table 3: Prediction accuracy by emotions.

6.2 Experiment 2

To illustrate that combining datasets has increased the range of emotions that the models can predict, we show a scatter plot of the predictions for each model in Figure 2. The colours in the plot represent the correct label prediction: red for positive, yellow for neutral, and blue for negative. The Y axis is the ID of the predicted sentence, and the X axis is the V score range. None of the models trained on a single dataset were able to categorise all three categories.

It can be seen that the All Model has the richest variety of emotions to predict and is better able to pick up subtle differences in emotions. Moreover, the All Model plot confirms that our threshold values for the V-dimension are reasonable, as the V-score seems to be divided into three categories between around 300 and 700.

	All	SEEC	Emo	Sem
Positive	0.494	0.411	0.0	0.494
Neutral	0.587	0.0	0.482	0.0
Negative	0.571	0.5	0.0	0.442
Average	0.551	0.304	0.161	0.312

Table 4: Accuracy of polarity emotions.

The prediction accuracy of each model for the three categories (positive, negative, and neutral) is examined in Table 4. In terms of prediction accuracy, the All Model has the highest accuracy, demonstrating that the BERT model was able to learn better when a combination of several emotion-based datasets was used. None of the models trained on a single dataset were able to categorise all three categories with consistent accuracy, as confirmed by the scatter plots in Figure 2.

7 Conclusion

The results of Experiments 1 and 2 in this study demonstrate the benefits of training with larger emotion-based datasets. By transforming these datasets using the Valence Arousal Dominance (VAD) framework, our findings suggest that it is possible to predict a wider range of emotional expressions. The results of the polarity analysis in Experiment 2 further support this conclusion.

As future work, it is expected that increasing the number of datasets used in training will result in improved accuracy of emotion prediction. The experiments conducted in this study also showed that it is possible to predict weak emotions, which are often overlooked by conventional sentiment analysis models.

References

- Muhammad Abdul-Mageed and Lyle Ungar. 2017. Emonet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of ACL'17*, pages 718–728.
- Ghadah Alwakid, Taha Osman, Mahmoud El Haj, Saad Alanazi, Mamoona Humayun, and

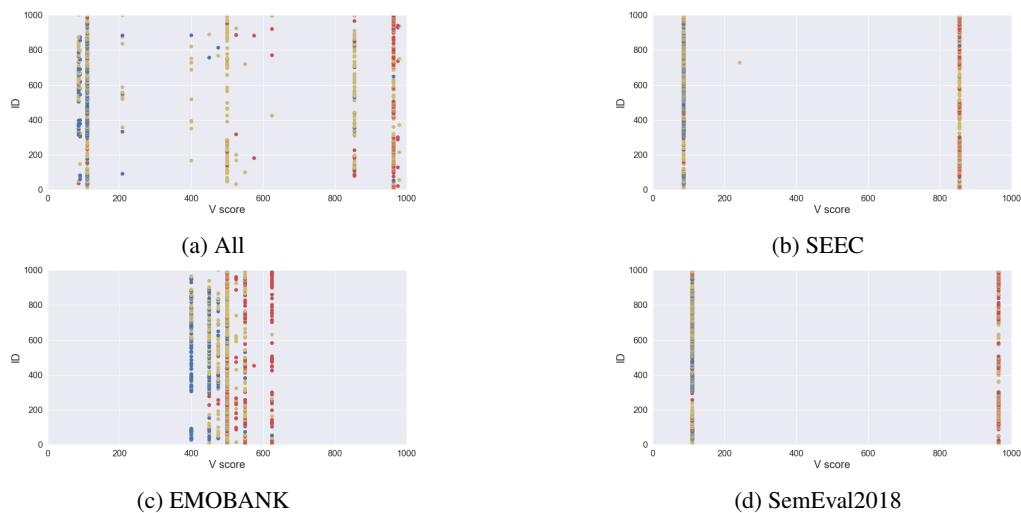


Figure 2: Predicted V scores.

Najm Us Sama. 2022. Muldasa: Multifactor lexical sentiment analysis of social-media content in nonstandard arabic social media. *Applied Sciences*, 12(8):3806.

Seven Buechel and Udo Hahn. 2017. Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis.

Robert. L. Chapman, Peter. Mark. Roget, et al. 1977. *Roget's international thesaurus*. Crowell.

P. Ekman. 1999. *Basic Emotions*. In *Dalgleish, Tim and Powers, M. J. (eds.), Handbook of Cognition and Emotion*, volume 1. Wiley.

Mahmoud El-Haj, Paul Edward Rayson, Steven Eric Young, Martin Walker, Andrew Moore, Vasiliki Athanasakou, and Thomas Schleicher. 2016. Learning tone and attribution for financial text mining.

Holger Hoffman, Andreas Scheck, Timo Schuster, Steffen Walter, Kerstin Limbrecht, Harald C. Traue, and Henrik Kessler. 2012. [Mapping discrete emotions into the dimensional space: An empirical approach](#). In *IEEE SMC'12*, pages 3316–3320.

Manasi Kulkarni and Pushpak Bhattacharyya. 2021. Retrofitting of pre-trained emotion words with vad-dimensions and the plutchik emotions. In *Proceedings of ICON'21*, pages 529–536.

Iker Luengo, Eva Navas, Igor Odriozola, Ibon Saratxaga, Inmaculada Hernaez, Inaki Sainz,

and Daniel Erro. 2010. Modified Itse-vad algorithm for applications requiring reduced silence frame misclassification. In *LREC*.

M. Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words.

S. Mohammad and F. Bravo-Marquez. 2017. *Wassa-2017 shared task on emotion intensity*.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. *Semeval-2018 task 1: Affect in tweets*.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *arXiv preprint cs/0409058*.

R. Plutchik. 1980. *Emotion: Theory, research, and experience*, volume Vol. 1, Theories of emotion. Academic Press.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. *Semeval-2017 task 4: Sentiment analysis in twitter*.

J. Russel. 1980. A circumplex model of affect.

J. Russel. 2003. Core affect and the psychological construction of emotion.

Hendrik Schuff, Jeremy Barnes, Julian Mohme, Sebastian Padó, and Roman Klinger. 2017. Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus.

Challenges and Solutions in Transliterating 19th Century Romanian Texts from the Transitional to the Latin Script

Marc Frincu

Department of Computer Science School of
Science and Technology Nottingham Trent
University, UK
marc.frincu@ntu.ac.uk

Simina Frincu and Marius E. Penteliuc

Department of Computer Science
Faculty of Mathematics and Computer
Science West University of Timisoara,
Romania
ioana.frincu@e-uvv.ro
marius.penteliuc@e-uvv.ro

Abstract

During the 19th century, the Romanian script has undergone a massive yet uneven transition from the Cyrillic to the current Latin alphabet. The amount of existing literature written in that script as well as the problems it poses for OCR and transliteration engines make the problem highly challenging from a Big Data perspective. In this paper, we discuss the issues and propose and test a machine-learning solution trained on small datasets using either transfer learning from Latin/Cyrillic or from scratch.

1 Introduction

Until the early 19th century Romanian texts were written in the Romanian Cyrillic Script (RCS) containing around 43 characters, a version of the script different from the standard Church Slavonic or Russian scripts. By the end of the 18th century, the first attempt to simplify the script to 38 letters comes from (Văcărescu, 1787). In 1823, to meet didactic purposes, I. H. Rădulescu highlights the same necessity for a reduced 30-letter script. Nonetheless, the reforms (to optimize or simplify the alphabet) proposed over time by different cultural figures ((Iorgovici, 1799), (Budai-Deleanu, 1812), (Diaconovici Loga, 1818), (Rădulescu, 1828) or (Pleșoianu, 1828)) remained until the official adoption in 1860 at the stage of individual and unofficial initiatives. The drive behind the change pertained also to the desire to reassert the Latin values of Roman origin of Romanian people, in the context of the sociopolitical events unfolding across Europe.

The alphabet transition did not occur abruptly (several versions coexisted between authors, publishing houses, editors, and regions) or simultaneously across the historical Romanian regions of Wallachia, Moldavia, and Transylvania (Cazimir, 2006). Yet, all these versions were based on the Simplified Modern RCS and a variable, increasingly higher in time, proportion of Latin letters.

The alphabet transition is extremely interesting for researchers studying the diachronic evolution of the language and encompasses thousands of typed manuscripts (some not digitized) written in various transitional script versions. Understanding these manuscripts starts with scanning, converting the scanned images into digital documents, reading the documents, and analyzing their content based on the researcher's objectives. The OCR process is the main driver behind digitization and it is here that existing software fails to recognize the Romanian Transitional Script (RTS), partly due to the quality of the original paper. Thus, Machine Learning (ML) models are better suited to handle different types of scanned documents and script versions (e.g. font type, publishing house, region). Tools like Transkribus (Miloni, 2020) and the open-source Tesseract (Smith, 2007) have been designed for such cases. However, the accuracy of the models depends on the volume and variety of training data. This turns the process into a Big Data problem where most data preparation is manually handled before training and testing the models.

The RTS digitization process consists of: (1) conversion to RTS characters (preserving the original text); and (2) interpretative phonetic transcription into Latin (transforming the original text into a version readable by modern researchers).

Following the CRISP-DM methodology (Wirth and Hipp, 2000) focusing on data understanding and preparation, we compare 2 approaches that lead to promising Tesseract models trained on few data and digitized to Latin/RTS.

2 Related Work

2.1 RTS Studies

Several studies (Cazimir, 2006), (Boerescu, 2014) refer to a formal “modernization” of the RCS after 1830. More precisely, the typographical Cyrillic capital letters were “carved” using the Latin-type

model, namely redesigned to resemble the Latin letters. Thus, the graphical overlay can be explained by the fact that some Latin capitals were identical in sound and meaning to Cyrillic ones (A, E, I, K, M, O, T). In contrast, others coincided graphically yet differed semantically (Cyrillic B for V, C for S, H for N, P for R, X for H). The purpose of this initiative, sometimes leading to surprising approaches (cf. Fig. 1), was to prepare the readers for the alphabet transition about to take place.

Two methods can be used to render a text written in the Cyrillic alphabet into Latin: transliteration or interpretative phonetic transcription. The first implies a one-to-one mapping (IRS, 1997), a character-by-character conversion, more precisely each Cyrillic letter to be replaced with one and the same Latin letter, irrespective of the context within the converted system. The latter demands an accurate determination of the phonetic values represented by the Cyrillic letters (Ursu, 1960). Both methods present disadvantages and are not entirely satisfying. The shortcomings of the transliteration method (the Latin script counts fewer letters than the Cyrillic one, therefore the same Latin letter with various diacritics attached to it can stand for two or even three Cyrillic letters) and the difficulties of the phonetic transcription lead to a hybrid approach and a composite solution.

2.2 Automated Transliteration and ML

Most works on automating the RTS transliteration were done by researchers in Rep. Moldova as the script was used both there and in Romania.

Boian et al. (2014) mention at least 7 versions for RTS, provide a first look into the challenges of transliterating RTS, and mention that except for one (for which they used a replacement), all RTS characters are available in Unicode (UTF-16). The reported percentages using the proprietary paid AB-BYY FineReader with and without training range between 63 and 95.4%.

Cojocaru et al. (2016) identify challenges when transliterating older scripts using OCR tools not supporting them. They mention the RTS versions and 3 existing fonts that cover the RTS characters, focusing on every script version starting from the RCS to the Moldavian Cyrillic Script in use in Rep. Moldova in the 20th century. Their approach targets ABBYY FineReader and experiments use both one-to-one mapping and rule-based context transliteration but they do not provide the number

of tested documents and errors only showing the upper limit of 96% in terms of accuracy without providing an error distribution plot or mean value.

Demidova and Burteva (2017) also focus on historical documents written in RTS. In addition to the previous paper, they briefly describe their transliteration module written in the Java language but do not present comprehensive results for their experiments. It is unclear if the module only transliterates already digitized documents or goes through the entire OCR process too. The reported accuracy is 99% without mentioning the dataset size.

Gifu and Plamada-Onofrei (2017) focus on creating a corpus of transliterated text to facilitate the automatic recognition and interpretative transcription from RTS to the modern Latin script.

While focusing on the older RCS and not on RTS the work of Burlacu and Rabus (2021) is interesting as it uses Transkribus, another online tool with limited free access that we considered. Their study involves handwritten manuscripts and the provider CER (Character Error Rate) is around 10%. We note here that Transkribus requires thousands of words for training its models (the authors used up to 30,900 words for one of their models) which calls for a significant upfront effort.

Compared to existing work using paid software and briefly discussing results, we focus on the open-source Tesseract Engine proposing a 2-phase automatic transliteration process: (1) to Latin/RTS characters followed by an interpretative phonetic transcription; (2) a corpus-based correction to improve the accuracy of the final text in Latin script.

Figure 1: Example of transitional characters invented and used in some of his texts by I. H. Rădulescu to visually ease the alphabet transition and familiarize readers with the Latin script (Cazimir, 2006).

3 Current Challenges

3.1 Processing

When dealing with large collections of historical books several preprocessing and processing challenges occur. Foremost, these documents must be digitized so that OCR and transliteration tools

can generate documents readable by present-day researchers (and the general public for that matter). This phase is largely manual and implies a significant amount of time and effort. Next, the ML model must be trained and validated on a relevant data sample covering the problems identified in Sec. 3.2. This process requires a manual transliteration of the training and validation data sets that will act as ground truth in the training and validation steps of the model. Finally, the best models need to be tested on a test data set which must also be manually transliterated to have a ground truth for automatically computing the errors. Our experiments have shown that the manual process takes around 30 minutes for 1 page with the time spent improving as users get accustomed to the RTS.

While a lot of manual transliteration is required, the computational and storage space also becomes an issue. Depending on the image format a scanned color page takes between 100 KB (jpeg) and ≈ 2 MB (tif) with the transliterated text file taking ≈ 2 KB. This means that a single book of 100 pages will occupy 10-200 MB. When it comes to thousands of books from the alphabet transition period storing all the data is a concern too. The Tesseract OCR process is fast taking between 0.18-0.59 secs per page while the training of a k-fold model ranges from 13.5-17.2 to 613-2,200 secs per fold times the number of folds and iterations (cf. Sec. 5).

3.2 OCR and Transliteration

All the titles printed between 1828-30 and 1860 used for the validation, training, and test phases have been selected by applying the “transitional alphabet” filter in the electronic catalogs of the libraries hosting rare/old book collections. The different degrees and types of paper alterations impact the ML-based OCR process and demand for additional processing of the images subject to further training. Hence, we have aimed at selecting scanned pages bearing a wide variety of physicochemical and a few physicommechanical types of age-related damage. These include (e.g., Fig. 3):

- 1) **Thick binding, ripped stitching, or broken spine** which led to poor quality scans, i.e. text deformations (crooked/bent text).
- 2) **Creases, folds, wrinkles, and undulation** due to humidity changes.
- 3) **Moisture halos, ink discoloration, foxing, burns, tearing, grease stains, glue residue.**
- 4) **Presence of post-printing elements**, e.g. sig-

natures, institutional stamps, inventory numbers, notes in pencil/soluble ink/pen, etc.

We have also considered printing aspects likely to make the OCR process more difficult, some of which needed to be tackled individually:

- 1) **Typesetting** using various inks (usually black or red), typefaces, and fonts (e.g. drop caps, enlarged and illustrated initial letters meant to mark the beginning of a book/chapter/section).
- 2) **Text visible from the verso** of the sheet due to thin physical support.
- 3) **Two-column versus single-column** printing approach, framed and/or manually underlined text.
- 4) **Glossing with marginal/interlinear notations**, either numbered or marked by typographical symbols and sometimes separated from the main text by a separator line.

4 Proposed Solution

The existing literature on RTS transliteration / phonetic transcription is lacking a clear description of the datasets used for training and testing and relies in some cases on paid software (cf. Sec. 2). We present our approach for testing and assessing two scenarios, using either a Latin or RTS baseline for training through transfer learning or from scratch the models in the open-source Tesseract 5.2.

4.1 Improving Transliteration Accuracy

Transliterating from RTS to Latin poses several challenges including character ambiguity (cf. Sec. 5) and phonetic transcription (rule-based approach depending on the subsequent characters). As Tesseract can only perform OCR the phonetic transcription must take place afterward and therefore its efficiency depends on the accuracy of the OCR process. This second step requires replacing the transliterated character with another single or group of characters based on context. E.g., ч is interpreted as: c if followed by e or i ; ce if followed by a ; ci otherwise (Cojocaru et al., 2016).

To assess Tesseract’s ability to accurately perform OCR we propose two approaches. Each uses a different baseline, Latin or RTS. The reason is that many documents have mixed Latin and RTS texts causing the phonetic transcription to fail as the text sections are neither automatically nor manually tagged with the script they use. For instance, the title can be in Latin, while the text itself is in RTS (cf. Fig. 4). In such a case, Latin c for instance is unnecessarily (and wrongly) phonetically analyzed

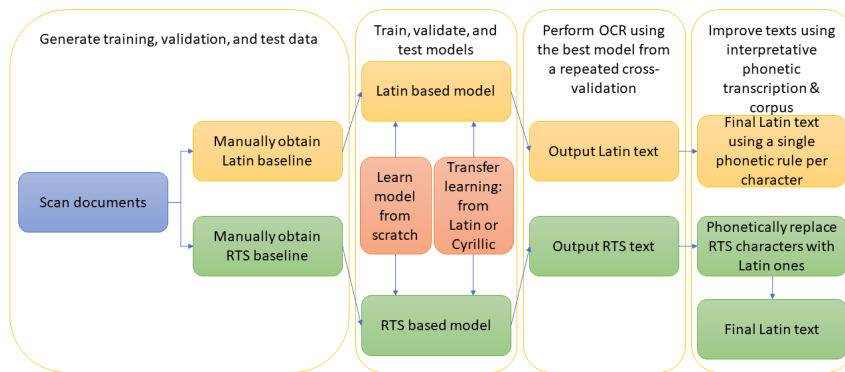


Figure 2: Overview of the two proposed approaches.

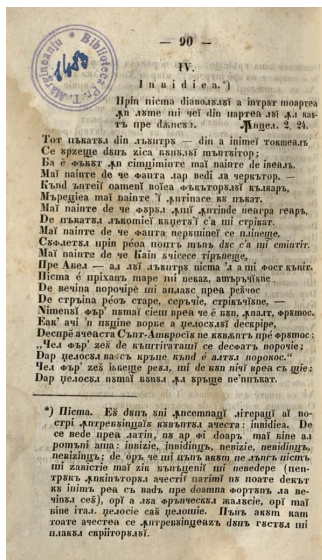


Figure 3: Glossed text from 1847 written in RTS with notation separated by a line and marked by an apostrophe. Also present moisture halo, deformed text, institutional stamp, and text visible from verso.

in the title. Due to constraints, for the Latin baseline (impossible to interpret Latin characters), the phonetic transcription is based on a single selected rule, e.g., ч → ci. The RTS approach focuses instead on Cyrillic but it too can misinterpret Cyrillic characters for Latin ones. The key difference is the output from Tesseract and the fact that the RTS approach performs the phonetic interpretation and transliteration in one step during the Latin conversion ignoring any Latin characters. A major issue during transliteration is the character similarity between scripts, e.g. Latin C and Cyrillic С – Latin S (cf. Sec. 2) which can be solved by providing the model with enough and varied training data.

Both texts are improved by using a corpus from the training and validation documents. At the moment, candidate words are selected based on the

Levenshtein distance (cf. Sec. 5) but other methods (e.g., based on n-grams) are possible.

5 Experiments

CER is a metric for assessing OCR quality. There is no consensus on what a good CER value is. Burlacu and Rabus (2021) mention a rate less than 5% (or <10% for a text to be manually corrected in a time less than that needed for manual transliteration), while (Halley, 2009) mentions 2% as a good result and 10% as average.

$$CER = (S + D + I)/N \tag{1}$$

where $S + D + I$ represents the Levenshtein distance and corresponds to the number of substitutions (S), deletions (D), and insertions (I) required to make two texts equal; and N is the length of the baseline (ground truth) text. Tesseract computes by default BCER (Bag of Characters Error Rate):

$$BCER = \sum_{i=1}^{no_{words}} \left(\frac{S_i + D_i + I_i}{N_i} \right) / no_{words} \tag{2}$$

It can be shown that $BCER \geq CER$. CER is a function of the overall quality and BCER penalizes text where the error is less uniformly distributed.

5.1 Setup

To test our approaches we collected a corpus of over 3,000 pages from distinct documents (1837-1861) from the Timisoara Central University Library. In this paper, we used a small subset of 30 pages of 24,148 characters (out of which 64.4% are Cyrillic). The Cyrillic characters' percentage per page was $61.8 \pm 9.6\%$. Each page was scanned and manually converted/transliterated (into RTS/Latin) to obtain two baselines. Unfortunately, the existing corpus from Gifu and Plamada-Onofrei (2017) does not



Figure 4: RTS text in which the title is written only with Latin characters. 1850 (top) and 1844 (bottom).

include the scanned pages making it unusable for our experiments. While small, our dataset allowed us to assess Tesseract’s potential to create good models from a few data. Tesseract uses LSTM deep network architecture. We trained our models either from scratch or starting from existing models through transfer learning (Latin or Cyrillic) and stopped the training after 10,000 iterations. One test page containing 745 characters (out of which 71.14% Cyrillic) was used.

Several model validation scenarios were used:

(S1): Initial 5-fold cross-validation of a randomly picked 15-page dataset for creating a model and using a single test page.

(S2-k): A repeated k-fold cross-validation for creating the model where $k \in \{3, 10, 29\}$. One page was omitted as it was unreadable by Tesseract.

We name the models for each baseline S1-L and S2-k-L, respectively S1-RTS and S2-k-RTS. Our aim is to assess if there are differences in CER when performing the ML-based conversion into RTS (followed by a Latin transliteration) or directly transliterating into Latin (Romanian). We also evaluated if using a corpus comprising the trained data can improve CER. We considered two cases, one containing a corpus from various regions and publishing houses, and one from Rădulescu’s publishing house. Color pages and their b/w counterparts were tested separately. As results were better for color pages we present exclusively these.

CER was computed using the Levenshtein distance (Eq. 1) after removing all spaces from baseline and transliterated texts. The BCER value was computed automatically by Tesseract.

5.2 Results

The test service and data are available online¹. Table 1 shows the results of our experiments. For repeated k-fold cross-validation we show the best results ($k=3$). As the number of folds increased both CER and BCER dropped indicating the sensitivity of our models to the small dataset. For Latin, the best model started from an existing Latin model enriched with our dataset and provided a CER=1.8 for S2-Lat. For RTS the best model was also one trained by enriching a Latin model and achieved a CER=17.7 for S2-3-RTS. The models starting from Cyrillic performed slightly worse for RTS. The reason for the high CER can be traced to the similarity of vocals in Cyrillic and Latin, e.g., a – a; e – e; i – i; o – o. As CER was computed based on the Unicode value it produced high values as most Cyrillic vocals were identified as Latin characters. Ignoring them reduces the number of wrongly classified characters by 52–59% depending on the base model. The RTS model trained from the Cyrillic base model performed slightly worse than the Latin-derived RTS model, partly due to wrongly classifying more Latin (e.g., *t*) characters. Improving these misclassifications would make the Cyrillic-derived model better. This would be ideal due to the non-existing phonetic transcription available for the Latin baseline. Overall, the Latin base model misidentified 52 characters compared to 54 by the Cyrillic-based one.

When using the training corpus to reduce CER for the test page we noticed that this happened only for a single model in the 5-fold and led to a 0.1% improvement. When using a model trained only for Rădulescu (2nd fold of a 3-fold) no CER improvement was noticed except when assuming that the corpus already contained all the words in the test page (0–2.3%). The reason is that the Levenshtein distance is unsuited for the task as it compares the words in terms of changes in characters not semantically. Even assuming a corpus containing the correct test page does not lead to a $CER = 0$ across the board as the OCR process can introduce additional erroneous words (cf. Sec. 3).

From a formal, script-related perspective, a typology of the recognition failure cases consists of: 1) Errors due to the graphic similarity between letters, accented letters mistaken for other letters, or for numbers resembling them visually, e.g., i – î,

¹<https://transitional-romanian-transliteration.azurewebsites.net/>

Target	Latin				RTS					
From	scratch		Latin		scratch		Latin		Cyrillic	
Scenario	S1	S2-3	S1	S2-3	S1	S2-3	S1	S2-3	S1	S2-3
CER %	–	10.6	2.5 ± 0.4	1.8	56.0 ± 7.3	19.4	27 ± 4.0	17.7	33 ± 2.0	19.6
BCER %	–	15.5	4.5 ± 1.3	8.2 ± 0.7	21.8 ± 6.4	20.7	13.9 ± 2.4	13.8	13.5 ± 1.8	15.5

Table 1: Test results for our two approaches including the model we started from, scenario, and error metrics.

$n - \pi$ (p), $m - \text{III}$ ($\$$), $\acute{i} - l$, $\acute{o} - \delta$, $k - \kappa$ (c/ch/k).

2) Errors caused by a lack of previous training. E.g. Greek symbols, and Latin script fragments.

3) Errors encountered in transliterating certain double consonants. It was noted that while double s and double n were 100% recognized, double l was always rendered faultily.

6 Conclusions

In this paper, we addressed the problem of transliterating 19th century Romanian texts. We proposed a solution based on Tesseract and demonstrated it on two targets: Latin and RTS. Initial results for Latin on a small dataset are very good but phonetically interpreting the text is challenging due to the mix of Latin and RTS phrases in some documents. Results for RTS indicate the need for a richer training dataset due to the similarity between Latin and Cyrillic characters. Future work will consider these aspects. We will also assess other methods for corpus-based text improvement such as n-grams and TF-IDF.

Acknowledgements

This work was supported by a grant of the Romanian Ministry of Research, Innovation and Digitization, CCCDI - UEFISCDI, project number PN-III-P2-2.1-PED-2021-0693, within PNCDI III.

References

- P. Boerescu. 2014. *About the History of Romanian Writing (Ro.)*. Editura Academiei Române.
- E. Boian, C. Ciubotaru, S. Cojocaru, A. Colesnicov, and L. Malahov. 2014. Cultural and historical heritage digitization, recognition and conservation. *Akademios*, 1(32):61–68.
- I. Budai-Deleanu. 1812. *The fundamentals of Romanian grammar (Ro.)*. În Buda Sau tipărit la Crăiasca Universităţii Tipografie.
- C. Burlacu and A. Rabus. 2021. [The digitization of documents written in the romanian cyrillic script by using transkribus: new perspectives \(ro.\)](#). *Diacronia*, (14):(1–10).
- S. Cazimir. 2006. *Transitional Alphabet 2nd ed. (Ro.)*. Humanitas.
- S. Cojocaru, L. Burteva, C. Ciubotaru, A. Colesnicov, V. Demidova, M. Ludmila, M. Petic, T. Bumbu, and S. Ungur. 2016. On technology for digitization of romanian historical heritage printed in the cyrillic script. In *Procs. of the Conference on Mathematical Foundations of Informatics*, pages 160–176.
- V. Demidova and L. Burteva. 2017. The digitization and presentation of the romanian transitional cyrillic script (ro.). *Akademios*, 1:24–29.
- C. Diaconovici Loga. 1818. *Orthography or the correct spelling to guide Romanian language writers (Ro.)*. În Crăiasca Typografie a Universitatii Ungariei.
- D. Gifu and M. Plamada-Onofrei. 2017. Developing a technology allowing (semi-) automatic interpretative transcription. In *TDDL/MDQual/Futurity@TPDL*.
- R. Halley. 2009. [How good can it get? analysing and improving ocr accuracy in large scale historic newspaper digitisation programs](#). Last accessed 01 December 2022.
- P. Iorgovici. 1799. *Observations on the Romanian language (Ro.)*.
- IRS. 1997. *Information and documentation. Transliteration of Cyrillic characters into Latin characters. Slavic and non-Slavic languages (Ro.)*. IRS.
- N. Miloni. 2020. *Automatic transcription of historical documents: Transkribus as a tool for libraries, archives and scholars*. Ph.D. thesis, Uppsala University.
- Gr. Pleşoianu. 1828. *Primer to facilitate learning for children (Ro.)*.
- I.H. Rădulescu. 1828. *Romanian Grammar (Ro.)*.
- R. Smith. 2007. [An overview of the tesseract ocr engine](#). In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 629–633.
- N.A. Ursu. 1960. The problem of interpreting romanian cyrillic texts written around 1800 (ro.). *Limba Română*, 3:33–46.
- I. Văcărescu. 1787. *Observations or considerations on the rules and regulations of Romanian grammar (Ro.)*.
- R. Wirth and J. Hipp. 2000. Crisp-dm: towards a standard process model for data mining. In *Practical application of knowledge discovery and data mining*, pages 29–40.

A variationist analysis of two French attitude expressions: *je pense* and *je crois*

Delin Deng
University of Florida
ddeng@ufl.edu

Abstract

Previous research has reported that *je pense* and *je crois* are interchangeable when used as attitude expressions (see, for example, Gosselin, 2015, 2018; Rendulić & Kanaan-Caillol, 2016; Angot, 2021). This paper conducts a variationist analysis of a corpus collected in Orléans, France, in 2008 to examine the variable use of *je pense* and *je crois* as attitude expressions among 10 French native speakers. The results of the regression analysis indicate that all three tested linguistic factors are significant, while no extralinguistic factors show significance in relation to this variable use. On one hand, we confirm that the variable use of *je pense* and *je crois* is not a change in progress in the apparent time. On the other hand, our analysis provides further insights into how the use of these variables is not only conditioned by semantic or pragmatic differences but also by the linguistic contexts in which the particles are used.

1 Introduction

There is a substantial body of literature on the French attitude expressions or personal opinion expressions *je pense* (I think) and *je crois* (I believe) (see, for example, Gosselin, 2015, 2018; Rendulić & Kanaan-Caillol, 2016; Angot, 2021). Various terms have been used to refer to them in the past decades.

Simons (2007: 1034) referred to verbs such as *see*, *hear*, *think*, *believe*, *discover*, and *know*, as clause-embedding verbs. According to him, “the

embedded clause carries the main point of the utterance, while the main clause serves some discourse function”. Gachet (2014: 147) posited, “the initial phrase (*je crois* ‘I believe’) is a peripheral clause, and the following clause is the main clause.” Gosselin (2018: 180) referred to them as “verbal expressions of personal opinion”. He argued that, instead of taking the traditional unitary point of view, each of them has “a specific meaning” and “are not always substitutable for one another”. Angot and Hansen (2021) analyzed *je pense*, emphasizing that it is a similar construction to *je crois* as pragmatic markers, which “fulfil both interpersonal, face-related functions and discourse-organizational functions” (Angot & Hansen, 2021: 1).

In this article, we aim to explore, using a quantitative method, the linguistic and extralinguistic factors that influence the choice between *je pense* and *je crois* in French native speech. The current study seeks to address the following questions: If we consider *je pense* and *je crois* to be functionally interchangeable as attitude verbs introducing an embedded clause, what sets them apart in terms of the linguistic environments in which they appear within an utterance? Is the choice between the two forms also conditioned by certain extralinguistic factors in native speech? Is it an ongoing change or a completed change?

Therefore, this article will be structured as follows: first, the long-stand puzzled set by attitude expressions will be reviewed and discussed. Second, methodology of the current work, including information on the corpus used, speakers, tokens, linguistic and extralinguistic factors to be examined

in this article as well as the statistical analysis, will be presented. Following this, the statistical results will be tabulated and discussed in detail. Lastly, a conclusion of the current work, its limitation as well as future implications will be laid out.

2 Puzzle

Attitude expressions, such as *je pense* and *je crois*, attracted so much attention as they challenge our existing understanding of the relations between propositions and their truth-values. Let us consider the following scenario:

I just arrived at the company and settled down at my desk. I have not chatted with anyone yet. At that time, John entered my office and asked me if Marie had already arrived at the office. Usually, Marie arrives at the office earlier than H el ene. So, I responded:

(1) Si H el ene est dans le bureau, je crois que Marie est d ej a arriv ee.

‘If H el ene is in the office, I believe that Marie has already arrived.’

(2) H el ene est dans le bureau.

‘H el ene is in the office.’

(3) Je crois que Marie est d ej a arriv ee.

‘I believe that Marie has already arrived.’

If we define:

p= H el ene est dans le bureau. ‘H el ene is in the office.’

q= Je crois que Marie est d ej a arriv ee. ‘I believe that Marie has already arrived.’

My sentence could be represented as:

$p \rightarrow q$

By modus ponens, if $p \rightarrow q$ is true, and p is true, then q must be true.

However, this poses some problems. ‘A person believes p’ is true only if ‘that person knows p’ is true. Nevertheless, given the scenario we provided, if I had just arrived at my office and had not chatted with anyone else, it would not be clear how I could know whether Marie had arrived. Therefore, it seems that we encounter a situation where modus ponens has led us from true premises to a false conclusion. So, how do we address this puzzle?

Some earlier studies proposed that expressions such as *je crois* play the role of a mitigator, which attenuates the certainty of a statement (See, for example, Benveniste, 1996; Borillo, 1982; Vet, 1994). Following this line of reasoning, our sentence *Je crois que Marie est d ej a arriv ee* ‘I believe that Marie has already arrived’ could be understood as indicating some degree of uncertainty. If this is indeed the case, then consider the following sentences:

(4) Patrick cro it que Marie est d ej a arriv ee.

‘Patrick believes that Marie has already arrived.’

(5) Patrick est convaincu que Marie est d ej a arriv ee.

‘Patrick is convinced that Marie has already arrived.’

Does (4) imply that Patrick is not sure of it? Clearly not. We would not deny that (4) would be compatible with (5). (4) implies that Patrick is convinced that Marie has already arrived. But why (3) implies that I am not sure of *Marie est d ej a arriv ee* ‘Marie has already arrived.’, but (4) implies that he is sure of *Marie est d ej a arriv ee* ‘Marie has already arrived.’? Why do they seem to be in contradiction?

Gosselin (2018: 182) highlighted that *croire* ‘believe’ “is a verb that only indicates that the speaker of the utterance does not presuppose the content of the complement”. He then argued that this could be explained by the *logic of conviction* proposed by Lenzen (2004). When a person believes p (p stands for any proposition), what that person really believes is not p itself but rather knowing p. In other words, this representation should not be expressed as B(a, p) but as B(a, K(a, p)) (a: person; p: proposition; B: believe; K: know). As Lenzen pointed out, “knowledge and conviction are subjectively indiscriminable in the sense that person a cannot tell apart whether she is ‘only’ convinced that p or whether she really knows that p” (Lenzen, 2004: 973). The same would apply to *je pense*. When a person ‘thinks’ p, they think that they know p.

So where do ‘he is sure of p’ and ‘I am not sure of p’ come from, as in (3) and (4)? As reasoned by Gosselin (2018: 183), “if a speaker uses non-factive epistemic expressions, like *je crois/suis certain(e)/persuadé(e)/convaincu(e) que*, ‘I believe/am sure/persuaded/convinced that’, it triggers an implicature from the utterance. The interpreter will think that if the speaker has used not just *p* or *je sais que p*, it is because she does not believe that she knows that *p* and therefore she is not really convinced that *p*, hence the systematic mitigation effect, which may seem contradictory to what the statement says literally.”

Now we have resolved our puzzle. However, another question arises: it appears that in oral French, native speakers use *je pense (que)* and *je crois (que)* in a quasi-interchangeable manner. Numerous previous studies have qualitatively discussed the semantic or pragmatic differences between the two. Is there any quantitative evidence that can shed light on their differences? Do native speakers tend to prefer one over the other in specific circumstances? If so, what are these circumstances? These are the questions that we aim to address with the current work.

3 Methodology

3.1 Corpus and data

The corpus we will use for the current study is ESLO 2 (Enquêtes Sociolinguistiques à Orléans: <http://eslo.huma-num.fr/index.php>, Baude and Dugua, 2011). It is an online corpus comprising sociolinguistic interviews with native speakers of French in Orléans, a city located approximately 120 km south of Paris. The variety of French spoken in Orléans is closer to the central French variety, which is considered to be accentless and closer to standard French. The ESLO 2 corpus was initiated in 2008 and is still under development. It includes various modules, ranging from interviews to questionnaires. For the current study, we will solely use the interview module, which consists of 81 interviews conducted in French. All interviews were

transcribed using Transcriber (Barras et al., 2001) and can be downloaded from the website. We imported all Transcriber files into Elan (2021) to identify the relevant occurrences and their surrounding linguistic environments for our final analysis.

3.2 Speakers

For this study, we randomly selected 10 speakers in ESLO 2 (5 females and 5 males). Table 1 provides detailed information on these 10 speakers, including their assigned ID (represented by two letters followed by one or two digits), gender, age at the time of the interview, and socioeconomic status (SES).

Speakers	Gender	Age	SES
QF28	m	58	high
MC59	m	81	low
GK11	m	31	high
BV1	m	23	low
BT17	m	28	middle
LX10	f	65	low
KC3	f	23	low
HT398	f	33	high
AN43	f	39	high
AJ38	f	21	middle

Table 1: Detailed information on 10 speakers.

From Table 1, we observe that the age range of the selected speakers is relatively representative, encompassing the younger, middle-aged, and older generations. Regarding SES, we categorized the speakers into three main groups based on the information provided in their ESLO profile: low (including blue-collar workers, manual workers, and the unemployed), middle (comprising technicians, supervisors, white-collar and office workers), and high (consisting of businesspeople, educated professionals, and intellectual workers). As indicated in Table 1, the distribution of SES among the 10 speakers is relatively balanced, thereby minimizing the potential for SES bias in our final statistical analysis.

3.3 Tokens

In total, we identified 190 occurrences of *je pense* and *je crois* (114 occurrences of *je pense* and 76 occurrences of *je crois*) in ESLO 2. However, the following cases are excluded from our final analysis:

1) Occurrences that appear in negation:

ex. 1 : donc y a pas je crois pas qu'y a un langage jeune euh orléanais

'so there is not I do not think that there is a youth language uh Orleanese'

ex. 2 : mais je pense pas euh

'but I do not think uh'

2) Occurrences that do not introduce an embedded clause:

ex 3 : euh sur Orléans je crois euh des dans le l'ha- l'habillement dans le dans le vêtement quoi les

'um in Orleans I believe um some in the clo- the clothing in the in the clothes what the'

Therefore, only 164 occurrences are included in our final analysis. Table 2 presents the detailed distribution of *je pense* and *je crois* in this study.

	<i>je pense</i>	<i>je crois</i>	Total
No. included	100	64	164
No. excluded	14	12	26
Total	114	76	190

Table 2: Distribution of *je pense* and *je crois*.

3.4 Linguistic factors

Table 3 presents the linguistic factors that might be relevant to choosing *je pense* and *je crois*. For each factor, we have at least two different levels (groups) to look at.

Factor	Levels
tense of the verb in the embedded clause	present
	future
	imperfect
	perfect
	pluperfect
status of the embedded clause	conditional
	judgment of reality
	judgment of the value

presence of <i>que</i>	present
	absent

Table 3: Linguistic factors to be examined.

As shown in Table 3, firstly, we will examine the tense of the verb in the embedded clause, while considering *je pense* and *je crois* as part of the matrix clause. We aim to determine if either form is more closely associated with a particular tense. Secondly, we will investigate the status of the embedded clause. In this regard, we will adopt the classification proposed by Gosselin (2018), which distinguishes between judgments of reality and judgments of value. As Gosselin (2018: 180) explains, "a judgment of reality states what the case is (it describes a situation), while a value of judgment consists of speaking well or ill of an individual or situation". Thirdly, we will analyze the presence of *que* before the embedded clause. While *que* is obligatory in written French when introducing the embedded clause, it is optional in oral French. With this factor, we aim to determine if either form shows a preference for the omission or retention of the particle *que*.

3.5 Extralinguistic factors

Table 4 presents the two extralinguistic factors to be examined in this study. For SES, we utilized the information provided by the corpus and classified the speakers into three groups: low, mid, and high SES. Regarding the age factor, we considered the age of the speaker at the time of the interview. Given the limited number of speakers (ten), we did not group them into different age categories. Instead, the age factor will be treated as a continuous variable for statistical purposes.

Factor	Levels
SES	low
	mid
	high
Age	continuous

Table 4: Extralinguistic factors to be examined.

3.6 Statistical analysis

For this article, we will use the mixed-effects regression model carried out in the R environment using Rbrul (Johnson, 2009). The model distinguishes the following levels for statistical significance: $p > 0.1$, not significant; $.05 < p < 0.1$, marginally significant; $p < .05$, significant; $p < .01$, very significant; $p < .001$, highly significant. For the results, the model provides one p -value for each predictor (the independent variable) to indicate if this predictor is statistically significant for the dependent variable. Meanwhile, it also provides the factor weight and log odds for each level of the predictor to indicate which level(s) favors/disfavors the chosen variable.

For our analysis, we look at the dependent variable, the attitude expression, at binary classifications of *je pense* vs. *je crois*. The fixed independent variables are both the linguistic and extralinguistic factors presented above. All fixed factors except for age are categorical. As we use the age of the speakers at the time of the interview, the age factor is thus continuous. To include the mixed-effects, we use participants as a random variable. For the modeling, we performed the one-level test.

Since “participants” is treated as a random variable, for the following section, we only provide the results for the fixed variables for further discussion. However, the detailed results for individual participants are provided in the appendix for readers’ reference.

4 Results & discussion

Table 5 presents the regression analysis results of *je pense/je crois*. Our results indicated that all three linguistic factors, the status of the embedded clause ($p=0.000733$; f.w.: value judgement: 0.752; judgement of reality: 0.248), tense of the verb in the embedded clause ($p=0.00231$; f.w.: conditional: >0.999 ; future: 0.792; present: 0.584; perfect: 0.51; imperfect: 0.341; plusperfect: <0.001) and presence of *que* ($p=0.0489$; f.w.: present: 0.598; absent: 0.402), are statistically significant for the choice

between *je pense* and *je crois*, while no social factors have been found to be statistically significant.

Among the three linguistic factors, the status of the embedded clause appears to be the most influential factor contributing to the choice between *je pense* and *je crois*. When the embedded clause represents a judgment of value, French native speakers are more inclined to use the form *je pense*, whereas when the embedded clause represents a judgment of reality, they are more likely to opt for *je crois*.

Regarding the tense of the verbs in the embedded clause, our results indicated that the conditional, future, and present tenses tend to favor the use of *je pense*, while the perfect, imperfect, and pluperfect tenses tend to favor the use of *je crois*. In other words, *je pense* is more commonly associated with present and future tenses, while *je crois* is more commonly associated with perfect tenses. *Je pense* is more likely to be used when referring to ongoing or future events, whereas *je crois* is more likely to be used when referring to past events. It is also noteworthy that the conditional tense is exclusively used with *je pense*, while the pluperfect tense is never used with this form.

Finally, regarding the presence of the particle *que* following these two attitude expressions, our results indicate that *que* is more likely to be present when the variant *je pense* is used, and more likely to be omitted when the form *je crois* is employed. In other words, native speakers of French tend to prefer using *je pense que* over *je crois que*.

<i>Je pense/Je crois</i>				
Input prob.	0.61			
Total no.	164			
Log. likelihood	-84.865			
	logodds	tokens	%	f.w.
Status	$p=0.000733$			
value	1.11	28	92.9	0.752
reality	-1.11	136	54.4	0.248
Tense	$p=0.00231$			
conditional	16.908	11	100.0	>0.999

future	1.339	15	86.7	0.792
present	0.341	96	61.5	0.584
perfect	0.038	20	50.0	0.51
imperfect	-0.661	19	36.8	0.341
pluperfect	-17.966	3	0.0	<0.001
Que	p=0.0489			
present	0.397	104	70.2	0.598
absent	-0.397	60	45.0	0.402
SES	Not significant			
High	0.328	65	69.2	[0.581]
Low	0.251	68	63.2	[0.562]
Middle	-0.579	31	38.7	[0.359]
Age	Not significant			
continuous				
+1	-0.005			
Speakers	Random			

Table 5: Regression analysis results of *je pense/je crois*.

While linguistic environment appears to be particularly influential in the choice between *je pense* and *je crois* for native speakers, none of the tested extralinguistic factors tested has been found to be significant for this choice. The use of either form is associated with specific age groups or SES groups, indicating that this variation between the two variants is not an ongoing change in contemporary French. Instead, it represents a completed change over time. The use of both forms has become widespread across all social classes in French.

5 Conclusion

In this study, in contrast to earlier qualitative studies, we conducted a mixed-effects regression analysis on data obtained from a corpus of oral French speech by native speakers. Our aim was to quantitatively examine two attitude expressions, *je pense* and *je crois*. We considered both linguistic and extralinguistic factors in our analysis. The results revealed that all three linguistic factors, namely the tense of the verb in the embedded clause, the status of the embedded clause, and the presence of the particle *que*, were found to be statistically significant in relation to the choice between the two

variants. These findings provide further evidence that, in addition to semantic and pragmatic differences, the linguistic context also plays a role in determining which variant speakers will choose. On the other hand, none of the extralinguistic factors were found to be significant in this choice, suggesting that the variable use of these expressions is not an ongoing change but rather a relatively stable feature of native speech.

However, it should be noted that the current study has limitations due to the small number of tokens analyzed, which means that the findings are not conclusive. Instead, this work can be seen as a pilot study, providing a starting point for further investigation.

In future studies, expanding the dataset by adding more data would be beneficial to conduct a more comprehensive analysis of the real-time use of *je pense* and *je crois* in French native speech. The ESLO corpus consists of two parts, ESLO 1 (1968-1974) and ESLO 2 (2008-), and the time interval between these two collections allows for a comparison of the changes in the use of these variables over a forty-year period. This comparative analysis would help determine if the observed variation is indeed a completed change in contemporary French, as well as shed light on any potential changes in the linguistic factors influencing the choice between the two variables over time.

Second, in our study, we only examined two social factors. With a larger dataset, it would be possible to incorporate additional factors, such as educational background or social network, to explore potential intergroup differences in the use of *je pense* and *je crois*.

Third, since our analysis was based on the interview module of the ESLO corpus, which predominantly represents informal speech, it would be valuable to investigate the use of *je pense* and *je crois* in other contexts, such as lectures, conferences, or casual conversations among family members. This would allow us to explore potential

variations in usage across different communicative settings.

Lastly, it would be beneficial to examine the use of *je pense* and *je crois* in the speech of individuals from Francophone countries other than France. Given that particles may undergo different stages of pragmaticalization in different regions, it is likely that usage patterns vary. Comparing the ESLO corpus with other corpora from diverse Francophone contexts would provide us with a broader understanding of this phenomenon.

References

- Angot, J. Epistemic and subjective expressions in French: the case of *je pense*, *je crois* and *je trouve*. Doctoral dissertation (The University of Manchester, 2021).
- Angot, J. and M. M. Hansen. “The meaning and functions of French *je pense* (que): a Constructionalist and interactional account.” In *Studies at the grammar-discourse interface*. (John Benjamins Publishing Company, 2021).
- Barras, C., E. Geoffrois, Z. Wu and M. Liberman. “Transcriber: development and use of a tool for assisting speech corpora production.” *Speech Communication*, 33(1-2) (2001), 5-22.
- Baude, O. and C. Dugua. “(Re) faire le corpus d’Orléans quarante ans après: quoi de neuf, linguiste?”, *Corpus*, (10) (2011), 99-118.
- Benveniste, É. *Problèmes de linguistiques générales I* (Paris: Gallimard, 1996).
- Borillo, A. “Deux aspects de la modalisation assertive: *croire* et *savoir*.” *Cahiers de grammaire*, 4 (1982), 5-38.
- ELAN (Version 6.2) [Computer software]. Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. (2021). Retrieved from: <https://archive.mpi.nl/tla/elan>
- Gachet, F. “Syntactic hypotheses about so-called ‘que-deletion’ in French.” In *Parenthesis and Ellipsis* (De Gruyter Mouton, 2014), 147-172.
- Gosselin, L. “L’expression de l’opinion personnelle: *Je crois/pense/trouve/considère/estime que p.*” *L’information grammaticale*, 144 (2015), 34-40.
- Gosselin, L. “French expressions of personal opinion: *je crois/ pense/ trouve/ estime/ considère que p.*” In *Epistemic Modalities and Evidentiality in Cross-Linguistic Perspective* (De Gruyter Mouton, 2018), 179-195.
- Hay, J. “Statistical analysis.” In Marianna Di Paolo/Malcah Yaeger-Dror (eds.), *Sociophonetics: a student’s guide* (Abingdon: Routledge, 2011), 198-214.
- Johnson, D. E. “Getting off the GoldVarb standard: Introducing Rbrul for mixed-effects variable rule analysis.” *Language and linguistics compass*, 3(1) (2009), 359-383.
- Lenzen, W. “Epistemic logic.” In *Handbook of Epistemology*. (Springer, Dordrecht., 2004), 963-983.
- Mooney, D. “Quantitative approaches for modelling variation and change: a case study of sociophonetic data from Occitan.” In *Manual of Romance Sociolinguistics*. (De Gruyter, 2018).
- Rendulić, N. and L. Kanaan-Caillol. “*Je crois que, je pense que*: valeurs et variation dans un corpus oral diachronique.” In *SHS Web of Conferences* (EDP Sciences, 2016), 27, 02014.
- Simons, M. “Observations on embedding verbs, evidentiality, and presupposition.” *Lingua*, 117(6) (2007), 1034-1056.
- Vet, C. “Savoir et Croire.” *Langue Française*, 102 (1994), 56-68.

A Appendix. Individual difference.

Participants (random)				
	intercept	tokens	p/p+c	f.w.
std dev	0.276	164	0.61	...
LX10	0.156	22	0.727	0.539
HT398	0.143	14	0.786	0.536
GK11	0.109	28	0.821	0.527
BT17	0.106	5	0.6	0.527
BV1	-0.003	18	0.667	0.499
MC59	-0.035	9	0.444	0.491
AN43	-0.072	8	0.5	0.482
AJ38	-0.104	26	0.346	0.474
KC3	-0.12	19	0.579	0.47
QF28	-0.186	15	0.467	0.454

Making Non-Normalized Content Retrievable – A Tagging Pipeline for a Corpus of Expert-Layperson Texts

Christian Lang and Ngoc Duyen Tanja Tu and Laura Zeidler

Leibniz Institute for the German Language

Mannheim, Germany

{lang, tu}@ids-mannheim.de

laura.zeidler@swhk.ids-mannheim.de

Abstract

Conventional terminology resources reach their limits when it comes to automatic content classification of texts in the domain of expert-layperson communication. This can be attributed to the fact that (non-normalized) language usage does not necessarily reflect the terminological elements stored in such resources. We present several strategies to extend a terminological resource with term-related elements in order to optimize automatic content classification of expert-layperson texts.

1 Introduction

One of many applications of Knowledge Organization Systems (KOS) is tagging texts to make them retrievable, cf. (Golub et al., 2019, p. 205). In our contribution, we describe the use of a KOS to process texts from the domain of expert-layperson communication – specifically, so-called language inquiries, i.e. questions that (supposed) laypeople ask linguistic experts about (German) language such as (1).

(1) Question: [...] *Muss bei... Kurs des Studienkreis_es... der Genitiv angezeigt werden, oder kann man 'Studienkreis' als undeklinierbaren Eigennamen einstufen [...]*? ([...] Does... course of the study group... need to display the genitive case, or can 'study group' be classified as an indeclinable proper name [...])?

Answer: *Im Deutschen werden Eigennamen grundsätzlich gebeugt. [...] Dies gilt auch in Ihrem Beispiel. [...]* (In German, proper names are always inflected. [...] This also applies to your example [...].)

Because language inquiries serve as a valuable primary source of authentic language data for a variety of linguistic research questions, cf. (Breindl, 2016), we plan to create a monitor corpus to make them accessible to the research community.

The core of this corpus is a collection of approx. 50,000 inquiries (and corresponding answers) sent by email to the language consulting service of a German publisher between 1999 and 2019.¹ The collection also contains additional metadata, such as the assignment of each question to a linguistic category (e.g. grammar, spelling, punctuation, etc.).

For optimal usability of the corpus by the research community, it is essential that researchers have access to the exact data points that are relevant to their research question. To make this possible, we identified and tagged elements in questions and answers that allow for the most precise content classification possible.

A first step in this process was terminological tagging, for which we utilized a KOS (see Section 2.1). However, as we show in Section 2.2, due to the nature of the data (expert-layperson communication), terminological tagging on its own is not sufficient. Therefore, in Section 4 we present strategies how to extend the KOS we use to meet the specific requirements of tagging texts in the domain of expert-layperson communication.

The extension of the KOS is a work in progress. Thus, we illustrate the strategies and their positive impact on the tagging process with individual example cases.

2 Tagging process

2.1 Terminological resource: WT

WT (Wissenschaftliche Terminologie)² is the terminological resource of the grammatical information system *grammis*.³ It is stored and maintained in an object-relational database. The resource – an

¹We will expand this core continuously with language inquiries received by Leibniz Institute for the German Language. In addition, we plan to extract language inquiries from other sources, including online sources, and add them to the corpus.

²A more exhaustive description of the resource can be found i.a. in (Suchowolec et al., 2019).

³<https://grammis.ids-mannheim.de>

onomasiologically-structured KOS that can be classified as a thesaurus according to Zeng’s taxonomy of KOS (Zeng, 2008, p. 161) – contains approx. 1,900 concepts from the domain of (German) grammar. As Figure 1 shows, various attributes, such as terms or explanatory texts, can be assigned to each concept. The concepts are linked to each other using three different semantic relations: (i) as hyperonyms and hyponyms (broader term (BT) and narrower term (NT)), (ii) as holonyms and meronyms (broader term partitive (BTP) and narrower term partitive (NTP)) and (iii) as non-hierarchical relatives (related terms (RT)), cf. (ANSI/NISO Z39.19-2005 (R2010), 2005)). Currently, the resource contains 2,961 German-language and 1,874 foreign-language terms.

While terms are not restricted to nouns in principle, WT has a strong bias towards nominal terms: Approx. 90% of WT’s elements are either single nouns or complex noun phrases.⁴

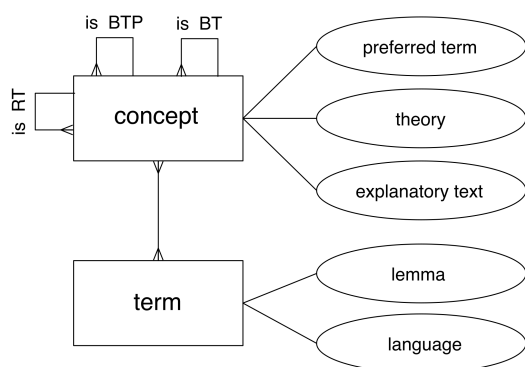


Figure 1: Data structure of WT, figure was first published in (Lang and Suchowolec, 2020, p. 31).

WT was also adapted into a SKOS vocabulary using the D2RQ platform, cf. (Suchowolec et al., 2019).

2.2 Terminological tagging

We used the terms of WT as the basis for a string-matching algorithm to tag specific keywords in our corpus. The algorithm operates as follows: First, we tokenized the data using spacy (Honnibal and Montani, 2017). Second, we applied three different lemmatizers, namely spacy, HanTa (Wartena, 2019) and GermaLemma.⁵ By using multiple lem-

⁴In their analysis of the linguistic properties of terms based on English terminological dictionaries from various technical fields, (Justeson and Katz, 1995, p. 83) found that depending on the domain, between 92% and 99% of terms are nouns or noun phrases.

⁵<https://github.com/WZBSocialScienceCenter/germalemma>

	questions (n=800)		
	Precision	Recall	F-Measure
uni	0.74	0.898	0.811
uni & bi	0.739	0.899	0.811
uni & tri	0.74	0.9	0.812
uni, bi & tri	0.739	0.901	0.812
	answers (n=300)		
	Precision	Recall	F-Measure
uni	0.691	0.849	0.762
uni & bi	0.69	0.858	0.765
uni & tri	0.691	0.849	0.762
uni, bi & tri	0.69	0.858	0.765

Table 1: Evaluation of string-matching algorithm. The evaluation was performed on a manually annotated gold standard consisting of 800 instances of linguistic inquiries and 300 instances of corresponding answers.

matizers, we tried to mitigate possible weaknesses in the performance of the individual tools regarding the lemmatization of low-frequent, specialized words, namely linguistic terms. Since spacy is a look-up lemmatizer for German, it is used as a baseline, i.e., if spacy lemmatizes successfully (based on spacy’s *out of vocabulary*-attribute), the lemma is adopted. If the lemmatization with spacy fails, we consult the results of the remaining two rule-based lemmatizers. If GermaLemma lemmatizes successfully, this result is adopted, otherwise we fall back on the lemmatization of HanTa.⁶ If all lemmatization attempts fail, i.e. if neither lemmatizer transforms the token in any way, the token itself is adopted. Finally, we used the terms in WT – preprocessed identically to the inquiries – as the basis for string-matching to identify and tag the terms in the language inquiries.

We evaluated the algorithm on a subset of the corpus. To this end, two linguists created a gold standard by manually annotating elements (up to 3-grams) they deemed to be terms (for example, *Deklination* (declension), *Kleinschreibung* (lower case), etc.) in a randomly selected subset of 1,100 data points (800 questions, 300 answers). Table 1 shows the results of the evaluation, i.e. string-matching algorithm vs. manually annotated gold standard.

The evaluation reveals problems in the tagging process. On the one hand, the precision value is comparatively low. A qualitative analysis of the

⁶We put GermaLemma first because this order proved to yield slightly better results in previous experiments.

elements falsely tagged as terms shows that these mainly are polysemous words that have both a technical (linguistic) and a general meaning, e.g. *Argument* (argument), *Thema* (topic). On the other hand, the comparatively high recall value turns out to be deceiving on closer inspection. About 32% of all data points were not tagged at all (neither automatically nor by the human annotators) because they did not contain any terms in the strict sense. Further, about 43% of data points contain either no terms or one of the two very broad terms *Satz* (sentence) and *Wort* (word) – which are un-suitable for precise content classification.⁷ This result is not surprising in view of the fact that the tagged data can be attributed to the field of expert-layperson communication. That is, elements of domain-specific language do appear, but – as the following section shows – not always in the form in which they are stored in a terminological resource such as WT. It thus becomes clear that a purely terminological tagging of the data cannot guarantee optimal retrievability.

2.3 Term-related elements

We find that the data points contain elements that, while not terminology in the strict sense, may crucially contribute to the classification of the questions and answers. We refer to these elements as term-related elements. Thus, in a follow-up step, the annotators marked all term-related elements in the 1,100 data points of the gold standard.⁸

A qualitative analysis reveals broadly speaking two types of term-related elements. Type 1 elements – which account for about 53% of all elements – are adjectives (12.2%) or verbs (41.3%), of which about 90% are derivations from a nominal term (e.g. *Komparation* > *komparieren/komparierbar* (comparison > (to) compare/comparable))⁹. Type 2 elements are nouns (46% of all term-related elements), of which almost 50% are compounds or nominal phrases that have at least one term as a component (e.g. *Genitivbezug* [genitive reference] or *paariges Komma* [paired

⁷The percentage of data points without terms (about 38%) and either without terms or with *Satz* (sentence) and *Wort* (word) (approx. 51%) is even higher if we consider only questions.

⁸In individual cases, it can be difficult to decide whether an element is a term or not. This classification always involves a degree of subjectivity.

⁹"Derivation from a nominal term" is not to be understood in the sense of a morphological analysis, but refers to the tendency of terms to be nouns.

comma]); another 34% of Type 2 elements are general language expressions (e.g. *Form* (form)).

If we include term-related elements, the proportion of untagged data points drops to 16% (compared to 32% when only terms are considered). In the case of data points that do not contain terms or term-related elements, linguistic examples play an important role (see Section 4.3). Although term-related elements are still insufficient to identify all questions and answers, the improvement is substantial and we believe the tagging process will benefit greatly from considering these elements.

The implementation differs depending on the type of term-related elements. While for the identification of some elements a mere adjustment in the tagging process is sufficient, for others an inclusion in WT as the KOS underlying the tagging process makes sense. For example, Type 2 compounds consisting of a term and one (or more) non-terms (e.g. *Kannkomma* (optional comma)) can be found by partial string-matching. Including these kinds of elements in WT is not particularly useful, especially since potentially infinite compositions of terms with other words exist. However, including Type 1 derivatives in WT will not only optimize the current tagging process, but also expand the future applications of the KOS.

3 Related work

A large number of domain-specific resources of various kinds exist that can act as potential linking points for an extension of WT.

For example, LingTermNet (Neumann-Schneider and Ziem, 2020), a frame-based resource of linguistic terms containing 73 frames and 257 terms. However, the terms included are mainly from the domain of conversational analysis – an area that is less relevant to our task. Additionally, LingTermNet includes only nouns, while we want to add non-nominal elements to our resource. The latter is also true for LiDo,¹⁰ a large relational database containing linguistic terms created by Christian Lehmann. While there are adjectives in the database, Lehmann postulates that based on conventions of scientific theory, terms should be appellatives (Lehmann, 1996, p. 4). LiDo, originally implemented in a relational database, has been converted to a Linked Data graph: LiDo RDF (Klimek et al., 2018) and is the base of OnLit, an ontology for linguistic terms

¹⁰<http://linguistik.uni-regensburg.de:8080/lido/Lido>

(Klimek et al., 2017).¹¹

Another approach is demonstrated by Medical WordNet, specifically for medical terms (Smith and Fellbaum, 2004), a resource that contains not only technical terms, but also medical vocabulary used by laypeople. Medical WordNet was partly built by extracting all medical terms from WordNet (Miller, 1995). WordNet is a large lexical database where among other things the semantic relation between senses of high-frequency English words is stored, either as a group of synonyms, i.e. the words refer to the same concept, or individual words.¹²

Accordingly, to extend WT, we could consider using GermaNet (Hamp and Feldweg, 1997, p. 9). While GermaNet allows different word classes to be linked (Hamp and Feldweg, 1997, p. 11), there is no noun-verb relation.¹³ For example, *Deklination* (declension) and *deklinieren* ((to) decline) are not linked to each other. For that reason, GermaNet does not seem to be ideal for a systematic extension of WT. Another possible reference point is the German wiktionary.¹⁴ We downloaded the German wiktionary dump from 21-Mar-2023 00:52.¹⁵ We extracted the titles from the wiktionary articles with a Python Package¹⁶ and checked if WT contains the title. This is true for 1,289 titles. In some of these articles there are derivations of nominal terms, as for example in the article of *Entlehnung* (loan), where the verb *entleihen* ((to) borrow) is listed. However, wiktionary is not domain-specific, so it is necessary to manually check whether the terms are listed in their linguistic meaning. Otherwise, it can happen that, for example, incorrect synonyms are extracted. Although some articles have the label "Linguistik" (linguistics) when a linguistic meaning is listed, not all do, such as the article for *Übersetzung* (translation).

None of the resources we considered have all the features necessary for the current task (systematic linking of nouns to other parts of speech; subject domain linguistics). Therefore, we turned to in-house resources to devise extension strategies.

¹¹OnLiT offers a term-termRelation property to specify the relation of "noun Term instances and adjective and verb Term instances" (Klimek et al., 2017, p. 48-49).

¹²<https://wordnet.princeton.edu/>

¹³<https://uni-tuebingen.de/fakultaeten/philosophische-fakultaet/fachbereiche/neuphilologie/seminar-fuer-sprachwissenschaft/arbeitsbereiche/allg-sprachwissenschaft-computerlinguistik/ressourcen/lexica/germanet-1/beschreibung/relationen/>

¹⁴<https://de.wiktionary.org/wiki/Wiktionary:Hauptseite>

¹⁵<https://dumps.wikimedia.org/dewiktionary/>

¹⁶<https://pypi.org/project/wiktionary-de-parser/>

4 Strategies for extending WT

With WT, we have our own comprehensive resource in which not only terms but also explanatory texts can be assigned as attributes to concepts. A total of approx. 600 terms have an explanatory text that can be used as a source for an extension. Moreover, the language inquiries themselves function as an extensive data base for finding relevant elements typically used by laypeople.

4.1 Extraction of Type 1 elements

We use (a) the terms and (b) the explanatory texts from WT to obtain Type 1 elements, i.e. derivations of nominal terms.

(a) For terms that are not linked to an explanatory text in WT, we take advantage of the fact that German is an inflectional language by applying a rule-based transformation of nominal terms in the WT into verbs and adjectives.¹⁷ We tested these approaches with terms ending in the German noun suffix *-ung*. We chose this suffix because an analysis of the 123 verbal and adjectival term-related elements of the gold standard showed that 69% are verbs that can be nominalized by suffixation with *-ung*.¹⁸

(a, 1) For compounds, we automatically iterate through all terms from WT, apply a compound splitter¹⁹ to the unigrams and filter for compounds that consist of a maximum of two elements. After that we replace *-ung* with the German verb suffix *-en* and concatenate the first constituent with the formed verb. For example, this produces *kleinschreiben* ((to) write in lowercase) for *Kleinschreibung* (lower case). Including the derived verb in the tagging process greatly increased the language inquiries found: *Kleinschreibung* yielded 1,806 language inquiries, *kleinschreiben* yielded 2,895 results (in 282 cases, both tags overlap).

(a, 2) For the remaining non-compound unigrams, we proceeded similarly, e.g. by deriving *steigern* ((to) compare) from the nominal term

¹⁷Rule-based approaches assume a regular derivational process, e.g. the nominalization of verbs with the suffix *-ung* or the adjectivization of verbs with the suffix *-bar*. If there is no regular relationship between noun and verb/adjective, other strategies must be applied.

¹⁸We also used two stemmers on the terms ending in *-ung*: while CISTEM (https://www.nltk.org/_modules/nltk/stem/cistem.html) does not correctly stem any of the terms, Snowball German Stemmer (https://www.nltk.org/_modules/nltk/stem/snowball.html) fails on 38% of the terms.

¹⁹<https://github.com/bminixhofer/nnsplit>

Steigerung (comparison). Also in this example, the integration of the verb into the tagging process leads to increased retrieval: *steigern* appears in 457 language inquiries, while *Steigerung* occurs in only 190 language inquiries (in 71 cases, both tags overlap).

(b) Next, we use the explanatory texts to find the derivations of nominal terms. We limit the search for the derivations to the explanatory texts, since in these the probability of finding true positives is very high. We perform the extraction by automatically searching for tokens in the explanatory texts that are similar to a term (with regards to spelling), have a certain suffix and belong to a certain word class (verb or adjective). For example, we search the tokenized, lemmatized and POS-tagged explanatory text of the term *Deklination* (declension) for lemmas that begin with the first three characters of the term (*dek*), end with a particular suffix, like *-ierbar* or *-ieren*, and are adjectives or verbs, depending on the suffix.²⁰ As a result, this produced *deklinerbar* (declinable) and *deklinieren* ((to) decline), which enabled the retrieval of 472 more language inquiries than with *Deklination* alone.

4.2 Extraction of Type 2 elements

As stated in Section 2.3, 46% of term-related elements are nouns. While approx. 50% of the 46% can be tagged by partial string-matching, a different strategy must be considered for the other half. This pertains, in particular, to general language expressions such as *Form* (form). Due to the gold standard annotations we have already a basis of term-related elements, which laypeople use instead of ‘proper’ linguistic terms.

These term-related elements usually occur with other words to paraphrase a linguistic term. Accordingly, we plan to perform a co-occurrence analysis on the language inquiries to analyze with which other words these elements occur frequently. We tried this approach on our gold standard data set and analyzed the co-occurrences of the token *Form*, among others, in more detail: it occurs frequently with the adjective *weiblich* (female) (164 times). We ascertained that questions containing these two words are questions about *Genus* (grammatical gender). Thus, using this methodology, we can link adjective-noun combinations to terms in WT, in

²⁰We have found that the character-matching should be limited to three characters, because there are terms whose derivations could not be matched otherwise, such as *Flexion* (inflection) and *flektieren* ((to) inflect).

this case the term-related elements *Form* and *weiblich* are linked to the concept *Genus*. This allows us to tag additional 91 language inquiries compared to tagging with *Genus* alone.

4.3 Extraction of examples

Terms and term-related elements do not always appear in language inquiries as stated in Section 2.3. However, in many cases an example is used in a language inquiry. Hence, on the one hand, we can extend WT with authentic examples extracted from the language inquiries, on the other hand, we can analyze the examples to identify patterns to tag them with specific terms. Therefore, language inquiries in which no terms or term-related elements are used can also be classified.

The following example of the terms *Getrenntschreibung* (separate spelling) and *Zusammenschreibung* (compound spelling) illustrates the approach: First, we clean the data by mapping all quotation marks to one quotation mark type. After that we extract the string(s) from a question that is between quotation marks, e.g. in (2), which concerns the correct spelling of "apple picking", *Apfel pflücken* (separate spelling) and *Apfelpflücken* (compound spelling) will be extracted.

(2) *Wie schreibt man "Apfel pflücken" oder "Apfelpflücken" [...]?* (How do you write "apple picking" or "applepicking" [...])?

The strings used in questions about separate and compound spelling are identical to each other if the whitespace is removed from the separate spelling variant, as demonstrated by *Apfel pflücken* and *Apfelpflücken* in (2). Based on this pattern, we can tag 214 language inquiries from our data with the terms *Getrenntschreibung* and *Zusammenschreibung*, of which only 52 questions contained the terms or term-related elements *Getrenntschreibung*, *Zusammenschreibung*, *getrenntschreiben/getrennt schreiben* or *zusammenschreiben/zusammen schreiben*.

5 Conclusion

In our contribution, we have described the challenges that arise when using a terminological resource to tag expert-layperson texts. We have described several strategies for extending the resource. As a result, the data structure (c.f. Fig. 1) will be extended by term-related elements and language

examples (patterns).

Based on the first promising results of both the KOS extension and the adjustments in the tagging process, we suggest the following pipeline for tagging the language inquiry corpus: (1) using the entries of the extended WT to detect terms as well as term-related elements (primarily verbs and adjectives), (2) partial string-matching to identify compounds containing at least one terminological or term-related element, (3) analyzing co-occurrences of term-related elements, (4) identifying typical example patterns. The next steps in optimizing the tagging process are to expand the rule-based extension beyond the cases already implemented and a systematic analysis of cases that cannot be covered by rule-based methods.

Scientific communication is assuming an increasingly more prominent role in everyday academia. This underlines the importance of creating resources and developing tools to machine process expert-layperson communication. This is why an extension of WT is a worthwhile endeavour.

References

- ANSI/NISO Z39.19-2005 (R2010). 2005. [ANSI/NISO Z39.19-2005 \(R2010\) Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies](#).
- Eva Breindl. 2016. [Sprachberatung im interaktiven Web](#). In Sven Staffeldt and Wolf Peter Klein, editors, *Die Kodifizierung der Sprache: Strukturen, Funktionen, Konsequenzen*, volume 17 of *WespA – Würzburger elektronische sprachwissenschaftliche Arbeiten*, pages 85–109. Univ. Würzburg, Institut für deutsche Philologie, Würzburg. OCLC: 959665919.
- Koraljka Golub, Rudi Schmiede, and Douglas Tudhope. 2019. [Recent applications of Knowledge Organization Systems: introduction to a special issue](#). *International Journal on Digital Libraries*, 20(3):205–207.
- Birgit Hamp and Helmut Feldweg. 1997. GermaNet - a Lexical-Semantic Net for German. In *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15, Madrid.
- Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.
- John Justeson and Slava Katz. 1995. [Technical terminology: Some linguistic properties and an algorithm for identification in text](#). *Natural Language Engineering*, 1:9–27.
- Bettina Klimek, John McCrae, Christian Lehmann, Christian Chiarcos, and Sebastian Hellmann. 2017. [Onlit: An ontology for linguistic terminology](#). pages 42–57.
- Bettina Klimek, Robert Schädlich, Dustin Kröger, Edwin Knese, and Benedikt Elßmann. 2018. [LiDo RDF: From a Relational Database to a Linked Data Graph of Linguistic Terms and Bibliographic Data](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 2429–2436, Miyazaki.
- Christian Lang and Karolina Suchowolec. 2020. [Wisensmanagement in der Praxis: Welchen Beitrag leistet deskriptive Terminologiearbeit?](#) In Barbara Ahrens, Morven Beaton-Thome, Monika Krein-Kühle, Ralph Krüger, Lisa Link, and Ursula Wiene, editors, *Interdependenzen und Innovationen in Translation und Fachkommunikation / Interdependence and Innovation in Translation, Interpreting and Specialised Communication*, pages 17–44. Frank & Timme, Berlin.
- Christian Lehmann. 1996. [Linguistische Terminologie als relationales Netz](#). In Clemens Knobloch and Burkhard Schaefer, editors, *Nomination — fachsprachlich und gemeinsprachlich*, pages 215–267. VS Verlag für Sozialwissenschaften, Wiesbaden.
- George A. Miller. 1995. [WordNet: a lexical database for English](#). *Communications of the ACM*, 38(11):39–41.
- Anastasia Neumann-Schneider and Alexander Ziem. 2020. [LingTermNet: Konzeption und Entwicklung eines FrameNet für linguistische Fachterminologie](#). In Christian Lang, Roman Schneider, Horst Schwinn, Karolina Suchowolec, and Angelika Wöllstein, editors, *Grammatik und Terminologie: Beiträge zur Ars Grammatica 2017*, number 82 in *Studien zur deutschen Sprache*, pages 105–128. Narr Francke Attempto, Tübingen. OCLC: on1142742225.
- Barry Smith and Christiane Fellbaum. 2004. [Medical WordNet: a new methodology for the construction and validation of information resources for consumer health](#). In *Proceedings of the 20th international conference on Computational Linguistics - COLING '04*, pages 371–382, Geneva, Switzerland. Association for Computational Linguistics.
- Karolina Suchowolec, Christian Lang, and Roman Schneider. 2019. [An empirically validated, onomasiologically structured, and linguistically motivated online terminology. re-designing scientific resources on german grammar](#). *International Journal on Digital Libraries*, 20(3):253–268.
- Christian Wartena. 2019. [A probabilistic morphology model for german lemmatization](#). *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, pages 40–49.
- Marcia Zeng. 2008. [Knowledge organization systems \(kos\)](#). *Knowledge Organization*, 35:160–182.

Posters

MG2P: An Empirical Study Of Multilingual Training for Manx G2P

Shubhanker Banerjee

ADAPT Centre

University of Galway

shubhanker.banerjee@adaptcentre.ie

Bharathi Raja Chakravarthi

ADAPT Centre University of Galway

bharathiraja.chakravarthi@

adaptcentre.ie

John P. McCrae

ADAPT Centre University of Galway

john.mccrae@adaptcentre.ie

Abstract

Neural networks have achieved state of the art results on grapheme-to-phoneme (G2P) conversion. In this paper we focus on the development of a G2P system for Manx, an extremely low-resourced language of the Goidelic branch of the Celtic family of languages. We preprocess the data using two different data augmentation techniques which we call DA1 and DA2 and carry out experiments with various model architectures to answer the question *What is the optimal choice of data augmentation, training strategy and model architecture for building G2P systems in extremely low-resourced scenarios?* The results demonstrate that multilingual training of the Transformer with DA1 augmented Manx dataset along with data from orthographically similar English and Welsh improve upon the phoneme error rate of Phonetisaurus, LSTM and IBM model 2 by 10.25%, 14.42% and 24.05% respectively.

1 Introduction

Grapheme-to-phoneme (G2P) conversion is the task of generating a phoneme sequence representative of the pronunciation of a given input word. This conversion can be thought of as a sequence mapping task where graphemes in the input word are mapped to phonemes in the output sequence. In recent years, there has been tremendous increase in the efficiency and sophistication of computer aided tools. As a result these tools have increasingly been utilized in all spheres of life. Specifically, Text-to-Speech (TTS) and Automatic Speech Recognition (ASR) tools have improved the accessibility of technology, more so for the disabled and the elderly.

G2P conversion is a critical component of TTS and ASR systems (Kim et al., 2002; Elias et al., 2021; Masumura et al., 2020). Pronunciation dictionaries can be used for building G2P systems, however such dictionaries have a limited coverage

over the vast vocabulary of any language. This necessitates the development of G2P systems that can map written language to its phonemic transcription.

The problem statement defined in this paper is closely related to the work done by Jyothi and Hasegawa-Johnson (2017). They propose the use of recurrent neural networks (RNNs) for tackling G2P conversion in low-resourced scenarios and devise three different alignment strategies which are used to align the grapheme and phoneme sequences. These aligned sequences are then used to train a sequence-to-sequence model composed of RNNs (Rumelhart et al., 1985). The proposed model is evaluated on three low-resourced languages Pashto, Tagalog and Lithuanian. In order to understand the impact of size of the dataset on performance they carry out experiments with datasets of three different sizes: 250, 500 and 1000 samples and as expected they show that larger datasets improve the performance of the model. The main difference between the problem proposed in this paper and their problem statement is the size of the dataset; the size of our Manx dataset (refer to Section 4) is approximately 60% smaller than their smallest dataset (250 samples), thus making the development of a G2P system for Manx more difficult.

Zhao et al. (2022) propose a noise controlled G2P system wherein they inject noisy data during the training phase to develop models that less sensitive to orthographic noise in the data. They report significant significant improvements in the word error rate (WER) on dict-based sources.

Li et al. (2022) propose a zero-shot G2P model that uses data from related languages during training. The related languages are selected using a k-nearest neighbour approach on a phylogenetic tree of the language family.

G2P systems are usually language specific and are dependent on the orthographic properties of

the language in consideration (Ager, 2008). There are challenges associated with the application of the rule-based or deep-learning-based G2P conversion methods for extremely low-resourced languages such as Manx. In these scenarios the linguistic expertise necessary to curate the grapheme-to-phoneme rules is often missing and this in turn makes the development of rule-based systems challenging. Furthermore, the development of deep learning based systems is dependent on annotated datasets which are also not available in extremely low-resourced scenarios. Even the results presented by Dong et al. (2022) where they sample 1000 pronunciations to simulate a low-resourced scenario is not representative of an extremely low-resourced language like Manx where very few data points are available to train the model (for details see Section 4).

In this paper, we study the impact of two different data augmentation strategies which we call DA1 and DA2 (for details see Section 3) as well as that of monolingual and multilingual training on the G2P conversion task. Specifically, we empirically analyze what is the optimal choice of data augmentation technique, training strategy and choice of model for G2P conversion of Manx, an extremely low-resourced language. We are particularly interested in how data from related languages can improve the performance in the multilingual training regime.

2 Related Works

G2P conversion has been an active area of research with a wide variety of methods being employed to tackle this problem (Taylor, 2005; Bisani and Ney, 2008; Rao et al., 2015; Chen, 2003; Novak et al., 2012; Dong et al., 2022). Braga et al. (2006) propose a rule-based system for G2P conversion of European Portuguese. The proposed system is intended as an unit of a larger TTS system. Their paper illustrates the G2P rules in European Portuguese and reports a very high phoneme accuracy rate of 98.80% achieved by the system. Deep learning based methods have achieved good performance on the G2P conversion task with LSTMs (Hochreiter and Schmidhuber, 1996) and Transformers (Vaswani et al., 2017) at the forefront of deep learning research in this area. Yolchuyeva et al. (2019) propose the use of the Transformer architecture for building a G2P conversion system for English. They train and evaluate the proposed

model on the CMUDict and NetTalk datasets and report low ($\sim 5\%$) Phoneme Error Rate (PER). Juzová et al. (2019) propose an encoder-decoder architecture composed of bi-LSTMs to tackle the G2P problem for English, Czech and Russian. They report high phoneme accuracy rates for all of the three languages. Dong et al. (2022) propose GBERT, a multi-layer Transformer encoder inspired by the BERT architecture (Kenton and Toutanova, 2019). Monolingual word lists with randomly masked graphemes (letters) are used to pre-train the GBERT encoder with the masked grapheme objective. The GBERT encoder is then trained/fine-tuned on the G2P conversion task with a Transformer decoder. Experiments have been carried out in the low and medium resourced scenarios and the results indicate the better performance achieved by masked grapheme pre-training.

The DA1 augmentation scheme proposed in this paper is closely related to the work done by Hammond (2021). They propose the use of LSTM (Hochreiter and Schmidhuber, 1996) to tackle G2P conversion for 10 low-resourced languages. Each of these languages has 800 word-pronunciation pairs available for training; in order to augment the training sets splitting of words based on unambiguous mapping of peripheral grapheme sequences to phoneme sequences is proposed. Multilingual training for G2P conversion of Manx in this paper was inspired by the work carried out by Vesik et al. (2020) where they propose the use of multilingual training of Transformers (Vaswani et al., 2017) on the G2P conversion task. They carry out experiments on 15 languages with relatively larger datasets of 4050 samples. The system was trained in a multilingual setting where each source grapheme sequence was prepended with the corresponding language identifier to allow the model to learn meaningful representations from the combined dataset while having the ability to discriminate amongst the languages during inference. The results show an improvement of over 50% in the phoneme and word error rates (PER and WER). We have also carried out experiments to empirically analyze the method proposed by Prabhu and Kann (2020) where they train a Transformer model jointly on grapheme-to-phoneme as well as phoneme-to-grapheme tasks i.e both the forward and the backward directions at each time step of the training. Their results indicate marginal improvement in performance on joint training. Novak et al.



Figure 1: DA1 applied to *braew* such that it is split into two grapheme sequences *b* and *raew*. The mapping of *raew* to *ræw* is independent of *b* and therefore is treated as a separate datapoint in addition to the original word i.e. *braew*. This split point is not based on linguistic rules but an observation of the grapheme and the phoneme sequences which shows that there is a direct correspondence between the phoneme *b* and the grapheme *b* and thus the split point at *b*.

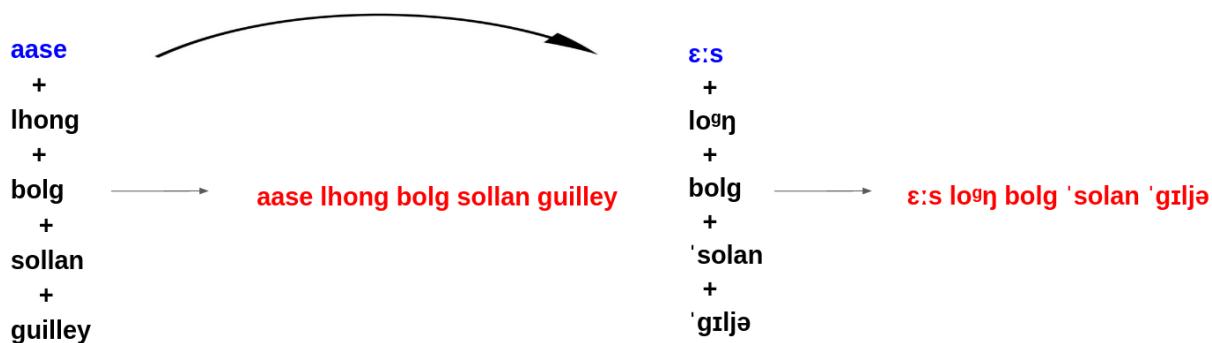


Figure 2: For *aase* we randomly sample 4 words *lhong*, *bolg*, *sollan* and *guilley* and concatenate them together to form the string *aase lhong bolg sollan guilley*, which is a new data point. The corresponding phonemic representations are also concatenated as illustrated in the figure.

(2016) introduced Phonetisaurus a joint n-gram based grapheme-to-phoneme toolkit built upon OpenFST framework¹. El-Hadi and Mhania (2017) carry out experiments on letter-to-sound mapping using Phonetisaurus and demonstrate good results thereby demonstrating its applicability to this task.

3 Data Augmentation

We introduce two data augmentation techniques namely, DA1 and DA2. The idea behind the DA1 augmentation scheme is that certain grapheme segments which are substrings of the original word can be mapped unambiguously to phoneme segments given that appropriate splitting points are found in the original word (see Figure 1 for details). There can be multiple such splitting points in a word leading to the creation of multiple such data points from one word-phoneme pair. The hypothesis is that creation of such subword level pronunciation pairs improves the learnability of the model with regards to the fine-grained grapheme-to-phoneme

rules.

In DA2 augmentation scheme for every word in the pronunciation list we randomly sample 4 other words from the word list and concatenate all the 5 words and correspondingly their 5 pronunciations (see Figure 2 for details). The resultant sequence-phoneme pair is now treated as a new datapoint and used in training. The hypothesis is that longer and more diverse sequences would help improve the performance of the model.

4 Dataset

The problem statement has been framed as a supervised learning problem and therefore a parallel word list comprising of words and their corresponding phonemic representations (pronunciations) is needed to train the model. In the multilingual training regime the idea is to leverage the phonetic and orthographic similarity of related languages to augment the Manx data available for training. Irish and Scottish Gaelic belong to the the same Goidelic language family as Manx and have a similar phonology (Paul, 2014), Welsh and English

¹<https://www.openfst.org>

Language	Train	Valid	Test
English	1,264	—	—
Irish	1,032	—	—
Welsh	512	—	—
Manx	77	34	28
Scottish Gaelic	86	—	—

Table 1: Split Statistics after Data augmentation

have an orthography similar to that of Manx (Gelb, 1968). Therefore, we collect pronunciation lists for English, Welsh, Scottish Gaelic and Irish. In order to collect the data required for the experiments, we use the Wikipron library (Lee et al., 2020) which allows the extraction of pronunciations from Wiktionary². It must be noted that during data collection we collect all available data points for Manx, Welsh, Irish and Scottish Gaelic. However, we limit the number of English samples to 1300 words. The reason behind doing so is to simulate situations where the main language (Manx in this case) as well as all related languages are low-resourced. Furthermore, we observe the presence of repeated entries in the English dataset. On removing these repeated entries we are left with 1264 words.

Initially, 106 Manx samples are collected for Manx using the Wikipron API. We then manually apply DA1 to these 106 words and observe that 33 word-pronunciation pairs can be split into two as illustrated in Figure 1 leading to the creation of 33 additional datapoints. Thus, a total of 139 grapheme-phoneme pairs are obtained after applying DA1. In order to compare DA1 and DA2 we then choose the same 33 words from the original pronunciation list and apply DA2 to each of these 33 word pronunciation pairs i.e for each of these 33 words we randomly choose 5 more words and concatenate them to the originally chosen word; the corresponding pronunciations are also concatenated. Thus, 139 samples are generated by applying the DA2 augmentation scheme. The Manx dataset obtained after the data augmentation has 139 samples and is split in the ratio of 80:20 train-test split. The train dataset is further split in the ratio of 70:30 train-validation split. The resultant dataset statistics are illustrated in Table 1. It illustrates the extremely low-resourced nature of Manx

and reinforces the previously mentioned challenges associated with building deep learning systems that are capable of mapping graphemes to phonemes with such few datapoints.

5 Background

5.1 IBM Model 2

IBM Model 2 is a translation model that was introduced by Brown et al. (1993) and is based on the noisy-channel model of parameter estimation (Weaver, 1949). It is important to note here that in this case the words are the source sequences and the corresponding pronunciations are the target sequences. The source sequences are translated into the target sequences according to a translation table and an alignment function which are learned from the data. For more details on IBM Model 2 we refer the reader to Brown et al. (1993).

5.2 LSTM

Recurrent Neural Networks (RNNs) are a class of neural networks that are capable of modelling time-distributed data sequences (Rumelhart, 1986). However, they suffer from the problem of vanishing gradients over a larger number of time steps (Basodi et al., 2020). Long Short-term Memory network (LSTM) first introduced by Hochreiter and Schmidhuber (1997) mitigate this problem by selectively retaining information over a larger number of time steps. LSTMs have achieved good performance across a wide variety of NLP tasks such as language modelling (Sundermeyer et al., 2012), sentiment classification (Wang et al., 2016), speech recognition (Graves et al., 2013) and named entity recognition (Jin et al., 2019). For further details on the gated architecture of a LSTM cell we refer the reader to Hochreiter and Schmidhuber (1997).

5.3 Phonetisaurus

Phonetisaurus is an open-source grapheme-to-phoneme converter based on the OpenFST frame-

²https://en.wiktionary.org/w/index.php?title=Category:Terms_with_IPA_pronunciation_by_language&from=W

work first introduced by [Novak et al. \(2016\)](#). It uses joint n-gram models to learn a mapping from graphemes to phonemes. The first step in the Phonetisaurus pipeline is the alignment of the source and the target sequences based on a modified form of the algorithm proposed by [Jiampoja-marn et al. \(2007\)](#). The next step involves training a n-gram language model which is then used to construct a Weighted Finite State Transducer (WFST) ([Novak et al., 2012](#)). The final step involves decoding using the WFST constructed in the previous step, the decoder finds the optimal phoneme sequence for a given input sequence of graphemes. For more details on the Phonetisaurus pipeline we refer the reader to [Novak et al. \(2016\)](#).

5.4 Transformer

The Transformer architecture first proposed by [Vaswani et al. \(2017\)](#) was introduced with the objective of mitigating the challenges associated with the recursive structure of sequence modelling neural architectures such as RNN and LSTM. The Transformer architecture is an encoder-decoder architecture with both the encoder and the decoder composed entirely of attention ([Bahdanau et al., 2015](#)) blocks. Transformer and modifications to its architecture such as BERT ([Devlin et al., 2018](#)) and GPT-3 ([Brown et al., 2020](#)) have achieved state-of-the-art results on various natural language processing tasks ([Patil et al., 2022](#); [Do and Phan, 2022](#); [Yang et al., 2022](#)). For further details on the Transformer architecture we refer the reader to [Vaswani et al. \(2017\)](#).

6 Experiments

As mentioned previously the development of rule-based systems for low-resourced languages such as Manx is challenging due to the absence of linguistic expertise. Concretely, there are three primary challenges:

- The curation of G2P rules for Manx often depends on the number of syllables in a word and whether the consonants are broad or slender ([Pickeral III, 1990](#)). Ascertaining these for a particular word requires specialist linguistic knowledge of Manx.
- The quality of a vowel depends on factors such as height of the tongue with relation to the jaw and horizontal position of the tongue in the mouth. Such variation in the quality of

a vowel leads to difference in pronunciation in different contexts ([Pickeral III, 1990](#)). As a result vowel letters often have one-to-many mappings with phonemes and thus the curation of rules mapping vowels to their corresponding phonemes is a linguistically involved task.

- Manx exhibits initial consonant mutation. The pronunciation of the initial consonant of a word alters depending on the morpho-syntactic context ([Hannahs, 2013](#)). Such alterations further complicate the curation of grapheme-to-phoneme rules for the language.

We carry out experiments with deep learning based methods and WFST based Phonetisaurus to empirically study their suitability for building G2P systems for Manx. The optimal hyperparameters are found by training on the train data and manual tuning on the validation set. 5 trials were conducted for hyperparameter search on the LSTM model using only Manx data during training, whereas the optimal hyperparameters for the Transformer model were found in 9 search trials using only Manx data. The test results have been reported in the form of mean and standard deviation of 5 evaluations on the test set using the optimal hyperparameters.

Data Augmentation	PER
No Data Augmentation	90.75 \pm 1.23
DA1	87.52 \pm 0.75
DA2	280.94 \pm 1.65

Table 2: Preliminary Results

6.1 Preliminary Experiments

We carry out preliminary experiments to study the impact of the two proposed data augmentation schemes on performance. Both DA1 and DA2 are applied to the original dataset independently and resultant datasets are used to train LSTM based sequence-to-sequence models for Manx G2P conversion. Furthermore, the unaugmented dataset is also used to train a model on the same task to establish a baseline. Phoneme error rate (PER) is used as the evaluation metric. It is a measure of the percentage of phonemes incorrectly generated by the model for each word. The results illustrated in Table 2 show that the performance significantly deteriorates with DA2 and marginal improvement

Model	LangID	gv	gv+ga+gd	gv+ga	gv+gd	gv+cy	gv+en	gv+cy+en
IBM 2	No	73.58±1.45	73.48±4.87	73.46±3.89	73.89±0.64	75.01±3.21	79.05±3.99	81.79±0.01
	Yes	73.58±1.45	73.48±4.87	73.46±3.89	73.89±0.64	75.01±3.21	79.05±3.99	81.79±0.01
LSTM	No	86.98±3.99	96.58±5.32	98.52±1.23	86.23±0.23	116.10±1.68	84.98±2.99	139.47±1.00
	Yes	70.89±2.09	62.00±1.99	64.96±2.43	70.89±1.78	112.98±3.56	65.82±4.56	73.39±5.32
Transformer	No	96.35±1.89	58.71±3.48	64.96±2.79	73.67±3.45	61.42±1.65	61.99±2.45	55.39±4.87
	Yes	73.89±1.00	59.14±2.67	64.01±0.24	70.49±0.98	58.86±6.25	62.13±1.12	49.53±0.01
Phonetisaurus	No	57.24±0.56	103.49±0.05	104.91±1.26	69.81±0.09	74.71±1.19	72.00±0.85	68.56±0.02

Table 3: PER without Language Identifiers

over the baseline is observed with DA1, thereby indicating the better performance of DA1 scheme on the G2P task. Thus, going forward all experiments are carried out with the DA1 augmented dataset.

6.2 Multilingual Training

The hypothesis is that training the models on the combined datasets would allow them to learn meaningful representations by leveraging the additional training data from related languages. However, this raises a question on the models’ ability to discriminate amongst languages during inference. The same grapheme might have same or different phoneme mappings across languages. To mitigate this problem, we prepend language specific identifiers to words and their phonemic representations. We hypothesize that adding these identifiers would facilitate the learning of language specific representations which in turn would allow the model to meaningfully utilize data from related languages to learn grapheme-to-phoneme rules while also enabling distinction amongst the languages during inference.

In order to study the validity of our hypotheses related to multilingual training and language identifiers we carry out experiments with IBM model 2, LSTM and the Transformer architecture. Multilingual models are trained on a Nvidia RTX2060 GPU using various subsets of the related languages both with and without language identifiers. These models are then evaluated on the Manx test data.

The results are illustrated in Table 3 and show that performance of the LSTM and the Transformer models trained on data with language identifiers is better than those trained without these identifiers. For the purpose of brevity these languages have been referred to by the following

Hyperparameter	Value
Number of Encoder & Decoder Blocks	2
Number of Attention Heads	2
Number of Training Epochs	200
Batch Size	16
Embedding Dimension	256
Maximum Sequence Length	256

Table 4: Training configuration of the best model (en+cy+gv)

ISO 693-1 language codes in Tables 3: Manx (*gv*), Irish (*ga*), Scottish Gaelic (*gd*), Welsh (*cy*) and English (*en*). No improvement in performance is observed with the addition of language identifiers in case IBM model 2. Furthermore, the Transformer model trained multilingually on English, Welsh and Manx data with language identifiers attains a PER of 49.53% and outperforms all other monolingual and multilingual models. The training configuration of this model is given in Table 4. It improves upon the PER (74.24%) of the baseline monolingual Transformer trained only on Manx data by a significant 24.71%.

6.3 Joint Training

The mappings from graphemes to phonemes (G2P) and from phonemes to graphemes (P2G) are monotonic relationships that proceed from left to right. We hypothesize that joint training of the model on both G2P and P2G tasks would facilitate the learning of the monotonic nature of these mappings. Furthermore, given that phonemes and graphemes have a bidirectional mapping between them, that is any given phoneme can be mapped to one or many graphemes and the vice-versa, we hypothesize that training the model to map a phoneme to a specific set of graphemes should introduce signals that drive the model towards optimal performance

on the G2P task.

$$\ell(\theta) = -\frac{1}{N} \sum_{i=1}^N \log P(T^i | S^i; \theta) + \sum_{j=1}^N \log P(S^j | T^j; \theta) \quad (1)$$

Thus, at each training step the model loss given by Eqn.1 is used to optimize the model parameters where S is the set of words and T is the set of corresponding phonemic sequences. As shown in Section 7, the model trained on the downsized English-Welsh dataset alongwith the Manx data has the best performance on the G2P task. In order to validate our hypothesis on joint learning, we use performance as a baseline and train a Transformer model jointly on the G2P and P2G tasks using the downsized English-Welsh data. The hyperparameters used during training are listed in Table 4. During evaluation we observe a PER of 71.45%. This result invalidates our hypothesis related to improvement of performance by introduction of the auxiliary P2G task during training.

6.4 Phonetisaurus

We carry out experiments with Phonetisaurus to assess its suitability for extremely low-resourced languages like Manx. We train the model on subsets of related languages along with the DA1 augmented Manx dataset and the results are presented in Table 3. The results indicate that the performance of Phonetisaurus in general is worse than the best performing model described in Section 7. This result further reinforces the optimality of multilingual training of Transformer to tackle G2P conversion in extremely low-resourced scenarios.

7 Ablation Study

As shown in Section 6.2, the best result is achieved by using data from English and Welsh alongside Manx. English and Welsh are orthographically similar to Manx and the size of the dataset (1,776 samples) is greater than that of the combined Irish and Scottish Gaelic dataset (1,118 samples). To ascertain the impact of orthographic similarity and size of the dataset on the performance we randomly sample 1,118 datapoints from the English-Welsh dataset. The hypothesis is that if orthographic similarity amongst the related languages and Manx is the dominant factor then the performance achieved by the model trained on the downsized English-Welsh dataset should be better than that achieved by training on the phonetically similar Irish-Scottish

Gaelic dataset of the same size. In order to validate our hypothesis we train a Transformer model on the downsized English-Welsh dataset with language identifiers using the training configuration demonstrated in Table 4. Then we evaluate the trained model on Manx test data and observe a PER of **47.94%**. Thus, the model trained on downsized English-Welsh data outperforms the Transformer model trained on the Irish-Scottish Gaelic (PER - 59.14%) dataset by 11.2% validating our initial hypothesis about the impact of orthographic similarity on performance of the system. Furthermore, it also marginally improves upon the performance of the model trained on the full English-Welsh dataset by 1.59%.

8 Computational Cost

The LSTM model used for preliminary experiments has 613,424 parameters whereas the transformer model used for multilingual training and joint training has 3,787,776 parameters. The average runtime of the LSTM model is 62ms per gradient step during training whereas for the Transformer architecture we observe an average runtime of 111 ms per gradient step during training. During inference, the transformer model took 15 ms per input instance and the LSTM had a runtime of 5ms per instance.

9 Error Analysis

We analyze the sequences generated by the best performing model described in Section 7 and observe that in 75% of the sequences, more than 50% of the errors were accounted for by the vowels. We observed that this is due to following two reasons primarily:

- The vowel sound is incorrectly classified altogether. Lhong should be transcribed to $l^{\text{h}}\text{o}^{\text{h}}\eta$, but is transcribed to $l^{\text{h}}\text{ɔ}^{\text{h}}\eta$.
- The quality of the generated vowel is incorrect. For example the vowel e in ane should be transcribed to $\text{ɛ}:\text{n}$, (Open-mid unrounded vowel), but it is transcribed to $\text{e}:\text{n}$ (Close-mid unrounded vowel).

10 Results

The preliminary results demonstrated in Table 2 show that the PER achieved by LSTM models across the augmented and the original datasets is not very low. This is primarily because these

models are trained only on extremely small Manx datasets which are not sufficient to train deep learning models. However, we empirically observe that multilingual training using related languages improves performance on the G2P task as shown by the results demonstrated in Table 3. The use of identifiers that enable the discrimination amongst languages during training have a positive impact on the performance of the model. Also, the optimality of Transformers for this task when they are trained on appropriate datasets is established. Furthermore, as observed in Section 7 orthographically similar languages have a greater impact on the performance of the model. This indicates that languages with similar writing systems when used in the multilingual training regime are more effective than phonetically similar languages. The experiments carried out using IBM model 2 show that there is no significant improvement in the performance of the model in the multilingual training regime. In order to validate our hypothesis as stated in Section 6.3 we conduct experiments by introducing an auxiliary P2G task during training. The results are significantly lower than those of the model described in Section 7 and invalidate our initial hypothesis; joint training on both tasks leads to catastrophic forgetting (Kirkpatrick et al., 2017) and therefore the performance of the model is suboptimal.

We also conduct experiments with Phonetisaurus to assess its applicability for this task. The result does not improve upon the performance of the multilingual model described in Section 7. Furthermore, as indicated by the results presented in Table 3, the performance of Phonetisaurus worsens when data from related languages is introduced during training. It must also be noted that the performance of the Phonetisaurus model trained only on the DA1 augmented Manx dataset is better than other monolingual models shown in Table 3. Finally, the PER of 47.94% achieved by the model trained on English-Welsh dataset is not optimally low, however the results indicate that design of better data augmentation schemes along with improved multilingual training mechanisms leave the scope open for development of G2P systems for Manx.

11 Conclusion

To conclude, we carry out experiments to identify the optimal training regime, model architecture and data augmentation scheme to build a G2P system

for Manx, an extremely low-resourced language. We propose the use of two augmentation schemes DA1 and DA2 to counter the low-resourced nature of Manx and empirically observe an improvement in performance when DA1 is applied to the original dataset. The results indicate that multilingual training of Transformer on data from orthographically similar languages in the presence of language identifiers outperforms all other monolingual as well as multilingual models. This is an interesting result and opens up avenues for application of other multilingual training methodologies for G2P conversion, especially for low-resourced languages where not a lot of training data is available.

12 Acknowledgement

Author Shubhanker Banerjee was supported by Science Foundation Ireland under Grant Agreement No. 13/RC/2106_P2 at the ADAPT SFI Research Centre at University Of Galway.

References

- Simon Ager. 2008. Omniglot-writing systems and languages of the world. Retrieved January, 27:2008.
- Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.
- Sunitha Basodi, Chunyan Ji, Haiping Zhang, and Yi Pan. 2020. Gradient amplification: An efficient way to train deep neural networks. *Big Data Mining and Analytics*, 3(3):196–207.
- Maximilian Bisani and Hermann Ney. 2008. Joint-sequence models for grapheme-to-phoneme conversion. *Speech communication*, 50(5):434–451.
- Daniela Braga, Luís Coelho, and Fernando Gil Vianna Resende. 2006. A rule-based grapheme-to-phone converter for TTS systems in European Portuguese. In *2006 International Telecommunications Symposium*, pages 328–333. IEEE.
- Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

- Stanley F. Chen. 2003. Conditional and joint models for grapheme-to-phoneme conversion. In *Proc. 8th European Conference on Speech Communication and Technology (Eurospeech 2003)*, pages 2033–2036.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Phuc Do and Truong HV Phan. 2022. Developing a BERT based triple classification model using knowledge graph embedding for question answering system. *Applied Intelligence*, 52(1):636–651.
- Lu Dong, Zhi-Qiang Guo, Chao-Hong Tan, Ya-Jun Hu, Yuan Jiang, and Zhen-Hua Ling. 2022. Neural grapheme-to-phoneme conversion with pre-trained grapheme models. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6202–6206. IEEE.
- Cherifi El-Hadi and Guerti Mhania. 2017. Phonetisaurus-based letter-to-sound transcription for Standard Arabic. In *2017 5th International Conference on Electrical Engineering-Boumerdes (ICEE-B)*, pages 1–4. IEEE.
- Isaac Elias, Heiga Zen, Jonathan Shen, Yu Zhang, Ye Jia, Ron J Weiss, and Yonghui Wu. 2021. Parallel tacotron: Non-autoregressive and controllable TTS. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5709–5713. IEEE.
- I.J. Gelb. 1968. Orthography studies, articles on new writing systems: William A. Smalley and others. *Helps for Translators, Volume VI* (Published by the United Bible Societies, London, in co-operation with the North-Holland Publishing Company, Amsterdam, 1964). VII + 173 pages. *Lingua*, 20:319–323.
- Alex Graves, Navdeep Jaitly, and Abdel-rahman Mohamed. 2013. Hybrid speech recognition with deep bidirectional LSTM. In *2013 IEEE workshop on automatic speech recognition and understanding*, pages 273–278. IEEE.
- Michael Hammond. 2021. Data augmentation for low-resource grapheme-to-phoneme mapping. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 126–130, Online. Association for Computational Linguistics.
- SJ Hannahs. 2013. Celtic initial mutation: pattern extraction and subcategorisation. *Word Structure*, 6(1):1–20.
- Sepp Hochreiter and Jürgen Schmidhuber. 1996. LSTM can solve hard long time lag problems. *Advances in neural information processing systems*, 9.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Sittichai Jiampojamarn, Grzegorz Kondrak, and Tarek Sherif. 2007. Applying Many-to-Many Alignments and Hidden Markov Models to Letter-to-Phoneme Conversion. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 372–379.
- Yanliang Jin, Jinfei Xie, Weisi Guo, Can Luo, Dijia Wu, and Rui Wang. 2019. LSTM-CRF neural network with gated self attention for Chinese NER. *IEEE Access*, 7:136694–136703.
- Markéta Juzová, Daniel Tihelka, and Jakub Vít. 2019. Unified Language-Independent DNN-Based G2P Converter. In *INTERSPEECH*, pages 2085–2089.
- Preethi Jyothi and Mark Hasegawa-Johnson. 2017. Low-resource grapheme-to-phoneme conversion using recurrent neural networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5030–5034. IEEE.
- Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.
- Byeongchang Kim, Gary Geunbae Lee, and Jong-Hyeok Lee. 2002. Morpheme-based grapheme to phoneme conversion using phonetic patterns and morphophonemic connectivity information. *ACM Transactions on Asian Language Information Processing (TALIP)*, 1(1):65–82.
- James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.
- Jackson L. Lee, Lucas F.E. Ashby, M. Elizabeth Garza, Yeonju Lee-Sikka, Sean Miller, Alan Wong, Arya D. McCarthy, and Kyle Gorman. 2020. Massively multilingual pronunciation modeling with WikiPron. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4223–4228, Marseille, France. European Language Resources Association.
- Xinjian Li, Florian Metze, David Mortensen, Shinji Watanabe, and Alan Black. 2022. Zero-shot learning for grapheme to phoneme conversion with language ensemble. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2106–2115, Dublin, Ireland. Association for Computational Linguistics.
- Ryo Masumura, Naoki Makishima, Mana Ihori, Akihiko Takashima, Tomohiro Tanaka, and Shota Orihashi. 2020. Phoneme-to-grapheme conversion based large-scale pre-training for end-to-end automatic speech recognition. In *INTERSPEECH*, pages 2822–2826.

- Josef R Novak, Nobuaki Minematsu, and Keikichi Hirose. 2012. WFST-based grapheme-to-phoneme conversion: Open source tools for alignment, model-building and decoding. In *Proceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing*, pages 45–49.
- Josef Robert Novak, Nobuaki Minematsu, and Keikichi Hirose. 2016. *Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the WFST framework*. *Natural Language Engineering*, 22(6):907–938.
- Priyadarshini Patil, Chandan Rao, Gokul Reddy, Riteesh Ram, and SM Meena. 2022. Extractive Text Summarization Using BERT. In *Proceedings of the 2nd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications*, pages 741–747. Springer.
- Russell Paul. 2014. *An Introduction to the Celtic Languages*. Longman Linguistics Library. Routledge.
- John J Pickeral III. 1990. A Preliminary Phonology of Manx. *Orbis*, 35:81–97.
- Nikhil Prabhu and Katharina Kann. 2020. *Frustratingly easy multilingual grapheme-to-phoneme conversion*. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 123–127, Online. Association for Computational Linguistics.
- Kanishka Rao, Fuchun Peng, Haşim Sak, and Françoise Beaufays. 2015. Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4225–4229. IEEE.
- David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1985. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science.
- De Rumelhart. 1986. Learning internal representations by error propagation. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, 1:318–362.
- Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. 2012. LSTM neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association*.
- Paul Taylor. 2005. Hidden Markov Models for grapheme to phoneme conversion. In *Ninth European Conference on Speech Communication and Technology*. Citeseer.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.
- Kaili Vesik, Muhammad Abdul-Mageed, and Miikka Silfverberg. 2020. *One model to pronounce them all: Multilingual grapheme-to-phoneme conversion with a transformer ensemble*. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 146–152, Online. Association for Computational Linguistics.
- Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 606–615.
- Warren Weaver. 1949. The mathematics of communication. *Scientific American*, 181(1):11–15.
- Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3081–3089.
- Sevinj Yolchuyeva, Géza Németh, and Bálint Gyires-Tóth. 2019. Transformer based grapheme-to-phoneme conversion. *Proc. Interspeech 2019*, pages 2095–2099.
- Chendong Zhao, Jianzong Wang, Xiaoyang Qu, Haoqian Wang, and Jing Xiao. 2022. r-g2p: Evaluating and Enhancing Robustness of Grapheme to Phoneme Conversion by Controlled Noise Introducing and Contextual Information Incorporation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6197–6201. IEEE.

Improving Graph-to-Text Generation Using Cycle Training

Fina Polat

University of
Amsterdam
f.yilmazpolat
@uva.nl

Ilaria Tiddi

Vrije Universiteit
Amsterdam
i.tiddi
@vu.nl

Paul Groth

University of
Amsterdam
p.t.groth
@uva.nl

Piek Vossen

Vrije Universiteit
Amsterdam
p.t.j.m.vossen
@vu.nl

Abstract

Natural Language Generation (NLG) from graph structured data is an important step for a number of tasks, including e.g. generating explanations, automated reporting, and conversational interfaces. Large generative language models are currently the state of the art for open ended NLG for graph data. However, these models can produce erroneous text (termed hallucinations). In this paper, we investigate the application of *cycle training* in order to reduce these errors. Cycle training involves alternating the generation of text from an input graph with the extraction of a knowledge graph where the model should ensure consistency between the extracted graph and the input graph. Our results show that cycle training improves performance on evaluation metrics (e.g., METEOR, DAE) that consider syntactic and semantic relations, and more in generally, that cycle training is useful to reduce erroneous output when generating text from graphs.

1 Introduction

Graph-to-Text generation (G2T) is a subtask of open-ended Natural Language Generation (NLG) that aims to create fluent natural language text describing an input graph, and is part of common NLG benchmarks (Gehrmann et al., 2021). G2T conversion is particularly of interest for open-ended generation tasks such as dialogue generation and generative question answering (Ribeiro et al., 2021; Trisedya and et al., 2019). Large generative language models are currently the state of the art for open ended NLG from graph data (Gehrmann et al., 2021). A major problem faced by these models is the output of non-sensical or unfaithful content to the provided input. This phenomenon is known as hallucination (Ji et al., 2022).

Figure 1 displays an example of Graph-to-Text conversion. The NLG model, a large language model (T5-small, Raffel et al. (2020)) is finetuned with a widely used benchmark corpus (WebNLG,

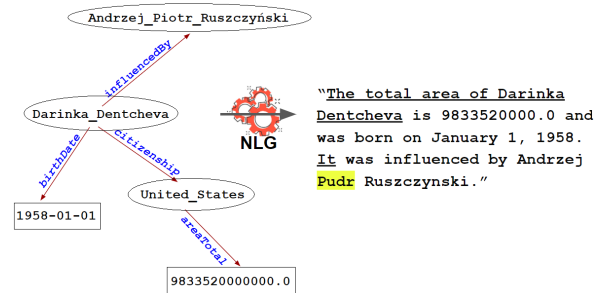


Figure 1: Graph-to-Text generation example with a hallucinatory verbalization.

Zhou and Lampouras (2020)), is asked to convert a graph taken from WebNLG. The output contains several errors. For example, *Darinka Dentcheva* is mentioned, as if she were a location, and attributed a total area. The generation continues with a proper verbalization of *birthdate*, but then again the model fails by referring *Darinka Dentcheva* with the pronoun *it*. Another mistake is the generation of an incorrect name. *Andrzej Piotr Ruszczyński* becomes *Andrzej Pudr Ruszczyński*.

Hallucinations are divided into two categories (Ji et al., 2022): intrinsic and extrinsic. In Figure 1, the intrinsic hallucinations are underlined, and the extrinsic hallucination is highlighted. Intrinsic hallucinations are the generation of output that contradicts the input graph, does not make sense, or contains some sort of commonsense violation. Extrinsic hallucinations are generations that cannot be verified by the source. Thus, the output can neither be supported nor contradicted by the input graph.

In this paper, we aim at addressing these problems by employing cycle training. Cycle training makes use of inverse tasks to add the model with additional signals. Here, the inverse task of G2T is Text-to-Graph (T2G) conversion where structures in the form of knowledge graphs are extracted from the text. In particular, we propose to use the T2G component of the cycle training to detect hallucinatory information in the generation by comparing

the extracted triples with the input triples. Additionally, combining G2T and T2G conversions is expected to improve the quality of the generated text and faithfulness of an NLG system because we hypothesize that cycle training would teach the NLG model to remain faithful to the input graph with the support of cycle consistency. Therefore, combining these two tasks is thought to improve the quality of the verbalization and reduce the hallucinatory generation. Our full code is available online.¹

The contributions of this paper are as follows:

1. An approach that employs cycle training to improve NLG faithfulness by reducing hallucinatory generation. Specifically, the approach introduces a T2G component to detect entity and relation mentions that are not part of the input graph.
2. A performance evaluation of this approach using three traditional lexical overlap metrics and two entailment evaluation methods used in the hallucination literature and show that the metrics with linguistic foundations (e.g. METEOR(+6%), DAE(+5%)) show significant improvement with cycle training.

2 Related Work

In recent years, there has been a paradigm shift in NLG. The shift stems from improvements in deep contextual language modeling and transfer learning (Ji and et al., 2020). NLG systems typically prioritize being coherent and discourse-related, disregarding control over generated content and its qualities such as faithfulness, factuality, freshness, and correctness. However, having control over the output is a major factor in NLG applications within industry (Leng and et al., 2020). Since cycle training reinforces the faithfulness of the NLG model and has the potential to detect extra information that is not part of the input, we relate our work to this controlability literature.

The state-of-the-art G2T generation results come from large generative models, but it is well known that these models are prone to hallucination. It is important to notice that all NLG tasks suffer from the hallucinatory text generation, and a control mechanism to solve this problem has not been found yet (Ji et al., 2022).

¹https://github.com/cltl-students/fina_polat_nlg_with_transformers.

Leveraging the fact that two functions are inverse of each other has been widely used in a variety of tasks in computer vision and machine translation (Godard et al., 2017; Sennrich et al., 2016). In the context of G2T, cycle training is used to address parallel data scarcity. Parallel graph-text data collection is difficult and costly. Therefore supervised approaches to both G2T and T2G conversions suffer from a shortage of domain-specific parallel graph-text data. Guo et al. (2020) and Schmitt et al. (2020) propose cycle training approach as an unsupervised learning solution when there is no or limited parallel data.

Guo et al. (2020) employ high-performing Named Entity Recognition (NER) tools such as Stanza (Qi and et al., 2020) to extract the entities and then build graphs with these automatically extracted entities. They train a G2T model called CycleGT using these automatically built graphs as the input graph in a cycle training regime. They test their unsupervised approach on parallel graph-text datasets such as WebNLG to compare their results with supervised approaches. We build on this work but instead of focusing on addressing the problem of data scarcity, we focus on the problem of hallucinations.

3 Approach

Our approach uses supervised cycle training with the objective of cycle consistency. Specifically, we employ CycleGT from Guo et al. (2020) and train it from scratch for five epochs. As our baseline, we use a pre-trained generative language model, the small version of T5, and finetune it for five epochs as well. For the training of CycleGT and the finetuning of the baseline T5, we use the WebNLG Dataset with the given train-test split. However, our approach is data and model agnostic and all components could be replaced with alternatives.

CycleGT is originally designed to address the parallel data scarcity and to be used as an unsupervised learning method when there is no or limited graph annotation. In the unsupervised setup, Guo et al. (2020) reduce the graph extraction task to relation prediction and rely on the Stanza NER module to extract the entities. Their results show that this approach works well to tackle parallel data scarcity. However, we are not interested in the unsupervised approach because we do not tackle the data scarcity problem, but instead we aim at less hallucinatory G2T generation.

As our objective is to improve the quality of the generated text by reducing/eliminating extrinsic hallucinations, supervision is essential for our case. We assume high-quality parallel graph-text data is given, and we rely on cycle consistency for improving generation quality, and T2G module for detecting extrinsic hallucinations. To the best of our knowledge, this is the first attempt to investigate cycle training in G2T for reducing/eliminating extrinsic hallucinations, reinforcing model faithfulness, and overall generation quality.

We compare the performance of CycleGT to the T5 baseline. All the experiments are run on a personal laptop. We now describe the data and models in more detail.

3.1 Data

WebNLG (Zhou and Lampouras, 2020) is a widely used G2T corpus that is created from DBpedia (Mendes and et al., 2011). DBpedia is a multilingual knowledge base that was built from various kinds of structured information contained in Wikipedia. This data is stored as RDF² triples, complies with Linked Data standards, and results in a high-quality dataset.³

3.2 Models

We choose T5 (Raffel et al., 2020) as the baseline pretrained language model, because it is state-of-the-art on the WebNLG dataset. Furthermore, T5 is a good representative sample of a generative large language model. We experiment with CycleGT because its G2T module is also based on T5 architecture that makes comparison easier. However, CycleGT does not exploit the pretrained language model but only utilize the architecture.

3.2.1 Baseline - T5

The “Text-to-Text Transfer Transformer” (or T5) is a unified framework that converts all text-based language problems into a text-to-text format (Raffel et al., 2020). The basic idea underlying the T5 model is to treat every textual task as a translation from input text to output text. In our case, the task consists in taking RDF triples as input, and producing a new text describing these triples as the output.

²Resource Description Framework: <https://www.w3.org/RDF/>

³https://gitlab.com/shimorina/webnlg-dataset/-/tree/master/release_v3.0

We finetune the small version of T5 model with the given train-test split of WebNLG for five epochs using Transformers library (Wolf and et al., 2020).

3.2.2 CycleGT

The G2T module of CycleGT transforms the graph to text. And, the T2G converts text to the graph by aligning each text with its back-translated version, and also each graph with its back-translated version. Since pretrained language models are shown to be effective on G2T conversions, Guo et al. (2020) use T5 (Raffel et al., 2020) architecture as the G2T component.

T2G produces a graph based on the given text. Guo et al. (2020) see relation extraction as the core problem in T2G conversion. In the supervised setup, T2G module of CycleGT directly uses the entities as they are given. Relations are predicted between every two pairs of entities with an LSTM-based Neural Network to form the edges in the graph. For our experiments, CycleGT is trained for five epochs in a supervised setup.

4 Evaluation

Considering the difficulty of quantifying hallucination, we use five different metrics for evaluation and divide them into two categories. The first category solely relies on lexical (n-gram) overlap while the second group is based on textual entailment.

4.1 Lexical Overlap Metrics

Lexical overlap metrics are widely used in NLG. The central idea behind these metrics is closeness. One of the simplest approaches is to leverage lexical features (n-grams) to calculate the similarity between the generation and the target text. We use BLEU (Papineni and et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005) as the lexical overlap metrics.

4.2 Entailment Metrics

Apart from well-established lexical overlap evaluation metrics, textual entailment models have been employed to evaluate the quality of automatically generated text. The entailment evaluation models are shaped around the idea that all information in the generated text should be entailed/inferred by the reference (gold) text.

For the evaluation of our NLG models, we employ two metrics that leverage entailment models: PARENT (Dhingra et al., 2019) and DAE (Goyal and Durrett, 2021).

Model	BLEU	ROUGE	METEOR	PARENT			DAE
				Precision	Recall	F1 score	
T5-small	19.6257	<u>0.5668</u>	0.4157	0.1910	<u>0.0976</u>	<u>0.0939</u>	0.2347
CycleGT	<u>20.9327</u>	0.5463	<u>0.4740</u>	<u>0.1980</u>	0.0894	0.0927	<u>0.2829</u>

Table 1: Graph-to-Text module evaluation scores.

4.2.1 PARENT

Lexical overlap metrics (BLEU, ROUGE, METEOR etc.) leverage the target text as the reference, and they do not take the input graph into account for the evaluation. However, it is common for a graph verbalization to have multiple plausible outputs from the same input.

Precision And Recall of Entailed N-grams from the Table, or PARENT, compares the generated text to the underlying graph as well as the reference text to improve evaluation. When computing precision, PARENT uses a union of the reference and the graph, to reward correct information missing from the reference. When computing recall, it uses an intersection of the reference and the graph, to ignore extra/incorrect information in the reference. The union and intersection are computed with the help of an entailment model to decide if an n-gram is entailed by the graph.

4.2.2 DAE

The DAE, or Dependency Arc Entailment, evaluation method is inspired by the downstream application of textual entailment models. Goyal and Durrett (2020) propose another formulation of the entailment that decomposes it at the level of dependency arcs. Rather than focusing on aggregate decisions, they instead ask whether the semantic relationship manifested by individual dependency arcs in the generated output is supported by the input. Arc entailment is a 2-class classification: entailed or not-entailed. This means that arcs that would be neutral or contradictory in the generic entailment formulation are considered non-entailed.

This approach views dependency arcs as semantic units that can be interpreted in isolation. Each arc is therefore judged independently based on whether the relation it implies is entailed by the reference sentence. A dependency arc in the generated sentence is assumed to be entailed by the reference if the semantic relationship between its head and child holds for the reference sentence. If the dependency relation does not hold for a head-child pair, then it is considered a factual error, and

the mismatched head-child span can be marked as the hallucinatory generation.

4.3 Human Evaluation: Qualitative Analysis

Automatic evaluation metrics struggle to deal with semantic or syntactic variations. Therefore, we need human judgment even though it is costly. For qualitative analysis, we sample 100 instances from the test set, and one annotator performs the annotations following a two step annotation scheme. First, we annotate whether the generation contains any hallucination, a binary decision. If the generation is hallucinatory, we add the hallucination type, one of the following classes: intrinsic, extrinsic, or both.

5 Results and Discussions

Due to the limited compute resources, we choose smaller models, and train or finetune them for just five epochs. Therefore, the performance of our models could not reach to the range in other NLG experiments. However, we observe noticeable improvement in METEOR and DAE scores. We now detail the results of our experiments.

5.1 Automatic Evaluation Results

In Table 1, we report the results of the automatic evaluation metrics. ROUGE and METEOR scores are reported in terms of F1 score. For readability, the highest scores are underlined.

The CycleGT model trained in cycle consistency outperforms the finetuned T5 model in precision-oriented metrics: +1,3070 BLEU score and +0,0070 PARENT-precision. However, the finetuned T5 model takes the lead in terms of ROUGE (+0,0205) and PARENT-recall (+0,0082) scores. Precision and recall results of PARENT are consistent with BLEU and ROUGE. This is expected because BLEU is a precision-oriented score while ROUGE is recall oriented.

It is notable that CycleGT gets higher scores in terms of METEOR (+0,0583) and DAE (+0,0482). Compared to the precision-oriented scores, the difference in METEOR and DAE is more significant.

Both METEOR and DAE are built on evaluation models with a linguistic backup. METEOR, for instance, not only compares the text as a direct string match but also exploits synonymy. For a linguistically sound comparison, it uses the Porter Stemmer and WordNet as lexical database. Similarly, DAE is empowered by a dependency parsing framework.

METEOR and DAE are both empowered by linguistic backup, and they are designed to be able to measure the quality of a generation on higher levels, e.g. semantics. The shortcoming of these models is that the linguistic enhancements are also built on sub-modules, off-the-shelf tools, and automatically created datasets that are known to be prone to error propagation. Regardless of their flaws, METEOR and DAE are more advanced evaluation methods enhanced with linguistic backup compared to their alternatives. We also argue that the higher performance of CycleGT in terms of METEOR and DAE is indicative that these metrics are more suitable to automatically judge the quality of a generation.

5.2 Evaluation of the T2G Component

The evaluation of the T2G module of CycleGT is important due to three reasons. First, we expect CycleGT model to generate better and less hallucinatory (at least on the extrinsic side) text because it is trained in cycle consistency. The second reason is that we employ the T2G module of CycleGT to detect extrinsic (not part of the input, but made up by the NLG model) hallucinations in the generation. Therefore, it is supposed to be able to extract all the information in the generated text. Finally, both modules (G2T & T2G) are supposed to be equally strong for getting the maximum benefit from cycle training.

T2G	F1 Score		% of predictions
	overall	partial	
CycleGT	0.1407	0.7873	32%

Table 2: Evaluation scores of the Text-to-Graph module.

In Table 2, we report the evaluation results of the CycleGT T2G module. F1 scores are micro averaged. The T2G module displays recall deficiency. The overall performance of the graph extraction module is pretty poor (0.14 F1 score). The module usually fails to make at least one prediction per instance. The maximum number of predictions is 1662 (32%) out of 5150 test instances. This means

that the model is unable to extract any triples from 68% of the test instances. However, it makes precise predictions when it does as indicated by the higher partial F1 score (0.78).

The poor performance of the T2G module of CycleGT reduces the robustness of cycle training. In order to enforce cycle consistency, a stronger T2G performance is necessary. Moreover, it is not possible to detect extrinsic hallucinations with this performance. Capturing extrinsic hallucinations would only be possible by a comparison between the input triples and the extracted triples. Therefore, it would be beneficial to aim at a better-performing triple extraction model to detect extrinsic hallucinations and reinforce cycle consistency.

5.3 Human Evaluation Results

Model	Only Intrinsic	Only Extrinsic	Both Int.&Ext
T5-small	11%	21%	20%
CycleGT	34%	18%	10%

Table 3: Qualitative Results.

Table 3 presents human evaluation results. This qualitative analysis confirms that CycleGT generates fewer extrinsic hallucinations. In our test sample, 18% of the CycleGT generations contain extrinsic hallucinations while the finetuned T5 model has 41%. Looking at the percentage of intrinsic hallucinations, the T5 model displays a better performance. On the one hand, we observe the generation of CycleGT mostly remains faithful to the input graph but contains wrong lexical associations (34%) with entities and their relations that occur as intrinsic hallucinations. On the other hand, we see that the finetuned T5 model makes more precise associations between entities and their relations but often makes up new entity names that were not part of the graph input (extrinsic hallucinations).

6 Conclusion

The use of generative models for NLG has led to improved performance, however, these models can still produce text with erroneous statements (i.e. hallucinations). In this paper, we show that combining G2T and T2G conversions in a cycle training setup helps such models improve the generated text conditioned on graph data. Automatic evaluation is one of the recognized obstacles for NLG.

To bypass the evaluation bottleneck, we exploited linguistics-enhanced evaluation methods such as METEOR and DAE. We find out that a more robust T2G module may help maximize the benefits of cycle training for NLG.

7 Acknowledgments

This work was partially supported by the European Union’s Horizon Europe research and innovation programme within the ENEXA project (grant Agreement no. 101070305).

References

- Satanjeev Banerjee and Alon Lavie. 2005. **METEOR: An automatic metric for MT evaluation with improved correlation with human judgments**. In *Proceedings of the ACL Workshop*, pages 65–72, Michigan. ACL.
- Bhuvan Dhingra, Manaal Faruqui, and et al. 2019. **Handling divergent reference texts when evaluating table-to-text generation**. In *Proceedings of the 57th Annual Meeting of the ACL*, pages 4884–4895, Florence, Italy. ACL.
- Sebastian Gehrmann, Tosin Adewumi, and et al. 2021. **The GEM benchmark: Natural language generation, its evaluation and metrics**. In *Proceedings of the 1st Workshop on NLG, Evaluation, and Metrics*, pages 96–120, online. ACL.
- Clément Godard, Oisín Mac Aodha, and Gabriel J Brostow. 2017. **Unsupervised monocular depth estimation with left-right consistency**. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279.
- Tanya Goyal and Greg Durrett. 2020. **Evaluating factuality in generation with dependency-level entailment**. In *Findings of the ACL: EMNLP 2020*, pages 3592–3603, Online. ACL.
- Tanya Goyal and Greg Durrett. 2021. **Annotating and modeling fine-grained factuality in summarization**. In *Proceedings of the 2021 NAACL: HLTs*, pages 1449–1462, Online. ACL.
- Qipeng Guo, Zhijing Jin, and et al. 2020. **CycleGT: Unsupervised graph-to-text and text-to-graph generation via cycle training**. In *Proceedings of the 3rd International Workshop on NLG from the Semantic Web*, pages 77–88, Dublin, Ireland (Virtual). ACL.
- Yangfeng Ji and et al. 2020. **The amazing world of neural language generation**. In *Proceedings of the 2020 Conference on EMNLP: Tutorial Abstracts*, pages 37–42, Online. ACL.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, and et al. 2022. **Survey of hallucination in natural language generation**. *ACM Comput. Surv.* Just Accepted.
- Yuanmin Leng and et al. 2020. **Controllable neural nlg: comparison of sota control strategies**. In *Proceedings of the 3rd International Workshop on NLG from the Semantic Web*, pages 34–39, Virtual. ACL.
- Chin-Yew Lin. 2004. **ROUGE: A package for automatic evaluation of summaries**. In *Text Summarization Branches Out*, pages 74–81, Ben, Spain. ACL.
- Pablo N Mendes and et al. 2011. **Dbpedia spotlight: shedding light on the web of documents**. In *Proceedings of the 7th international conference on semantic systems*, pages 1–8.
- Kishore Papineni and et al. 2002. **Bleu: A method for automatic evaluation of machine translation**. In *Proceedings of the 40th Annual Meeting on ACL, ACL ’02*, page 311–318, USA. ACL.
- Peng Qi and et al. 2020. **Stanza: A python natural language processing toolkit for many human languages**. In *Proceedings of the 58th Annual Meeting of the ACL: System Demonstrations*, pages 101–108, Online. ACL.
- Colin Raffel, Noam Shazeer, and et al. 2020. **Exploring the limits of transfer learning with a unified text-to-text transformer**. *Journal of ML Research*, 21(140):1–67.
- Leonardo F. R. Ribeiro, Martin Schmitt, and et al. 2021. **Investigating pretrained language models for graph-to-text generation**. In *Proceedings of the 3rd Workshop on NLP for Conversational AI*, pages 211–227, Online. ACL.
- Martin Schmitt, Sahand Sharifzadeh, Volker Tresp, and Hinrich Schütze. 2020. **An unsupervised joint system for text generation from knowledge graphs and semantic parsing**. In *Proceedings of the 2020 EMNLP*, pages 7117–7130, Online. ACL.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. **Improving neural machine translation models with monolingual data**. In *Proceedings of the 54th Annual Meeting of the ACL (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. ACL.
- Bayu Distiawan Trisedya and et al. 2019. **Neural relation extraction for knowledge base enrichment**. In *Proceedings of the 57th Annual Meeting of ACL*, pages 229–240, Florence, Italy. ACL.
- Thomas Wolf and et al. 2020. **Transformers: State-of-the-art natural language processing**. In *Proceedings of the 2020 Conference on EMNLP: System Demonstrations*, pages 38–45, Online. ACL.
- Giulio Zhou and Gerasimos Lampouras. 2020. **WebNLG challenge 2020: Language agnostic delexicalisation for multilingual RDF-to-text generation**. In *Proceedings of the 3rd International Workshop on NLG from the Semantic Web*, pages 186–191, online. ACL.

FinAraT5: A text to text model for financial Arabic text understanding and generation.

Nadhem Zmandar and Mahmoud El-Haj and Paul Rayson

UCREL NLP group,
School of Computing and Communications,
Lancaster University, UK

Abstract

The financial industry generates a significant amount of multilingual data, and there is a pressing need for better multilingual NLP models for tasks such as summarisation, structure detection, and causal detection in the financial domain. However, there are currently no pre-trained finance-specific Arabic language models available. To address this need, we continue the pre-training of AraT5 to create FinAraT5, the first pre-trained Arabic language model specifically designed for financial use cases, trained on a large Arabic financial communication corpus consisting of annual and quarterly reports and press releases. We hypothesise that FinAraT5 would perform better than AraT5 on financial domain tasks. We demonstrate this through research on a publicly available discriminative task (translated from English), and a generative task from a novel summarisation dataset called FinAraSum. Our results show FinAraT5 is highly competitive with state-of-the-art models such as mT5, AraBART, BERT, and the original AraT5 on Arabic language understanding and generation tasks.

1 Introduction

Pre-trained language models are a hot topic in Natural Language Processing. Despite their success, most are trained on English or multilingual datasets. Leveraging the vast amount of unlabeled data available online, they provide an efficient way to pre-train continuous word representations that can be fine-tuned for a downstream task, along with their contextualization at the sentence level. Generally, pre-trained models are trained on massive corpora using GPUs or recently TPUs. Most follow the architecture proposed by (Vaswani et al., 2017). Sequence-to-sequence is the best architecture for abstractive models, and abstractive models are very efficient for news summarisation and text paraphrasing. Unlike extractive summarisation, abstractive approaches are not restricted to the input words (Rush et al., 2015; Chopra et al., 2016).

Arabic is a very rich language with few resources, and significantly fewer language models compared to English and other Latin languages. Arabic remains understudied in the Natural Language Processing (NLP) community. In addition, Arabic NLP and generation tasks have proven to be very challenging to tackle. Most Arabic language models are mainly encoder only and are not field-specific (Antoun et al., 2020).

The middle eastern stock exchanges have an increasing market cap motivated by oil and gas companies, real estate companies and especially investment companies (e.g. kingdom holding). Therefore, the middle eastern markets are gaining in popularity among western investors, especially with the evolution of jurisdiction in the UAE through the free trade zone and the flexibility of investment in a Gulf-listed company. In addition, the Tadawul Saudi Exchange is the ninth most significant stock market among the 67 members of the World Federation of Exchanges by market capitalization of listed companies (approximately US\$2.6 trillion on 30 June 2021) and is the dominant market in the Gulf Cooperation Council (GCC). The successful IPO of Saudi Aramco demonstrates this. Tadawul is also included in the MSCI, FTSE Russell and S&P Emerging Market indices. It is the third largest stock market amongst its emerging market peers. It is an affiliate member of the International Organization of Securities Commissions (IOSCO), the World Federation of Exchanges (WFE), and the Arab Federation of Exchanges (AFE). These facts point to the increasing importance and scale of textual financial data in Arabic, which needs to be followed by an advance in Arabic NLP covering finance and investment-related tasks. Therefore, we propose the training of a monolingual Arabic T5 model customized for financial corpora.

We present FinAraT5, based on araT5, as a continuation of pre-training of araT5 on a large collected monolingual financial Arabic corpus. Un-

like previously released Arabic BERT different versions, FinAraT5 is adapted for both generative and discriminative tasks. We evaluate the pre-trained model on financial sentiment analysis and financial news summarisation on a novel Arabic news summarisation dataset, FinAraSum, that we collected ourselves. This work aims to meet the need for a monolingual financial text-to-text model for the Arabic language since no previous public model existed. One issue with the past work targeting Arabic abstractive summarisation is the evaluation of such models on highly extractive datasets. The primary available Arabic extractive datasets are ANT Corpus (Chouigui et al., 2017) and KALIMAT (El-Haj and Koulali, 2013). Therefore in this study, we prepare our customized highly abstractive financial summarisation dataset to suit the financial model we created. Our contributions through this research paper are summarised as follows:

- We present the first pre-trained Arabic text-to-text financial language model pre-trained on financial narratives Arabic corpus. The model features 220 Million parameters and is trained on 25 GB of PDF text for 45 days using a google cloud TPU V3.8. The model is suitable for generative and discriminative tasks.
- We describe the steps to collect, convert, preprocess and clean a financial narratives corpus covering different middle eastern stock exchanges.
- We present the collection and creation of FinAraSum, a highly abstractive financial and economic news dataset which are an Arabic equivalent of OrangeSum (Kamal Eddine et al., 2021) and Xsum (Narayan et al., 2018)
- We evaluate FinAraT5 on discriminative and generative tasks and show that it produces promising results.
- We compare FinAraT5 with different versions of multilingual T5 to prove the importance of training monolingual language models.
- We show that FinAraT5 achieves state-of-the-art results on the small Arabic benchmark we created. It outperforms Bert based model, multilingual text-to-text models and some general-purpose Arabic models.

- All our models are integrated into a hugging face repository to facilitate replicability and reuse.

2 Background and Related work

2.1 T5 transformer

The T5 (Raffel et al., 2019a) text-to-text transformer is a sequence-to-sequence (encoder-decoder) language model pre-trained on a multi-task mixture of unsupervised and supervised tasks for which each task is converted into a text-to-text format. T5 works well on various tasks by prepending a different prefix to the input corresponding to each task (e.g., for translation: translate English to German; for summarisation: summarize:). It is configured for 4096 maximum input tokens. However, the model is based on relative position embeddings, which allows it to scale to longer input sequences. Because of the complexity $O(n^2)$ of the Transformer's self-attention mechanism, such scaling increases memory consumption exponentially. The idea of a unified Transformer framework for different tasks was introduced by (Raffel et al., 2019a). The T5 framework treats all generative and discriminative tasks as a text-to-text problem. This enabled a more efficient transfer learning approach. In addition, Google researchers recently extended the T5 model to multilingualism by releasing mT5 (Xue et al., 2021), a multilingual version of T5. In this work, we will also test the portability of mT5 to the Arabic language and explore its performance on Arabic financial tasks, for the first time.

Several models trained for seq2seq models were previously released. Seq2seq models connect the left encoder and the right decoder part of the transformer with attention to enable the model to produce output. A Seq2Seq model achieves this by using the following scheme: Input tokens-> embeddings-> encoder-> decoder-> output tokens. Among the commonly used seq2seq models is the BART model, which was pre-trained on several languages such as French (Kamal Eddine et al., 2021) and English (Lewis et al., 2020). In addition, there is a multilingual version of BART (Liu et al., 2020).

2.2 Arabic Pre-trained Language models

Since the emergence of transformer models, a number of Arabic LMs has been developed. AraBERT (Antoun et al., 2020) was trained with the same architecture as BERT (Vaswani et al.,

2017) and used the BERT Base configuration. AraBERT is trained on 23GB of Arabic text, making approximately 70M sentences and 3B words from Arabic Wikipedia, the Open Source International dataset (OSIAN) (Zeroual et al., 2019), and (El-Khair, 2016) Corpus (1.5B words). Antoun et al. compared the performance of AraBERT to multilingual BERT from Google and other state-of-the-art models. The results prove that araBERT achieves state-of-the-art performance on most tested Arabic NLP tasks. ARBERT (Abdul-Mageed et al., 2021) is a large-scale pre-trained masked language model for Modern Standard Arabic. To train ARBERT, Abdul-Mageed et al. used the same architecture as BERT Base: 12 attention layers. It has approximately 163M parameters and was trained on a 61GB collection of Arabic datasets.

AraBART (Kamal Eddine et al., 2022b) is the first Arabic sequence-to-sequence model where the encoder and the decoder are trained end-to-end. It is based on BART. AraBART follows the architecture of BART Base which has 6 encoder and 6 decoder layers and 768 hidden dimensions. AraBART has 139M parameters and achieved state of art results on multiple abstractive summarisation datasets. araT5 (Nagoudi et al., 2022) created the first Arabic text to text model (araT5). They released three powerful Arabic text-to-text Transformer versions. For evaluation, they used an existing benchmark for Arabic language understanding and introduced a new benchmark for Arabic language generation (ARGEN).

JABER and SABER: Junior and Senior Arabic BERT (Ghaddar et al., 2021) found that most of the released Arabic BERT models were under-trained and therefore developed JABER and SABER, Junior and Senior Arabic BERT models. Experimental results show that their models achieve state-of-the-art performances on ALUE, a new benchmark for Arabic Language Understanding Evaluation.

2.3 Financial pre-trained language models:

Finbert: is the first BERT model pre-trained on financial narrative text. It is trained on a 4.9B tokens corpus composed of Corporate Reports 10-K and 10-Q (2.5B tokens), Earnings Call Transcripts (1.3B tokens), and Analyst Reports (1.1B tokens). Finbert is fine-tuned for three use cases: a

sentiment classification task, ESG classification task and forward-looking statement (FLS) FinBERT. Their fine-tuned FinBERT models are available on Huggingface’s transformers library¹. This model achieves superior performance on financial sentiment classification tasks. (Yang et al., 2020)

3 Training Corpus Description

Training a transformer model needs a large corpus in plain text because of the large number of parameters in the model’s architecture. There is no available public financial corpus covering financial statements in Arabic. Hence, we also created the training corpus ourselves. We aggregated two corpora of different orders of magnitude to train the models.

3.1 Financial Reports

In this section, we describe in detail our approach to collecting large-scale financial text in Arabic. The task is challenging, as financial reports are not readily available or centralised in one location.

Data Acquisition We collected several types of financial documents from different middle eastern markets: auditor reports, earning announcements, accounting documents, quarterly reports (Q1, Q2, Q3, Q4), annual reports and management board reports. A total of 30,000 PDF files were collected to form our source data. The total size of PDF files collected is around 25Gb.

We focused on major stock exchanges in the middle east to collect our corpus. Our data is collected from the following Arab markets: KSA exchange: TASI (Tadawul All Share Index) and NOMU (Saudi Parallel Market Growth parallel market), UAE (Dubai Financial Market (DFM), Abu Dhabi index), Kuwait (Boursa Kuwait), Oman (Muscat Stock Exchange), Qatar (Qatar Stock exchange) and Bahrain (Bahrain stock exchange).

The corpus is constituted as a diverse set of documents from different sectors and covers several categories. We have more than 35 categories in this corpus (E.g. financial services, Banking, insurance, telecommunication, oil and gas, energy, real estate, and utilities). We did not include the Egyptian financial disclosures since their data was not freely available. For other North African markets,

¹<https://huggingface.co/yiyanghkust/finbert-tone>

such as Morocco and Tunisia, companies communicate mainly in French rather than Arabic.

Table 1 describes the corpus in detail by providing summary statistics about the different indexes used in this corpus.

3.2 PDF to Text process

A significant constraint is the nature of the documents which are scanned PDF, contain old Arabic fonts or a lot of noise. In addition, the use of Arabic numerals and a lot of tabular data made the task of converting to text files very complex.

We selected the pro version of the sejda app, but firstly used a PDF2Text algorithm to convert our PDF reports to plain text files. If the conversion did not work, we used their Arabic OCR solution. The Arabic OCR inverts the order of words from left to right, hence this has to be corrected. Among the 30,000 collected reports, 24,000 were used in the process. We passed them through a PDF-to-text script in several batches. Converting as PDF2text worked very well for many reports. The success rate was more than 40%. Some scanned docs were converted but generated ASCII code files, meaning the conversion script cannot detect the content.

For the others, we used the OCR tool of sejda². On average, 10 PDF files took around one hour to be OCRed. The OCR operation took more than eight days in total, including the post-processing. Although the OCR solution of sejda is less efficient than we would like, it has an acceptable success rate given the poor quality of the report files. Finally, we performed a manual check to verify that all the files had the minimum required Arabic structure for our pre-training process. We manually deleted all the badly converted files. Further significant challenges during the data construction and data conversion process include the following aspects.

PDF2Text One of the common issues we observed from applying OCR on Arabic-written PDF files were repeated characters or additional spaces between the characters of one word (all the words are written with spaces) or concatenated words (not separated by spaces). This is reported to be a common issue for OCR in Arabic, especially if the quality of data is not good.

Memory Management Producing such a large-scale corpus is very time-consuming; hence we divided the whole task into small tasks. It took around three months to construct the corpus,

from web scraping until the last cleaned and pre-processed files are used in training.

OCR Low success rate for Arabic and especially a very long processing time given there was no possibility for parallel execution.

3.3 Newswires

In addition to our financial and board reports corpus, we selected more than 30,000 financial and economic news items from a leading news Arabic website. This helps to make our training corpus more diverse and enables coverage of several topics and styles of writing. All the corpus text is written in Modern Standard Arabic.

3.4 Cleaning

Once converted from PDF to text, we cleaned the text in order to be ready for the training. We used farasa³ for segmentation. We read files in chunks and applied our cleaning pipeline. This process started by removing all diacritics, HTML elements and their attributes, all special characters, and English alphabets and digits. We also removed tatweel characters, which are used regularly in Arabic writing. We reduced repeated characters to single characters, removed links and long words (longer than 15 chars). We used (Alyafeai and Saeed, 2020) to prepare our cleaning and preprocessing pipeline.

4 FinAraT5: Our financial text-to-text model

FinAraT5 is the first financial Arabic language model designed for text generation and text understanding. It is trained using a text-to-text approach. Our model is based on araT5 (Nagoudi et al., 2022), a pre-trained Arabic text-to-text model. It is the first financial Arabic model pre-trained in an encoder-decoder manner.

4.1 Architecture

We use the BASE architecture of T5 encoder-decoder (Raffel et al., 2019a), with 12 encoder layers and 12 decoder layers. Both the encoder and decoder have 12 attention heads and 768 hidden units. In total, therefore, FinAraT5 Base is an encoder-decoder with 220M parameters.

4.2 Vocabulary

Because we are continuing the pre-training of araT5, we opted for using the same vocabulary

²<https://www.sejda.com/ocr-pdf>

³<https://farasa.qcri.org/segmentation/>

Index	Tasi	Nomu	Dubai	AD	Kuwait	Qatar	Oman	Bahrain
# companies	223		178	73	163	47	111	42
MKT cap	3158.57		294.83		105.98	165.39	13.00	24.60
Time range	2003-2021		2009-2021		2012-2021	2010-2021	2015-2021	2014-2021
# reports	19651		3338		3192	2454	23	536
# sectors	21		11		13	7	2	6

Table 1: Statistics for the financial pre-training corpus. This table shows correct figures as at July 2022 from different sources such as statista.com. The columns represent the different indexes used. The rows describe the number of listed companies included in the report, market caps in US billion dollars, time range of the corpus, number of reports collected and the number of sectors included in the corpus. AD stands for Abu Dhabi stock exchange.

model used to train araT5 by Nagoudi et al., which was created using SentencePiece (Kudo and Richardson, 2018) which encodes text as WordPiece tokens (Bostrom and Durrett, 2020) with 110K WordPieces. Hence, our vocabulary model has a size of 110,000.

4.3 Training details

Pre-Training: We pre-train FinAraT5 on a TPU V-3.8 (with 8 cores) offered by Google cloud, with a learning rate of 0.001. We used the Adam optimizer (Kingma and Ba, 2014) and fix the batch size to 100,000 tokens. We set the maximum input and target sequence length to 512 sequences. We continued the training of the araT5 MSA base for additional 500,000 steps. We started from step 1 million, where the arat5 was stopped. In total, we pre-train FinAraT5 for 1.5 million steps⁴. The pre-training took around 40 days on the google cloud platform.

Pre-training TASK T5 was pre-trained on a mixture of supervised (mask language modelling) and unsupervised tasks. AraT5 was pre-trained using an unsupervised task. Therefore we use the same pre-training strategy as araT5, which is an unsupervised learning task trained on a raw plain text of financial qualitative data in Arabic. We cloned the architecture of T5 directly from the T5 repository⁵. We defined the task and performed the training using the t5 library⁶, which enables us to perform the training using Tensorflow and get a Mesh TensorFlow Transformer.

⁴We note that the English T5Base (Raffel et al., 2019b) was trained only for 512K steps

⁵<https://github.com/google-research/text-to-text-transfer-transformer>

⁶<https://pypi.org/project/t5/>

5 FinAraBen: Financial Arabic benchmark

To evaluate any pre-trained models, we need to compare them against a benchmark task. Unfortunately, there are no public financial datasets in Arabic that could be used in this study. In fact, in the case of Arabic finance texts, labelled datasets are very scarce resources. Thus, we created a new benchmark for the financial Arabic language called FinAraBen which includes two datasets: financial text summarisation and financial sentiment analysis. The first was collected, cleaned and created by ourselves. The second was translated from a previously released dataset in English.

5.1 FinAraSum dataset

The FinAraSum dataset was inspired by the XSum dataset and OrangeSum dataset. It was created by scraping the “Arabyia asswak” website⁷. Alarabya is a large Saudi information media with 21.0M visitors per month. It publishes in Arabic and English, covering the MENA region. We decided to create our own Arabic financial news dataset to solve the issue of the need for more open sources of NLP datasets. The choice was to create a dataset adapted to abstractive summarisation, which is news headline generation. This enables testing the efficiency of the pretrained model by testing the generative component of the model, which is itself a challenging task in NLP.

Motivation: We followed the collection procedure described by (Narayan et al., 2018) and (Kamal Eddine et al., 2021) who presented Xsum and OrangeSum respectively, which are highly abstractive datasets. We present the financial Arabic version of Xsum, which is more abstractive.

⁷<https://www.alarabiya.net/aswaq>

Collection Process: We collected the newswires from “Al Arabiya Asswak” website⁸. The choice of this news source is motivated by the fact that it is the largest news website in the middle east, with 21M monthly visitors. Alarabya has specialized financial and economic journalists writing several articles daily covering the region’s financial news. They mainly use Modern Standard Arabic. The collected dataset covers seven categories: financial markets, economics, real estate, energy, economy, tourism and special stories. We collected all the available news articles covering a decade from 2012 to 2021.

Statistics about the FinAraSum: Table 2 compares FinAraSum with the previously released dataset such as CNN, DailyMail, NY Times, OrangeSum and XSum datasets. Our dataset is smaller than Xsum, CNN, NYT, and Daily Mail but larger than the OrangeSum title and OrangeSum abstract. Table 2 shows that our dataset comprises 44,900 newswires in the training split. The article body and the title are 238.3 and 9 words in length on average, respectively. The dataset was very clean and did not require any specific post-processing. Table 3 shows that our dataset is more abstractive than the previously released one, making it a very challenging task for our financial pretrained model. There are 37.8% novel unigrams in the FinAraSum Gold summaries, compared with 35.76% in Xsum, 26.54% in OrangeSum title, 30.03% in OrangeSum Abstract, 16.75% in CNN, 17.03% in DailyMail, and 22.64% in NY Times. Similar results are reported for Bigrams, Trigrams and 4-grams. This proves that FinAraSum is more abstractive than previously released datasets.

Split FinArasum train/val/testing We randomly split the dataset into train, validation, and test splits. The test set is composed of 2,500 news articles. The validation is composed of 1,500, with the remainder for training.

5.2 Financial Sentiment Analysis Dataset

Currently, to the best of our knowledge, there are no available financial sentiment analysis corpora in the Arabic language. For our experiments, we used the FinancialPhrase dataset⁹. The dataset was collected by (Malo et al., 2013). This release of the financial phrase bank covers a collection of 4,840 sentences.

⁸<https://www.alarabiya.net/aswaq>

⁹https://huggingface.co/datasets/financial_phrasebank

The selected collection of phrases was annotated by 16 people with adequate background knowledge of financial markets. We used sentences with more than 50 per cent agreement. To pre-process the classification dataset, we separated it into inputs and labels. The inputs are financial-related sentences, and the labels are sentiments (positive, neutral, negative). Then we encoded our labels as follows ‘positive’: 0, ‘neutral’:1, ‘negative’:2. We then split our dataset into training (80%) and testing (20%), and we ensured that our split respected a normal distribution of our labels. The training and testing datasets’ length are 3,876 and 970, respectively.

6 Experiments and Results

6.1 Financial Text Summarisation

The task of headline generation was addressed several times in past summarisation challenges, such as the Document Understanding Conferences (DUC) for 2002, 2003 and 2004.

Technical decision Usually, the summarisation script would set the loss function as the rouge score. In this study, we changed the loss function to the Bert score using the multilingual BERT checkpoint. Therefore, we could monitor the evolution of the Bertscore loss function in real time on the training and validation split using the Weights and Biases AI tool¹⁰. In addition, we used early stopping and took the best checkpoint on the validation split. We use the multilingual version of the BERT language model. This choice is justified by the highly abstractive nature of our dataset. Before this decision, we tried to train our models by minimizing the loss function of rouge and Bleu scores. However, Bertscore was the best choice and performed very well on the validation dataset. We used the original implementation of BertScore¹¹. Bertscore calculates the similarity of the contextual embeddings of the system and reference summaries. We set our evaluation process to be executed at every step. For this work, we trained mT5 small, base, and large. We were unable to train the mT5 Xlarge due to memory limitations. We also trained arat5 small, arat5 base, araBart large and bert2bert base. For BERT2BERT, we followed the methodology proposed by Rothe et al.. We created a sequence-to-sequence model whose encoder

¹⁰<https://wandb.ai>

¹¹We use the official implementation https://github.com/Tiiiger/bert_score

Dataset	Train/Val/Test	Avg Doc Length		Avg Summary length		Vocab Size	
		words	Sentence	words	Sentence	Docs	Sum
CNN	90.3/1.22/1.09	760.50	33.98	45.70	3.58	34	89
Daily mail	197/12.15/10.40	653.33	29.33	54.65	3.86	564	180
NYT	590/32.73/32.73	800.04	35.55	45.54	2.44	1233	293
Xsum	204/11.33/11.33	431.07	19.77	23.26	1.00	399	81
Orangesum title	30.6/1.5/1.5	315.31	10.87	11.42	1.00	483	43
Orangesum Abstract	21.4/1.5/1.5	350	12.06	32.12	1.43	420	71
FinAraSum(ours)	44.90/1.5/2.5	238.3	10.15	9.0	1.0	492	46

Table 2: Sizes (column 2) are given in thousands of documents. Document and summary lengths are in words. Vocab sizes are in thousands of tokens as reported in (Kamal Eddine et al., 2021)

Dataset	% of novel n-grams in gold summary				LEAD		
	Unigrams	Bigrams	Trigrams	4-grams	R-1	R-2	R-L
CNN	16.75	54.33	72.42	80.37	29.15	11.13	25.95
Daily mail	17.03	53.78	72.14	80.28	40.68	18.36	37.25
NYT	22.64	55.59	71.93	80.16	31.85	15.86	23.75
Xsum	35.76	83.45	95.50	98.49	16.30	1.61	11.95
Orangesum title	26.54	66.70	84.18	91.12	19.84	08.11	16.13
Orangesum Abstract	30.03	67.15	81.94	88.3	22.21	07.00	15.48
FinAraSum(ours)	37.8	73.6	89.0	95.2	18.30	07.5	14.79

Table 3: Degree of abstractivity of FinAraSum compared with that of other datasets, as reported in (Narayan et al., 2018) and (Kamal Eddine et al., 2021). It can be observed that FinAraSum is more abstractive than XSum and OrangeSum and traditional summarisation datasets.

and decoder parameters are multilingual uncased Bert base model¹². We will oblige the mbert model to work as an encoder and a decoder to generate the summary. To obtain the reported results, we fine-tuned all pretrained models for 22 epochs with train and validation data, and we used a learning rate that warmed up to $5e-5$ with a batch size of 8. LEAD-1 baseline is included, a competitive extractive baseline for news summarisation by extracting the first sentence. We report BERTScore (Zhang et al., 2020), BARTScore (Yuan et al., 2021), Bleurt score (Sellam et al., 2020), Meteor (Banerjee and Lavie, 2005), Rouge (Lin, 2004), infolm score (Colombo et al., 2021) and Frugal score (Kamal Eddine et al., 2022a). Frugal 1 uses a tiny bert base mover scorer¹³. Frugal 2 uses a tiny deberta bertscore¹⁴.

Table 4 benchmarks the performance of the models fine-tuned on the headline generation task. Fi-

nAraT5 shows very promising results compared to multilingual versions of mT5, especially with Base and Small models. It outperformed all the small and base models. This confirms the importance of pre-training monolingual models. Finally, all T5-based models outperform BERT2BERT by a significant margin.

Table 5 reports results of infoLM score (Colombo et al., 2021) on FinAraSum test split. This score calculates the mathematical distribution of the reference and candidate sentences then it calculates the mathematical distance between the two distributions. The less the distance is, the better the result is. We report different mathematical distances. The authors claim that regarding fluency and text structure, FisherRao distance works better.

We also report about rouge metrics. We report ROUGE-1, ROUGE-2 and ROUGE-L f1- scores (Lin, 2004). The original google implementation of rouge does not support the Arabic language. Instead, we used another implementation¹⁸. This table is for informational purposes only because

¹²<https://huggingface.co/bert-base-multilingual-uncased>

¹³https://huggingface.co/moussaKam/frugalscore_tiny_deberta_bert-score

¹⁴https://huggingface.co/moussaKam/frugalscore_tiny_deberta_bert-score

¹⁸https://github.com/ARBML/rouge_score_ar

	title generation					
	BE score	BA score	Frugal 1	Frugal 2	Bleurt	meteor
lead	72.66	44.51	85.10	86.30	-15.00	27.08
mT5 small	79.17	62.48	91.50	89.30	5.90	32.43
araT5 small	79.68	63.33	91.65	89.40	6.70	33.84
bert2bert base	75.50	56.27	91.26	89.20	-1.57	18.25
mT5 base	79.03	62.44	91.46	89.30	5.51	31.27
araT5 base	80.21	64.37	92.04	89.50	8.29	35.18
finaraT5 base(ours)	80.46	64.66	92.04	89.52	8.76	36.08
mT5 large	80.32	64.54	92.04	89.45	9.42	35.47
araBART Large	80.35	64.67	92.30	89.55	9.50	35.18

Table 4: Results on FinAraSum test split. BE Score stands for Bert score which uses uncased multilingual bert checkpoint. BA score stands for Bart score and uses the mbart checkpoint¹⁵. Macro F1 score averages are computed over all datasets. Frugal 1 uses a tiny bert base mover scorer¹⁶. Frugal 2 uses a tiny deberta bertscore¹⁷

	kl	alpha	beta	ab	renyi	l1	l2	l_infinity	fisher_rao
lead	-8.829	-4.252	6.993	9.256	2.206	1.893	0.285	0.134	2.887
mT5 small	-8.165	-4.090	6.705	8.258	2.053	1.861	0.292	0.144	2.832
mT5 base	-8.294	-4.120	6.830	8.387	2.086	1.867	0.295	0.145	2.842
mT5 large	-8.370	-4.123	6.880	8.462	2.089	1.867	0.297	0.147	2.845
araBART	-8.669	-4.157	7.125	8.777	2.136	1.870	0.300	0.147	2.858
araT5 small	-8.387	-4.104	6.858	8.484	2.067	1.863	0.297	0.149	2.840
araT5 base	-8.376	-4.093	6.809	8.501	2.059	1.859	0.296	0.147	2.835
finaraT5 base	-8.334	-4.077	6.789	8.408	2.041	1.856	0.295	0.146	2.830

Table 5: Reporting Results of infoLM (Colombo et al., 2021) on FinAraSum test split. The authors of InfoLM claim that it is a flexible metric and it can adapt to different criteria using different measures of information. KL stands for kl divergence between the reference and hypothesis distribution. alpha and beta stand for alpha and beta divergence between the reference and hypothesis distribution. Renyi stands for renyi divergence between the reference and hypothesis distribution. l1 and l2 and l_infinity stands for three versions of norm distances between the reference and hypothesis distribution. FisherRao is the distance between the reference and hypothesis distribution. Finally, the authors claim that regarding fluency and text structure, FisherRao distance works better

MODEL	rouge1	rouge2	rougeL
lead	23.21	9.55	21.02
mT5 small	37.91	20.02	35.93
araT5 small	39.31	21.33	37.24
bert2bert	24.34	9.10	23.08
mT5 base	37.35	19.45	35.31
araT5 base	40.91	22.49	38.71
finaraT5 base	41.74	23.19	39.61
mT5 large	41.17	23.14	38.99
araBART	41.38	23.19	39.34

Table 6: Results on FinAraSum in terms of ROUGE-1 (R1), ROUGE-2 (R2), ROUGE-L (RL)

rouge variants are based on n-gram-form matching and have no sense of semantic similarity (Kamal Eddine et al., 2021).

In Table 7, we report the degree of novel ngrams introduced per model on the generated summaries on the test dataset. We can see that FinAraT5 introduces on average 28.8% , 64.5% , 82.6% , 91.5% of novel unigrams, Bigrams, Trigrams and 4-grams respectively, in its summaries for the title generation task. These scores are superior to other models. We can deduce that FinAraT5 and araT5 base are more abstractive than other models, especially multilingual T5. Bert2Bert is an exception since it generates some random words. This may be justified by the fact that it is not a native encoder-decoder model.

We followed the method proposed by (Rothe et al., 2020). We calculated the percentage of repetition and the average length of the generated summary. The repetition rate is the rate of summaries including at least one word from the most frequent 400 words from the corpus. Results are detailed in Table 8. For repetitions, the less redundant models, closest to the ground truth, are araBART and MT5 large. The use of auto-generative models on abstractive datasets increases the risk of repetition. Our model FinAraT5 shows less repetition on this summarisation dataset than other models. This is a good sign of the quality and novelty of the generated text. Bert2Bert is the only model redundant with 15.76% of repetitions. The architecture of the model justifies this. In addition, this model generated more tokens on average. This is consistent with previous results. All the other models generate nine tokens coherently with the gold summaries' length.

6.2 Discriminative task: Financial sentiment prediction:

In order to further test the model we performed training on a discriminative task. We can use either encoder-only models or encode-decoder models. In the second, the input sequence is passed to both the encoder and the decoder and we add a classification head to the representation of the sequence of tokens. text-to-Text models can perform discriminative tasks

Training details: we fine-tuned the models for 20 epochs with a learning rate of $2e-5$. We set the batch size to be 32 and the max sequence length to 128.

Evaluation: Table 9 shows the results of the sentiment analysis task. We report only the models with a base architecture. FinAraT5 performed the best on the test split. We can conclude that the monolingual financial text model could perform well on generative and discriminative tasks.

6.3 Discussion

Multilingual vs. Monolingual Models The empirical results show the better performance of dedicated monolingual language models compared to multilingual models (multilingual T5 versions: 110 languages) of the same size (base). The FinAraT5 model benefits from the previously pre-trained araT5 on a large Arabic MSA corpus. In addition, it specialises in the financial context by being trained on a large financial narrative corpus. This improved performance could be explained by the quality of the data collected from different financial reports and financial newswires in Arabic. **Transfer Learning:** Multilingual models do not learn very well on some downstream tasks. Our monitoring of the evolution of bertscore using wandb.ai show that multilingual models do not improve significantly during training. They have a flat curve during the fine-tuning process compared to the monolingual models. mT5 models may suffer from capacity issues.

Abstractiveness: We manually evaluate our text-to-text models' ability to generate good quality financial context MSA text. Our qualitative analysis shows that the FinAraT5 is very powerful in summarising news and in generative tasks in general. It has a compelling ability to abstract and paraphrase the input. It introduces advanced grammatical Arabic structures, such as using question marks, exclamations, and oratorical questions. In addition,

Model	% of novel n-grams in system generated summary			
	Unigrams	Bigrams	Trigrams	4-grams
Gold	37.1	73.1	88.8	95.1
bert2bert	34.2	77.3	95.4	97.3
mT5 small	22.1	52.8	71.1	82.0
araT5 small	27.5	62.2	80.4	90.0
mT5 base	23.7	54.2	72.6	83.7
araT5 base	28.3	63.9	82.4	91.5
FinAraT5 base(ours)	28.8	64.5	82.6	91.5
mT5 large	26.3	60.8	79.5	88.8
araBART large	25.6	60.0	79.2	89.0

Table 7: Proportion of novel n-grams in the generated summaries on the test dataset using different models .

	Length	Repetition %
Gold	9.04	0.52
mT5_small	9.27	4.44
araT5_small	9.28	5.64
bert2bert	10.03	15.76
mT5_base	9.05	2.64
araT5_base.txt	9.08	3.64
finarat5_base	9.05	3.48
mT5_large	8.92	1.2
araBART large	8.71	1.04

Table 8: Summary statistics: Sequence length generated by models on the Test dataset and percentage of word repetition in the summary among the most common 400 words in the dataset

MODEL	arabert	mT5	araT5	finarat5
accuracy	0.9246	0.9246	0.9362	0.9449

Table 9: Sentiment analysis task on the test split

we see good use of commas, which is crucial in Arabic, enabling emphasis on some words. Finally, we can see that different versions of Arabic T5 generate content that has approximately the same meaning using different structures. In conclusion, we can see that our models are able to generate syntactically correct summaries in Arabic.

Evaluation methods: Three main types of metrics are used to measure the similarity between two sets of data: model-based, n-gram, and statistical-based. Model-based metrics use models to estimate the similarity between two sets of data. N-gram metrics measure the similarity between two data sets by counting the number of n-grams or phrases appearing in both data sets. Statistical-based metrics use statistical models to estimate the similarity between two data sets.

Grammatical: We manually analysed system summary generated examples. The generated text is syntactically correct, and the spelling is also correct. It is also in line with the general topic of the corpus. The method allowed the generation of coherent text and has succeeded in fully synthesising suitable Arabic financial text.

7 Conclusion And Future Work

We presented FinAraT5, a domain-specific skilled text-to-text model for financial Arabic text understanding and generation. We trained the model on a large dataset of Arabic financial texts which we collected and cleaned ourselves. Then we evaluated the model’s performance on a new benchmark that we created. The results showed that FinAraT5 could model and generate coherent and accurate texts in the Arabic financial domain, outperforming strong baselines and demonstrating its ability to be a good benchmark as a language model for

financial Arabic. Overall, we claim that FinAraT5 represents a significant step forward in the development of practical natural language processing tools for financial Arabic, which is at the moment still less well represented in previous research, and we believe it has the potential to be fine-tuned on several other downstream tasks (machine translation, summarisation, and information retrieval). Our next step is to perform a large-scale human evaluation task on Mechanical Turk.

8 Acknowledgements

We gratefully acknowledge support from Lancaster University to provide access to the high-end computing GPU cluster. We thank also the Google TensorFlow Research Cloud TFRC¹⁹ program for the free access to Cloud TPUs V3.8 which was crucial for the pre-training process. In addition, we thank Google Cloud research team for the 1,000 USD GCP credits²⁰ to perform this research. We also acknowledge the AraT5 team for their help and for sharing their model checkpoints.

References

- Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. **ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic**. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.
- Zaid Alyafeai and Maged Saeed. 2020. tkseem: A pre-processing library for arabic. <https://github.com/ARBML/tnkeeh>.
- Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. **AraBERT: Transformer-based model for Arabic language understanding**. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.
- Satanjeev Banerjee and Alon Lavie. 2005. **METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments**. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- Kaj Bostrom and Greg Durrett. 2020. **Byte pair encoding is suboptimal for language model pretraining**.
- Sumit Chopra, Michael Auli, and Alexander M. Rush. 2016. **Abstractive sentence summarization with attentive recurrent neural networks**. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98, San Diego, California. Association for Computational Linguistics.
- Amina Chouigui, Oussama Ben Khiroun, and Bilel Elayeb. 2017. **Ant corpus: An arabic news text collection for textual classification**. In *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, pages 135–142.
- Pierre Colombo, Chloe Clave, and Pablo Piantanida. 2021. **Infomn: A new metric to evaluate summarization & data2text generation**. In *AAAI Conference on Artificial Intelligence*.
- Mahmoud El-Haj and Rim Koulali. 2013. **Kalimat a multipurpose arabic corpus**.
- Ibrahim Abu El-Khair. 2016. **1.5 billion words arabic corpus**. *CoRR*, abs/1611.04033.
- Abbas Ghaddar, Yimeng Wu, Ahmad Rashid, Khalil Bibi, Mehdi Rezagholizadeh, Chao Xing, Yasheng Wang, Duan Xinyu, Zhefeng Wang, Baoxing Huai, Xin Jiang, Qun Liu, and Philippe Langlais. 2021. **JABER: junior arabic bert**. *CoRR*, abs/2112.04329.
- Moussa Kamal Eddine, Guokan Shang, Antoine Tixier, and Michalis Vazirgiannis. 2022a. **FrugalScore: Learning cheaper, lighter and faster evaluation metrics for automatic text generation**. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1305–1318, Dublin, Ireland. Association for Computational Linguistics.
- Moussa Kamal Eddine, Antoine Tixier, and Michalis Vazirgiannis. 2021. **BARThez: a skilled pretrained French sequence-to-sequence model**. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9369–9390, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Moussa Kamal Eddine, Nadi Tomeh, Nizar Habash, Joseph Le Roux, and Michalis Vazirgiannis. 2022b. **AraBART: a pretrained Arabic sequence-to-sequence model for abstractive summarization**. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 31–42, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Diederik P. Kingma and Jimmy Ba. 2014. **Adam: A method for stochastic optimization**.

¹⁹<https://sites.research.google/trc/about/>

²⁰<https://cloud.google.com/edu/researchers>

- Taku Kudo and John Richardson. 2018. [SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. [BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- Chin-Yew Lin. 2004. [ROUGE: A Package for Automatic Evaluation of Summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. [Multilingual denoising pre-training for neural machine translation](#). *Transactions of the Association for Computational Linguistics*, 8:726–742.
- Pekka Malo, Ankur Sinha, Pyry Takala, Pekka J. Korhonen, and Jyrki Wallenius. 2013. [Good debt or bad debt: Detecting semantic orientations in economic texts](#). *CoRR*, abs/1307.5336.
- El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022. [AraT5: Text-to-text transformers for Arabic language generation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 628–647, Dublin, Ireland. Association for Computational Linguistics.
- Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. [Don’t give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019a. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *CoRR*, abs/1910.10683.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019b. [Exploring the limits of transfer learning with a unified text-to-text transformer](#).
- Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. [Leveraging pre-trained checkpoints for sequence generation tasks](#). *Transactions of the Association for Computational Linguistics*, 8:264–280.
- Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. [A neural attention model for abstractive sentence summarization](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.
- Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. [Bleurt: Learning robust metrics for text generation](#). In *Proceedings of ACL*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). *CoRR*, abs/1706.03762.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. [mT5: A massively multilingual pre-trained text-to-text transformer](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.
- Yi Yang, Mark Christopher Siy UY, and Allen Huang. 2020. [Finbert: A pretrained language model for financial communications](#).
- Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. [Bartscore: Evaluating generated text as text generation](#).
- Imad Zeroual, Dirk Goldhahn, Thomas Eckart, and Abdelhak Lakhouaja. 2019. [OSIAN: Open source international Arabic news corpus - preparation and integration into the CLARIN-infrastructure](#). In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 175–182, Florence, Italy. Association for Computational Linguistics.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [BERTScore: Evaluating Text Generation with BERT](#). *arXiv:1904.09675 [cs]*. ArXiv: 1904.09675.

Modeling and Comparison of Narrative Domain Knowledge with Shallow Ontologies

Franziska Pannach

Göttingen Centre for Digital Humanities

University of Göttingen
franziska.pannach@
uni-goettingen.de

Theresa Blaschke

Marburg Center for
Digital Culture and Infrastructure
Philipps-Universität Marburg
blaschkt@
students.uni-marburg.de

Abstract

Ancient myths have fascinated scholars and laymen for centuries. In comparatistic efforts, classical scholars try to detect and interpret variations between versions of the same myth. We present a way to structure the underlying background information in myth variants. The background knowledge of twelve different versions of the popular myth *Orpheus and Eurydice* has been modeled in individual shallow ontologies that allow inter- and intra-myth comparison.

1 Introduction

The story of *Orpheus and Eurydice* is one of the most popular Greek myths with a long tradition of re-use and re-adaptation. Each of the variants of the myth uses certain elements of the narrative while leaving others out. One aspect of comparing those variants is to investigate not only the plot, but also which assumptions we can make about the circumstances in which the myth takes place, i.e. the background knowledge about the world it takes place in. Projects like Wikidata or Mythoskop¹ combine information from different sources and give a good overview of how characters and concepts are connected. However, investigating the difference between narratives, especially contradictory information, is an interesting research objective in itself.

Consider the following example: In most variants, *Orpheus* manages to reach the netherworld and is allowed to take *Eurydice* with him. But once he turns around to look at her, he loses her forever. However, why do we still consider it a variant of the same myth if *Orpheus* reaches the surface without turning around and is hence successful in bringing *Eurydice* back from the dead [6, L.1-14], [4]? The reason is, that we know that both variants concern the same characters and the circumstances are overall the same (e.g. *Eurydice* is in the netherworld.

¹<https://mythoskop.de/>

Orpheus has nothing but his musical talent to convince the inhabitants of the netherworld to release her.) In other words, the background is the same or at the very least similar. Additionally, the similarity of ancient mythical plots has already been studied thoroughly, e.g. by Bowra [1] or Marlow [8].

Hence, we focus this work on the question “Who is who and what is what?” and not “What happens?”

Comparatistic efforts of mythological narratives are still conducted mainly manually. In this paper, we demonstrate how we approach the comparison of the background information in mythological (and other) narrative domains in a manner that results in re-usable, machine-readable domain ontologies, and how we can use them to compare variants of the same myth.

2 Related Work

Nakasone and Ishizukua [10] use Rhetorical Structure Theory (RST) as a basis for a generic ontology model that focusses on storytelling paradigms.

In his work on the narrative formalism of Vladimir Propp [14], Peinado et al. [13] uses ontologies for automatic fairy tale generation. Ciotti [2] uses character-centric domain ontologies and highlights their importance in the Digital Humanities and the field of digital narratology.

Most digital analyses of narration focus on texts. Xu et al. [16] propose a model that uses ontologies and human annotation to capture narration on digitized artifacts, such as vase paintings, and other cultural heritage objects.

Re-tellings of folktales, similar to myth variants in this work, have been studied by [7]. Their story networks represent ancestral relationships between folktale variants, such as “Little Red Riding Hood”. However, they do not focus on the content of the tales.

For the mythical domain, the Mythoskop

project² presents a knowledge graph focussed on the relationship and genealogy of characters of the Greek mythology. The VAST (Values across space & time) project presents a semantic knowledge graph³ of annotations on “past of values”, including *Peace* or *Justice*. Their sources include Greek tragedies, among others.

3 Data

For this project, we use twelve myth variants of the myth of *Orpheus and Eurydice* from various antique sources. A complete list of sources and their abbreviations can be found in the project repository. The variants span a considerable time, with the earliest source approx. 400 BCE (Plato, Symposium) and the latest 875–1075 CE (Mythographus Vaticanus). The data consist of a number of statements per myth variant that form one narrative sequence describing the plot. They have been derived by domain experts of classical studies according to the hylistic approach [18][17].

This approach was developed specifically to extract and analyze narrative structures from mythological sources. It has been applied to different temporal and geographical backgrounds, such as ancient Mesopotamia, ancient Greece, or Egypt [3].

The individual statements in each sequence are derived from the original text of the source, e.g. a Greek poem, but they are not re-tellings of the story nor direct quotes from a translation, as the examples in Section 4 illustrate. Each sequence of statements was extracted by one or more domain experts, and reviewed, discussed, and agreed-upon within the research group.

The sequences of statements that describe a myth variant include two coarse types of elements: 1. statements concerning the background or circumstantial knowledge (durative) and 2. narrative statements that form the plot (single-point). We can distinguish these types of sequence elements by their truth values over the sequence.

For instance, the statement “Eurydice is the wife of Orpheus” is *true* at all times during the narrative sequence, while “Orpheus turns around” is *true* only once, at one point in the sequence. Consequentially, “Eurydice is dead” is *true* after she was killed by a snake, so only over a part of the sequence.

To compare narrative domains, i.e. all circumstantial and background knowledge available from a source about a myth, which are the basis of a myth variants, we only consider statements that are *true* over the entire sequence. In hylistic terms [18], those are considered *durative-constant*. Statements that are only *true* before or after a certain event, e.g. “Eurydice is dead.”, depend on the context of the narrative sequence (*durative-initial* or *-resultative*). Therefore, we do not consider them as parts of the overall background knowledge.

Each of the twelve sequences corresponding to one myth variant contains one or more of those statements, i.e. statements that hold true over the course of the entire variant (*durative-constant* statements). Those statements are *assertions* we can make about the domain knowledge, i.e. the world in which a plot takes place.

The statements describing the background knowledge were originally in German, but translated for this paper.

4 Domain modeling

We demonstrate the domain modeling approach using two variants of the myth of *Orpheus and Eurydice*. The English translation of the source text is shown below in Examples 1 and 2. The sequences of statements describing the plot and the narrative background knowledge are derived by experts in diverse mythological studies according to the hylistic approach [18].⁴ From both texts, we can derive background information that holds true in the respective variant. Table 1 shows which assertions can be made from the information in the sequences. Those assertions form the ground truth, the *a priori* knowledge for the ontology modeling process. According to the hylistic approach, we only consider statements that are relevant to the *Orpheus and Eurydice* myth. Statements like “Linus is Orpheus’ brother” are not considered, since they pertain to a different myth.

- (1) “But Orpheus, son of Oeagrus, they sent back with failure from Hades, showing him only a wraith of the woman for whom he came; her real self they would not bestow, for he was accounted to have gone upon a coward’s quest, too like the minstrel that he was, and to have lacked the spirit to die as Alcestis did for

²<https://mythoskop.de/>

³<https://ontology.vast-project.eu/>

⁴<https://www.uni-goettingen.de/en/556429.html>

the sake of love, when he contrived the means of entering Hades alive. Wherefore they laid upon him the penalty he deserved, and caused him to meet his death.”⁵

(2) “Now Calliope bore to Oeagrus or, nominally, to Apollo, a son Linus, whom Hercules slew; and another son, Orpheus, who practised minstrelsy and by his songs moved stones and trees. And when his wife Eurydice died, bitten by a snake, he went down to Hades, being fain to bring her up, and he persuaded Pluto to send her up. The god promised to do so, if on the way Orpheus would not turn round until he should be come to his own house. But he disobeyed and turning round beheld his wife;”⁶

Background information was collected for all twelve variants of the myth of *Orpheus and Eurydice*. Subsequently, a small controlled vocabulary⁷ specifically for the myth was created that allows matching of concepts, such as consort/wife/female spouse → wife. The concepts are given in German (*skos:prefLabel*) and English (*skos:altLabel*). The vocabulary also includes definitions (*skos:definition*) for the interpretation of the concepts, e.g. the definition of the concept *son* would be “direct male descendant of a person”.

While matching synonyms for the target languages, German and English, is a fairly straightforward task to automate, e.g. using WordNet [9] and GermaNet [5], the controlled vocabulary allowed us to create the ontologies more uniformly. Using controlled vocabulary for classes and relationships also helps to compare ontologies visually or by manual inspection. Additionally, the controlled vocabulary can be extended and re-used for other myths that contain similar concepts. Using those concepts, a set of twelve shallow ontologies were constructed following the guidelines outlined by

⁵Plato Symp. 179d <http://www.perseus.tufts.edu/hopper/text?doc=Perseus:text:1999.01.0174:text=Sym.:section=179d&highlight=Orpheus>

⁶Apollod. Lib. 1.3 <http://www.perseus.tufts.edu/hopper/text?doc=Perseus:text:1999.01.0022:text=Library:book=1:chapter=3&highlight=Orpheus>

⁷Controlled Vocabulary and ontologies in TTL-format are available under: <https://gitlab.gwdg.de/franziska.pannach/hylva> under a creative commons license.

Noy and McGuinness [11]. Important concepts are represented as ontology classes, such as terms for spouses or descendants, geographic concepts, or concepts of arts and music. Narrative characters like *Orpheus*, geographic locations and specialised concepts such as *Kitharodie* are individuals of the ontology. The resulting class hierarchy is shallow in the sense that only important higher-class concepts are modeled (e.g. wife → spouse → person).

In this regard, the ontologies are to our knowledge the only machine-readable and re-usable source of *source-specific* background knowledge for the individual sources. Figure 1 shows an example of a shallow domain ontology for the *Orpheus* myth in Apollodorus’ library.

The information in each ontology corresponds to one myth variant and one source. The information is not combined into a single ontology for two reasons. Firstly, background or circumstantial information between ontologies may be contradictory, e.g. with regard to a characters ancestry. Secondly, if one statement is missing in one source, but is present in another we cannot make assumptions about the truth value of the information in the first source. For example, “Orpheus is the beloved son of Oiagros.” implies *loves*(Oiagros,Orpheus) to be *true*. If another source does not contain that information, we cannot assume it to be *true* or *false*.

Object properties in the domain ontologies contain all relations that are not *isA*-relations derived from the background statements. These contain information such as spousal relationships or locations, e.g. *isIn*(Person, Location). Object properties have role restrictions for domain and range, depending on the classes they apply to.

Each ontology has translations of class concepts, object properties and data properties in German and English (*skos:altLabel*).

Public semantic sources such as Wikidata contain information on narrative characters, but they do not distinguish between source-specific and general information. For instance, the Wikidata entry on *Orpheus*⁸ states that his occupation is poet and writer, and that he was killed by Maenad. In the myth variants studied for this project, we can only derive that his profession was that of a musician, more specifically that he was a minstrel who practised *Kitharodie* (κίθαρωδία). The manner of his death is discussed in multiple variants, where it is stated as ‘being killed and torn to pieces by the

⁸<https://www.wikidata.org/wiki/Q174353>

Table 1: Background information from two myth variants

Plato	Apollodorus
Orpheus is the son of Oeagrus.	Orpheus is the son of Calliope and Oeagrus.
Eurydice is in Hades.	Eurydice is the wife of Orpheus.
	Eurydice is in Hades.
	Orpheus practises minstrelsy. (κithαρῳδία)

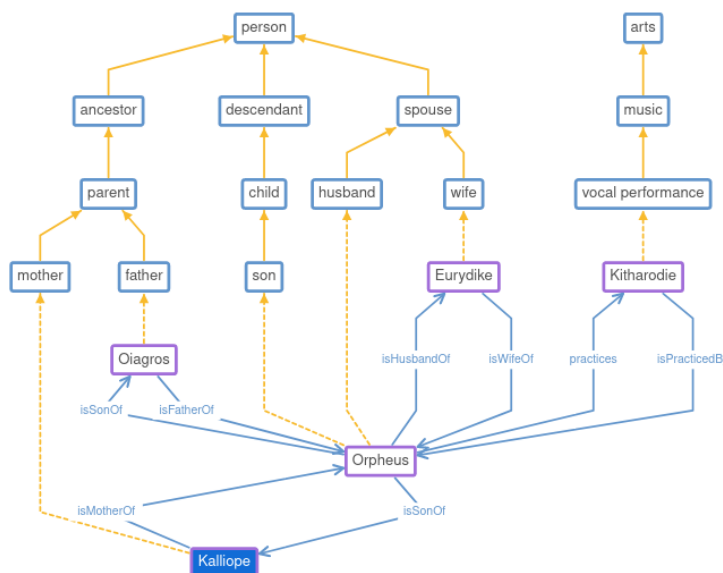


Figure 1: Ontology of ‘Orpheus and Eurydice’ concepts in Apollodorus’ library

Thracian women’.

Therefore, we cannot investigate or compare different views on the character *Orpheus* using resources like Wikidata. However, we link the information in the domain ontologies with the corresponding concepts in Wikidata via Wikidata ID, and Pleiades⁹ ID in case of geographic locations.

5 Domain Comparison

The resulting domain ontologies can be used to compare the domains, i.e. the background information we have about the characters and the setting within the narrative variant. We can do so by applying two measures: Firstly, we can compare classes of the ontology. This answers the question ‘Which general concepts are present in this narrative variant?’ This way, we can interpret the background information in Apollodorus’ library, as shown in Figure 1, as ‘some people who are related to each other either by marriage or ancestry’, and ‘some music presented in the form of song’. Since the controlled vocabulary was created during the ontology modeling process, we can match classes easily.

⁹<https://pleiades.stoa.org/>

We define class overlap as:

$$CO = \frac{|C_1 \cap C_2|}{|C_1 \cup C_2|}, \quad (1)$$

where C_i is the set of classes of ontology i . Secondly, we can map the individuals of the ontologies to answer questions like “Who appears in this story?” and “Who is this story about?”. Figure 1 shows the characters *Orpheus*, *Eurydice*, *Calliope*, and *Oeagrus* as individuals. We match individuals iteratively by: name, alias, and WikidataID or PleiadesID if the individual is a geographic entity. The node for *Orpheus* in the example ontology in Figure 1 has the most in- and out-going relations, represented as arrows. Graphically, he is the most ‘connected’ character in the domain ontology, we can derive that he is most likely the main character. We define the individual overlap as:

$$IO = \frac{|I_1 \cap I_2|}{|I_1 \cup I_2|}, \quad (2)$$

where I_i is the set of individuals of ontology i . Furthermore, the characters in different mythological sources can be compared using relations (object properties) between two or more characters or con-

cepts. This way, a degree of similarity between two different characters can be estimated, e.g. if both characters are sons of a father who is a king, or if both characters love a woman who is located in the netherworld. For the twelve variants of *Orpheus and Eurydice* such a measure was implemented but yielded few interesting results because the characters were either already matched due to name, alias or Wikidata ID, or too dissimilar to be compared in a meaningful way (e.g. *Hades* and *Kitharodie*).

6 Results

The twelve ontologies are freely available in TTL-format for download.¹⁰ Figure 2 shows the class and individual overlaps across variants. We see that for both class and individual overlap, the variant P_6 (Pausanias 9) is very dissimilar to the other variants.

We can match our results against information that is available about the sources. e.g. Pausanias only briefly mentions the story of Orpheus and Eurydice in his travel report [12].¹¹

On the level of ontology individuals, IO, the pair Plato-Hermesianax has the highest overlap score of 0.6. The highest score for the class overlap is 0.65 between the domain ontologies derived from statements based on the Mythographus Vaticanus and Apollodorus' Library.

In Figure 2c and 2d, we highlight only the closest matches between variants (without the self-matches on the matrix diagonal). Neither time of creation of the sources nor their geographic origin, e.g. Roman or Greek, seem to correspond to the similarity of the domain descriptions.

7 Discussion

The statements about the background information are based on the source texts of the original versions of a myth variant. To extract these is not a matter of simple NLP technique. Especially the decision on the truth value (single-point or durative) of a statement needs to be based on the source texts and made by informed experts on the material. This means that the extraction of these statements happens manually which is time-consuming. The construction of the ontology based on the background information, on the other hand, can be as-

¹⁰<https://gitlab.gwdg.de/franziska.pannach/hylva> Creative Commons license (CC-BY 4.0)

¹¹<http://www.perseus.tufts.edu/hopper/text?doc=Paus.+9.30.6&fromdoc=Perseus%3Atext%3A1999.01.0160>

sisted semi-automatically using simple rules, e.g. for *isA*-relationships. Common concepts, such as geographical concepts and entities, are available in common thesauri and semantic web resources such as Wikidata. Their freely available data could be re-used for our purposes. However, to link them in shallow ontologies instead of creating them as classes might not always be the best option. For instance, we suggest modeling locations with a distinction between mythological (e.g. Hades) and real – past or present – locations (e.g. Macedonia). In this sense, the class distance (in our case the depth to the lowest common ancestor (LCA)) in the shallow ontology between Hades and Macedonia would have a value of two. If we applied Wikidata classes, those two concepts would not share a meaningful common ancestor beyond *Wikidata metaclass*.

As discussed at the end of Section 5, we do not report similarity measures for relationships (object properties) for the myth variants studied in this paper. However, this measure is interesting for inter-myth comparison, where different characters with similar features appear. It can also serve useful to compare re-use of mythological storytelling in modern fiction, e.g. comparing the myth of Persephone to Ginny Weasleys story in Harry Potter and the Chamber of secrets [15]. Furthermore, the stylistic analysis and the comparison of narrative domain knowledge using shallow ontologies can be applied to other fictional genres as well, e.g. the study of folktales or comparison of different character representations in fanfiction, among others. We leave these efforts for future studies.

When studying modern texts in well-resourced languages, such as German or English, the extraction of sequences and subsequent ontology modelling could be assisted by automation through NLP methods, such as named entity recognition and semantic role labelling. With a larger number of texts and corresponding sequences, it would also be possible to automatically identify candidate statements from text. However, the creation of final sequences and knowledge bases, like the ones presented here, will most likely continue to include some form of manual work.

8 Acknowledgements

This work is funded by the DFG as part of the Myth-Research Group 2064 STRATA. The authors would like to thank Dr. Balbina Bäbler for her

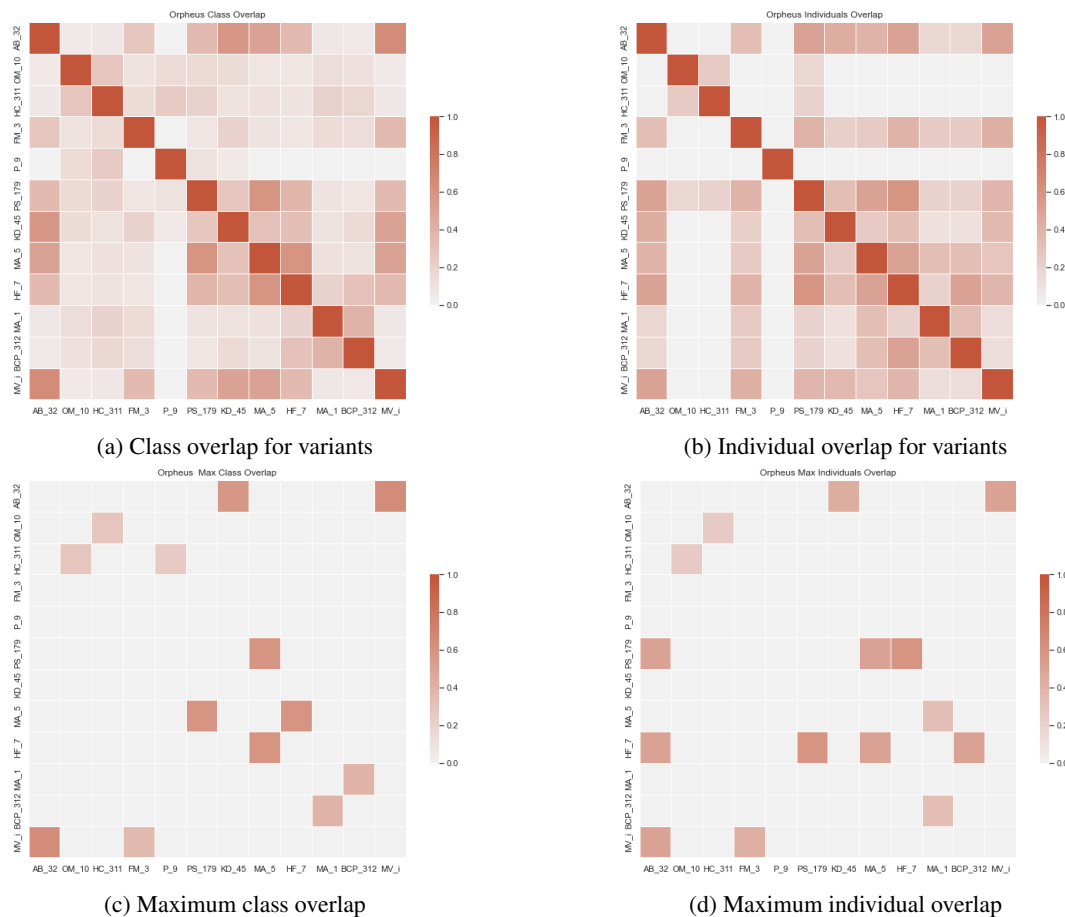


Figure 2: Overlap for variants of *Orpheus* journey to the netherworld

helpful insights and comments.

References

- [1] C. M. Bowra. 1952. *Orpheus and Eurydice*. *The Classical Quarterly*, 2(3-4):113–126.
- [2] Fabio Ciotti. 2016. *Toward a formal ontology for narrative*. *MATLIT: Materialidades da Literatura*, 4(1):29–44.
- [3] Gösta Ingvar Gabriel, Brit Kärger, Annette Zgoll, and Christian Zgoll, editors. 2021. *Was vom Himmel kommt: Stoffanalytische Zugänge zu antiken Mythen aus Mesopotamien, Ägypten, Griechenland und Rom*. De Gruyter, Berlin, Boston.
- [4] Fritz Graf and Ernst Badian. *Eurydice*.
- [5] Birgit Hamp and Helmut Feldweg. 1997. Germanet-a lexical-semantic net for german. In *Automatic information extraction and building of lexical semantic resources for NLP applications*.
- [6] Hermesianax. *Fragment 7*.
- [7] Folgert Karsdorp and Antal Van den Bosch. 2016. The structure and evolution of story networks. *Royal Society open science*, 3(6):160071.
- [8] A. N. Marlow. 1954. *Orpheus in Ancient Literature*. *Music and Letters*, XXXV(4):361–369.
- [9] George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.
- [10] Arturo Nakasone and Mitsuru Ishizuka. 2006. Storytelling ontology model using RST. In *2006 IEEE/WIC/ACM International Conference on Intelligent Agent Technology*, pages 163–169. IEEE.
- [11] Natalya F Noy, Deborah L McGuinness, et al. 2001. Ontology development 101: A guide to creating your first ontology.
- [12] Henry Arderne Ormerod, Richard Ernest Wycherley, William Henry Samuel Jones, et al. 1918. *Pausanias description of Greece*, volume 1. W. Heineemann.
- [13] Federico Peinado, Pablo Gervás, and Belén Díaz-Agudo. 2004. A description logic ontology for fairy tale generation. In *Procs. of the Workshop on Language Resources for Linguistic Creativity, LREC*, volume 4, pages 56–61.
- [14] Vladimir Propp. 1968. *Morphology of the Folktale*, volume 10. University of Texas Press.

- [15] Richard A Spencer. 2015. *Harry Potter and the Classical World: Greek and Roman Allusions in JK Rowling's Modern Epic*. McFarland, Jefferson.
- [16] Lei Xu, Albert Merono-Penuela, Zhisheng Huang, and Frank Van Harmelen. 2017. An ontology model for narrative image annotation in the field of cultural heritage. In *Proceedings of the 2nd Workshop on Humanities in the Semantic web (WHiSe)*, pages 15–26.
- [17] Annette Zgoll and Christian Zgoll, editors. 2020. *Mythische Sphärenwechsel: Methodisch neue Zugänge zu antiken Mythen in Orient und Okzident*. De Gruyter, Berlin, Boston.
- [18] Christian Zgoll. 2019. *Tractatus mythologicus: Theorie und Methodik zur Erforschung von Mythen als Grundlegung einer allgemeinen, transmedialen und komparatistischen Stoffwissenschaft*. De Gruyter, Berlin, Boston.

A new learner language data set for the study of English for Specific Purposes at university level

Cyriel Mallart¹, Andrew Simpkin², Rémi Venant³, Nicolas Ballier⁴,
Bernardo Stearns⁵, Jen Yu Li¹, Thomas Gaillat¹

¹LIDILE, Université Rennes 2

²School of Mathematics, Statistics and Applied Mathematics, University of Galway

³LIUM, Université du Mans

⁴CLILLAC-ARP, Université Paris Cité

⁵Insight, Data Science Institute, University of Galway

Abstract

This paper presents the release of a new data set for the study of English as a second language (L2), which is specialised in specific academic domains. The corpus includes 671 texts written by university students of different academic domains. All learners and their CEFR levels had to respond to the same task prompt eliciting language related to a domain. The data set includes structured textual data with rich Universal-Dependency linguistic annotation and metadata. It is available online in the CONLL-U format and can be exploited in several types of NLP tasks related to English L2 analysis.

1 Introduction

This paper reports on the release of the Corpus for the Study of Foreign Languages Applied to a Specialty (CELVA.Sp)¹, a new data set for the study of learner English. Learner corpora have been a topic for research for more than 30 years. They lend themselves to statistical methods for different types of analyses including Contrastive Interlanguage Analysis (CIA), error or linguistic complexity analysis or proficiency assessment. Today, a number of applications rely on learner corpora for modelling tasks. Output models are subsequently exploited in data processing pipelines tuned for specific language learning objectives. Learner corpora have turned out to be an essential resource for Computer-Aided Language Learning (CALL) systems.

¹Corpus d'Etude des Langues Vivantes Appliquées à une Spécialité. Available from the Huma-Num Nakala repository located at <https://nakala.fr/10.34847/nkl.41d57kb0>, DOI 10.34847/nkl.41d57kb0

In this context, it is essential to use data sets that have been collected with accuracy in controlled environments so as to ensure quality and experimental validity. English learner corpora have benefited from a lot of attention, resulting in the availability of several large corpora such as the Cambridge Learner Corpus (CLC) (Yannakoudakis et al., 2011), the EF™ CAMbridge DATabase (EFCAMDAT) (Geertzen et al., 2013) or the International Corpus of Learner English (ICLE) (Granger et al., 2020). In spite of their sizes, these corpora may suffer from one or more possible limitations such as limited access to raw data files, lack or unclear validity of proficiency annotation, lack of rich behavioural learning metadata. These limitations stem from the fact that learner corpus collection requires a lot of resources in terms of man/woman hours. Collecting such data means identifying learners willing to provide writings or oral recordings together with personal information regarding the learning behaviour, all of this while respecting privacy as required by the European GDPR directive. As a result, access to free, accessible and rich English L2 data sets is not so simple as it may appear. In addition, the aforementioned corpora tend to focus on learners by way of general English writing tasks. As a result, it is difficult to make comparisons between learners of different study domains such as medicine, pharmacy, computer science or sports.

Our proposal is to deliver an English L2 data set designed for the study of L2 English writing skills at university level and across ten different academic domains. We provide writings produced by 671 learners of six levels of proficiency. Learners' metadata are included and inform researchers on the learners' backgrounds and their behaviour

in learning English, e.g. exposure to English media, reading attitude, language trips and secondary school focus on advanced English classes. This data set is available in an interoperable format allowing automatic processing methods.

2 Related work

A number of learner English writing corpora exist on the commercial market. The International Corpus of Learner English (ICLE) version 3 is certainly one of the main resources in this field. It includes 9,529 long essays written by learners of twenty-six L1s and associated with educational metadata. It is also possible to apply for a non-commercial user licence for access to its exploration interface. The Cambridge Learner Corpus is commercial in its full version, but it includes a publicly released subset made up of exam scripts taken by candidates of the First Certificate in English (FCE). This subset includes 1,244 scripts together with proficiency marks and error annotation but it lacks metadata concerning the exam takers.

In the realm of non-commercial data corpora, the EFCAMDAT corpus is a collection of learner writings which have been classified in terms of proficiency levels. Its 1,180,309 scripts make it the biggest learner corpus of its kind as far as we know. It comes with some learner metadata such as learner nationality, EFTM proficiency levels, lesson units, task topics and grades. The learners' backgrounds are unknown and the evaluation of proficiency annotation is not reported in the paper.

Some learner corpora specifically focus on university students. The University of Pittsburgh English Language Institute Corpus (PELIC) (Juffs et al., 2020) focuses on university students and provides 46,230 scripts split into many different generic writing task topics. The NUS Corpus of Learner English (NUCLE) (Dahlmeier et al., 2013) is made up of about 1,400 essays, including error annotation, written by university students. Likewise the ASAG corpus (Tack et al., 2017) provides short texts written by third-level students as short answers to general-English questions. The corpus includes a subset of 299 writings that were graded according to the CEFR levels.

The aforementioned corpora rely on data that come from learners of English of unknown academic fields. The writing prompts were designed to fit all possible types of students and thus were

not necessarily linked to the field of studies. Yet, at university level, there is a need to study how learners of English for Specific Purposes (ESP) construct their linguistic knowledge in relation to their future professional domain. In this respect, the Varieties of English for Specific Purposes dAtabase (VESPA) (Paquot et al., 2022) provides more than 900 long essays written by learners of different L1 and different academic fields. This type of data is very useful to help explore and compare learner linguistic profiles across several domains.

We propose a more modest ESP corpus. Its main difference is that it relies on a single prompt designed to elicit domain-specific writings of the same genre and discourse types. This allows for comparisons between the writings of students of different academic fields. The texts are 200- to 300-words long and reflect a typical writing requirement set by language teachers in class. In addition, the writings are associated with learning-behaviour metadata and learner proficiency.

3 Corpus design

3.1 Data collection and task

The corpus includes learner texts in L2 English collected in two French universities of the same city. The learners were mostly French students between 2018 and 2020 at undergraduate level, ranging from first to third year.

The data was collected via a MOODLE Database² (Dougiamas and Taylor, 2003) designed specifically for this purpose. It can be installed on any MOODLE server for further collection in other educational environments.

The corpus texts were collected during class under the supervision of university language teachers trained on the collection protocol. It includes recommended metadata (Gilquin, 2015; Callies, 2015) about the characteristics of the subjects such as domain of studies, age, number of years studying the L2 and their learning behaviours such as frequency of exposure to L2 and travelling to L2 countries. Database fields were defined to control the possible values that could be entered, hence avoiding too much variability in categorical data names. The corpus data were then be exported as a UTF8 .csv file for further processing.

In terms of task, the learners were required to conduct a writing task with one and the same

²The MOODLE package is available from Gitlab URL

prompt. It required the description of an experiment/discovery/invention/technology/technique of their domain followed by their opinion on the impact of the described concept. The prompt was chosen as it allowed each learner to elaborate text dedicated to their own domain while ensuring the same text genre and discourse type. The learners had 45 minutes to complete the task.

Prior to recording their texts and learner profiles, learners were also requested to carry out the Dialang³ test (Alderson and Huhta, 2005). For practical reasons related to test taking duration in class, only the written module of the test was used with the exception of the "Placement test" screen and the "Self-assessment- writing" screen. In other terms, only the 30 cloze questions were used.

3.2 Data cleaning

After collecting the data, some records were discarded. These include the records where no email address is known, which is due to database tests. Duplicates, that is, records that contain exactly the same text from the same student but at two different times, were reduced to a single occurrence with the earliest date set as the submission date. Finally, we removed records in which the student wrote in Spanish or German while declaring that their L2 was English, and the samples in which the text was shorter than 10 words.

Some records were cleaned. The texts written by the students were cleared of all HTML formatting, while conserving the original paragraph structure. We simplified a variable that previously contained the names of advanced language sections followed by a student into a binary one. It now stores whether the student followed an advanced language curriculum in the past or not. Dates were set to a uniform format.

3.3 Data pseudonymization

In order to comply with the GDPR guidelines, the data were pseudonymized and learner-identifying information removed. Identifying information covers name, email address, age and level of studies. Other metadata relevant to the learning behaviour, and that do not allow for identification of an individual student, were kept, such as L1, number of years studying the L2, reading frequency, exposure to the language or number of trips taken

³see <https://dialangweb.lancaster.ac.uk/>

in an English-speaking country. Learners who answered negatively to whether they consented to the use or distribution of their data were also removed.

Each learner is represented by a secure encoding of their email address, created through an HMAC algorithm (Bellare et al., 1996) that uses a SHA256 cryptographic hash function. This algorithm encodes the email address of the student to a unique 64 letters and digits long pseudonym. This choice ensures unicity of the pseudonym. A secure SHA256 encoding of the email address requires a secret key, known only to the curators of the data set. Indeed, one pseudonym represents one student only. This will allow following the progression of a given student across time or tasks in the future. Should a participant revoke their consent to having their data used, the curators of the data set are capable of finding the records of this individual to remove them from the data set. This complies with the GDPR's guidelines on the right to request the rectification (and erasure) of personal data.

Beyond the metadata, learners may also disclose personal information in their writings. We replaced names with a placeholder, "Alex Dupont", instead of other methods such as initials or special symbols in order to stay as faithful as possible to the original language used by the student.

3.4 Linguistic annotation

In addition to plain text, the data set also contains linguistic information relying on the framework of Universal Dependencies (de Marneffe et al., 2021). The annotations notably include Universal Dependency tagged part-of-speech, lemmas of tokens, and morphological features such as case, number, gender, etc. These were obtained with the UDPipe pipeline (Straka et al., 2016) using the English model trained on the GUM corpus⁴ (Zeldes, 2017) as it was shown to be very reliable for POS and dependency annotation on L1 and L2 (Kyle et al., 2022). Evaluation of annotation accuracy was not conducted on these data .

4 Data set description

4.1 Metadata and text descriptions

The data set includes 671 writings from French-L1 learners and made up of 215 words on average (SD = 116.35) as shown in Figure 1. The writings are spread over ten different academic fields taught in

⁴english-gum-ud-2.5-191206

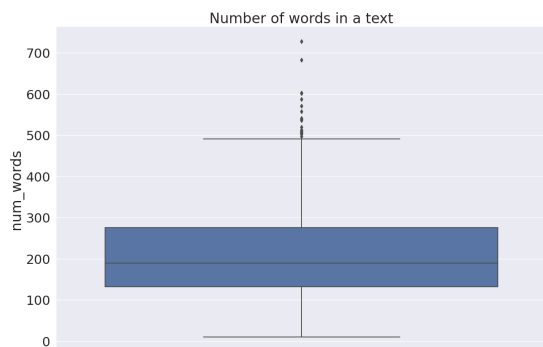


Figure 1: Distribution of the number of words per text

the universities of the city. Table 1 provides a detailed view of the data. Note that the imbalance is due to the domains in which data-collecting teachers were involved.

Domains	texts
Media Studies	199
Earth and Life Sciences	109
Medicine	96
Pharmacy	82
Computer Science and Electronics	65
Physics and Chemistry	40
Education Sciences	38
Science and Technology of Sport and Exercise	38
Mathematics	2
Social Sciences and Humanities	2

Table 1: Distribution of the number of texts per academic domain

All the writings are linked to the CEFR levels obtained by the learners in the DIALANG test. Figure 2 shows the distribution of texts per CEFR level. Interestingly, the number of words increases as CEFR levels increase except for the top C2 level. C2 learners seem to deflate their writing volume, maybe in favour of better pragmatic efficacy in discourse complexity and coherence. Figure 3 shows the variations of the number of words per level, giving an insight into the writing productivity of the learners. The metadata and the texts are all included in the same CSV file. The linguistic information about all the textual elements is included in a separate data file as described in Section 4.2. Both files are indexed with the pseudonymized identifier as described in Section 3.3.

4.2 Data formats

The data set adopts the CONLL-U format as part of a CSV file. More specifically, each CONLL-

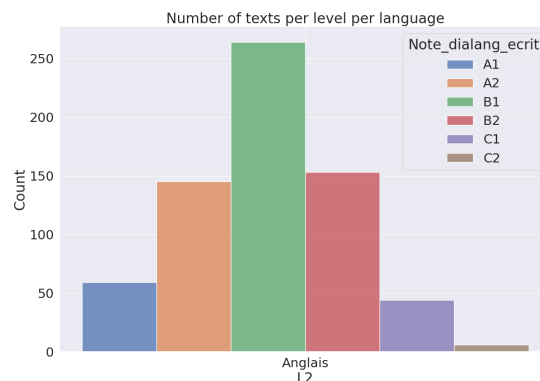


Figure 2: Distribution of the number of texts per CEFR level

U representation is formatted as a string, and for each text a single string is stored in the *conllu_text* column of the CSV. In this format each text is associated with a multi-layer representation of its linguistic annotation. For instance, each token is assigned the following information:

- FORM,
- LEMMA,
- UPOS,
- XPOS,
- FEATS (List of morphological features),
- HEAD (Head of the word dependency governor),
- DEPREL (Universal dependency relation to the HEAD),
- DEPS (A list of head-dependency relations pairs),
- MISC (Any other annotation such as givenness)⁵.

Thanks to the encoded dependency information, the files can subsequently be visualized with the CoNLL-U Viewer⁶ or queried with tools such as Grew-match (Amblard et al., 2022).

In addition, we added the metadata to the files. The metadata are accounted for with categorical and numerical variables named in French. They are:

- *Nb_annees_L2*: Number of years studying L2 English
- *L1*: Native language
- *Domaine_de_specialite*: Academic domain of the learner
- *Sejours_duree_semaines*: Total number of weeks spent in English speaking countries

⁵See <https://universaldependencies.org/format.html> for detailed information

⁶Available at <https://universaldependencies.org/conllu-viewer.html>

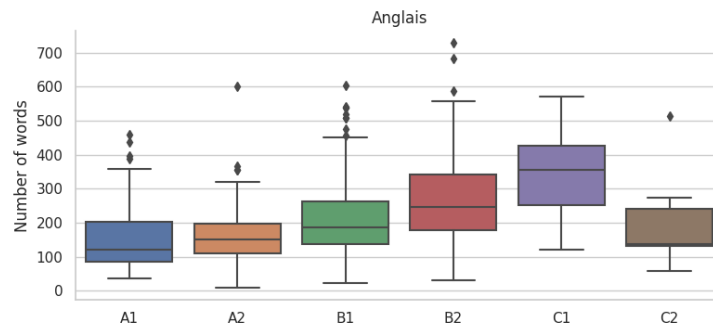


Figure 3: Distributions of texts according to their number of words and the CEFR levels of the learners

- *Sejours_frequence*: Number of trips
- *Lang_exposition*: Out-of-class exposure to L2 English (movies, radio ...)
- *Note_dialang_ecrit*: CEFR class with the DI-ALANG test
- *Lecture_regularity*: Reading frequency (daily, weekly, montly)
- *autre_langue*: Other L2 being learnt
- *tache_ecrit*: Identifier of writing task (only one)
- *Texte_etudiant*: Texts written by students
- *Date_ajout*: Date of writing
- *pseudo*: Pseudonymised ID of learner

5 Exploitation of the data set

This data set may be exploited in a wide array of tasks. ESP corpora play an important role in the field of academic language research as they help identify L2 developmental patterns linked to a specialised domain. They can thus support course material design with adapted content depending on academic profiles. Such data are useful for the design of Intelligent Computer-assisted Language Learning (ICALL) systems. These systems rely on supervised learning approaches that use learner corpora for error detection (Tetreault et al., 2018) or CEFR classification (Yannakoudakis et al., 2018; Gaillat et al., 2021) or language feature visualization (Gaillat et al., 2023).

Researchers involved in the ESP field will find the corpus useful for linguistic exploration and its potential for multidimensional analysis combining learning behaviour information with fine-grained linguistic annotation. In this respect, the CELVA.Sp data set can be exploited with a the Grew-match tool which provides for linguistic queries. Note that, thanks to the data and metadata formats, it is possible to sub-sample the data in order to obtain balanced datasets.

The data set could also be used in supervised learning tasks as it offers well-structured data. Traditional methods of machine learning such as logistic regression, support vector machines, random forests or gradient tree boosting require a large amount of tabular data. The CELVA.Sp data set provides tabular metadata, with little work required to create either tabular bag-of-words (Harris, 1954) features from the raw text or more complex dependency or morphological features from the linguistic annotations. More recent deep learning methods, such as convolutional neural networks (Kim, 2014), recurrent neural networks (LeCun et al., 2015) and transformer-based neural networks (including BERT (Devlin et al., 2019) and chatGPT⁷), require an unprecedented amount of data to train. However, the power of these models lies in the fact that they can be pre-trained on vast amounts of unannotated data from various sources, and then fine-tuned on a precise natural language task using task-relevant data. (Zhang et al., 2021) trained a BERT model on a task of textual entailment using the RTE dataset (Dagan et al., 2006) which consists of only 2,500 training data samples. The model achieved a 69.5 F1 score without any optimization. Our data set fits within this paradigm, with enough annotated learner data to fine-tune state-of-the-art deep learning models and leverage the predicting power of those models for tasks such as CEFR level prediction, or error modelling.

We intend to exploit this corpus as part of a Computer-Assisted Language Learning (CALL) system dedicated to the automatic analysis of learner language at university level. The corpus will be used to model learner proficiency across different academic domains. The system will display linguistic feature visualizations within the

⁷<https://openai.com/blog/chatgpt>

MOODLE system.

Further data enrichment is also planned. The corpus texts will be annotated by six language-certification experts following CEFR guidelines and inter-rater agreement will be evaluated. The final corpus will include texts of other L2s than English, including German, Swedish and Spanish. Keylog information recorded at time of writing will also be included. More writing tasks will be added for learners of all levels to ensure genre variety. The corpus will be available online.

6 Credits

We wish to thank all the language teachers who helped in collecting the data. This project is funded by the French National Research Agency ANR-22-CE38-0015-01



References

- J. Charles Alderson and Ari Huhta. 2005. [The development of a suite of computer-based diagnostic tests based on the Common European Framework](#). *Language Testing*, 22(3):301–320.
- Maxime Amblard, Bruno Guillaume, Siyana Pavlova, and Guy Perrier. 2022. [Graph querying for semantic annotations](#). In *Proceedings of the 18th Joint ACL - ISO Workshop on Interoperable Semantic Annotation within LREC2022*, pages 95–101. European Language Resources Association.
- Mihir Bellare, Ran Canetti, and Hugo Krawczyk. 1996. [Keying Hash Functions for Message Authentication](#). In *Advances in Cryptology*, pages 1–15. Springer.
- Marcus Callies. 2015. [Learner corpus methodology](#). In *The Cambridge Handbook of Learner Corpus Research*, Cambridge Handbooks in Language and Linguistics, pages 35–56. Cambridge University Press.
- Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. [The PASCAL Recognising Textual Entailment Challenge](#). volume 3944, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg. Book Title: Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment Series Title: Lecture Notes in Computer Science.
- Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a Large Annotated Corpus of Learner English: The NUS Corpus of Learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186. Association for Computational Linguistics.
- Martin Dougiamas and Peter Taylor. 2003. Moodle: Using Learning Communities to Create an Open Source Course Management System. In *Proceedings of the EDMEDIA 2003 Conference*, pages 171–178. Association for the Advancement of Computing in Education.
- Thomas Gaillat, Antoine Lafontaine, and Anas Knefati. 2023. [Visualizing Linguistic Complexity and Proficiency in Learner English Writings](#). *CALICO Journal*, 40(2):178–197.
- Thomas Gaillat, Andrew Simpkin, Nicolas Ballier, Bernardo Stearns, Annanda Sousa, Manon Bouyé, and Manel Zarrouk. 2021. [Predicting CEFR levels in learners of English: The use of microsystem criterial features in a machine learning approach](#). *ReCALL*, 34(2). Publisher: Cambridge University Press.
- Jeroen Geertzen, Theodora Alexopoulou, and Anna Korhonen. 2013. Automatic Linguistic Annotation of Large Scale L2 Databases: The EF-Cambridge Open Language Database (EFCamDat). In *Selected Proceedings of the 2012 Second Language Research Forum: Building Bridges between Disciplines*. Cascadia Press.
- Gaëtanelle Gilquin. 2015. From design to collection of learner corpora. In *The Cambridge Handbook of Learner Corpus Research*, Cambridge Handbooks in Language and Linguistics, pages 9–34. Cambridge University Press.
- Sylviane Granger, Maïté Dupont, Fanny Meunier, Hubert Naets, and Magali Paquot. 2020. *The International Corpus of Learner English. Version 3*. Presses universitaires de Louvain.
- Zellig S. Harris. 1954. [Distributional structure](#). *WORD*, 10(2-3):146–162.
- Alan Juffs, Na-Rae Han, and Ben Naismith. 2020. [The University of Pittsburgh English Language Institute Corpus \(PELIC\)](#).
- Yoon Kim. 2014. [Convolutional neural networks for sentence classification](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751. Association for Computational Linguistics.

- Kristopher Kyle, Masaki Eguchi, Aaron Miller, and Theodore Sither. 2022. [A Dependency Treebank of Spoken Second Language English](#). In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 39–45, Seattle, Washington. Association for Computational Linguistics.
- Yann LeCun, Y. Bengio, and Geoffrey Hinton. 2015. [Deep learning](#). *Nature*, 521:436–44.
- Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. [Universal Dependencies](#). *Computational Linguistics*, (2):255–308.
- Magali Paquot, Tove Larsson, Hilde Hasselgård, Signe O. Ebeling, Damien De Meyere, Larry Valentin, Natalia J. Laso, Isabel Verdaguer, and Sanne van Vuuren. 2022. [The Varieties of English for Specific Purposes dAtabase \(VESPA\): Towards a multi-L1 and multi-register learner corpus of disciplinary writing](#). *Research in Corpus Linguistics*, 10(2):1–15.
- Milan Straka, Jan Hajič, and Jana Straková. 2016. [UD-Pipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 4290–4297. European Language Resources Association.
- Anaïs Tack, Thomas François, Sophie Roekhaut, and Cédric Fairon. 2017. [Human and Automated CEFR-based Grading of Short Answers](#). In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 169–179. Association for Computational Linguistics.
- Joel Tetreault, Jill Burstein, Ekaterina Kochmar, Claudia Leacock, and Helen Yannakoudakis, editors. 2018. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, New Orleans, Louisiana.
- H. Yannakoudakis, Øe Andersen, A. Geranpayeh, T. Briscoe, and D. Nicholls. 2018. [Developing an automated writing placement system for ESL learners](#). Accepted: 2019-02-16T00:31:04Z Publisher: Informa UK Limited.
- Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. [A new dataset and method for automatically grading ESOL texts](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, pages 180–189. Association for Computational Linguistics.
- Amir Zeldes. 2017. [The gum corpus: Creating multilayer resources in the classroom](#). *Language Resources and Evaluation*, 51(3):581–612.
- Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q. Weinberger, and Yoav Artzi. 2021. [Revisiting Few-sample BERT Fine-tuning](#).

Grumpiness ambivalently relates to negative and positive emotions in ironic Austrian German text data

Andreas Baumann^{1,♣} and Nicole Bausch^{1,◇} and Juliane Benson^{1,◇} and Sarah Bloos^{1,◇}
and Nikoletta Jablonczay^{2,♣} and Thomas Kirchmair^{2,♣} and Emilie Sitter^{1,◇}

¹University of Vienna, Faculty of Philological and Cultural Studies

²University of Vienna, Faculty of Social Sciences

♣{andreas.baumann,nikoletta.jablonczay,
thomas.kirchmair}@univie.ac.at

◇{a01615138,a12240767,a12110417,a01648097}@unet.univie.ac.at

Abstract

We present a quantitative analysis of grumpiness as expressed in Austrian German text data. Based on a sample of annotated texts, we examine to what extent grumpiness relates to emotional properties and stylistic features. We show that grumpiness is mostly related to emotional configurations characteristic of anger but that grumpiness can alternatively signal positive emotions in ironic contexts.

1 Introduction

Grumpiness is one of the notorious characteristics of Austrian culture. With far-reaching consequences: Vienna dropped to the final position¹ in the category ‘friendliness’ in a recent expat city ranking.² The issue with grumpiness is, however, more intricate than one would think. In linguistics and cultural studies, grumpiness was shown to be vaguely related to verbal aggression (Havryliv, 2017) and even thought to be associated with positive characteristics like sense of humor. Grumpiness is seen as a kind of charm, adding to the city’s unique character and identity (Creath, 1995; Chen and Wu, 2019).

Despite its socio-cultural relevance, research on the topic lacks a systematic and quantitative assessment of which emotions Austrian grumpiness actually relates to. In this contribution, we conduct a statistical analysis of emotional and stylistic associates of grumpiness. Our analysis is based on a sample of texts written in Austrian German that were annotated and enriched with respect to various emotional and stylistic properties. We demonstrate that grumpiness results from a complex interaction of emotional features and irony, and that grumpiness does not exclusively signal negative emotions.

¹<https://www.derstandard.at/story/2000141285183/>

²<https://www.internations.org/expat-insider/>

2 Background

According to research in cognitive psychology, grumpiness is an emotional state that is often associated with dissatisfaction, annoyance, bad temper, and irritation (Barker et al., 2020; Brosschot et al., 2010; Dietvorst et al., 2021). As such, grumpiness can be a temporary state of mind, caused by factors such as lack of sleep, stress, or physical discomfort (Deonna and Teroni, 2009), or it can be a more persistent aspect of someone’s personality.

Dimensional models of emotion allow for a characterization of emotional states along several axes, most often valence (ranging from negative to positive), arousal (ranging from calm to aroused), and dominance (ranging from submissive to dominant) (Russell, 1980; Calvo and Mac Kim, 2013), often referred to as VAD model.

Considering grumpiness from the perspective of the VAD model, the emotional state is considered likelier to be negative, because it is associated with unpleasant experiences. Grumpy people tend to focus on the negative aspects of their experiences and may have difficulties finding pleasure or enjoyment in everyday activities (Watson and Clark, 1984). In terms of arousal, the judgement is less clear. Grumpy people may feel tired or sluggish and less motivated or interested in their surroundings. In an experiment on facial expressions, grumpiness was shown to be associated with relatively low arousal (Barker et al., 2020). However, they may also experience moments of increased arousal, e.g., when they become agitated or frustrated by a particular situation (Dietvorst et al., 2021).

As far as dominance is concerned, grumpiness could be potentially associated with a sense of powerlessness or frustration, and hence submissive emotions (Leach and Weick, 2018). On the other hand, grumpiness is related to anger, which is characterized by low valence, high arousal, and high dominance (Calvo and Mac Kim, 2013). Thus,

it would be interesting to see where exactly grumpiness is located in the VAD space.

How emotional states like grumpiness are intertwined with texts like poetry, literature or, more recently, the vast amount of text data produced on social media has become a field of interdisciplinary interest. For this purpose, also data science and the digital humanities are constantly working on new modelling techniques mainly using techniques from NLP like keyword detection or lexica to predictive modelling, there has been a shift to more sophisticated, state-of-the-art neural networks.

What they all share is the search for the best combination of stylistic, structural and semantic features to determine the emotions or ‘tone’ of interest. The solution depends mainly on the data and goal. For the detection of ironic comments for example, besides using standard features like word count or PoS distributions (Alm et al., 2005), it has proven useful to include interjections, punctuation, capitalization, use of first-person pronouns, repetitions, negations or even labelled emoticons as features (Ortega-Bueno et al., 2018). It was also indicated by Reyes et al. (2012) that special linguistic features like morphosyntactic ambiguity — linked with lesser syntactic complexity — are useful for inferring irony as well. This is relevant because irony and grumpiness show a distinct connection: irony is often used to soften an angry remark or criticism, with the speaker appearing to be more in control (Dews et al., 1995).

Diving deeper into the matter, Van Hee (2017) shows that lexical features like character and punctuation flooding in tweets (e.g. in words like ‘Looovv’) outperformed word n-grams in irony detection next to structural and sentiment features like tags, valence or polarity scores. Nonetheless, the best results were yielded when combining all three feature-sets. The author concludes that certain features suit certain ‘types’ of irony.

The addition of stylistic features in general does statistically improve the overall performance of emotion detection models (Malheiro et al., 2016) but they do not seem to work equally well alone, and they don’t have an effect as high as semantic features. Hence, it makes sense to take stylistic features into account when investigating grumpiness manifested in text data.

3 Data

3.1 Annotation

We based our analysis on the Million Posts corpus (Schabus et al., 2017). It consists of postings taken from the user forum of the Austrian news website <http://derstandard.at>. Texts represent a sample of the Austrian variety of German. This user forum is a suitable resource for studying grumpiness as it accommodates a large population of users with diverse political views (mostly excluding strong right-wing attitudes) so that topics are typically discussed vividly and emotionally (note, though that the forum is moderated, hence hate-postings do not get published if they are detected). About 3500 of the texts in the corpus have been already labeled with respect to sentiment (pos/neu/neg; three categorical labels per posting). We computed average sentiment ratings for each posting and found that only 69 texts in the data set show a positive sentiment. To create a balanced sample, we sampled a roughly equal amount of neutral and negative texts and ended up with a stratified sample of 200 texts in total.

Subsequently, texts were annotated with respect to five characteristics: a rousal, dominance, abstractness, irony, and grumpiness. Annotators were asked to judge the texts with respect to these characteristics based on a five-point Likert scale. All texts were labeled by three annotators each. All annotators (some of which are authors of this paper) were students speaking German as their first language (they received course credit and no monetary compensation for their labeling efforts). Annotators were provided with the parent posting (if it existed) and the title of the news article postings related to as additional context.

We computed Cronbach’s α to assess inter-annotator agreement. Apart from abstractness with $\alpha = 0.31$, inter-annotator agreement was sufficiently high³ (arousal: $\alpha = 0.67$; dominance: $\alpha = 0.69$; irony: $\alpha = 0.78$; grumpiness: $\alpha = 0.75$) and comparable with the quality of the ratings in the Million Posts Corpus (Schabus et al., 2017). Notably, the relatively high inter-annotator agreement for grumpiness was reassuring for our study (see Figure 1).

³Values of Cronbach’s α greater than 0.8 are considered to be good, values between 0.7 and 0.8 are considered to be acceptable, and values below 0.5 are interpreted as unacceptable (Li et al., 2016; Streiner, 2003).

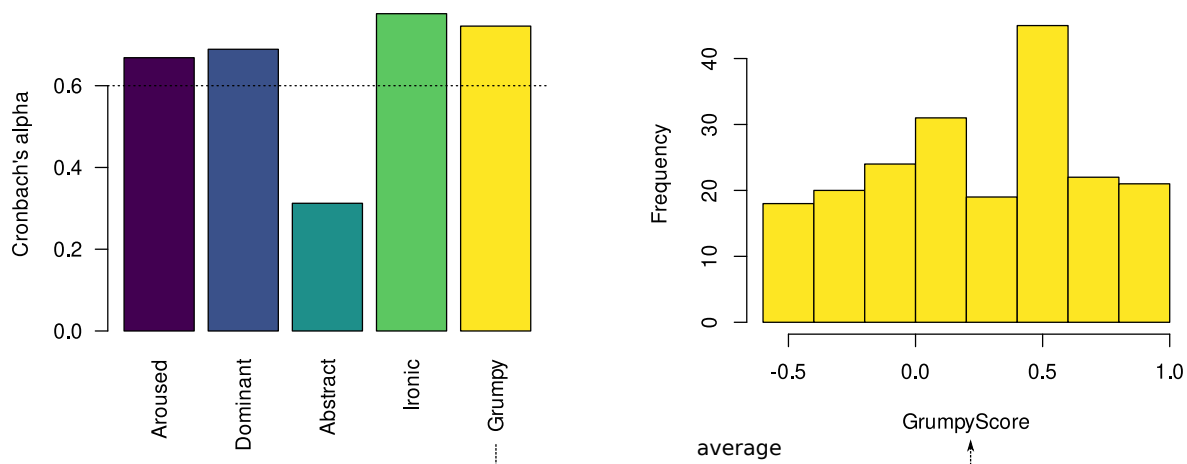


Figure 1: Left: inter-annotator agreement (Cronbach’s α) for all annotated features (acceptability threshold shown by dashed line). For each text, GrumpyScore was computed as average of all grumpiness ratings. Right: distribution of GrumpyScore in the sample of 200 texts.

3.2 Emotional features

In a next step, the average of all annotator ratings was computed for each characteristic and each text to obtain overall scores (ArousedScore, DominantScore, AbstractScore, IronicScore, GrumpyScore). All scores including sentiment taken from the Million Posts Corpus (SentiScore) were subsequently scaled to the interval $[-1, 1]$ in such a way that 0 corresponds to a neutral score. The histogram in Figure 1 shows that GrumpyScore is fairly equally distributed across the interval $[-0.5, 1.0]$. That is, the texts in the sample were classified as rather grumpy on average (despite the sample being balanced with respect for sentiment).

3.3 Stylistic features

In order to capture potential stylistic correlates of grumpiness, we derived a range of linguistic variables. First, we used the Flair PoS tagger to compute the fraction of Nouns, Verbs and Adjectives for each text. Second, we counted the number of Colons, Periods, ExclamationMarks, and QuestionMarks, as well as the number of happy (:), sad :(or :/), and blinking (;) emoticons (HappyEmoticon, SadEmoticon, BlinkEmoticon, respectively). Finally, we retrieved TextLength measured as the number of characters, as well as TypeTokenRatio to include a proxy for lexical diversity.

4 Analysis

4.1 Emotional and stylistic features

What is the relative impact of emotional and stylistic features on grumpiness? To shed light on this

question, we first computed a linear (Gaussian) regression model in which GrumpyScore depends on all other 18 features described in the previous section. We used the per text computed reciprocal of the standard deviation of the grumpiness ratings as weights in the model, so that texts with a more accurate GrumpyScore are weighted higher in the model. The resulting model shows a reasonably high goodness of fit at $R^2 = 0.68$ (and a fairly symmetric residual distribution), indicating that grumpiness is characterized well by the emotional and stylistic features at hand.

Since much information in the data about the outcome is shared among the 18 predictors, we employed AIC-driven top-down model nesting to optimize the previously computed linear model. The resulting model (which scores the lowest AIC and $R^2 = 0.67$) features eight predictors, five of which show statistically non-trivial effects: grumpiness is associated with high arousal, high dominance, negative sentiment, and, to a lesser extent, irony. Thus, the linear model suggests grumpiness to be associated with anger (which is itself characterized by high arousal, high dominance and low valence). Interestingly, the number of verbs shows a particularly strong positive impact on the outcome variable. See Table 1 for a breakdown.

To get insights into the ranking of the predictors, we computed relative variable importance based on the AIC scores of all sub-models of the maximal model featuring 18 predictors (Burnham and Anderson, 2004). More specifically, we derived Akaike weights for all sub-models and, for each predictor, computed relative variable importance as

Predictor	Coef.	SE	t value
(Intercept)	0.32	0.33	0.95
ArousedScore	0.36	0.06	5.56
DominantScore	0.45	0.08	5.55
IronicScore	0.11	0.04	2.64
TypeTokenRatio	-0.62	0.35	-1.77
SentiScore	-0.28	0.04	-7.29
SadEmoticon	0.14	0.10	1.40
Adjectives	0.32	0.19	1.74
Verbs	0.76	0.21	3.56

Table 1: Effects on GrumpyScore in the optimal linear model. Bold indicates statistically non-trivial effects at a 5% significance level.

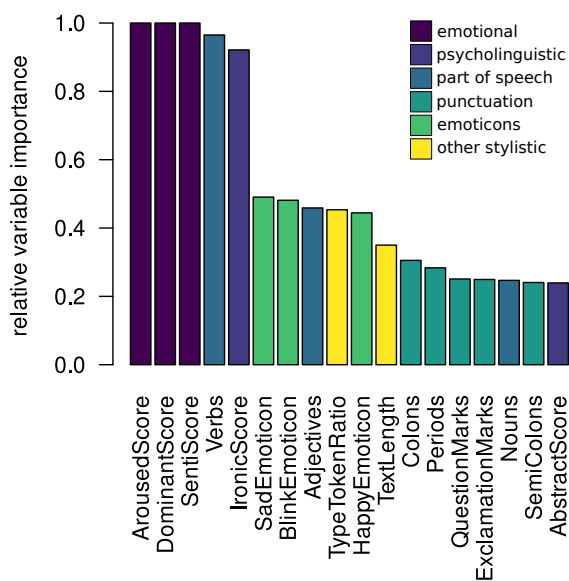


Figure 2: Relative variable importance based on multiple linear regression models.

the sum of Akaike weights of all models in which that predictor is present. The ranking is shown in Figure 2. There seem to be three different groups: emotional features and the number of verbs are most important for inferring grumpiness. The predictors in the second group are only roughly half as important. Interestingly, the group shows all emoticon counts. The remaining predictors (mostly punctuation, but also the number of nouns) display the lowest relevance for inferring grumpiness.

4.2 Grumpiness in the VAD space

The significant effects of all emotional predictors in the model make clear that grumpiness is unsurprisingly associated with specific emotional aspects. To explore the location of grumpiness in the emotional space spanned by valence, arousal,

and dominance, we used generalized additive models (GAM) (Wood, 2006). Here, GrumpyScore is predicted by three interacting variables SentiScore, ArousedScore, DominantScore). The interaction was implemented as a smooth tensor-product term (number of knots $k = 5$). Due to the distribution of GrumpyScore (Figure 1, right), we used a Gaussian link function. Again, reciprocal standard deviations of GrumpyScore were used as weights like in the linear model.

The model is visualized in the upper panel of Figure 3. It displays the valence-arousal space for four different dominance bins. Light colors (yellow) indicate a stronger association with grumpiness than dark colors (purple). It can be seen that grumpiness increases with dominance (in line with the linear model), and that grumpiness is associated with high arousal and low valence, i.e., it is co-located with emotional categories like anger. This particularly holds true for submissive scenarios but is weakened as dominance increases. High dominance apparently allows for a slightly more positive association with grumpiness.

4.3 Interaction with irony

In the linear model, the significant effect of irony is particularly interesting. We computed a second GAM, but this time GrumpyScore was predicted by SentiScore, ArousedScore, and IronicScore in order to assess the effect of irony of the location of grumpiness in the valence-arousal space. The result is shown in the lower panel of Figure 3. In line with the linear model, the effect of irony is weaker than that of dominance (overall, the plotted surface does not become substantially lighter).

Interestingly, if irony is low (first plot) grumpiness is relatively strictly confined to the negative and aroused region of the emotional space. However, if irony scores high, there are relatively high associations of grumpiness with negative *and* positive regions, while (valence-wise) neutral regions show diminished grumpiness. This indicates that grumpiness is highly ambivalent in ironic settings: grumpiness could either correspond to angry contexts but also to joyful ones (but not to indifferent contexts).

5 Discussion and conclusion

In this paper, we presented a quantitative analysis of linguistically represented grumpiness based on a sample of texts that were annotated for various

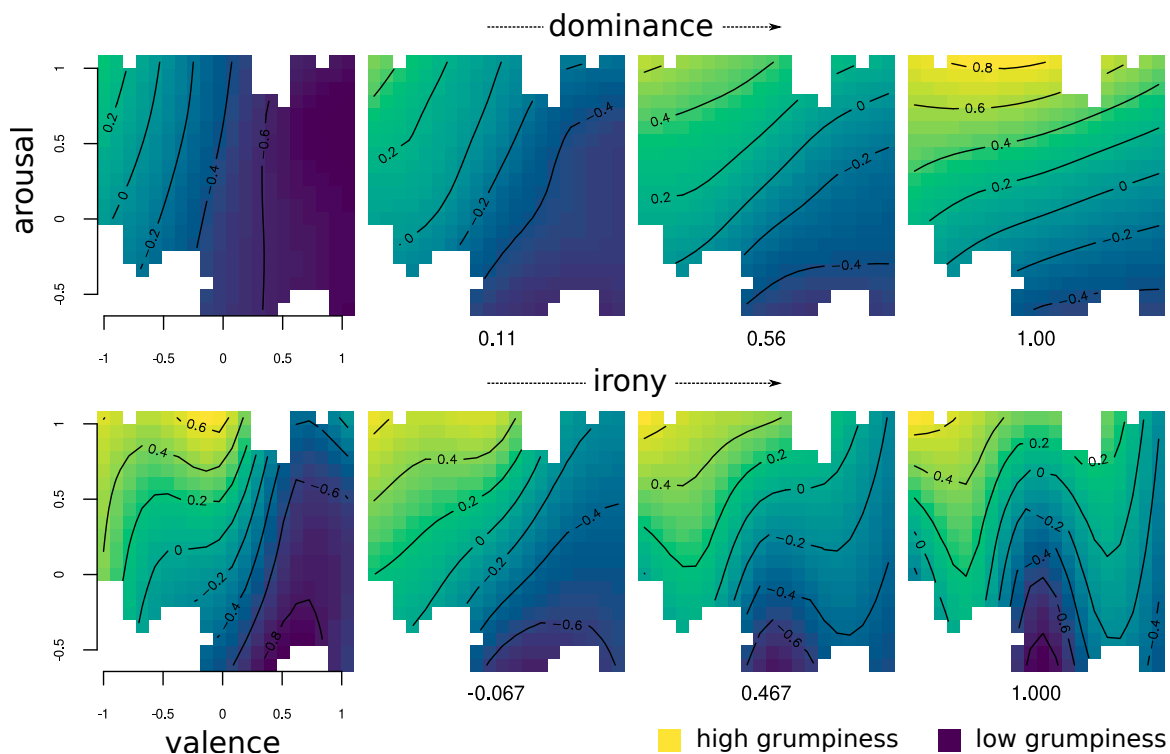


Figure 3: Valence-arousal spaces modulated by dominance (upper panel) and irony (lower panel), based on GAMs.

emotional properties and enriched with stylistic information. The main result of our analysis is that, overall, grumpiness is associated with anger. However, this clear association does not hold in ironic contexts. Here, grumpiness can relate to either negative or positive emotions. Interestingly, this means that knowledge of whether or not a text is ironic (i.e., a certain sensitivity with respect to irony) does not suffice to categorize grumpy utterances. Individuals require additional information to decode the emotional state underlying a grumpy utterance.

This result is in line with the observation that Austrian grumpiness can signal humor as well. (Creath, 1995; Chen and Wu, 2019; Havryliv, 2017). Whether or not this intricate relationship between emotion, grumpiness, and irony is responsible for the fact that Viennese people tend to be perceived as unfriendly as suggested by surveys among expats (see footnote 1 and 2), still needs to be looked at more closely.

Another result of our modeling analysis is that grumpiness seems to be associated with an extensive usage of verbs (as opposed to nouns and adjectives). Given that verbs are typically less concrete than other lexical categories, this result seems surprising at first sight. Nominal style is typical of less aroused genres like legal or scientific texts, while verbal style is generally represented more strongly

in everyday speech (Radovanovic, 2001). Either way, the results point at the relevance of stylistic cues when inferring emotional states from text.

It is evident that our study is subject to limitations. For one, the number of texts as well as the number of annotations per text is not large. However, inter-annotator agreement was sufficiently high (in particular as far as grumpiness is concerned) and the fact that our models show statistically robust effects, high goodness of fit, and relatively small standard errors despite the small sample size is reassuring. In addition to a larger number of texts (and annotators), potential follow-up studies would need to take different genres into account. Clearly, considering spoken corpora would be most relevant in this regard (however, forum postings represent an already relatively informal genre).

Finally, it would be interesting to see to what extent grumpiness ratings from raters with different social, linguistic, or geographic backgrounds deviate from each other. This would help to shed light on how linguistically expressed grumpiness is perceived cross-culturally.

Supplementary materials

The analysis can be reproduced in the following project on Posit Cloud: <https://posit.cloud/content/5527995>. The processed data set of all aggregated

scores our analysis is based on is available at <https://phaidra.univie.ac.at/o:1634249>. A supplementary analysis involving several emotional lexica can be found here: <https://phaidra.univie.ac.at/o:1634258>.

Acknowledgements

We would like to thank Jennie Atwell, Ines Konnerth, Maximilian Moser, Marja Nikolcic, Sarah Sulollari, Lale Tüver, and Sebastian Wegerer for additional help with data annotation.

References

- Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: machine learning for text-based emotion prediction. In *Proc. of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 579–586.
- Megan S Barker, Emma M Bidstrup, Gail A Robinson, and Nicole L Nelson. 2020. “grumpy” or “furious”? arousal of emotion labels influences judgments of facial expressions. *Plos one*, 15(7):e0235390.
- Jos F Brosschot, Bart Verkuil, and Julian F Thayer. 2010. Conscious and unconscious perseverative cognition: is a large part of prolonged physiological activity due to unconscious stress? *Journal of Psychosomatic Research*, 69(4):407–416.
- Kenneth P Burnham and David R Anderson. 2004. Multimodel inference: understanding aic and bic in model selection. *Sociological methods & research*, 33(2):261–304.
- Rafael A Calvo and Sunghwan Mac Kim. 2013. Emotions in text: dimensional and categorical models. *Computational Intelligence*, 29(3):527–543.
- Yeong-Shyang Chen and Shou-Tsung Wu. 2019. Social networking practices of viennese coffeehouse culture and intangible heritage tourism. *Journal of Tourism and Cultural Change*, 17(2):186–207.
- Richard Creath. 1995. From königsberg to vienna: Coffa on the rise of modern semantics. *Dialogue: Canadian Philosophical Review/Revue canadienne de philosophie*, 34(1):113–118.
- Julien A Deonna and Fabrice Teroni. 2009. Taking affective explanations to heart. *Social Science Information*, 48(3):359–377.
- Shelly Dews, Joan Kaplan, and Ellen Winner. 1995. Why not say it directly? the social functions of irony. *Discourse Processes*, 19(3):347–367.
- Evelien Dietvorst, Marieke Hiemstra, Dominique Maciejewski, Eeske van Roekel, Tom Ter Bogt, Manon Hillegers, and Loes Keijsers. 2021. Grumpy or depressed? disentangling typically developing adolescent mood from prodromal depression using experience sampling methods. *Journal of Adolescence*, 88:25–35.
- Oksana Havryliv. 2017. Verbale Aggression: das Spektrum der Funktionen. *Linguistik Online*, 82(3).
- Stefan Leach and Mario Weick. 2018. From grumpy to cheerful (and back): How power impacts mood in and across different contexts. *Journal of Experimental Social Psychology*, 79:107–114.
- Junyi Jessy Li, Bridget O’Daniel, Yi Wu, Wenli Zhao, and Ani Nenkova. 2016. Improving the annotation of sentence specificity. In *Proc. of the 10th International Conference on Language Resources and Evaluation*, pages 3921–3927, Portorož, Slovenia.
- Ricardo Malheiro, Renato Panda, Paulo Gomes, and Rui Pedro Paiva. 2016. Emotionally-relevant features for classification and regression of music lyrics. *IEEE Transactions on Affective Computing*, 9(2):240–254.
- Reynier Ortega-Bueno, Carlos E Muniz-Cuza, José E Medina Pagola, and Paolo Rosso. 2018. Uo upv: Deep linguistic humor detection in spanish social media. In *Proc. of the 3rd Workshop on Evaluation of Human Language Technologies for Iberian languages*, pages 204–213.
- Milorad Radovanovic. 2001. On nominal and verbal style: cultures or languages in contact? *International Journal of the Society of Language*, 151(3):41–48.
- Antonio Reyes, Paolo Rosso, and Davide Buscaldi. 2012. From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering*, 74:1–12.
- James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161–1178.
- Dietmar Schabus, Marcin Skowron, and Martin Trapp. 2017. One million posts: A data set of german online discussions. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1241–1244.
- David L. Streiner. 2003. Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, 80:99–103.
- Cynthia Van Hee. 2017. *Can machines sense irony?: exploring automatic irony detection on social media*. Ph.D. thesis, Ghent University.
- David Watson and Lee A Clark. 1984. Negative affectivity: the disposition to experience aversive emotional states. *Psychological Bulletin*, 96(3):465.
- Simon N Wood. 2006. *Generalized additive models: an introduction with R*. CRC.

Orbis Annotator: An Open Source Toolkit for the Efficient Annotation and Refinement of Text Corpora

Norman Süsstrunk and Andreas Fraefel and Albert Weichselbraun
 Fachhochschule Graubünden
 Chur, Switzerland
 {first.lastname}@fhgr.ch

Adrian M.P. Braşoveanu
 Modul University and
 Modul Technology GmbH
 Vienna, Austria
 adrian.brasoveanu@
 modul.ac.at

Abstract

Annotated language data plays an important role in training, fine-tuning and evaluating natural language processing components. Nevertheless, manually annotating language data is still a cumbersome task.

This paper presents the Orbis Annotator framework, a user-friendly, easy to install, web-based software that supports users in efficiently annotating language data. Orbis Annotator supports standard and collaborative workflows, reuse of language resources through corpus versioning, and provides built-in tools for assessing corpus quality. In addition, it offers an API which enables the use of different clients (e.g., web-based, command line, etc.) and the use of third-party tools that accelerate the annotation process by pre-annotating corpora.

The paper concludes with an evaluation that compares its features to other open-source annotation frameworks and the description of two use cases that outline its use in more sophisticated settings.

1 Introduction

With the emergence of deep neural networks, unsupervised pre-training on massive datasets has gained in importance. Although pre-trained language models require a considerably lower number of training examples when compared to the early deep learning models, these models still benefit tremendously from further fine-tuning on labelled data. Gold standard corpora play a pivotal role in adapting models to concrete tasks, and in evaluating model performance. This is particularly true when considering the rise of machine learning approaches in research and industry.

Creating annotated gold standard corpora is still a labor-intensive task, although many toolkits such as Annotation Study¹, BRAT² (Brat Rapid Anno-

tation Tool; Stenetorp et al. (2012)), Prodigy³, Do-canno⁴, Gate Teamware⁵, and INCEpTION⁶ that support the annotation process exist.

But even with specialized tools, annotators lose valuable time with marking annotation spans and assigning them to the corresponding annotations. Drawing upon automatically generated silver standard annotations, has the potential to significantly improve efficiency. More sophisticated annotation tools support pre-annotating text, and in some cases even online learning, which ensures that human feedback (e.g., corrections of machine-generated annotation annotations) is leveraged for improving the automated pre-annotation process.

Unfortunately, many solutions are either difficult to install, lack vital functionality such as support for pre-annotated corpora, collaborative workflows and computation of corpus statistics (e.g., the inter-rater agreement), or are only available under commercial licenses.

Orbis Annotator addresses these shortcomings and builds upon prior work by providing a solution which

- is easy to install and use
- integrates tightly with machine learning approaches, that provide silver-standard annotations
- allows refining and improving existing corpora
- supports collaborative annotation processes
- increases annotator efficiency through (optional) pre-annotations, keyboard shortcuts and mouse actions (i.e., it supports both keyboard-centric and mouse-centric annotators)

³<https://prodi.gy>

⁴<https://github.com/doccano/doccano>

⁵<https://gate.ac.uk/teamware>

⁶<https://github.com/inception-project/inception>

¹<https://annotation-study.org>

²<https://brat.nlplab.org>

In addition, Orbis Annotator will be coupled with the next version of the Orbis Visual Benchmarking Platform (github.com/orbis-eval) which will bundle the creation of gold standards with a suite of explainable benchmarking tools that supports evaluating human and machine annotators on the created datasets.

The presented research, therefore, provides the following contributions:

1. the introduction of Orbis Annotator, a text annotation framework that is easy to use and considerably improves the efficiency of creating gold standards;
2. an overview and comparison of existing open-source annotation tools,
3. the presentation of two use cases (machine-based corpus pre-annotation of custom entity types, and corpus migration to a new knowledge graph) that demonstrate how Orbis Annotator has been successfully deployed in real-world settings.

The rest of this paper is organized as follows: Section 2 provides an overview of related work. Afterwards, Section 3 introduces *Orbis Annotator*. Section 4 discusses the strengths and weaknesses of Orbis Annotator based on two use cases, compares it to related frameworks, and outlines the gains in productivity achieved by drawing upon the system. The paper closes with the conclusions and an outlook presented in Section 5.

2 Related Work

Deep Learning requires large text collections for unsupervised training. Depending on the chosen learning tasks, unsupervised training might be complemented with fine-tuning on annotated data to help in improving systems' performance. This has led to an increase in the number of annotation tools developed in the past five years, as can be seen by examining the papers accepted at leading natural language processing and machine learning conferences such as ACL, EMNLP, CoNLL, COLING, LREC, etc. Therefore, the following discussion on related research had to be narrowed to a limited number of papers. The criteria used in this paper were: (i) historical significance (e.g., tools supported by larger number of users who are still popular within the academia and industry); (ii) availability (e.g., published in open-source repositories

or free to use); (iii) ease of use (i.e., tools can be installed and operated without specialized training and in-depth knowledge of their implementation); and (iv) support for current NLP trends (e.g., if the tools support machine-aided annotation generation mechanisms like active learning).

Readers interested in a comprehensive survey on annotation tools, may refer to a recent overview paper by [Neves and Seva \(2021\)](#) that surveyed 78 tools and provides a detailed comparison of 15 of them. Although their survey is mostly focused on the domain of bioinformatics, it also includes well-known general tools such as BRAT, ezTag and Prodigy. Nevertheless, none of the tools included was able to cover all the needs of the survey's authors.

Perhaps the oldest, and best known software in the space is GATE ([Cunningham, 2002](#)) which started as a single annotator tool in the late 1990s and morphed into a collaborative tool called GATE-Teamware ([Bontcheva et al., 2013](#)) a decade ago. GATE was created for multiple span annotations and turned out to be ideal for tasks like tokenization, named entity recognition (NER), sentiment analysis, dependency parsing (DP), part-of-speech tagging (POS), and coreference resolution (CR).

UIMA (Unstructured Information Management Architecture; [Ferrucci and Lally \(2004\)](#)) is a generalized annotation architecture that supports interoperability. Various annotation toolkits such as DKPro WSD ([Miller et al., 2013](#)) and TextAnnotator ([Abrami et al., 2020](#)) are built around UIMA's philosophy.

BRAT ([Stenetorp et al., 2012](#)) gained some traction a decade ago, but was eventually abandoned. BRAT can be used for similar tasks as GATE. WebAnno ([Yimam et al., 2013](#)) builds directly on top of the BRAT functionality. More recent tools such as Apletny ([Nghiem and Ananiadou, 2018](#)), ActiveAnno ([Wiechmann et al., 2021](#)) and Paladin ([Nghiem et al., 2021](#)) adapt WebAnno's functionality to new active learning use cases.

The Stanford CoreNLP ([Manning et al., 2014](#)) toolkit supports the creation of custom annotators, and provides a regular expression-based mechanism (RegexNER) for pre-annotating documents. CoreNLP was the first annotator widely used for Deep Learning tasks, and its description in [Manning et al. \(2014\)](#) provides good definitions for the supported annotation tasks.

In addition to domain-specific tools (e.g., for the

medical and finance domain), many frameworks that have been tailored towards specific text annotation tasks exist. Yedda (Yang et al., 2018), for instance, was built for annotating specialized entity types (e.g., events). TAG (Forbes et al., 2018) is optimized towards showcasing complex relations between sentences and documents. ALIGNMEET (Polák et al., 2022) and EZCAT (Guibon et al., 2022) focus on annotating meetings and conversations and support a wide array of languages, symbols, and emojis. Ellogon (Ntogramatzis et al., 2022) annotates moral values and arguments. Textinator (Kalpakchi and Boye, 2022) was created for internationalization and language evolution use cases. Semantic storytelling (Raring et al., 2022) is another use case that led to the development of a specialized tool.

AWOCATo (Daudert, 2020) is a recent tool that supports various annotation formats. Although not used for creating annotations, Spicy Salmon (Fäth and Chiarcos, 2022) deserves mentioning, since it provides an interface for converting between 50 different annotation formats. An early attempt towards interoperable annotations was NIF (Hellmann et al., 2012), an RDF-based language for producing customized annotation, although it is primarily used within the European data spaces.

Inception⁷ (Klie et al., 2018) builds upon UIMA’s interoperability concepts and WebAnno’s annotation functionalities. Inception offers several new concepts, like recommender algorithms that help improve annotation efficiency, and advanced customization capabilities.

Some open-source annotation tools that stand out include Argilla⁸ and Docanno⁹. Since they are produced collaboratively under open licenses, these tools have a wider reach than the academic ones. Argilla supports active learning through its HuggingFace integration, provides a simple API, and has recently gained a significant following. Docanno offers collaborative editing, REST APIs and emoji support. Another famous but proprietary tool, Prodigy¹⁰, was introduced by the Explosion team that created Spacy. Also powered by active learning, Prodigy offers classic text annotation features, supports A/B testing, and zero-shot prompts.

While not necessarily direct competitors to Orbis or other annotation solutions, instrumentation and

explainability tools such as MLFlow¹¹, Weights and Biases¹² and neptune.ai, also deserve attention since their APIs allow for quick and easy instrumentation of AI components that train upon annotated corpora. An overview of these tools can be found in Braşoveanu and Andonie (2022).

3 Method

Several years ago, we started developing a benchmarking ecosystem after an early study about named entity linking evaluations (Braşoveanu et al., 2018) showcased a significant number of errors in existing gold standards and knowledge graphs. The initial version of Orbis (Odoni et al., 2018) was the first step in this direction. The first version only focused on named entity linking (NEL) evaluations, but later versions included support for content extraction evaluations (Weichselbraun et al., 2020), NER and basic slot filling evaluations. In time, it became clear that focusing only on the visual evaluation issue was not enough, and that there was a need for integrated platforms that support both the annotation and evaluation workflows. The Orbis Annotator, the tool presented in this paper, is focused on annotation workflows. Since this tool represents both a reimplement and a significant expansion upon the previous generation, it was named Orbis 2. The design of the current version is modular (e.g., backend, frontend, or corpus exporter components are already included).

Major barriers towards deploying specialized software for annotating complex corpora are the software’s availability (i.e., whether it is free to use or requires licenses), skill and effort required for setting up the software, and time necessary for using it efficiently. Many state-of-the-art solutions are either limited in terms of functionality, freedom of use, or are really difficult to setup and operate. Orbis Annotator aims at addressing these shortcomings by bundling all necessary components into a docker container, and providing an efficient, intuitive Web-based workflow that covers its basic functionality and does not require any prior training. In addition, Orbis Annotator supports more complex workflows through its data model (Section 3.1) and backend API (Section 3.2). The software has been released under the Apache 2.0 license and is available on Github¹³ for download.

⁷<https://inception-project.github.io/publications/>

⁸<https://github.com/argilla-io/argilla>

⁹<https://github.com/doccano/doccano>

¹⁰<https://prodi.gy/>

¹¹<https://mlflow.org/>

¹²<https://wandb.ai/site>

¹³<https://github.com/orbis-eval/orbis2-frontend>

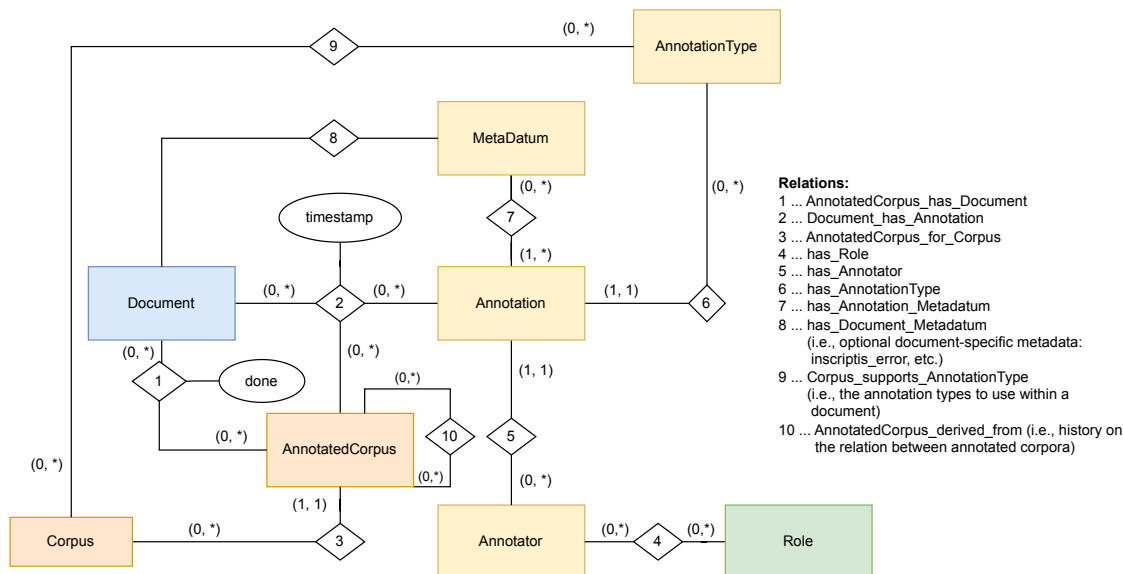


Figure 1: Entity Relationship model of the Orbis database (the attributes of the entities have been excluded, due to limited space)

3.1 Orbis data model

Orbis stores corpora, documents, annotations, and metadata (e.g., annotators, corpus versions, etc.) in a relational PostgreSQL¹⁴ database. Its data model supports corpus and annotation versioning, atomic real-time updates and the export and import to popular formats such as JSON, Excel and NIF.

Use case studies and analysis of existing annotation and benchmarking suites yielded the following requirements for the Orbis data model:

1. *Interoperable*: Although Orbis does not aim at introducing another annotation format, its data model is required to support importing and exporting existing formats without information loss.
2. *Reusable*: Orbis promotes reuse of existing corpora by refining and improving them. This requirement comprises use cases such as using human annotators to promote automatically annotated silver standards to gold standards, updating corpora to newer versions of the knowledge base (e.g. DBpedia 2015-10 to a more recent version), and improving upon existing gold standard annotations.
3. *Multi-user capable*: Orbis supports groups of annotators that collaboratively add, correct and improve annotations. The data model records individual contributions, and supports

multiple task designs (e.g., annotators working independently, versus collaborative settings).

4. *Workflow agnostic*: The data model shall enable multiple workflows with different levels of complexity (e.g., manual annotation by a single annotator, by multiple annotators; machine learning for pre-annotating corpora with silver standard annotations; hybrid workflows that combine machine and human annotators).
5. *Process metrics oriented*: The data model supports computing process metrics on individual annotators (e.g., throughput in terms of documents and number of annotations), and shared metrics (e.g., different kinds of inter-rater agreement).

Figure 1 provides the Entity Relationship model of the Orbis database.

Central element of the model is an *AnnotatedCorpus* which represents a certain version of a *Corpus* with all its documents, annotations and metadata. Importing a corpus creates a *Corpus* entity and the corresponding *AnnotatedCorpus*, which might either be empty (if an unannotated corpus has been imported) or contain initial annotations (e.g., from a gold standard, automated annotators, etc.) alongside the documents. Each *AnnotatedCorpus* consists of *Documents* and the corresponding *Annotations*. Orbis also records the *AnnotationType*, the *Annotator* and optional *MetaDatum* for all

¹⁴<https://www.postgresql.org>

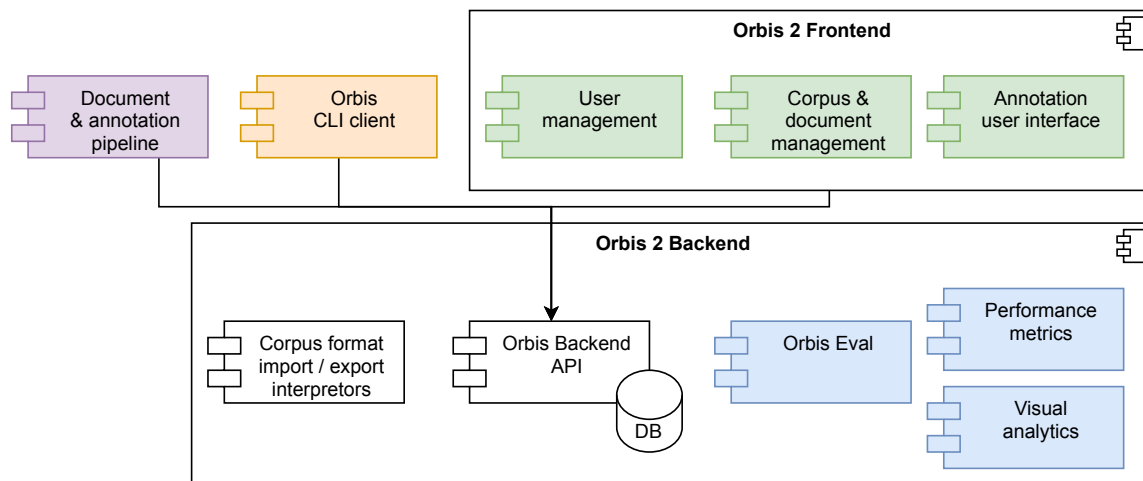


Figure 2: Overview of Orbis 2 architecture which outlines important frontend and backend components. The Orbis API also allows interaction with third-party pipelines and the Orbis Command Line (CLI) client.

annotations. In addition, Orbis implements user management and access control via the relations between *Annotators* and their respective *Roles*.

The relation between *Corpus* and *Annotation-Type* allows specifying the set of annotation types to use within a corpus, and the *derivedFrom* relation enables tracking the relationship between different corpus versions. The chosen data model also allows tracking changes between *AnnotatedCorpus* entities (e.g., gold standard annotations, annotations provided by different persons, machine-generated annotations, etc.) which represent different corpus versions. These versions may be derived from

- gold standard labels which have been provided with the corpus;
- automated approaches such as named entity linking, named entity recognition and sentiment analysis which provide silver standard labels for evaluations or to accelerate manual annotation processes;
- manual annotations provided by annotators. Depending on the use case requirements, annotators might work on the same or different *AnnotatedCorpora* (i.e., produce common or separate corpus versions).

Orbis also supports computing standard metrics such as precision, recall, F1-measure and inter-rater agreement between these versions (Section 3.5).

3.2 Orbis backend

Figure 2 outlines how Orbis exposes its data model through a publicly available backend API. The Orbis backend API currently supports (i) the Orbis

Annotator frontend used for annotating and refining corpora, (ii) the Orbis command line interface (CLI) client which focuses on performing evaluations and computing metrics, and (iii) integrating custom document and annotation pipelines which can add new documents to existing corpora, and manipulate corpus annotations (e.g., to provide silver standard annotations). As outlined in Section 4.2, the machine aided pre-annotations may be used to further enhance the efficiency of human annotators.

The backend also contains interpreters for corpus formats such as NIF, JSON and Excel which allow native consumption and production of these formats through the Orbis API. These interpreters are essential for compatibility with publicly available corpora, other annotation frontends, and existing software libraries such as SpaCy.

Future versions of Orbis Annotator will tightly integrate with the Orbis Explainable Benchmarking framework which will enable performing evaluations, and drill-down analyses on top of the created corpora.

3.3 Orbis Annotator frontend

The following design goals led to the development of the Orbis Annotator frontend: (i) the user interface should be intuitive and responsive, (ii) changes (i.e., added, modified and deleted annotations) should be automatically serialized to prevent data-loss, (iii) the interface should contain usability optimizations that are tailored towards annotator efficiency and support both mouse- and keyboard-centred workflows.



Figure 3: Visualization of the rendered tree structure in Orbis Annotator. The borders were added to illustrate the underlying tree-structure, and are invisible in the Orbis Annotator interface. The border color is used to indicate whether elements are annotated (yellow) or unannotated (grey).

3.3.1 Responsiveness and real-time updates

Converting the list of annotations into a tree using the nested set algorithms yields a tree structure from a list of annotations with start and end indices. The obtained tree structure offers several advantages:

1. It provides a more efficient way to query, retrieve and modify annotations, especially when dealing with large numbers of annotations;
2. the tree structure simplifies the rendering process by providing a clear hierarchy of the annotations;
3. it also allows for easier management of annotations, including sorting, filtering and adding or removing annotations in the text.

Figure 3 visualizes how the annotation tree is rendered into an HTML document. Boxes with a yellow border indicate the annotations rendered from the tree structure. Grey borders outline text blocks between annotations and line breaks.

Figure 4 illustrates the rendering of the document shown in Figure 3 within the Orbis Annotator user interface. Edits by annotators trigger calls to the Orbis API which ensures that changes are serialized in real-time.

3.4 Usability optimizations

Orbis supports both mouse- and keyboard-centred workflows. The mouse-centred workflow allows

users to perform annotation tasks without any use of the keyboard. The keyboard-centred workflow is currently in beta.

3.5 Corpus metrics

The current version of Orbis Annotator implements the following corpus quality metrics which may be computed through the Orbis evaluation command line client.

1. *Average F1 measure*: The average F1 measure computes the F1 metric between n annotators, to assess the amount of agreement between them.

$$\bar{F}_1 = \frac{1}{n \cdot (n-1)} \sum_i^n \sum_{j \neq i}^n F_1(i, j) \quad (1)$$

2. *Modified Kappa*: The modified Kappa metric is based on the Fleiss' Kappa but does not correct for random agreement since it is usually negligible for corpus annotation tasks. It is computed by obtaining the average probability (P_i) of agreement among raters for each annotation i . Equation 3 shows the computation of P_i for annotation i based on the number of total raters n_i for that particular annotation and the number of raters considering it to be valid ($n_{i,vd}$) and invalid ($n_{i,-vd}$).

$$P_i = \frac{\sum_{j \in \{vd, -vd\}} n_{ij} (n_{ij} - 1)}{n_i (n_i - 1)} \quad (2)$$

$$\kappa^* = \frac{1}{n} \sum_{i=1}^n P_i \quad (3)$$

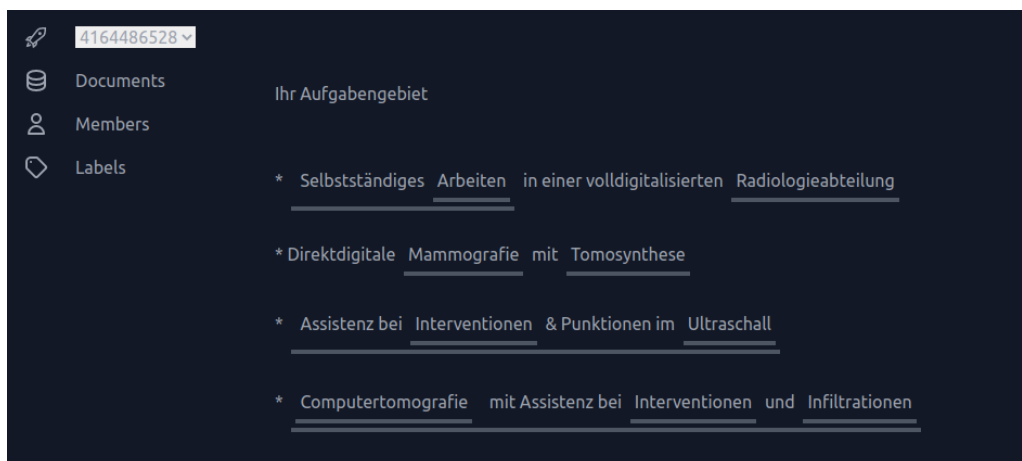


Figure 4: Orbis Annotator user interface for a given document with overlapping annotations.

Future version of Orbis Annotator will fully integrate with the Orbis Evaluation framework, which will allow conducting comprehensive evaluations and visual analytics on all annotated datasets.

3.6 Extensibility

Orbis Annotator includes the basic functionality required for uploading, annotating, evaluating and downloading corpora. In addition, it supports more complex use cases, such as automatically pre-annotating corpora through its API.

Future Orbis Annotator versions, will provide a plugin framework which allows extending both its user interface and API. Bundling these plugins in docker images that also include dependencies will provide additional functionality which is accessible to any user capable of starting a docker image and working with a web browser. Pre-configured docker images with automatic annotators such as SpaCy¹⁵, DBpedia Spotlight¹⁶ or Recognize (Weichselbraun et al., 2019b), for instance, can enrich Orbis Annotator with active learning support.

4 Evaluation

The following section performs a qualitative evaluation which compares Orbis Annotator to other open-source annotation tools (Section 4.1), and presents its application to two sophisticated real-world use cases (Section 4.2).

4.1 Comparison of Annotation Frameworks

The following comparison of annotation frameworks focuses on open-source software, that is still under active development.

¹⁵<https://spacy.io/>

¹⁶<https://www.dbpedia-spotlight.org/>

We excluded proprietary tools, since they are limited in transparency, customizability, and interoperability with other software. Moreover, commercial tools often require payment of high licensing fees, which are a significant barrier for researchers with limited resources or those who require extensive customization or experimentation with the software. Commercial solutions are, therefore, not considered in the comparison.

The comparison also excludes software which might not be maintained any more. As criteria for assessing a software’s maintenance status, we investigated its code repository and excluded tools that haven’t received any fixes or updates within the last two years, as we wanted to focus on systems that are still actively developed. This constraint led to the exclusions of Callisto¹⁷, CoSACT¹⁸ and Gate Teamware¹⁹.

We assess popular annotation tools based on the following criteria:

Custom Types: The ability to define custom annotation types in an annotation tool is essential for adapting annotation tools to new domains and use cases. Custom annotation types enable domain-specific annotations that capture the unique features and nuances of the data being annotated, improving the accuracy of downstream analyses. Furthermore, the ability to define custom annotation types enables collaboration and reproducibility by allowing researchers to use a standardized annotation schema. Overall, custom annotation types are crucial for achieving high-quality annotations and advancing scientific research.

¹⁷<https://mitre.github.io/callisto/>

¹⁸<https://github.com/TDaudert/CoSACT>

¹⁹<https://gate.ac.uk/teamware/>

Table 1: Comparison of popular open-source annotation tools.

	Nested Annotations	Custom types	Machine-aided annotations	Metrics	Multi User	Easy setup	License
Orbis Annotator	⊕	⊕	⊕	⊕	⊕	⊕ (Docker)	Apache 2.0
Argilla	⊕	⊕	⊕	⊕	⊕	⊕ (Docker)	Apache 2.0
Doccano	-	⊕	⊕	⊕	⊕	⊕ (Docker)	MIT
TagEditor	-	⊕	-	-	-	- (EXE-file)	MIT
Inception	-	⊕	⊕	⊕	⊕	- (Runnable Jar)	Apache 2.0
Annotation Studio	⊕	⊕	-	-	⊕	- (multi-step setup)	GPL 2.0
BRAT	-	⊕	-	-	⊕	- (Installer-Script)	MIT

Machine-Aided Annotations: Due to the sheer volume of data that needs to be annotated, machine-aided automatic annotations have become increasingly important recently. Machine learning algorithms can assist human annotators by automatically suggesting annotations for a given input based on pre-existing labelled data. This can significantly reduce the time and cost associated with manual annotation.

Multi-User: Multi-user-support in an annotation-tool is crucial for collaborative annotation projects in scientific research. With the ability to support multiple users, teams can work together to complete annotations more efficiently and effectively. This feature enables team members to view and edit annotations made by others, fostering collaboration and enhancing the accuracy and completeness of the annotations. Additionally, multi-user-support can provide a platform for experts to review and validate annotations made by less experienced annotators, improving the quality of the annotations.

Nested Annotations: Often, named entities are not linear but rather nested (i.e., a single entity can contain other entities). For instance, the mention “Barack Obama” refers to a person, but is nested within the mention “Barack Obama’s administration” which points to an organization. Being able to annotate such nested annotations is crucial for accurately capturing the complexity of named entities in text. Annotating nested entities can improve the quality of the corpus and the performance of named entity recognition systems trained on it, as they can learn to recognize more complex named entity structures.

Easy Setup: Ease of setup is an essential factor to consider. With the increasing complexity of NLP and machine learning models, researchers require efficient and user-friendly tools to streamline their work. Single-platform executables were generally excluded, as we wanted to focus on tools

for a larger audience. Software that is difficult to set up and configure can pose significant barriers to adoption, hindering the progress of research. In contrast, tools that are easy to set up and use can save researchers valuable time and effort, allowing them to focus on their research questions and hypotheses. Additionally, software with straightforward setup processes can encourage collaboration and community-building, as they make it easier for researchers to share their work and replicate experiments.

License Type: Open-source tools have revolutionized the fields of natural language processing (NLP) and machine learning research by providing researchers with accessible and customizable software. The use of open-source software has contributed towards increasing the reproducibility and transparency of research, since code and data are freely available for inspection and modification. In addition, open-source tools facilitate collaboration and community-building, by enabling researchers to share resources, expertise, and best practices.

Table 1 summarizes the evaluation results. The ⊕ symbol indicates that a criterion has been fully fulfilled, a minus refers to missing or only partially met criteria.

Support for nested annotations, machine-added annotations and corpus metrics are the areas that are most often neglected in the compared tools. Both Argilla and Orbis excel in these areas. In addition, future versions of Orbis Annotator will offer a tight integration with the Orbis Visual Benchmarking framework which will allow performing comprehensive evaluations of the created datasets and enable features designed toward improving the explainability of benchmarking results, such as drill-down analyses and aids for visualizing and interpreting evaluation results.

4.2 Use cases

This section discusses the use of the Orbis Annotator in two sophisticated real-world use cases which have significantly benefited from its development.

4.2.1 Machine-aided corpus annotation with non-standard, complex entity types

The first use cases showcased how machine-aided pre-annotations of complex entity types can lead to significant productivity gains of human annotators.

This use case design has been triggered by an applied research project in which the industry partner used a custom composite entity type to represent employee skills. This custom type combines a noun which specifies the skill's topic (e.g., Python) with a verb that indicates the skill's scope (e.g., programming). The composite skill type, therefore, enables a much more fine-grained distinction of a skill's required depth and direction (e.g., knowledge versus application or use). The skill scope may range from a shallow understanding ("knowing Python"), to different levels of practical experience ("programming Python", "debugging Python"), and the expertise required to actually teach a skill ("teaching Python").

Initially, human annotators identified these skills manually in real-time job posting feeds. They then copied sentences mentioning skills into a Google spreadsheet and provided a list of topic+scope tuples for these sentences.

The low productivity of the described process triggered the development of Orbis Annotator and migration to the machine-aided processes outlined in Figure 5. A machine learning pipeline splits job announcements into sentences, and then identifies sentences that are likely to contain composite skills. Afterward, an entity linking component provides a silver standard of annotated skill topics and skill scopes, which is then fed into the Orbis Annotator. Domain experts validate, extend and correct the provided silver standard annotations, creating a corpus of gold standard annotations, and the corresponding composite skills required for the industry partner's skill database. The annotation pipeline also queries the Orbis API for feedback on corrected annotations that is then used for enhancing the pipeline's machine learning components. The new process has considerably improved the productivity of the human annotators and helped in identifying over 80,000 different composite skills.

4.2.2 Knowledge Graphs migration

Knowledge graphs (KG) such as DBpedia and Wikidata have considerably grown recently (Hogan et al., 2021). Consequently, named entities that haven't been available in earlier KG versions (i.e., so called *nil entities*), are often present in more recent graphs. The issue of nil entities is particularly important when evaluating machine learning components with older gold standards. The Reuters 128 corpus, for instance, has been published in 2014 (Röder et al., 2014) and consequently misses entities that haven't been available in DBpedia at annotation time (Brasoveanu et al., 2018).

Also, shifts in a graph's popularity or the need to collaborate with partners that rely on a specific KG may trigger the need to migrate to either a newer KG version or even to another KG (e.g., from DBpedia to Wikidata).

Orbis supports such use cases by recording the history between annotated corpora. It, therefore, supports comparative evaluations and the computation of standard metrics which outline the differences between these annotated corpus versions. Orbis' corpus versioning also tracks relations between corpora, making changes more traceable and explicit (Weichselbraun et al., 2019a).

Figure 6 outlines a semi-automatic process for efficiently translating a language resource to a new KG. An automatic KG translation component aims at linking existing entities to the new KG. Depending on the involved KGs either knowledge rich approaches (e.g., based on owl:sameAs links between the KGs) or named entity linking might be deployed at this stage. Afterwards, a named entity recognition component enriches the corpus with candidate entities. Human annotators create a new version of the gold standard by correcting the automatically generated silver standard annotations. Finally, feedback on these corrections is leveraged for improving the machine learning components used in this process.

5 Outlook and Conclusions

This paper introduces the Orbis Annotator framework, a user-friendly, easy to install software that supports users in efficiently annotating language data. Orbis Annotator supports standard use cases through a pre-configured docker image and supports advanced setups through its API. Orbis Annotator also supports use cases that require tracking corpus versions and changes between these ver-

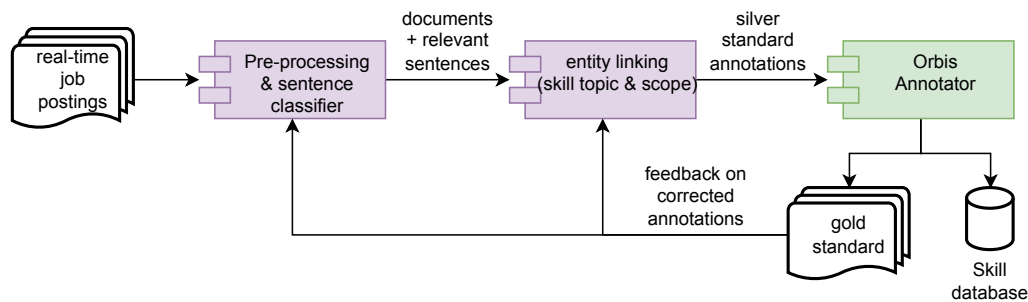


Figure 5: Machine-aided corpus pre-annotation with human feedback.

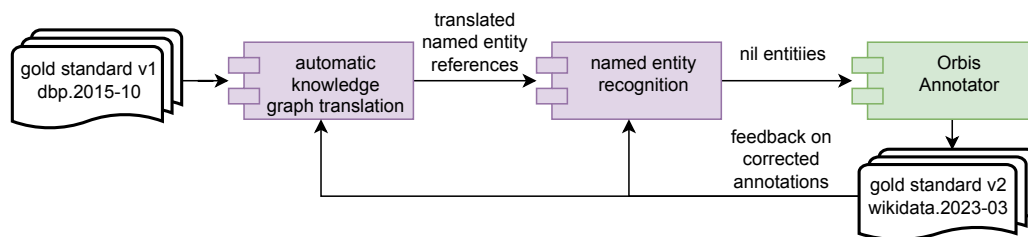


Figure 6: Migrate gold standards to a new knowledge graph or knowledge graph version.

sions. In addition, it aids researchers in tracking and assessing annotator reliability by computing corpus metrics such as inter-rater agreement.

Future work will focus on improving annotator efficiency (e.g., by adding support for additional workflows), and will integrate Orbis Annotator with the Orbis Visual Benchmarking framework. This will enable researchers to conduct evaluations of human, machine and hybrid annotators from within the Orbis Web Interface and to draw upon tools that help in explaining evaluation results such as drill-down analysis and visualizations. Orbis is currently built around JSON, NIF and CSV formats, but since many other formats are used within the research community, we aim at considerably increasing the number of supported formats by integrating software such as Spicy Salmon (Fäth and Chiarcos, 2022) into the toolkit.

Acknowledgements

The research presented in this paper has been conducted within the Future of Work project (<https://www.fhgr.ch/future-of-work>) funded by InnoSuisse. Adrian M.P. Brasoveanu has been partially funded by the Vienna Science and Technology Fund (WWTF) [10.47379/ICT20096] and SDG-HUB (FFG, G.A. No. 892212). The authors would like to thank Philipp Kuntschik, Alexander van Schie, and Marc-Alexander Iten who created an annotation system that has inspired the creation of

Orbis Annotator. We would also like to express our gratitude towards Fabian Odoni who has authored an early version of the Orbis visual benchmarking platform.

References

- Giuseppe Abrami, Manuel Stoeckel, and Alexander Mehler. 2020. *Textannotator: A UIMA based tool for the simultaneous and collaborative annotation of texts*. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 891–900. European Language Resources Association.
- Kalina Bontcheva, Hamish Cunningham, Ian Roberts, Angus Roberts, Valentin Tablan, Niraj Aswani, and Genevieve Gorrell. 2013. *GATE teamware: a web-based, collaborative text annotation framework*. *Lang. Resour. Evaluation*, 47(4):1007–1029.
- Adrian Brasoveanu, Giuseppe Rizzo, Philipp Kuntschik, Albert Weichselbraun, and Lyndon J. B. Nixon. 2018. *Framing named entity linking error types*. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).
- Adrian M.P. Braşoveanu and Răzvan Andonie. 2022. *Visualizing and explaining language models*. In *Integrating Artificial Intelligence and Visualization for Visual Knowledge Discovery*, pages 213–237. Springer.
- Hamish Cunningham. 2002. *Gate, a general architecture for text engineering*. *Comput. Humanit.*, 36(2):223–254.

- Tobias Daudert. 2020. [A web-based collaborative annotation and consolidation tool](#). In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 7053–7059. European Language Resources Association.
- Christian Fäth and Christian Chiarcos. 2022. [Spicy salmon: Converting between 50+ annotation formats with fintan, pepper, salt and powla](#). In *Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference, LDL@LREC 2022, Marseille, France, June 20-25, 2022*, pages 61–68. European Language Resources Association.
- David A. Ferrucci and Adam Lally. 2004. [UIMA: an architectural approach to unstructured information processing in the corporate research environment](#). *Nat. Lang. Eng.*, 10(3-4):327–348.
- Angus G. Forbes, Kristine Lee, Gus Hahn-Powell, Marco Antonio Valenzuela-Escárcega, and Mihai Surdeanu. 2018. [Text annotation graphs: Annotating complex natural language phenomena](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).
- Gaël Guibon, Luce Lefevre, Matthieu Labeau, and Chloé Clavel. 2022. [EZCAT: an easy conversation annotation tool](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 1788–1797. European Language Resources Association.
- Sebastian Hellmann, Jens Lehmann, Sören Auer, and Marcus Nitzschke. 2012. [NIF combinator: Combining NLP tool output](#). In *Knowledge Engineering and Knowledge Management - 18th International Conference, EKAW 2012, Galway City, Ireland, October 8-12, 2012. Proceedings*, volume 7603 of *Lecture Notes in Computer Science*, pages 446–449. Springer.
- Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d’Amato, Gerard de Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. 2021. [Knowledge Graphs](#). Synthesis Lectures on Data, Semantics, and Knowledge. Morgan & Claypool Publishers.
- Dmytro Kalpakchi and Johan Boye. 2022. [Textinator: an internationalized tool for annotation and human evaluation in natural language processing and generation](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 856–866. European Language Resources Association.
- Jan-Christoph Klie, Michael Bugert, Beto Boulosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. [The inception platform: Machine-assisted and knowledge-oriented interactive annotation](#). In *COLING 2018, The 27th International Conference on Computational Linguistics: System Demonstrations, Santa Fe, New Mexico, August 20-26, 2018*, pages 5–9. Association for Computational Linguistics.
- Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. [The stanford corenlp natural language processing toolkit](#). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, System Demonstrations*, pages 55–60. The Association for Computer Linguistics.
- Tristan Miller, Nicolai Erbs, Hans-Peter Zorn, Torsten Zesch, and Iryna Gurevych. 2013. [Dkpro WSD: A generalized uima-based framework for word sense disambiguation](#). In *51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, Proceedings of the Conference System Demonstrations, 4-9 August 2013, Sofia, Bulgaria*, pages 37–42. The Association for Computer Linguistics.
- Mariana Neves and Jurica Seva. 2021. [An extensive review of tools for manual annotation of documents](#). *Briefings Bioinform.*, 22(1):146–163.
- Minh-Quoc Nghiem and Sophia Ananiadou. 2018. [Aplenty: annotation tool for creating high-quality datasets using active and proactive learning](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*, pages 108–113. Association for Computational Linguistics.
- Minh-Quoc Nghiem, Paul Baylis, and Sophia Ananiadou. 2021. [Paladin: an annotation tool based on active and proactive learning](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, EACL 2021, Online, April 19-23, 2021*, pages 238–243. Association for Computational Linguistics.
- Alexandros Fotios Ntogramatzis, Anna Gradou, Georgios Petasis, and Marko Kokol. 2022. [The ellogon web annotation tool: Annotating moral values and arguments](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 3442–3450. European Language Resources Association.
- Fabian Odoni, Philipp Kuntschik, Adrian M. P. Brasoveanu, and Albert Weichselbraun. 2018. [On the importance of drill-down analysis for assessing gold standards and named entity linking performance](#). In *Proceedings of the 14th International Conference on Semantic Systems, SEMANTiCS 2018, Vienna, Austria, September 10-13, 2018*, volume 137 of *Procedia Computer Science*, pages 33–42. Elsevier.

- Peter Polák, Muskaan Singh, Anna Nedoluzhko, and Ondrej Bojar. 2022. [ALIGNMEET: A comprehensive tool for meeting annotation, alignment, and evaluation](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 1771–1779. European Language Resources Association.
- Michael Raring, Malte Ostendorff, and Georg Rehm. 2022. [Semantic relations between text segments for semantic storytelling: Annotation tool - dataset - evaluation](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 4923–4932. European Language Resources Association.
- Michael Röder, Ricardo Usbeck, Sebastian Hellmann, Daniel Gerber, and Andreas Both. 2014. [N³ - A collection of datasets for named entity recognition and disambiguation in the NLP interchange format](#). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014*, pages 3529–3533. European Language Resources Association (ELRA).
- Pontus Stenetorp, Sampo Pyysalo, Goran Topic, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. [brat: a web-based tool for nlp-assisted text annotation](#). In *EACL 2012, 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, April 23-27, 2012*, pages 102–107. The Association for Computer Linguistics.
- Albert Weichselbraun, Adrian M. P. Brasoveanu, Philipp Kuntschik, and Lyndon J. B. Nixon. 2019a. [Improving named entity linking corpora quality](#). In *Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP 2019, Varna, Bulgaria, September 2-4, 2019*, pages 1328–1337. INCOMA Ltd.
- Albert Weichselbraun, Adrian M. P. Brasoveanu, Roger Waldvogel, and Fabian Odoni. 2020. [Harvest - an open source toolkit for extracting posts and post meta-data from web forums](#). In *IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, WI/IAT 2020, Melbourne, Australia, December 14-17, 2020*, pages 438–444. IEEE.
- Albert Weichselbraun, Philipp Kuntschik, and Adrian M. P. Brasoveanu. 2019b. [Name variants for improving entity discovery and linking](#). In *2nd Conference on Language, Data and Knowledge, LDK 2019, May 20-23, 2019, Leipzig, Germany*, volume 70 of *OASICS*, pages 14:1–14:15. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.
- Max Wiechmann, Seid Muhie Yimam, and Chris Biemann. 2021. [Activeanno: General-purpose document-level annotation tool with active learning integration](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 99–105. Association for Computational Linguistics.
- Jie Yang, Yue Zhang, Linwei Li, and Xingxuan Li. 2018. [YEDDA: A lightweight collaborative text span annotation tool](#). In *Proceedings of ACL 2018, Melbourne, Australia, July 15-20, 2018, System Demonstrations*, pages 31–36. Association for Computational Linguistics.
- Seid Muhie Yimam, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. 2013. [Webanno: A flexible, web-based and visually supported system for distributed annotations](#). In *51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, Proceedings of the Conference System Demonstrations, 4-9 August 2013, Sofia, Bulgaria*, pages 1–6. The Association for Computer Linguistics.

Open-Source Thesaurus Development for Under-Resourced Languages: a Welsh Case Study

Nouran Khallaf¹, Elin Arfon², Mo El-Haj¹, Jonathan Morris², Dawn Knight², Paul Rayson¹,

¹Lancaster University, ²Cardiff University
 {n.khallaf1,m.el-haj,p.rayson}@lancaster.ac.uk,
 {ArfonE,morrisj17,knightd5}@cardiff.ac.uk

Tymaa Hammouda³ and Mustafa Jarrar³

³Birzeit University
 {1171779@student.birzeit.edu,mjarrar@birzeit.edu}

Abstract

This paper introduces an open-access, user-friendly online thesaurus for the Welsh language, aimed at enriching digital resources for Welsh speakers and learners. Utilising advances in Natural Language Processing (NLP), our approach combines pre-existing word embeddings, a Welsh semantic tagger, and human evaluation to establish related terms. In this case, an initial list of 250 words was expanded by adding 6,953 synonyms provided by linguists, creating a more extensive foundation for building the gold-standards. With this expanded list, when a user queries a particular word, the thesaurus presents all of its synonyms, allowing them to choose from a wider range of options. This is especially helpful when a user is unsure of the exact word they want to use or wants to explore different ways to express a concept. The resulting thesaurus offers a comprehensive, reliable resource for Welsh language users, fostering enhanced communication and expression. Our work promotes Welsh NLP and showcases NLP's potential to support under-resourced languages. The thesaurus will be accessible via a bilingual website, and the accompanying Python code will be available in a bilingual, public GitHub repository, and it will be available as a web service. Our approach presents a more efficient, cost-effective method for thesaurus creation, with potential applicability to other under-resourced languages.

1 Introduction

The Welsh language is a critical component of Welsh cultural identity and heritage. The latest (2021) census reports that 538,300 people aged three and over consider themselves to be speakers of the language, which corresponds to 17.8% of the population¹. Despite its importance, Welsh language users face significant challenges in accessing digital resources, particularly when it comes to

¹Welsh Language in Wales (Census 2021) <https://www.gov.wales/welsh-language-wales-census-2021-html>

reference tools such as a thesaurus. This is a significant barrier to the promotion and preservation of the Welsh language, as it limits the ability of users to effectively communicate and express themselves in Welsh. While there are some Welsh language thesauri currently available, such as The Gweiadur project², which is still in beta, these resources are limited in scope and do not provide the level of functionality that users need to fully utilise the Welsh language. As such, the development of a comprehensive Welsh language thesaurus is essential for the promotion and preservation of the Welsh language, and to enable Welsh language users to communicate effectively and express themselves in their native language.

Currently, the creation of a comprehensive Welsh language thesaurus involves significant manual effort, with lexicographers and linguists required to curate the content and ensure its accuracy. This process is time-consuming, expensive, and often reliant on the availability of skilled professionals. By leveraging recent developments in NLP and word embeddings, we can create a thesaurus for Welsh that is faster, more cost-effective, and more scalable. Word embeddings provide a way to identify and group words based on their meaning and usage, allowing for the automated creation of a network of related words. This significantly reduces the need for human intervention, enabling us to create a comprehensive Welsh language thesaurus that can be easily updated and maintained over time. In this way, our approach has the potential to significantly enhance the availability of digital resources for Welsh language users, facilitating effective communication and expression in Welsh.

This paper presents the development of an open-access, freely available online thesaurus for the Welsh language, which aims to enhance digital resources available to Welsh speakers and learners

²<https://www.gweiadur.com/thesaurus>

(financed by Welsh Government). Our approach leverages recent advances in NLP, using preexisting word embeddings to identify related words, a Welsh semantic tagger (Piao et al., 2017) and human evaluators to refine the similarities. This innovative methodology has shown success with more widely spoken languages, such as French (Hazem and Daille, 2018), and our work represents an important contribution to under-resourced languages such as Welsh, where the availability of digital resources is limited. The resulting thesaurus provides a comprehensive and reliable resource for Welsh language users, enabling more effective communication and expression in Welsh. In addition, our methodology has the potential to be applied to other under-resourced languages, offering a more automated and cost-effective approach to thesaurus compilation. This paper contributes to the advancement of Welsh language NLP and demonstrates the potential for NLP methods to benefit under-resourced languages.

Recent developments in NLP have enabled the creation of word embeddings, which involve transforming words in a corpus (collection of speech) to vectors. Words that are similar in meaning or association are mapped to a similar location in the vector space, allowing for the identification of related words and the creation of a network of related words. For the language user, this represents a valuable resource that goes beyond traditional thesauri, as it enables them to discover and explore a wider range of related words and concepts.

In our project, we used pre-existing word embeddings for Welsh (Corcoran et al., 2021) to find similar words, providing a starting point for the development of our Welsh language thesaurus. However, to ensure the accuracy and relevance of the thesaurus, we further refined the similarities using the Welsh Semantic Tagger, which helped to ensure that similar words belong to the same Part-of-Speech (POS) and the same semantic field as the original. This process will enable us to create a comprehensive and reliable resource for Welsh language users.

The resulting thesaurus will be publicly available as a fully bilingual and user-friendly website. Additionally, the accompanying python code will be available through a bilingual, public-facing GitHub repository, enabling other researchers to build on our work and further improve Welsh language NLP. In this way, our work will contribute to the ad-

vancement of Welsh language NLP and provide an additional valuable resource for Welsh language users (El-Haj et al., 2022a,b; Ezeani et al., 2022; Morris et al., 2022).

2 Related Work

2.1 Low Resourced Languages: The Importance of Welsh Language and Technology

Welsh is an official language in Wales and current legislation places responsibilities on certain public bodies to provide bilingual services, including digital resources³. However, the availability of such resources for Welsh language users arguably remains limited, particularly when it comes to reference tools such as a thesaurus.

The Welsh government has made efforts to safeguard and promote the use of the Welsh language (Carlin and Chr st, 2016), but the uptake of Welsh language websites and e-services remains relatively low (Cunliffe et al., 2013). One reason for this may be the assumption that the language used in such resources will be too complicated. However, guidelines exist for creating easy-to-read documents in Welsh, including the use of everyday words rather than specialised terminology and a neutral register (Arthur and Williams, 2019; Williams, 1999).

The work presented in this paper aims to contribute to the digital infrastructure of the Welsh language, by developing an open-access, freely available online thesaurus for Welsh speakers and learners alike, including the introduction of Welsh Language Standards which place requirements on public institutions to provide fully bilingual web content (Carlin and Chr st, 2016).

The resulting thesaurus will complement the suite of Welsh language technologies, making it easier for content creators and Welsh readers to communicate effectively in Welsh. Additionally, the thesaurus will be of use to Welsh-medium educators and learners, who can use it as a pedagogical tool to better understand the nuances of the Welsh language. In addition, the work contributes to the advancement of Welsh language NLP and demonstrates the potential for NLP methods to benefit under-resourced languages. By leveraging the power of technology, we can help make the

³Welsh Language Standards www.welshlanguagecommissioner.wales/public-organisations/welsh-language-standards

Welsh language more accessible and easier to use for Welsh speakers and learners alike.

2.2 Semantic Field Annotation

In terms of thesaurus compilation for low resourced languages, we can benefit from linguistic knowledge already embedded in any existing taxonomies or ontologies if they are available, and in the case of Welsh, one such key resource is the UCREL Semantic Analysis System (USAS)⁴. Originally developed for English text (Rayson et al., 2004), a similar system was subsequently created for Welsh during the CorCenCC project⁵ (Piao et al., 2018). USAS is a knowledge based annotation system, drawing on lexicons of single words and multiword expressions (MWEs) that have been manually created or checked by native speakers, to provide lists of potential coarse-grained word senses for each word or MWE. The USAS tagger then uses a variety of disambiguation methods to select the most likely meaning in context, employing a set of 232 semantic fields for its labelling of semantic tags or concepts⁶. For Welsh, the tagger achieves coverage of 91.78% in text, thus providing a wide set of information linking words to others that share the same conceptual category, in this case, via the semantic field tagging.

2.3 Thesaurus Creation

Creating a thesaurus involves compiling a list of related terms organised by the meaning of the words. There are several methods for creating a thesaurus, including manual and automated methods.

Manual methods involve human experts compiling lists of related terms based on their knowledge of the subject area. These experts may use a variety of sources, such as domain-specific dictionaries, thesauri, and other reference materials to identify related terms. This method is time-consuming but can produce high-quality thesauri (Aitchison et al., 2000).

Automated methods use either statistical algorithms or NLP techniques to identify relationships between words. This method depends on using large corpora to identify related terms based on their co-occurrence patterns in the corpus. This method is faster than manual methods but the results can be less accurate (Manning et al., 2008).

⁴<https://ucrel.lancs.ac.uk/usas/>

⁵<https://corcenc.org/>

⁶<https://ucrel.lancs.ac.uk/usas/USASemanticTagset.pdf>

There are several NLP approaches that can be used for the creation of thesauri. Distributional semantics, semantic clustering, semantic role labelling, graph-based algorithms, and other techniques such as Latent Semantic Analysis (LSA) (Turney, 2007), Latent Dirichlet Allocation (LDA), and word embeddings are all effective methods for identifying relationships between words and grouping them based on their semantic meaning.

One example of an NLP approach to thesaurus creation is the use of distributional semantics, which models the meaning of a word based on the distribution of its context words in a large corpus. This approach has been used to create a variety of thesauri in different languages, including English (Turney, 2007). Another semantic clustering algorithm that group words together based on their semantic similarity. As such, the WordNet thesaurus was created using this method, where words are organised into synsets (sets of synonyms) based on their meanings (Fellbaum, 1998). Semantic Role Labelling (SRL), which identifies the roles that words play in a sentence is another method for word grouping. For example, the WordNet Domains thesaurus was created using SRL, where the roles played by nouns in a corpus of texts were used to identify the semantic domains of the words (Magnini et al., 2000).

Hybrid methods combine manual and automated methods, using human experts to validate the results of automated algorithms. This method can produce high-quality, more efficient and cost-effective thesauri than relying solely on manual methods. Nonetheless, the use of NLP techniques for thesaurus creation has shown promise in creating comprehensive and accurate thesauri.

Latest NLP techniques that have been used for thesaurus creation include Word Embeddings. Landthaler et al. (2018) proposed a method for extending existing thesauri by leveraging word embeddings and the intersection method. Their approach involved using word embeddings to identify candidate synonyms for each entry in an existing thesaurus, and then intersecting these candidates with the existing synonym sets to identify and validate new synonyms. The authors evaluated their method on an existing thesaurus of human resources management terms and demonstrated that their method significantly improved the coverage and precision of the thesaurus, while maintaining its consistency and coherence.

Our approach utilises recent developments in NLP to identify related words by using pre-existing Welsh word embeddings. We further refine these similarities through a Welsh semantic tagger and human evaluators to create a reliable and comprehensive resource for users of the Welsh language. Our method has been tested on existing dictionaries and graph-based thesauri, and is described in detail in the rest of the paper.

3 Words lists description

In order to build and evaluate our thesaurus for Welsh, we began by creating gold-standard synonyms for a list of 250 words. This list was comprised of 84 NOUN lemmas, 84 VERB lemmas (excluding conjugated verbs), and 82 ADJECTIVE lemmas, all taken from a frequency list of Welsh words (Knight et al., 2020).

We started by obtaining a list of 500 most frequent Welsh words from the Welsh National Corpus (Knight, 2020; Knight et al., 2021), specifically from the Yr-Amliadur.pdf document available on the CorCenCC website (Knight et al., 2020). From this list, we selected roughly equal numbers of nouns, adjectives, and verbs, excluding any duplicates or conjugated verbs.

To ensure a diverse selection of words for our gold standard, we included items from both the beginning and final parts of the list. We also included a number of homophones. This approach allowed us to capture a range of word types and usage contexts, including less common words that may be important for Welsh language users but are not frequently encountered in everyday language.

Our aim in building this gold-standard was to provide a reliable set of synonyms that we could use to evaluate the performance of our thesaurus-building methods. By establishing a solid foundation of gold-standard synonyms, we could measure the accuracy and usefulness of our thesaurus, and identify areas for improvement as we refined our approach.

4 Experiment 1: Welsh word embeddings

Word embeddings are widely used in NLP and machine learning tasks to capture the semantic and syntactic properties of words in a continuous vector space. FastText is a popular method for training word embeddings that can handle out-of-vocabulary words and subword information using character n-grams (Grave et al., 2018). The

two pre-trained Welsh word embeddings used in this experiment were the FastText embeddings trained on a large Welsh corpus from Wikipedia and the fine-tuned FastText (Fine-Tuned-FastText) embeddings using the Welsh Wikipedia as well as the Welsh National Corpus along with 9 other resources (92,963,671 words) (Corcoran et al., 2021).

To evaluate the performance of the word embeddings, we used the gold-standard synonyms generated by Welsh speakers as the reference. We compared the generated synonyms for each word in the gold-standard with the synonyms generated by the two word embeddings.

We used the FastText embeddings and fine-tuned-FastText to generate the 10 nearest (most related) words to each input word on our 250-word list. The resulting list of nearest words for the example word “pobl” is shown in Table 1, along with their translations. Based on the Table 1, it is clear that the fine-tuned FastText approach yielded better results than the standard FastText approach in terms of identifying the most related words to the Welsh word “pobl”. The most related words generated by the fine-tuned FastText approach were very close in meaning to the original word, as indicated by their high similarity scores ranging from 0.733 to 0.468. In contrast, the most related words generated by the standard FastText approach had lower similarity scores ranging from 0.629 to 0.504.

An important point to consider is that the nearest words may include antonyms of the input word, as the embeddings are based on the behaviour of the word in various contexts. This process allowed us to leverage the power of FastText embeddings to quickly and automatically generate potential synonyms for each word on our list.

To refine the word embedding results, we used the Python Multilingual UCREL Semantic Analysis System (PyMUSAS)⁷, which retains Welsh language resources and methods originally included in an earlier Java version developed during the CorCenCC project (Piao et al., 2018). The PyMUSAS tagger assigns a set of fine-grained semantic tags to each word based on its POS (assigned by the CyTag Welsh POS tagger also created during the CorCenCC project), morphological features, and semantic field. We selected a subset of the generated fastText words for each original word based on matching the semantic tags and removing matching lemmas. This can be done by comparing the

⁷<https://pypi.org/project/pymusas/>

FastText			Fine-Tuned-FastText		
Vector	Most Rel.	Trans.	Vector	Most Rel.	Trans.
0.629	bobl	people	0.733	pobol	people
0.556	rhai	some	0.641	bobl	people
0.546	phobl	people	0.554	phobl	people
0.530	LHDTQ	LGBTQ	0.551	bobol	people
0.528	pobol	people	0.551	rhywun	someone
0.522	cleiantiaid	clients	0.540	trigolion	inhabitants
0.515	ifanc	young	0.498	pawb	everyone
0.515	bod	being	0.482	dinasyddion	citizens
0.514	trwy'r	through/by the	0.480	plant	children
0.504	Ogleddwyr	North Walians	0.468	pobl'	people

Table 1: 10 most related words to the Welsh word ‘pobl’

semantic tags of the generated words with the semantic tag(s) of the original word and selecting the ones that share the same tag(s). Table 2 compares the performance of FastText embeddings and Fine-Tuned-FastText word embeddings in finding synonyms for the Welsh word “pobl”. While not all the FastText embeddings share the POS tag and seven synonyms do not share the PyMUSAS tag with the original word. Seven synonyms share both the POS tag and PyMUSAS tag of the original word, and three of the new synonyms share the same lemma with the original word. Based on this analysis, it appears that the Fine-Tuned-FastText word embeddings perform better than the FastText embeddings in terms of producing synonyms that share the same POS speech tag and PyMUSAS tag as the original word. Therefore, it may be worth exploring different techniques for refining word embeddings’ results, such as using a lemmatiser and semantic tagging.

In this experiment, when the semantic tagger produced a $Z99$ for a word that was not in its lexicon, the approach taken was to remove only the matched lemmas from the list rather than eliminating the $Z99$ words to avoid a very short list.

After applying lemmatisation and removing words that share the same lemma as the original word, the number of data entries was reduced from 2490 to 2047 for the FastText model, with an average of fewer than 9 synonyms per word. For the fine-tuned FastText model, the number was reduced to 1776, with an average of 7 synonyms per word. The lemmas that exactly match the original word lemma are in bold font in Table 2.

Next, by selecting only the words that share the same PyMUSAS semantic tag but do not share the

lemma, the number of entries further reduced to 132 for the FastText model and 173 for the Fine-Tuned-FastText model. This means that some of the data did not have any synonyms that share the same semantic PyMUSAS tag [S2, People].

This process of selecting synonyms that share the same semantic tag but not the same lemma can be useful in reducing redundancy and increasing the diversity of the synonyms list. It can also help in avoiding circular dependencies and improving the quality of the generated data. However, it is important to note that this process may also result in a loss of some relevant synonyms that do not share the same semantic tag as the original word. Therefore, it is essential to carefully evaluate the trade-offs and choose the appropriate method.

Once we have selected a subset of the generated words based on semantic similarity and lemma dissimilarity, we can match them with the gold-standard user input by comparing the words and their order with the user-generated synonyms. This will allow us to evaluate the quality and relevance of the generated synonyms and identify any discrepancies or inconsistencies with the user input.

5 Experiment 2: Analysis to create gold-standard

The objective of this study was to analyse the input provided by Welsh speakers in generating synonyms for a pre-compiled list of 250 Welsh words. The study aimed to create a gold-standard list of synonyms for these words based on the input of seven paid evaluators for each word. The seven evaluators were native speakers of Welsh, either in the final year of an undergraduate Welsh degree programme or postgraduate students with experi-

FastText				Fine-Tuned-FastText			
Word	Lemma	POS	PyMUSAS	Word	lemma	POS	PyMUSAS
bobl	pobl	E	S2	pobol	pobol	E	S2
rhai	rhai	unk	A13.5	bobl	pobl	E	S2
phobl	pobl	E	S2	phobl	pobl	E	S2
LHDTQ	LHDTQ	E	Z99	bobol	pobol	E	S2
pobol	pobol	E	S2	rhywun	rhywun	E	Z8mfc
cleiantiaid	cleiantiaid	unk	Z99	trigolion	trigolyn	E	H4/S2mf
ifanc	ifanc	Ans	T3-	pawb	pawb	unk	Z8/N5.1+c
bod	bod	B	A3+, Z5	dinasyddion	dinesydd	E	G1.1/S2mf
trwyr	trwyr	unk	Z99	plant	plentyn	E	S2mf/T3
Ogleddwyr	Ogleddwyr	E	Z99	pobl	pobl	E	S2

Table 2: Comparison of FastText and Fine-Tuned-FastText Word Embeddings in Generating Synonyms for the Welsh Word ‘pobl’

ence of writing in Welsh. They were asked to provide up to ten synonyms for each word, and the order in which they presented the synonyms was determined individually.

To create the gold-standard list, the study conducted comparison experiments to match the agreement of synonyms and their POS across the evaluators, as well as the agreement of the ordering of the presented synonyms. Additionally, the ordering of the synonyms provided by the evaluators was compared against the frequency of these words in the CorCenCC corpus frequency.

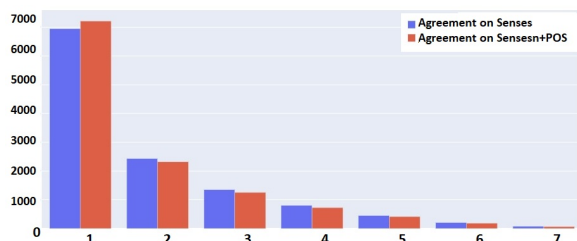


Figure 1: Sense versus POS agreement across the participants

Table 3 presents information on the level of agreement among annotators on the senses of words, with and without their part of speech tags. The table shows that for the majority of the words (4517 out of 6953), only one annotator suggested the sense of the word. As the number of annotators in agreement increases, the number of words decreases, indicating that agreement among annotators is less common for most of the words. For instance, only 64 words had seven annotators in agreement on their sense and part of speech tag. Figure 1 represents the relationship between sense

and POS agreement across participants.

The overall aim of the study was to provide a reliable and standardised list of synonyms for Welsh words that could be used in NLP applications. By analysing the input of multiple evaluators and creating a gold-standard list based on their input, the study aimed to ensure that the list was comprehensive and accurate. Furthermore, the study aimed to provide insights into the agreement and variability among Welsh speakers in generating synonyms, as well as the relationship between the ordering of synonyms and their frequency in the corpus. This experiment can be used to further develop NLP applications for Welsh language processing tasks and improve the accuracy and relevance of the generated synonyms.

Agreements	Sense	Sense & POS
only 1	4517	4895
at least 2	2436	2323
at least 3	1354	1257
at least 4	808	729
at least 5	454	416
at least 6	209	186
at least 7	80	64

Table 3: Gold-standard Words Agreements.

Agreements: the number of annotators in agreement. Sense: number of agreements on senses. Sense & POS: number of agreements on senses and their part of speech tags

The gold-standard synonyms provided by the seven participants were ordered based on the mean position of each synonym across all participants. For instance, the word ‘pobl’ had 31 unique synonyms suggested by the participants, as shown in Figure 2. To quantify the variability or fluctua-

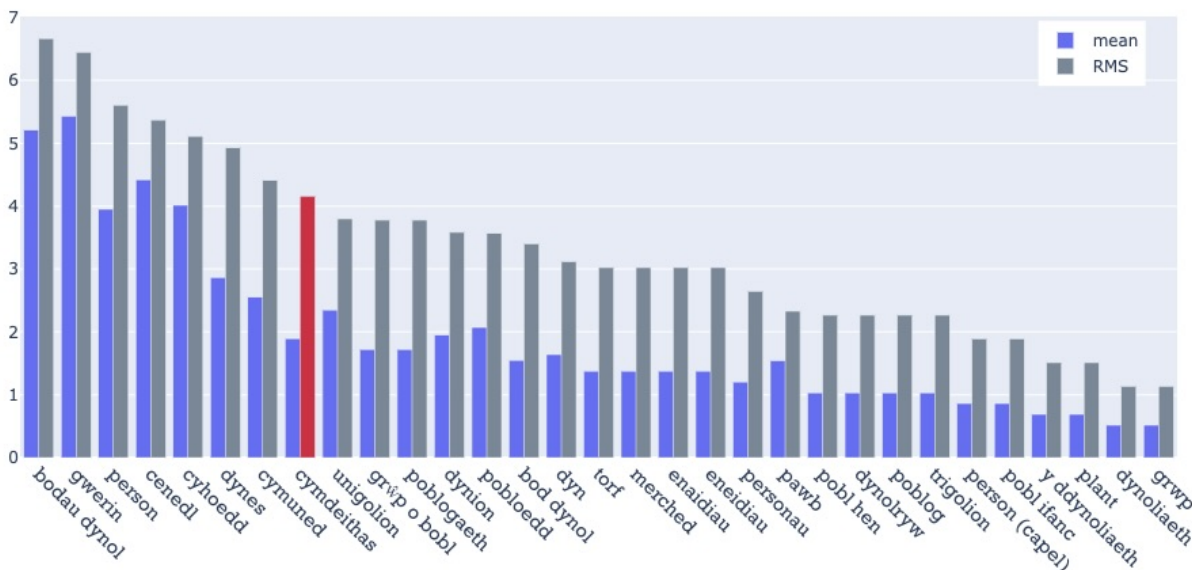


Figure 2: Mean versus RMS for the word ‘pobl’ senses

tion in a set of values, we used the Root Mean Square (RMS), which is a mathematical measure commonly used in various fields, including language processing. Let a set of n values be denoted by

$$x_1, x_2, \dots, x_n$$

Then, the RMS can be computed as:

$$x_{RMS} = \sqrt{\frac{1}{n} (x_1^2 + x_2^2 + x_3^2 \dots + x_n^2)}$$

By using RMS to reorder synonyms, words suggested by a single participant but in a higher position will be given more weight than words suggested by multiple participants but in lower positions. This is because RMS takes into account the variability of the data and gives more weight to values that are farther from the mean.

In the specific example of the word “cymdeithas” [red column] shown in Figure 2, the word was introduced by only one speaker but was in a higher position when RMS was used to reorder the synonyms. This indicates that the word was used more frequently or prominently by the participant who suggested it, and thus should be given more weight in the final output.

Using RMS to reorder synonyms can be a useful technique to ensure that the most relevant and frequently used words are given priority, even if they are suggested by fewer participants. This approach can help to produce a more accurate and representative list of synonyms for Welsh words, which can be valuable for various NLP applications.

In this case the 250 list of words was expanded by adding 6953 synonyms from linguists, and we now have a more extensive words to build the gold-standards. With this expanded list of words, when a user queries a particular word, the thesaurus can now present all of its synonyms as well, allowing the search to see and choose from a wider range of options. This can be especially useful when a user is unsure of the exact word they want to use, or when they want to explore different ways to express a particular concept.

One thing to keep in mind is that not all synonyms are interchangeable in every context, and some synonyms may have different connotations. Therefore, it was important to consider the context in which each synonym is used and to provide additional information or context as needed to help users choose the most appropriate synonym for their particular situation. This will be done by extracting an example for each word from CorCenCC corpus (Knight, 2020; Knight et al., 2021).

6 Experiment 3: Graph-based Approach

For our next experiment utilising existing dictionaries and thesauri, we developed a web tool for validating Welsh synonyms based on a graph-based algorithm as described by Ghanem et al. (2023)⁸. This algorithm constructs a graph at level k from a set of translation or synonymy pairs and consid-

⁸<https://portal.sina.birzeit.edu/synonyms/>

Figure 3: Synonyms Web Tool

ers all cyclic paths as candidate synonyms. The algorithm then calculates a fuzzy value for each candidate synonym to determine its likelihood of being a member of a synset.

Figure 3 depicts the tool⁹, which features several bilingual dictionaries, including the Welsh-English Dictionary by Hawke and the Welsh WordNet¹⁰, that we uploaded to the tool. It accepts a set of synonyms and validates them using this algorithm.

Table 4 displays the assessment of the linguists' synonyms in comparison to the tool's outcomes using three evaluation metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the Jaccard coefficient.

MAE and RMSE are numerical prediction accuracy metrics used to determine the ranking difference between the linguists' synonyms and the tool's results, with the synonyms represented by numerical vectors. The MAE measures the average absolute difference between the predicted and actual values, while the RMSE measures the square root of the average squared differences between the predicted and actual values. The evaluation resulted in MAE values ranging from 10.02 to 28.79 and RMSE values ranging from 13.05 to 35.38. Linguist 4 at level 3 exhibited the lowest MAE and RMSE values of 19.26 and 23.36, respectively, signifying the highest level of synonym prediction accuracy. Overall, the performance of all linguists and word-embeddings (WE) was superior at level 3 than at level 2.

The Jaccard coefficient calculates the similarity between two sets, ranging from 1 for identical sets to 0 for completely dissimilar sets. If a synonym is not found in the tool, it is ranked at the end of the synset and labeled as "out of vocabulary". Consequently, we must measure the overlap between the tool's identified synonyms and the input synset using the Jaccard coefficient. The comparison outcomes varied from 0.34 to 0.83, with linguist 4 at

⁹<https://portal.sina.birzeit.edu/synonyms/>

¹⁰<https://datainnovation.cardiff.ac.uk/is/wecy/access.html>

Linguist	Level	Jaccard	MAE	RMSE
1	2	0.77	28.55	35.00
	3	0.80	28.79	35.38
2	2	0.75	22.34	27.18
	3	0.77	22.34	27.40
3	2	0.74	24.27	30.01
	3	0.76	23.89	29.74
4	2	0.81	19.11	23.15
	3	0.83	19.26	23.36
5	2	0.75	24.06	29.53
	3	0.77	23.83	29.39
6	2	0.52	14.54	18.15
	3	0.54	14.86	18.51
7	2	0.34	10.02	13.05
	3	0.35	10.03	13.07
WE	2	0.43	27.69	33.79
	3	0.45	28.74	35.03

Table 4: Evaluation of linguists' synonyms against the dictionaries and WordNet results

level 3 exhibiting the highest Jaccard coefficient value of 0.83, indicating a high degree of similarity between their synonyms and the reference set.

Overall, the evaluation results indicate significant variation in the quality of the linguists' synonyms, with linguist 4 at level 3 demonstrating the best performance across all three evaluation metrics.

This experiment provides valuable insights into the effectiveness of multilingual extraction methods in generating related words in Welsh, while also highlighting the strengths and limitations of different techniques and linguists. These findings can further inform the development of more precise and comprehensive thesauri and word embeddings for Welsh language processing tasks.

7 Conclusion and Future Work

In this paper, we presented our approach to creating a comprehensive thesaurus for Welsh using a combination of existing resources and novel techniques. We demonstrated the effectiveness of our

approach through a series of experiments and evaluations, and showed that our thesaurus outperformed existing Welsh-language resources in generating related words. Our approach leverages the power of FastText embeddings and semantic tagging to generate candidate synonyms, and RMS reordering to identify the most relevant and frequently used words.

However, our work is not without limitations. While we aimed to create a comprehensive and accurate thesaurus, it is possible that our resource is still incomplete or may contain errors. To further refine and improve the thesaurus in the future, we will enlist the help of human evaluators who are fluent in Welsh. Specifically, we will use a pre-existing platform to crowd-source human participants to evaluate the resource. This will help ensure that the thesaurus is relevant, accurate, and meets the needs of its users, enhancing its value and utility for Welsh speakers and learners.

By combining or comparing the results of the three experiments, we can gain a deeper understanding of how to optimise our approach and further refine the thesaurus. Specifically, we can identify areas for improvement and investigate how to address potential limitations or errors in the resource.

Overall, our work contributes to the growing body of research on NLP and machine learning for under-resourced languages, and demonstrates the potential of using novel techniques and approaches to create valuable resources for these languages. We hope that our work will inspire further research and development in this area, and that our thesaurus will be a useful tool for Welsh speakers, learners, and researchers alike.

Acknowledgements

We would like to express our gratitude to the evaluators who participated in this study, all of whom are native Welsh speakers. Their invaluable contributions and dedication to the project made this research possible. This research was funded by the Welsh Government, under the Grant ‘Using Word Embeddings to Create an Interactive Thesaurus of Contemporary Welsh’.

8 Ethics

The payment provided to the evaluators was in accordance with the UK’s national minimum wage regulations, to ensure that it meets or exceeds the

standard wage requirements.

References

- Jean Aitchison, Alan Gilchrist, and David Bawden. 2000. *Thesaurus Construction and Use: A Practical Manual*, 4th edition. Aslib, London.
- Rudy Arthur and Hywel TP Williams. 2019. The human geography of twitter: Quantifying regional identity and inter-region communication in england and wales. *PloS one*, 14(4):e0214466.
- Patrick Carlin and Diarmait Mac Giolla Chríost. 2016. A standard for language? policy, territory, and constitutionality in a devolving wales. In *Sociolinguistics in Wales*, pages 93–119. Springer.
- Padraig Corcoran, Geraint Palmer, Laura Arman, Dawn Knight, and Irena Spasić. 2021. [Creating welsh language word embeddings](#). *Applied Sciences*, 11(15):6896.
- Daniel Cunliffe, Delyth Morris, and Cynog Prys. 2013. Young bilinguals’ language behaviour in social networking sites: The use of welsh on facebook. *Journal of Computer-Mediated Communication*, 18(3):339–361.
- Mahmoud El-Haj, Ignatius Ezeani, Jonathan Morris, and Dawn Knight. 2022a. Creation of an evaluation corpus and baseline evaluation scores for welsh text summarisation. In *Proceedings of the 4th Celtic Language Technology Workshop within LREC2022*, pages 14–21.
- Mahmoud El-Haj, Ignatius Ezeani, Jonathan Morris, and Dawn Knight. 2022b. Welsh summaries correlation between rouge and human evaluation. In *Proceedings of the 4th Celtic Language Technology Workshop within LREC2022*, pages 14–21.
- Ignatius Ezeani, Mahmoud El-Haj, Jonathan Morris, and Dawn Knight. 2022. [Introducing the Welsh text summarisation dataset and baseline systems](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5097–5106, Marseille, France. European Language Resources Association.
- Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT press.
- Sana Ghanem, Mustafa Jarrar, Radi Jarrar, and Ibrahim Bounhas. 2023. A benchmark and scoring algorithm for enriching arabic synonyms. *The 12th International Global Wordnet Conference (GWC2023)*.
- Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

- Amir Hazem and Béatrice Daille. 2018. Word embedding approach for synonym extraction of multi-word terms. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- Dawn Knight. 2020. Corcencc: Corpws cenedlaethol cymraeg cyfoes—the national corpus of contemporary welsh. *Oxford Text Archive Core Collection*.
- Dawn Knight, Fernando Loizides, Steven Neale, Laurence Anthony, and Irena Spasić. 2021. Developing computational infrastructure for the corcencc corpus: the national corpus of contemporary welsh. *Language Resources and Evaluation*, 55:789–816.
- Dawn Knight, Steve Morris, Beth Tovey-Walsh, and Tess Fitzpatrick. 2020. [Yr amliadur: Frequency lists for contemporary welsh](#).
- Jörg Landthaler, Bernhard Watzl, Dominik Huth, Daniel Braun, and Florian Matthes. 2018. Extending thesauri using word embeddings and the intersection method. In *Proceedings of the 10th International Conference on Knowledge Engineering and Ontology Development*, pages 159–166. SciTePress.
- Bernardo Magnini, Carlo Strapparava, Piotr Pezik, and Ido Dagan. 2000. Integrating subject field codes in wordnet. *Computational Linguistics*, 26(2):199–227.
- Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.
- Jonathan Morris, Ignatius Ezeani, Ianto Gruffydd, Katharine Young, Lynne Davies, Mahmoud El-Haj, and Dawn Knight. 2022. Welsh automatic text summarisation. In *Wales Academic Symposium on Language Technologies*. Banolfan Bedwyr.
- Scott Piao, Paul Rayson, Dawn Knight, and Gareth Watkins. 2018. [Towards a Welsh semantic annotation system](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Scott Piao, Paul Rayson, Dawn Knight, Gareth Watkins, and Kevin Donnelly. 2017. Towards a welsh semantic tagger: creating lexicons for a resource poor language. In *Proceedings of Corpus Linguistics*.
- Paul Rayson, Dawn Archer, Scott Piao, and Tony McEnery. 2004. The UCREL Semantic Analysis System. In *Proceedings of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks in association with 4th International Conference on Language Resources and Evaluation (LREC 2004)*, pages 7–12.
- Peter D Turney. 2007. Measuring semantic similarity by latent relational analysis. *The Journal of Computer and System Sciences*, 73(4):389–401.
- Cen Williams. 1999. *Cymraeg Clir: Canllawiau Iaith*. Bangor: Gwynedd Council, Welsh Language Board and Canolfan Bedwyr.

ISO LMF 24613-6: A Revised Syntax Semantics Module for the Lexical Markup Framework

Francesca Frontini, Anas Fahad Khan
Cnr-Istituto di Linguistica
Computazionale “Antonio Zampolli”
Pisa, Italy
name.surname@ilc.cnr.it

Laurent Romary
Inria - Scientific Information
and Culture Directorate
Paris, France
laurent.romary@inria.fr

Abstract

The Lexical Markup Framework (LMF) is a meta-model for representing data in monolingual and multilingual lexical databases with a view to its use in computer applications. The "new LMF" replaces the old LMF standard, ISO 24613:2008, and is being published as a multi-part standard. This short paper introduces one of these new parts, ISO 24613-6, namely the Syntax and Semantics (SynSem) module. The SynSem module allows for the description of syntactic and semantic properties of lexemes, as well as the complex interactions between them. While the new standard remains faithful to (and backwards compatible with) the syntax and semantics coverage of the previous model, the new standard clarifies and simplifies it in a few places, which will be illustrated.

1 Introduction

The Lexical Markup Framework (LMF) is undoubtedly one of the most influential lexical standards of the last two decades. First published in 2008 by the International Standards Organization (ISO) as **ISO standard 24613:2008** it was intended as a “standardized framework for the construction of computational lexicons” (Francopoulo, 2013). LMF was developed with a special focus on two different kinds of lexicon, namely, digital born electronic lexicons specifically intended for use by Natural Language Processing applications, so called *NLP dictionaries*, as well as for electronic versions of print dictionaries, or more generally lexicons primarily intended for human consumption, so called *Machine Readable Dictionaries* (MRD). The original LMF, ISO 24613:2008, contained, two modules for syntax and semantics, respectively, whose scope, taken together, was to provide means of representing the syntactic and semantic argument structure of individual lexical entries. The approach taken by the original committee tasked with drafting LMF was a theory agnostic one which identified a nucleus of elements that were generic enough

to allow for the modelling of syntax, semantics and their interface without any particular theoretical bias. After its publication in 2008, LMF came to be used by a variety of different organisations and in a number of national and international projects¹. In particular, the syntactic and semantics models were extensively used in projects such as PAROLE and SIMPLE (Ruimy et al.; Lenci et al.) as well as being the basis for other models of the syntax/semantics interface in lexical resources, such as the W3C OntoLex Syntax and Semantics Module².

After a detailed review of the original standard, however, the decision was made in 2015 to revise LMF and, what’s more, to make it a multi-part standard with each part being published separately (as distinguished from the old LMF standard which was published in a single part but which contained separate modules as sub-parts). This new multi-part version of LMF is currently being developed within the standardisation sub-committee ISO TC 37/SC 4/WG 4 (to which the authors of the current article are all contributing), with the first five parts of the new version having already been published, and other parts at an advanced stage of completion. The current paper is dedicated to ISO 24613-6, a soon-to-be published part of the revised LMF standard dealing with Syntax and Semantics (henceforth **SynSem**), two areas which as we mentioned above were previously covered by separate modules in the old LMF. SynSem stays true to the overall approach of ISO 24613:2008, but some simplifications/modifications were introduced. In what follows, we shall begin by placing SynSem in the context of the new multipart LMF, and providing an update as to its current status. Then we shall describe the constituent parts of the standard: Syntax,

¹Searching for "LMF" in the CLARIN Virtual Language Observatory gives a good indication on resources and also tools using the 2008 model (<https://vlo.clarin.eu/?l&q=lmf>).

²https://www.w3.org/community/ontolex/wiki/Syntax_and_Semantics_Module

Semantics, and SynSem interface. Finally we shall provide some details as to its serialisation.

2 An Overview of the New Multipart LMF

Following (Romary et al., 2019) we provide a list of the new LMF parts in the present section along with their current status.

ISO 24613-1:2019 Language resource management — Lexical markup framework (LMF) — Part 1: Core model: This module defines the basic classes required to model a baseline lexicon and is a pre-requisite for the use of the other classes. **Status:** *Published in 2019 it is now being further revised to make it even easier to use.*

ISO 24613-2:2020 Language resource management — Lexical markup framework (LMF) — Part 2: Machine-readable dictionary (MRD) model: Contains components providing a deeper specification of lexical description encapsulated within the core model. **Status:** *Published in 2020.*

ISO 24613-3:2021 Language resource management — Lexical markup framework (LMF) — Part 3: Etymological extension: A completely new addition to the LMF meta-model covering etymological and diachronic information. This part makes etymologies, etymological links and etymons first class citizens. See (Khan and Bowers, 2020) for more details. **Status:** *Published in 2021.*

ISO 24613-4:2021 Language resource management — Lexical markup framework (LMF) — Part 4: TEI serialization: A TEI serialisation of the other parts of the model which aims to make both TEI and LMF fully compatible and which leverages the knowledge and makes use of the established practices of the TEI community in dealing with lexicographic resources. **Status:** *Published in 2021.*

ISO 24613-5:2022 Language resource management — Lexical markup framework (LMF) — Part 5: Lexical base exchange (LBX) serialization: Another XML serialisation. **Status:** *Published in 2022.*

ISO/CD 24613-6 Language resource management — Lexical markup framework (LMF) — Part 6: Syntax and Semantics. **Status:** *A candidate for an ISO Draft International Standard (DIS) ballot.*

3 The New SynSem module

Figure 1 gives the SynSem class diagram. The classes in white (*LexicalEntry*, *Sense* and *SenseRelation*) are inherited from the LMF core (Part 1), while the salmon-pink coloured classes are newly defined in Part 6. Notably, Part 6 introduces two important new classes which provide the means to describe both the *Syntactic Behaviour* of entries and the *Predicative Representation* of senses as well as allowing for the specification of connections between the two. The main difference with respect to ISO 24613:2008 is the absence of the previously defined *Synset* class. Indeed the semantic module of the prior version of LMF contained elements that were entirely dedicated to the modelling of WordNet-like lexicons. However, this was not judged to be necessary in the current standard since the *Sense* and *SenseRelation* classes can be used instead.

Another crucial difference with respect to the former version of LMF is the lack of a *feat* class, formerly used to make up for specific elements which a lexicographer may want to introduce but which were not generic enough to be included in the model. In the old model, class arguments could be specified as pairs of attributes of the specific tag *feat*: *att* would contain the name of the attribute, and *val* the value. In the new model, attributes can be added as needed; in Figure 3 for example a *SemanticArgument* can be specified in terms of *type* and *restriction*. Generally speaking – and here guided by the same principle already introduced for other parts – only the core features of the syntax and semantics interface are described in the present UML based standardisation, however the user can extend the model to add other features.

Regarding the modelling of syntax in Figure 1, a *LexicalEntry* may have one or more instances of *SyntacticBehaviour*, associated with separate *SubcategorizationFrame* instances, each described with *SyntacticArgument*. As for the modelling of semantics, it applies to senses. The *Sense* class, which is specified in the core package, is aggregated in the *LexicalEntry* class. A *PredicativeRepresentation* serves to connect a *Sense* with one or more instances of *SemanticPredicate*, which are described in terms of *SemanticArgument* instances. Linking between syntax and semantics is done by the *SynSemArgMap* component, which links a *SemanticArgument* with a *SyntacticArgument*.

In modelling semantics, allowance is made in

Part 6 for drawing from other relevant standards. In particular **ISO 24617-4:2014 (en) - Language resource management — Semantic annotation framework (SemAF) — Part 4: Semantic roles (SemAF-SR)** provides a background terminology and methodology for designing a semantic role scheme in a coherent way, based upon the work carried out in the LIRICS projects ((Petukhova and Bunt, 2008)). The examples provided in this paper illustrate the use of such roles without providing a normative list thereof.

3.1 Examples

In this section we will illustrate Part 6 in more detail by means of an exhaustive example (Figure 3), drawn from the Parole Simple CLIPS Italian lexicon³. The example contains two lexical entries, the Italian verb *costruire* ('to build') and the deverbal noun *costruzione* ('a building'). For simplicity's sake, in this example each entry has just one sense (though many are possible), each linked to a separate *PredicativeRepresentation*, but these are in turn linked to just one *SemanticPredicate* (PREDcostruire-1). The predicate is described with its two arguments to which are added semantic roles, and restrictions (the latter represented by types in the SIMPLE ontology (Del Gratta et al., 2015)). From the syntactic point of view, a *SyntacticBehaviour* element links the *LexicalEntry* to a *SubcategorizationFrame SCFtxa*, representing the transitive construction, which is in turn described by its two syntactic arguments (subject and object). A *SynSemCorrespondence* component (of type *ISObivalent*) allows for a mapping between each pair of syntactic/semantic arguments. In this rather straightforward case, the subject maps onto the agent and the object onto the patient. Finally a further diagram (Figure 2) illustrates how syntactic alternations can be represented. In the example, which represents the anti-causative syntactic alternation, a *SubcategorizationFrameSet* has been created to connect two *SubcategorizationFrames* that can be subject to alternation, as in the case of the transitive and intransitive in verbs such as *bollire* ('boil')⁴. The *SynArgMap* class can also be used to represent the link between syntactic arguments: in this case the representation tells us that the object in the transitive construction becomes the subject

of the intransitive one.

3.2 Serialisation

We designed the serialisation of ISO LMF 24613-6 as an extension of the TEI guidelines⁵. In doing so, we wanted to achieve the following objectives:

- Maintain coherence with the overall serialisation framework for LMF which has already set out a dedicated TEI subset covering parts 1, 2 and 3 within the ISO LMF 24613-4 standard;
- Benefit from the TEI specification language ODD ("One Document Does it all") which provides a flexible framework compatible with literate programming principles and which allows for the generation of both schemas (DtD, RelaxNG, W3C) and documentation from a single specification document;
- Integrate the specific development of LMF syntax and semantic descriptions within a broader lexicographic landscape in which the TEI guidelines have been widely adopted (also within the framework of the TEI Lex 0 initiative⁶) for maintaining sustainable lexical resources, which are thus FAIR by construction.

More precisely, we integrated SynSem components at three specific places within the standard structure of a TEI lexical entry:

- We added a <syntacticBehaviour> element to the possible grammatical descriptions associated with a lemma (within the TEI <gramGrp> element) that points to a sub-categorisation frame (see below);
- The content of the TEI <sense> element was expanded to contain a <predicativeRepresentation> element with references to a semantic predicate and possible syntactic-semantic correspondences;
- We extended the general intermediate of a TEI document to allow <subcategorizationFrame>, <SemanticPredicate> and <SynSemCorrespondence> elements to occur freely and be referred to from <syntacticBehaviour> and <predicativeRepresentation> within entries.

³<http://hdl.handle.net/20.500.11752/ILC-88>

⁴This example works in English ("I boil the water/The water boils")

⁵<https://tei-c.org/guidelines/>

⁶<https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html>

All content relating to the serialisation of 24613-6 is available from the DARIAH WG on lexical resources⁷.

4 Conclusion

The new ISO LMF 24613-6 will soon be available as a published standard. Resources encoded in the previous model are easily converted to the new one, which remains overall backward compatible. Another crucial task will involve developing user-friendly conversion methodologies for other commonly used formats, particularly OntoLexLemon, by defining convenient crosswalks. This would, among other things, provide an easy way to go from tree based TEI-XML representations to RDF-based graph-like representations, thus potentially contributing to the extension of the Linguistic Linked Open Data Cloud.

Acknowledgements

The work described in this paper was carried out as part of the activities of the CLARIN-IT national consortium; aspects concerning the link between the ISO standard and other formats were also explored as part of the activities of the COST Action NexusLinguarum – “European network for Web-centered linguistic data science” (CA18209), supported by COST (European Cooperation in Science and Technology) www.cost.eu.

References

- Riccardo Del Gratta, Francesca Frontini, Anas Fahad Khan, and Monica Monachini. 2015. **SIMPLE-LOD**. ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa.
- Gil Francopoulo. 2013. *LMF lexical markup framework*. John Wiley & Sons.
- Fahad Khan and Jack Bowers. 2020. Towards a lexical standard for the representation of etymological data. *Convegno annuale dell'Associazione per l'Informatica Umanistica e la Cultura Digitale*.
- Alessandro Lenci, Nuria Bel, Federica Busa, Nicoletta Calzolari, Elisabetta Gola, Monica Monachini, Antoine Ogonowski, Ivonne Peters, Wim Peters, Nilda Ruimy, Marta Villegas, and Antonio Zampolli. **SIMPLE: A General Framework For The Development Of Multilingual Lexicons**. 13(4):249–263.
- Volha Petukhova and Harry Bunt. 2008. **LIRICS semantic role annotation: Design and evaluation of a set of data categories**. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).
- Laurent Romary, Mohamed Khemakhem, Fahad Khan, Jack Bowers, Nicoletta Calzolari, Monte George, Mandy Pet, and Piotr Bański. 2019. Lmf reloaded. *arXiv preprint arXiv:1906.02136*.
- Nilda Ruimy, Ornella Corazzieri, Elisabetta Gola, Antonietta Spanu, Nicoletta Calzolari, and Antonio Zampolli. **LE-PAROLE Project: The Italian Syntactic Lexicon**. In *EURALEX '98*.

⁷GitHub project under <https://github.com/DARIAH-ERIC/lexicalresources/tree/master/Schemas/LMF%20SynSem%20Specification>

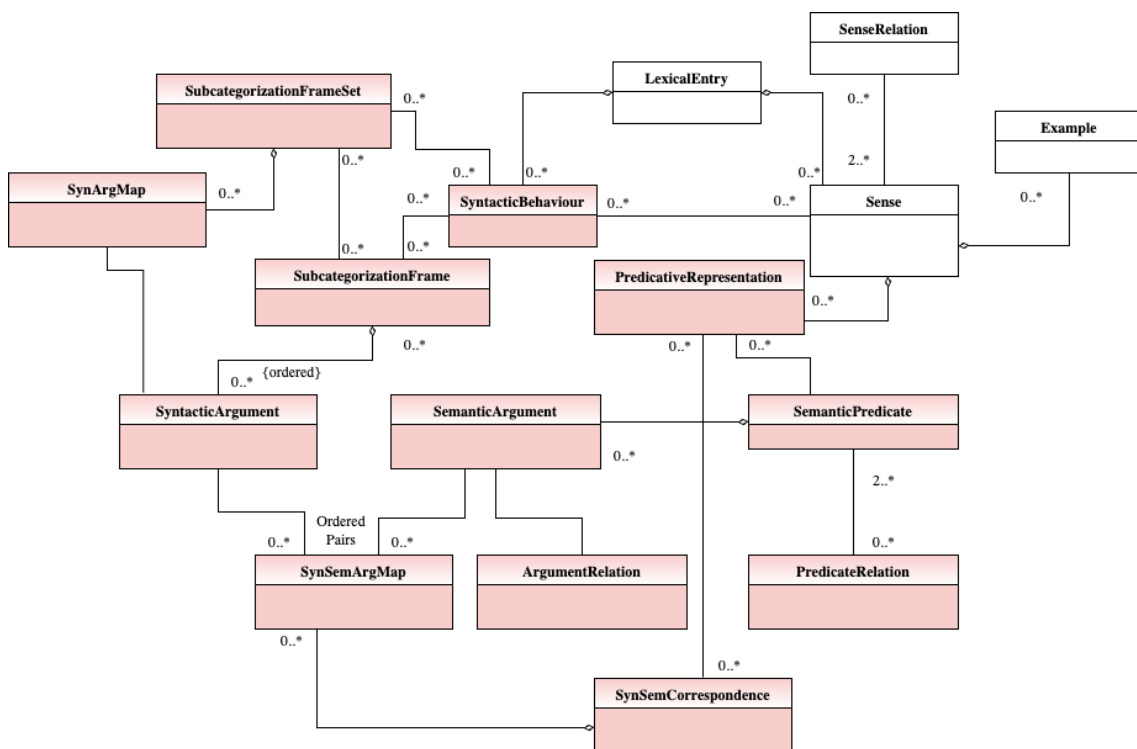


Figure 1: Synsem Module - Class diagram.

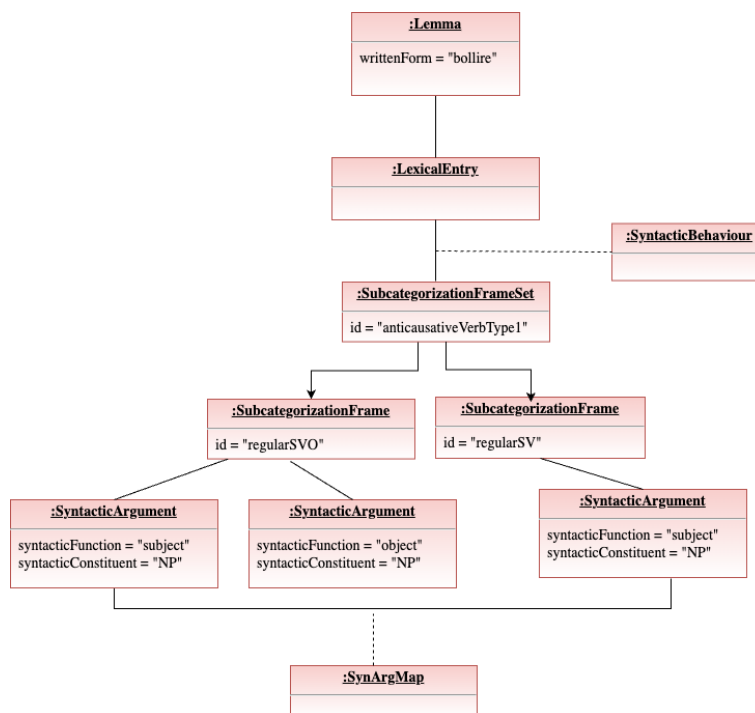


Figure 2: "Bollire" ("boil") syntactic alternation.

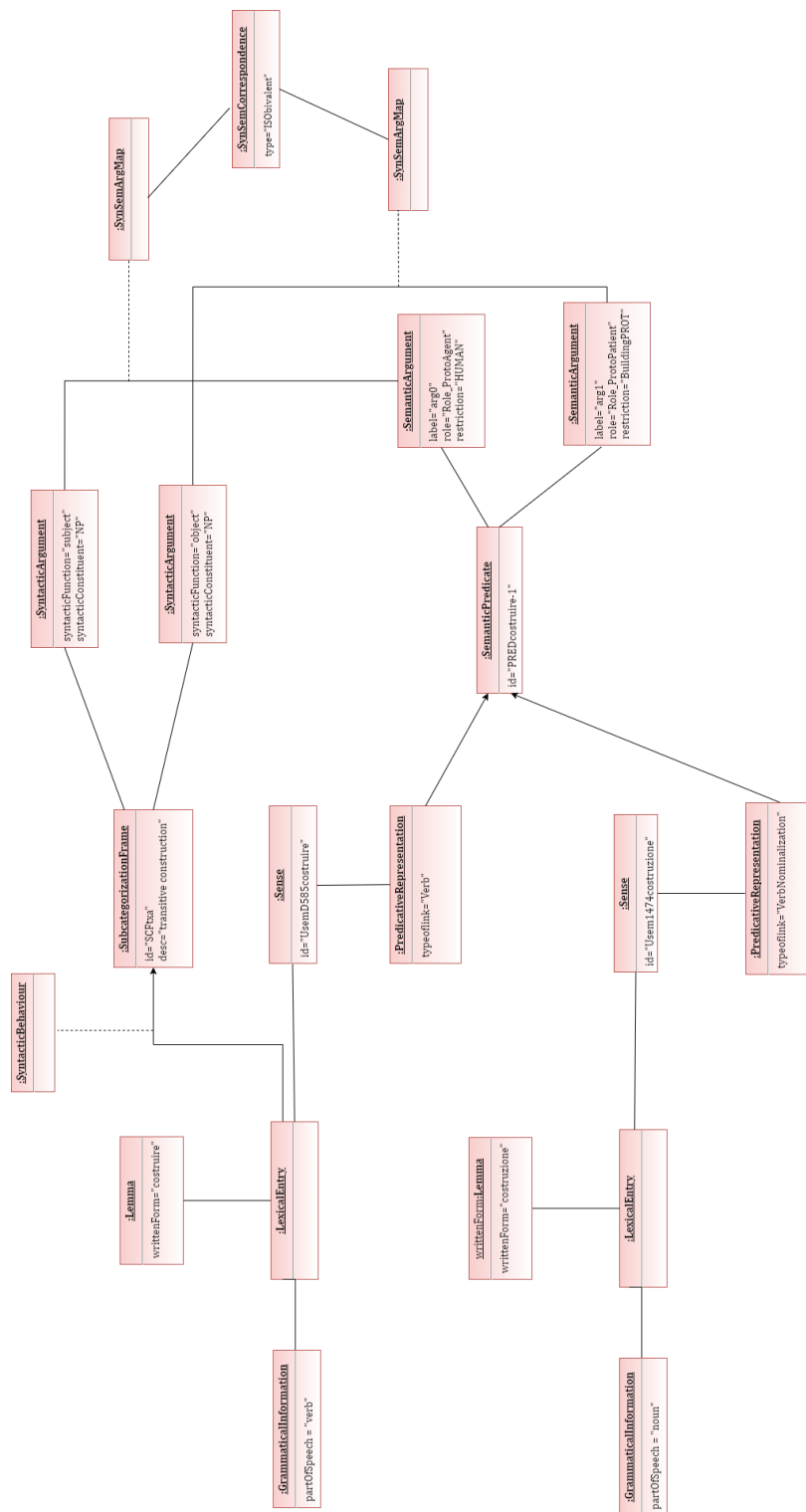


Figure 3: Costuire / costruzione (build/building) in Parole Simple CLIPS.

Word in context task for the Slovene language

Timotej Knez and Slavko Žitnik

University of Ljubljana, Faculty of Computer and Information Science

{timotej.knez, slavko.zitnik}@fri.uni-lj.si

Abstract

In natural language, it is important to understand which meaning of a word is used based on its context. For this reason, a Word in Context task was designed where the model is presented with two sentences containing the same target word. The goal of the model is to recognise if the same sense of the word is used in both sentences. Over the years, many models for solving this task in the English language have been proposed. However, research on the Word-in-Context (WiC) task for the Slovene language has been limited by the lack of annotated data available in the Slovene language. In this paper, we construct a new Slovenian corpus for the WiC task that will enable future research in this area. The constructed corpus is comparable in size to the widely used WiC corpus in the SuperGLUE task. We also perform some tests using simple algorithms to validate the usability of the corpus.

1 Introduction

The Slovenian language, like many other languages, contains numerous words with multiple meanings. For instance, words like "gol" (naked/goal) and "klop" (tick/bench) can have different interpretations in various sentences. The ambiguity of such a word poses a challenge for many NLP tasks, as the models need to recognise the intended meaning based on the context. The goal of the Word-in-Context (WiC) task is to help the embedding models learn to recognise the context and differentiate between different meanings. The task is formulated such that a model receives a pair of sentences that both contain the same target word. The model needs to then recognise whether the same meaning of the two words is used in both sentences. The WiC task is also included in the SuperGLUE benchmark (Wang et al., 2019). Solving this task for the Slovene language is limited by the lack of appropriately annotated datasets containing Slovene sentences. As part of one of the possible

student projects in the natural language processing course at the Faculty for Computer and Information science at the University of Ljubljana, the students annotated a small number of sentences for the WiC task and used them to try and solve the task for the Slovene language. In this paper, we combined their manually annotated sentences into a single dataset that can be used for the Slovene Word in Context task. We also included a larger number of automatically annotated examples to help train models that might require a larger amount of data. We also used a number of simple models for the WiC task to demonstrate the usability of the constructed corpus. We compared the results achieved on our dataset to the results achieved with the same algorithms on the English dataset. We found that our dataset is somewhat more challenging than the English one due to some words with multiple similar meanings. The dataset is published in the Clarin.si repository¹.

2 Related work

The goal of this paper is to enable the Word-in-Context (WiC) task in the Slovene language. The Word-in-Context task was described by Wang et al. (Wang et al., 2019) as part of the SuperGLUE benchmark. The task is defined as a binary classification, where the model is presented with two sentences that contain a common homonym. The goal is for the model to recognise whether the same meaning of the target word is used in both sentences.

2.1 Datasets for the Word-in-Context task

The most commonly used dataset for the Word-in-Context task is the WiC dataset (Pilehvar and Camacho-Collados, 2018), provided by the SuperGLUE benchmark. The dataset contains around 7500 sentence pairs compiled from WordNet, Wiktionary, and VerbNet. Recently a larger version

¹<http://hdl.handle.net/11356/1781>

of the dataset was published under the name XL-WiC (Raganato et al., 2020) which in addition to the English sentence pairs from (Pilehvar and Camacho-Collados, 2018), contains sentences from multiple other languages. The dataset contains training sets in three additional languages (German, French, and Italian) and validation and test sets in 12 additional languages. The goal of the dataset is to support cross-lingual inference. The sentence pairs were extracted from wiktionary and the multilingual WordNet.

A related dataset for the Finnish, Croatian, and Slovene languages was presented by Wand et al. (Armendariz et al., 2019). The dataset is designed for the word similarity in context task where we need to predict the semantic similarity between two different words based on the context presented in two sentences. They constructed the dataset by manually annotating sentence pairs based on how similar the two words are.

2.2 Models for solving the WiC task

El-Gedawy (El-Gedawy, 2013) presented a method for determining the meaning of Arabic words based on their context. They construct a dataset from WordNet. To improve the results, they provide the model with the most frequent words that appear when searching the sentence on Google and Bing search engines. This way the model gets information about the context of the sentence. The classification is performed by computing similarity between observed terms and terms from all word senses. The model manages to achieve an f-score of 80%. They also recognise, that removing stop words increases model performance.

Another approach for the task was proposed by Pal et al. (Pal et al., 2013). They use a model combining the bag-of-words approach with a Modified Lesk algorithm. The bag-of-words model is used to find the meaning of the ambiguous word. They construct a bag for each sense of the word. The sentence with removed stop words is compared to the words in each of the bags to determine the most likely sense. The Modified Lesk algorithm is used to detect word sense without supervision. While on its own it does not provide good performance, it improves the results when used in combination with the bag-of-words approach. The bag-of-words alone achieves 66% F-score, while the addition of the Modified Lesk algorithm improves the F-score to 85%.

Another interesting approach for word sense disambiguation was presented by Chaplot and Salakhutdinov (Chaplot and Salakhutdinov, 2018). The approach detects the topics that appear in the entire text instead of relying solely on the sentence the word is located in. The senses of the words are predicted based on the topics that appear in the document. The topic detection is performed using the Latent Dirichlet Allocation (LDA).

3 Dataset construction

In this section, we present an explanation of our pipeline for constructing a WiC corpus. The corpus was compiled from six student projects, where each group prepared a small dataset for the word-in-context task. As all groups followed a similar methodology, we present the combined process. An overview of the pipeline is depicted in Figure 1.

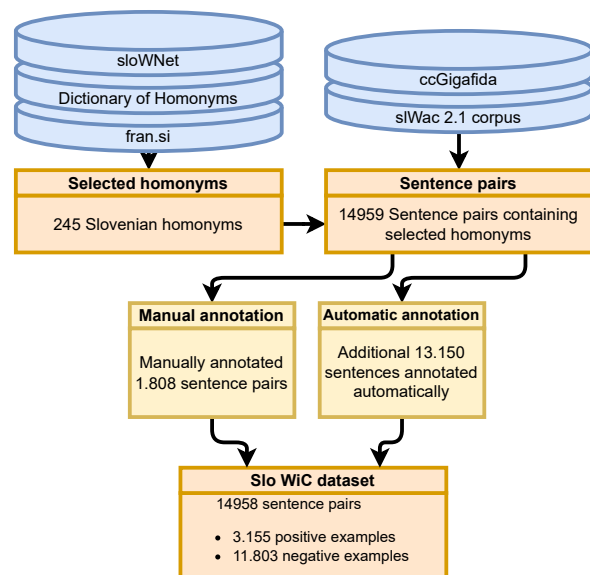


Figure 1: An illustration of the pipeline for constructing the Slovenian WiC dataset.

The first step in constructing the Slovenian corpus for the Word in Context task is to gather a list of homonyms to be included in our corpus. We gathered the homonyms from the Slovene dictionary of Homonyms (Bálint, 1997), Slovene wordnet (Fišer, 2015), and by scraping the Slovene dictionary website *Fran.si*. Once we had the interesting words to include in the dataset, we collected the sentences where the selected homonyms appear in different contexts. The sentences were gathered by searching the ccGigafida corpus (Logar et al., 2013) for the selected homonyms. The ccGigafida is a large corpus of Slovenian text. One group gathered the

sentences from the sIWaC-Slovene web corpus.

Once the sentences were gathered, we need to annotate them to be used as training examples. We used a combination of manual annotations and automatic annotations computed by multiple machine-learning models. The process of manual annotation was performed in a few different ways by different groups. Most of the corpus was annotated by first constructing sentence pairs and manually annotating them with a label that shows whether the target word is used in the same sense in both sentences. On the other hand, one group first annotated a number of sentences with the senses of the target word. After that, they formed pairs of annotated sentences to get combinations of the different senses.

In addition to the manually annotated sentence pairs we also prepared some automatically labelled sentence pairs. The labels for these pairs were computed by clustering the sentences based on multiple algorithms. We used contextualized word embeddings computed by the BERT model, sentence embeddings based on Glove and Word2Vec embeddings, and bags of words. The labels were then determined by observing the similarity between both sentences. This approach produces some errors in annotations. To combat that we discarded the sentence pairs where the similarity scores were close to the threshold and only kept the pairs with very high and very low similarity. We manually analyzed a random sample of the automatically annotated corpus and found that the relations have 76% accuracy.

3.1 Dataset structure

For using the constructed corpus, it is important to understand its structure and parameters. As described in Section 3, a part of the corpus was annotated manually, while the other part contains automatically generated annotations. Altogether there are 7855 sentence pairs annotated manually and 7103 sentence pairs with only automatic annotations. Another important piece of information is how many times the same sentence can occur in the dataset. A large majority of the sentences appear in no more than four different sentence pairs. While some of the sentences appear in multiple sentence pairs, a large majority of the sentences appear in only a single sentence pair. 74% of all sentence pairs in the dataset contains only sentences that do not appear in any other sentence pair.

For training, it is important that the dataset is

Table 1: Comparison of the size of our word in context dataset and the English WiC dataset.

Corpus	Sentence pairs
English WiC - Train	5428
English WiC - Val	638
English WiC - Test	1400
English WiC - Sum	7466
Slo WiC - Manual	1808
Slo WiC - Automatic	13150
Slo WiC - Sum	14959

not too imbalanced. To check that, we analyzed the distribution of both classes. The manually labelled portion of our dataset contains 1200 sentence pairs (66.4%) that have the same meaning in both sentences and 608 sentence pairs (33.6%) with different meanings. In the entire corpus, there are 11803 sentence pairs (78.9%) with the same meaning and 3155 sentence pairs (21.1%) with different meanings. We found that the classes are a bit imbalanced; however, we believe that the level of imbalance is acceptable. Because of the imbalance we used the AUC measure in our tests instead of the classification accuracy.

3.2 Comparison to the WiC dataset

We compare our Slovenian word in context dataset to the widely used English WiC dataset (Pilehvar and Camacho-Collados, 2018). When taking into account all of the annotated sentence pairs in our dataset including the automatically labelled examples, our dataset contains 14959 sentence pairs, which is larger than the English WiC dataset which contains 7466 sentence pairs. However, the automatically labelled examples might not be useful in all use cases as they might contain errors. Because of that the more appropriate comparison would be to observe the manually annotated part of our dataset, which contains 1808 sentence pairs. We present the size comparison of both corpora in Table 1.

Another important metric is the number of homonyms captured in the dataset. The English WiC dataset compares 2345 unique words. While our Slovenian WiC only contains 245 unique homonyms. That is because we include a larger number of sentence pairs for each homonym. We present the number of unique homonyms contained in each part of the two datasets in Table 2.

Table 2: Comparison of the number of homonyms contained in our word in context dataset and the English WiC dataset.

Corpus	Homonyms
English WiC - Train	1265
English WiC - Val	599
English WiC - Test	1184
English WiC - Combined	2345
Slo WiC - Manual	228
Slo WiC - Automatic	240
Slo WiC - Combined	245

4 Word in context models

Once we constructed the Slovenian Word in Context dataset, we can use it to train a WiC model. We constructed several models for solving the Word in Context task.

4.1 Clustering based prediction

The main approach that we used is based on clustering the sentences together. The goal is that we compute a contextual embedding of both sentences that captures the context in which the words are used. After that, we compute the distance between the embeddings to determine if the contexts are similar. For that, we need to determine a threshold similarity value based on the training data. Here we are working under the assumption that when a homonym is used in the same context, its sense will also be the same and vice versa.

For computing the distance between sentence embeddings we used cosine similarity. We tested multiple different methods for generating sentence embeddings to represent the context of each target word. A potential problem with this approach is that the assumption that when the word is used in different contexts its meaning will also be different might not always hold. On the other hand, the approach has a large advantage in that it is unsupervised and only requires training data to determine the similarity threshold.

4.2 Bag-of-words algorithm

To establish a baseline for our results, we utilized the Bag-of-words technique as a basic and straightforward approach. To implement this method, we utilized sentences that had already been stripped of stopwords. We kept track of the words that were in close proximity to the target word and represented them as a single large vector. By tallying the num-

ber of times these words appeared, we generated a vector for each sentence. To determine whether a target word was used similarly in two given sentences, we measured the cosine similarity between their respective vectors and applied a thresholding technique. Our Bag-of-words method takes the following parameters into account:

- Window size: This determines how many adjacent words around the target word will be used as context.
- Cosine distance threshold: If the cosine similarity between two vectors exceeds this predetermined threshold, the pair is deemed to have the same context.

4.3 The Simplified Lesk algorithm

We experimented with a simplified version of the Lesk algorithm as another method for solving the WiC task. For this algorithm, we used the sentences from our dataset with the stopwords removed. The Simplified Lesk algorithm works by comparing the sentence with a sample sentence with a known meaning. For the sample sentences we used the entire Dictionary of Standard Slovene Language (SSKJ) from a Github repository². We computed the overlap between the lemma forms of the words that occurred in the sentences and the words in dictionary glosses of different meanings. During the preprocessing step, we stored the glosses in a dictionary based on the target words for efficient search. We also precomputed the lemmas of the words in glosses so that we could compare them with our sentence pairs. We used the CLASSLA pipeline (Ljubešić and Dobrovoljc, 2019) for extracting the lemma forms of all words used by this algorithm. This approach is especially interesting as it determines the meaning of the target word in each sentence and not only if the words in both sentences have the same meaning.

4.4 Pretrained language models

In recent years, many natural language tasks rely on using large pretrained language models like Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) for computing token embeddings. The main advantage of such models compared to using precomputed token embeddings is that they produce contextualized token embeddings which capture not only the information about

²<https://github.com/van123helsing/SSKJ>

the token but also about its context. Because of this, such models are very useful for differentiating between different meanings of the same word. Once we had the embeddings, we compared them using cosine distance to determine if the words are likely used in the same context. The architecture of the approach is shown in Figure 2

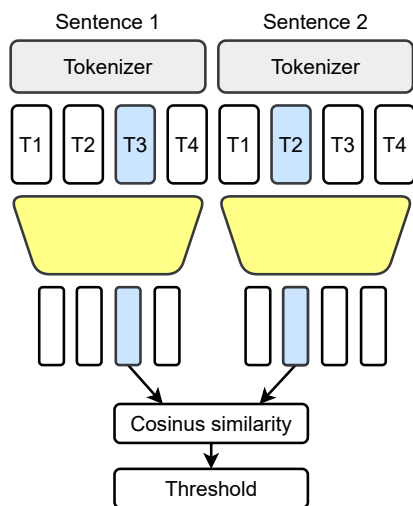


Figure 2: Architecture of the clustering model based on a pretrained language model.

In our tests, we used multiple pretrained BERT networks that are able to analyze Slovene text to produce contextualized embeddings of the target word in each sentence. The first network that we used is the Multilingual BERT model that was trained on 102 languages including Slovene. The second pretrained language model that we used is the CroSloEngual BERT (Ulčar and Robnik-Šikonja, 2020) which was trained on Croatian, Slovene, and English languages. The final pretrained language model that we used is the SloBERTa (Ulčar and Robnik-Šikonja, 2021) which was trained on just Slovene text. The multilingual models here have the advantage of being trained on a larger amount of data; however, that also means that they might not be well fitted to the Slovene language. On the other hand, SloBERTa is well fitted to the Slovene language but was trained on a much smaller corpus.

5 Results

We tested the presented methods for detecting if the same sense of the target word is used in both sentences in a sentence pair. The methods based on cosine similarity provide a score that needs to be compared with a threshold value. Instead of

Table 3: The area under the curve scores of all tested algorithms. We also include scores on the English dataset for the best-performing multilingual approaches for comparison.

Embedding method	Slo AUC	Eng AUC
Random baseline	50%	50%
Bag-of-words	56.1%	
CroSloEngual BERT	68.9%	71.7%
Multilingual BERT	65.6%	68.5%
SloBERTa	55.5%	
Simplified Lex	58.7%	

determining a single threshold value, we decided to evaluate the algorithms by observing the area under the ROC curve as we change the threshold. The curves are shown in Figure 3. The simplified Lesk algorithm provides classifications instead of some likelihood scores that could be compared to the threshold. Because of that, its performance is denoted by an x in Figure 3. We computed the AUC scores of all algorithms and presented them in Table 3. We also tested the best-performing algorithms on the English dataset (Pilehvar and Camacho-Collados, 2018) for comparison.

All of the models were tested on the manually annotated part of the Slovene WiC corpus. We did not use the automatically generated part of the corpus as the proposed models do not benefit from a larger dataset and we wanted the results to be as accurate as possible.

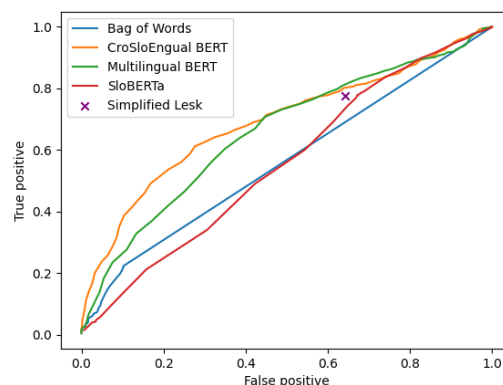


Figure 3: ROC curves of the predictions by the tested algorithms.

We found that the Simplified Lex algorithm achieved similar results as cosine similarity using the BERT embeddings. As expected the bag-of-words algorithm achieved worse results. The results are not directly comparable to the results achieved by previous research as the models were

tested on a different dataset.

5.1 Discussion

When using the clustering models, we are assuming that when two contexts of a word are different, the meaning of the word will be different as well. This assumption is somewhat problematic as the same meaning of a word might be used in multiple different contexts. In this case, the distance between the sentence embeddings might be large even though the meaning of the target word is the same. This aspect is improved by the Lesk algorithm, which compares the sentence to all known meanings of the word, which means that even if the two sentences fall under different clusters, they might get assigned the same meaning.

We also compared the scores achieved on the Slovene dataset to the ones achieved by the same algorithms on the English dataset. We found that the algorithms perform better when used on English data. The reason for this is likely that we included a number of words that have multiple very similar meanings that might be used in the same context. We believe that difficult words like this make the dataset better as they teach the model to differentiate between similar meanings.

Acknowledgment

The work presented in this paper was done as part of student projects in the Natural language processing course at the Faculty of computer and information science at the University of Ljubljana. We combined work done by the following students: Anže Luzar, Anže Tomažin, Blaž Beličič, Marko Ivanovski, Matej Miočič, Matej Kalc, Zala Erič, Miha Debenjak, Denis Derenda Cizel, David Miškič, Kim Ana Badovinac, Sabina Matjašič, Nejc Velikonja, Jure Tič, and Sandra Vizlar.

A part of this research was financially supported by the Slovenian Research Agency in the young researchers grant.

References

- Carlos Santos Armendariz, Matthew Purver, Matej Ulčar, Senja Pollak, Nikola Ljubešič, Marko Robnik-Šikonja, Mark Granroth-Wilding, and Kristiina Vaik. 2019. Cosimlex: A resource for evaluating graded word similarity in context. *arXiv preprint arXiv:1912.05320*.
- Júlia Bálint. 1997. *Slovar slovenskih homonimov: na podlagi gesel Slovarja slovenskega knjižnega jezika*. Znanstveni Institut Filozofske Fakultete.
- Devendra Singh Chaplot and Ruslan Salakhutdinov. 2018. Knowledge-based word sense disambiguation using topic models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Madeeh Nayer El-Gedawy. 2013. Using fuzzifiers to solve word sense ambiguity in arabic language. *International Journal of Computer Applications*, 79(2).
- Darja Fišer. 2015. [Semantic lexicon of slovene sloWNet 3.1](#). Slovenian language resource repository CLARIN.SI.
- Nikola Ljubešič and Kaja Dobrovoljc. 2019. What does neural bring? analysing improvements in morphosyntactic annotation and lemmatisation of slovenian, croatian and serbian. In *Proceedings of the 7th workshop on balto-slavic natural language processing*, pages 29–34.
- Nataša Logar, Tomaž Erjavec, Simon Krek, Miha Grčar, and Peter Holozan. 2013. [Written corpus ccGigafida 1.0](#). Slovenian language resource repository CLARIN.SI.
- Alok Ranjan Pal, Anirban Kundu, Abhay Singh, Raj Shekhar, Kunal Sinha, et al. 2013. An approach to word sense disambiguation combining modified lesk and bag-of-words. *Comput. Sci. Inform. Technol.*, 3:517–524.
- Mohammad Taher Pilehvar and José Camacho-Collados. 2018. Wic: 10,000 example pairs for evaluating context-sensitive representations. *arXiv preprint arXiv:1808.09121*, 6.
- Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. Xl-wic: A multilingual benchmark for evaluating semantic contextualization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. The Association for Computational Linguistics.
- Matej Ulčar and Marko Robnik-Šikonja. 2020. [CroSloEngual BERT 1.1](#). Slovenian language resource repository CLARIN.SI.
- Matej Ulčar and Marko Robnik-Šikonja. 2021. [Slovenian RoBERTa contextual embeddings model: SloBERTa 2.0](#). Slovenian language resource repository CLARIN.SI.
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Large Vocabulary Continuous Speech Recognition for Nepali Language using CNN and Transformer

Shishir Paudel and Bal Krishna Bal and Dhiraj Shrestha

Information and Language Processing Research Lab

Kathmandu University, Dhulikhel, Nepal

shishirpaulofficial@gmail.com

bal@ku.edu.np

dhiraj@ku.edu.np

Abstract

Despite the availability of various algorithms for speech recognition, their performance for low resource languages like Nepali is suboptimal. The Transformer architecture is a state-of-the-art NLP deep learning algorithm that uses self-attention to model temporal context information. Although it has shown promising results for English ASR systems, its performance for Nepali has not been extensively explored. This work implements an end to end CNN-Transformer based ASR system to explore the potential of Transformer for building an ASR for the Nepali language. The study used around 159K datasets extracted from openSLR which was further complemented with original recordings that incorporated sentences representing different tenses, grammatical persons, inflections, direct-indirect speech, level of honorifics, etc to address the grammatical structures of the Nepali language. The end to end CNN-Transformer architecture was trained with varying size of datasets, epochs and parameter tuning. The best resulting model achieved a CER of 11.14%.

1 Introduction

Automatic speech recognition (ASR) systems have gained significant importance in recent years due to its wide range of applications, such as virtual assistants, voice command interfaces, automated customer service systems, transcription services etc. Traditionally, ASR systems were built using separate acoustic, language, and pronunciation modules (Jelinek, 1976) and relied on statistical methods such as Hidden Markov model (HMM) and Gaussian Mixture model (GMM). However, such systems required forcefully aligned data, and had limited ability to model complex phenomena such as coarticulation, speaker variability, context etc., (Rabiner and Juang, 1993). In recent years, ASR systems have shifted towards end to end deep neural network (DNN) models that can directly map

speech signals to text without entailing separate modeling of different linguistic features.

Some of the prominent deep neural architecture that can be used to build ASR systems include Convolution Neural Network (CNN), Recurrent Neural Network (RNN) and Transformer. CNNs are efficient in learning local patterns such as spectral or temporal patterns and are mostly employed to extract non-linear features from audio signals. On the other hand RNNs are used to address the temporal relation using the feedback connections and internal status. A problem with RNNs is that they suffer from vanishing gradient problem. Variants of RNN like Long Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997), Gated Recurrent Unit (GRU) and bidirectional LSTM (BiLSTM) try to alleviate the issue however are slow to train and computationally demanding due to their sequential nature. In addition, they are not able to capture the long term dependencies efficiently as the vanishing gradient problem still persists (Zeyer et al., 2019). These shortcomings are solved by the Transformer that employs a multi-headed attention mechanism to compute self-attention. The self attention in Transformer allows each segment in the input to reference every other in the input to capture the long term dependencies (Vaswani et al., 2017). Further the multi head attention allows multiple self attention to be computed simultaneously on different segments of the input that significantly makes the training faster along with capturing of the context for longer sentences (Chernyshov et al., 2021; Dai et al., 2019; Kleinebrahm et al., 2020).

Despite the evolution of ASR systems and deep neural architectures, research is mainly prioritized for prominent languages like English and Mandarin, while for low resource languages like Nepali, ASR systems haven't been explored to that extent (Banjara et al., 2020). Only a few research materials and ASR products based on Nepali language exist today. Further, research carried out in Nepali

ASR include implementation and study of traditional methods such as HMM, RNN etc., while the implementation of recent architecture such as Transformers is largely missing. An efficient ASR based on the Nepali language can be applied to automate various data input systems in different sectors in Nepal, such as banks, hospitals, governmental offices, etc., that could help reduce errors and increase efficiency, ultimately saving time and improving the quality of service.

Nepali is an Indo-Aryan language spoken by 44.64% of the Nepali population and is written using the Devanagari script, which is phonetic (Bal, 2004; Khanal, 2019). Nepali language incorporates a complex system of noun, adjective and verb inflections. Nouns have a system of gender, case and number. Nouns can be inflected to reflect singular or plural, and can be adjusted to seven cases (Bal, 2004) Adjectives in Nepali occur before the noun they modify and they must correspond to gender, case, and number of the noun. Verbs inflect to show contrasts for the grammatical persons, singular/plural, tenses, gender of a subject, grades of honorifics etc.

In this work, we present an end-to-end CNN-Transformer model for Nepali ASR and study the potential of Transformer for low resource languages with the available Nepali datasets. We incorporate variations in the grammar structures, speaking rate, and accent during the training process to enhance the model's ability to generalize over unseen data. Our study sheds light on the performance of contemporary ASR systems for low resource languages and highlights the potential for further research in this area.

2 Past Work

Nepali speech recognition system is one of the least covered topics considering its essence. However, there exist some significant works carried out in Nepali ASR systems. One of the earliest works include a Nepali ASR proposed by Prajapati et al., 2008 that implemented an Ear model based on the human auditory system. Likewise, a HMM based model that was used for processing the Mel Frequency Cepstral Coefficients (MFCC) features from the audio signals was presented by Gajurel et al., 2017. In recent years, deep learning based ASRs were also researched by several authors. Regmi et al., 2019 presented a Nepali ASR based on CNN, RNN and connectionist temporal

classification (CTC) combination. The model was trained on a 2 hour Nepali speech data, where, the CNN was used for extracting the MFCC features while RNN was used for processing the sequential data after feature extraction and CTC for decoding. A total of 67 Nepali characters were used to decode the final text and the model provided a Character Error Rate (CER) of 52% on test data.

Similarly, Banjara et al., 2020 also experimented the combination of CNN, with various RNNs on Nepali dataset. However, compared to Regmi et al., 2019, a larger dataset corpora was used which was collected from OpenSLR¹ namely slr43 and slr54 consisting of 158,113 Nepali utterances. From their experiment, the best resulting CNN-GRU-CTC model achieved a CER of 23.72% which is almost half compared to the prior. Their result showed that GRU performs better than normal RNN due to the reduced severity of issues like vanishing gradient in GRU, while also highlighting the significance of a larger dataset in improving RNNs' performance. Likewise, an end to end CNN-BiLSTM-CTC architecture based model was presented by Regmi and Bal, 2021. The authors also used the slr43, slr54 Nepali data corpus which are openly accessible from OpenSLR for training the model. A total of 129 Nepali characters were used for decoding. Their BiLSTM based model provided a CER of 10.3% on test data. The authors also reported that the training for 20 epochs of the CNN-BiLSTM-CTC model required around 8 days.

From the literature study, we found that all the existing researches in Nepali speech recognition have predominantly used traditional statistical methods such as HMM and deep neural networks such as RNN and its variants like LSTM, GRU and BiLSTM. Remarkably, none of these studies have utilized the Transformer model since it is a recent deep neural architecture. Therefore, in this study we aim to explore the possibilities of Transformer model in Nepali language speech recognition by implementing an end to end CNN-Transformer architecture and compare with the existing DNN implementations.

3 Datasets

Nepali speech datasets are not abundantly available. We collected two freely accessible Nepali speech data set corpora namely "SLR43" and "SLR54" provided by the openSLR.org. The first corpus con-

¹<https://openslr.org/>

sisted of 2064 utterances collected from 18 female speakers that were mostly of longer length sentences while the second consisted of 157k speech data with mostly shorter length sentences. Further, 6031 original recordings corpus generated as a part of this study named Nep_DS were also added to the collected datasets. Nep_DS consists of various Nepali phrases scraped from Nepali language based websites such as ekantipur², setopati³, hamro patro⁴, "Nepali Me" etc., and also several sentences from English websites like BBC translated to Nepali using the google translate api⁵. In addition, we also added several Nepali sentences with varying lengths that address grammatical structures in Nepali language such as different tenses, inflections, grammatical persons, direct and indirect speech, honorifics, etc. The sentences were checked for errors and recorded using Samsung M51 mobile phone, involving 5 speakers. Subsequently, the collected datasets were preprocessed that involved first converting the audio files from .flac to .wav, followed by downsampling the audio from 48 Khz to 16 Khz. The purpose of this pre-processing step was to minimize the computational cost during training. Furthermore, the vocabulary was generated consisting of a total of 119 unique Nepali characters along with 8 additional characters extracted from the text of datasets. The characters were then indexed from 0 to 126, which is shown in Figure 1.

0	-	19	ए	38	ढ	57	व	76	ॆ	95	ज	114	८
1	#	20	ऐ	39	ण	58	श	77	ॆ	96	ड	115	९
2	<	21	आ	40	त	59	ष	78	ॆ	97	ढ	116	०
3	>	22	आ	41	थ	60	स	79	ॆ	98	फ	117	१
4	@	23	आ	42	द	61	ह	80	ॆ	99	स	118	२
5	!	24	आ	43	ध	62	ॆ	81	ॆ	100	ऋ	119	३
6	"	25	क	44	न	63	ा	82	ॆ	101	ॆ	120	४
7	;	26	ख	45	न	64	ा	83	ॆ	102	ॆ	121	५
8	:	27	ग	46	प	65	ऽ	84	ॆ	103	ॆ	122	६
9	?	28	घ	47	फ	66	ॆ	85	ॆ	104	ॆ	123	७
10	!	29	ड	48	ब	67	ॆ	86	ॆ	105	ॆ	124	८
11	!	30	च	49	भ	68	ी	87	ॆ	106	ॆ	125	९
12	!	31	छ	50	म	69	ी	88	ॆ	107	ॆ	126	०
13	!	32	ज	51	य	70	ी	89	ॆ	108	ॆ	127	१
14	!	33	झ	52	र	71	ी	90	ॆ	109	ॆ	128	२
15	!	34	ञ	53	ॆ	72	ी	91	ॆ	110	ॆ	129	३
16	!	35	ट	54	ल	73	ी	92	ॆ	111	ॆ	130	४
17	!	36	ठ	55	ळ	74	ी	93	ॆ	112	ॆ	131	५
18	!	37	ड	56	ळ	75	ी	94	ॆ	113	ॆ	132	६

Figure 1: Nepali Characters used in proposed ASR

4 Architecture of CNN-Transformer

The proposed Nepali ASR system includes an end-to-end CNN-Transformer architecture as illustrated

²<https://ekantipur.com/>
³<https://www.setopati.com/>
⁴<https://www.hamropatro.com/news>
⁵<https://pypi.org/project/googletrans/>

in Figure 2. At first, audio is transformed into spectrogram using short time fourier transform (STFT). The CNN then processes the spectrogram frames to extract high-level spectral features and these extracted audio feature maps are passed to Transformer. The encoder receives a sequence of feature vectors produced by the CNN and transforms it into a fixed-length vector representation. This is accomplished through a series of self-attention and feed forward layers. Self-attention allows the encoder to attend to different parts of the input sequence, depending on their relevance for the current context (Zeyer et al., 2019). The self-attention mechanism is applied multiple times, with each layer building on the output of the previous layer. The output of the final layer is a fixed length vector representation that summarizes the most important information in the input sequence. This vector representation is fed into the decoder to generate the corresponding text output. The decoder uses self-attention to attend to the previously generated output characters while incorporating information from the input sequence using encoder-decoder attention and generates raw discrete representation. The softmax function in the decoder transforms the raw output discrete vectors into a probability distribution over the 128 Nepali output characters. The character with highest probability is given as text output.

During training, the masking mechanism of the Transformer ensures that only relevant parts of the input sequence are attended. Likewise, masking also prevents the model from attending to future tokens during training, ultimately preventing the model from overfitting (Vaswani et al., 2017). Overall, the combination of the CNN and transformer allows the ASR system to effectively capture both low-level spectral features and high-level temporal dependencies in the input audio signal, which is important for accurate speech recognition.

For the implementation, we have used 3 stacks of 1-D CNNs with each having 64 hidden layers, 11 filter size. The opt for 1 D CNN is to minimize the computation cost, and to handle data acquired from varying sources (Kiranyaz et al., 2021) Likewise, the employed transformer consists of encoder and decoder layer as the one suggested by (Vaswani et al., 2017) while the parameters of the transformer are varied in the experiment to optimize its performance for Nepali dataset. The CNN and Transformer were implemented in python language using the Keras library over TensorFlow platform.

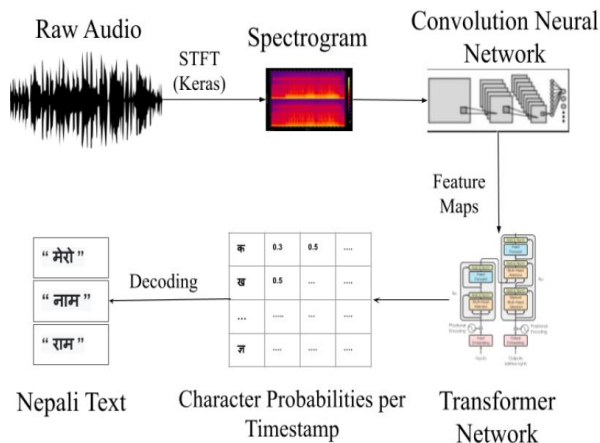


Figure 2: Architecture of the proposed CNN-Transformer ASR

5 Experiments

5.1 Experimental Setup

A total of 14 model training experiments were conducted in two sets to test the potential of the Transformer for recognizing the Nepali speech. The first experiment set involved training the Transformer model on three different Nepali speech datasets: "SLR43", "SLR54", and "Nep_Ds" keeping the training as well as Transformer's parameter values consistent to 200 hidden layers, 2 attention head, 400 FFN, 4 encoders, and single decoder while learning rate was kept 0.001. Different combinations of these datasets were used in the experiments, and alterations were made to the data split ratio and batch size of the training and validation data. The best resulting configuration from the first set was used in the second set, where additional alterations were made to the training parameter i.e., learning rate and Transformer parameters i.e., numbers of attention head, encoder, hidden layer and feed forward network (FFN). The experiments were carried out on two different machines: Machine 1, which had an Intel i9 processor, RTX 3080Ti GPU, and 32 GB of RAM, and Machine 2, which had an RTX 2060 GPU with other specifications remaining the same. Each trained model was evaluated using CER (Character Error Rate) to analyze the best configuration.

5.2 Experimental Result

In the first experiment set, when the model was trained with a smaller dataset i.e. "SLR43" for 105 epochs, the model overfitted. For training with larger dataset, we introduced early stopping and saving with checkpoints in order to stop the

training upon no progress and retain the model with best accuracy. The model performed well when the larger dataset was used i.e. "slr53". For "Nep_DS" as well the model produced a satisfactory result on unseen data. Moreover, the best result was achieved when the all the three corpora i.e: "SLR43", "SLR54" and "Nep_DS" were combined with data split ratio kept at 90:10 rather than 80:20 and batch size kept at 64/4 where the obtained CER was 13.97%. This shows that the CNN-Transformer performs better when the dataset has a higher number of examples for training. Furthermore, the model training speed increased when the batch size of the training data was increased although the performance of the model did not improve. The results from the first experiment set are summarized in Table 1.

In the second experiment set, the learning rate (LR) was altered from 0.001 to 0.0095 and then to 0.00001, while the transformer's parameters such as number of attention heads was altered between 2, 4 and 8. Similarly the number of encoders was increased from 4 to 8 and the number of FFN was altered from 400 to 800. After 6 different training sessions with such variations in parameters of training and Transformer we found that the model was able to achieve the least CER value i.e: 11.14% when the learning rate was 0.001 and the attention head was increased from 2 to 4, while no progress was seen when changing other parameters. Besides, the training with all three corpora merged together (166K datasets) required around 72 hours for 105 epochs on a RTX 2060 GPU based system while it only took around 12.5 hours on RTX 3080Ti based system. The results from the second experiment set are shown in Table 2.

Some of the predictions outputted by the best resulting model on the sample test data is presented in Table 3 which reveals that the model was accurate in most transcriptions. While the majority of the predictions were accurate, a few minor errors were observed, specifically in outputting the corresponding word for numeric utterances. For instance, the numeric sound "२००६" (English translation: "2006") was predicted as "दुई हजार छ" (English translation: "Two thousand and six"). Similarly, the word "आकाशवाणीबाट" was predicted as "आकाशवाणी बाट". Nevertheless, it should be noted that these errors can be neglected as the pronunciation in the predictions precisely matches the reference in both cases.

Expt.	Data	Data Split	Batch Size	Train Data	Test Data	Avg_Epoch_Time(sec)		CER
						Machine1	Machine2	
Ex 1	SLR43	80:20	64/4	1651	413	5.23	39.43	86.98%
Ex 2	SLR54	80:20	64/4	126324	31581	358.75	2337.86	22.77%
Ex 3	Nep_DS	80:20	64/4	4825	1206	12.54	98.7	47.38%
Ex 4	SLR43+54	80:20	64/4	127975	31994	362.36	2440.76	16.57%
Ex 5	SLR43+54+Nep_DS	80:20	64/4	132800	33200	377.51	2658.34	14.74%
Ex 6	SLR43+54+Nep_DS	90:10	64/4	149400	16600	385.02	2586.12	13.97%
Ex 7	SLR43+54+Nep_DS	90:10	128/32	149400	16600	336.44	2498.12	15.34%
Ex 8	SLR43+54+Nep_DS	90:10	256/64	149400	16600	348.37	2434.46	15.89%

Table 1: CNN-Transformer performance results from first experiment set on different datasets, data split ratio and batch size.

Expt.	Attention Head	Encoders	Hidden Layer	FFN	LR	Avg_Epoch_Time(sec)		CER
						Machine1	Machine2	
Ex 9	2	4	200	400	0.00095	417.62	2550.3	15.66%
Ex 10	2	4	200	400	0.00001	364.34	2431.54	16.35%
Ex 11	4	4	200	400	0.001	432.01	2464.23	11.14%
Ex 12	8	4	200	400	0.001	604.68	2888.57	15.71%
Ex 13	4	8	200	400	0.001	533.54	2449.08	16.53%
Ex 14	4	4	200	800	0.001	464.38	2454.29	13.74%

Table 2: CNN-Transformer performance results from second experiment set on various parameter tunings

S.No	Reference	Prediction
1	सुमात्राको टापुमा रहेको इन्डोनेसियाली राष्ट्रिय निकुञ्ज	सुमात्राको टापुमा रहेको तीन इन्डोनेसियाली राष्ट्रिय निकुञ्ज
2	पञ्चमी शब्दले दुई वटा कुरा जनाउँछ	पञ्चमी शब्दले दुईवटा कुरा जनाउँछ
3	२००६ मा उनले	दुई हजार छ मा उनले
4	संसारको पाँचौं अग्लो हिमाल मनासलु यही क्षेत्रमा पर्छ	संसारको पाँचौं अग्लो हिमाल मनासलु यही क्षेत्रमा पर्छ
5	गीतहरूलाई आकाशवाणीबाट प्रसारित	गीतहरूलाई आकाशवाणी बाट प्रसारित

Table 3: Model’s predictions on sample test data

6 Discussions

After several experiments and parameter tunings, the proposed CNN-Transformer achieved a CER value of 11.14% for a combined SLR 43, SLR54 and Nep_DS dataset. Table 4 presents the comparison of our model with other deep learning architecture based Nepali speech recognition systems available in the previous literature. In the previous researches, CNN-RNN-CTC implemented by Regmi et al., 2019 achieved a CER of 52% for

a small dataset while similar architecture implemented by Banjara et al., 2020 achieved a CER of 23.72% for a larger dataset with around 159K utterances. Similarly, BiLSTM-CTC based model implemented by Regmi and Bal, 2021 provided a CER of 10.3% for the same dataset used by Banjara et al., 2020. From the comparison, it is evident that the CNN-Transformer model proposed in our study outperforms most of the past CNN-RNN-CTC based implementations in terms of CER when trained on a large dataset. Besides, the performance of our model is slightly lower but comparable to the best CER value from the previous researches which was achieved by Regmi and Bal, 2021 with the similar size of dataset using CNN-BiLSTM. Nevertheless, our proposed CNN-Transformer model required only about 14 hour for 20 epochs of training on RTX 2060 GPU which is almost 14 times less than the reported training time for CNN-BiLSTM model presented by Regmi and Bal, 2021 which required 8 days for 20 epochs on RTX 2060 GPU when trained with similar size dataset. As a whole, it can be revealed that Transformer has the ability to recognise Nepali speech as accurately as other state of the art RNN based implementations, while the training time it takes is exceptionally less than RNN and its variants.

Papers	Model	Dataset	Dataset Size	CER(%)
Regmi et al., 2019	CNN-RNN-CTC		2 Hours	52
Banjara et al., 2020	CNN-RNN-CTC	SLR 43+54	159K	23.72
Regmi and Bal, 2021	BiLSTM-CTC	SLR 43+54	159K	10.3
This study	CNN-Transformer	SLR 43+54 +Nep_DS	166K	11.14

Table 4: Comparison of the proposed CNN-Transformer model with other deep neural based Nepali ASR

7 Conclusion

In conclusion, this study explored various algorithms used in Nepali ASR. Further, we implemented the Transformer architecture in combination with CNN to build an ASR for Nepali language. Various experiments were conducted to analyze the performance of the CNN-Transformer model on different Nepali datasets with several parameter tunings. The training and validation datasets were extracted from openSLR and augmented with 6031 original speech recordings developed for this study named "Nep_DS". The best resulting CNN-Transformer model obtained an accuracy of 11.14% CER on test data, outperforming many RNN based Nepali ASR in terms of both accuracy and training speed.

Data Availability

The "Nep_DS" corpus generated in this study will be made publicly available at <https://ilprl.ku.edu.np/> upon the publication of this work.

References

- B. K. Bal. 2004. [Structure of nepali grammar](#). *PAN Localization Working Papers*, pages 332–396.
- Janardan Banjara, Kaushal Raj Mishra, Jayshree Rathi, Karuna Karki, and Subarna Shakya. 2020. [Nepali speech recognition using cnn and sequence models](#). In *2020 IEEE International Conference on Machine Learning and Applied Network Technologies (ICMLANT)*, pages 180–185. IEEE.
- Artem Chernyshov, Valentin Klimov, Anita Balandina, and Boris Shchukin. 2021. [The application of transformer model architecture for the dependency parsing task](#). *Procedia Computer Science*, 190:142–145.
- Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. [Transformer-xl: Attentive language models beyond a fixed-length context](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988.
- Avaas Gajurel, Manish K. Ssarma, Anup Pokhrel, and Basanta Joshi. 2017. [Hmm based isolated word nepali speech recognition](#). In *2017 International Conference on Machine Learning and Cybernetics*, volume 1, pages 71–76.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. [Long short-term memory](#). *Neural computation*, 9(8):1735–1780.
- Frederick Jelinek. 1976. [Continuous speech recognition by statistical methods](#). *Proceedings of the IEEE*, 64(4):532–556.
- Rajendra Khanal. 2019. [Linguistic geography of nepalese languages](#). *The Third Pole: Journal of Geography Education*, 18:45–54.
- Serkan Kiranyaz, Onur Avci, Osama Abdeljaber, Turker Ince, Moncef Gabbouj, and Daniel Inman. 2021. [1d convolutional neural networks and applications: A survey](#). *Mechanical Systems and Signal Processing*, 151.
- Max Kleinebrahm, Jacopo Torriti, Russell McKenna, Alessandro Ardone, and Wolf Fichtner. 2020. [Using neural networks to model long-term dependencies in occupancy behavior](#). *Working Paper Series in Production and Energy*.
- C. Prajapati, J. Nyoupane, J. Shrestha, and S. Jha. 2008. [Nepali speech recognition](#).
- Lawrence Rabiner and Biing-Hwang Juang. 1993. *Fundamentals of speech recognition*. Prentice-Hall, Englewood Cliffs, NJ.
- Paribes Regmi, Arjun Dahal, and Basanta Joshi. 2019. [Nepali speech recognition using rnn-ctc model](#). *International Journal of Computer Applications*, 178(31):1–6.
- Sunil Regmi and Bal K. Bal. 2021. [An end-to-end speech recognition for the nepali language](#). In *Proceedings of the 18th International Conference on Natural Language Processing*, pages 180–185.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010.
- Albert Zeyer, Parnia Bahar, Kazuki Irie, Ralf Schlüter, and Hermann Ney. 2019. [A comparison of transformer and lstm encoder decoder models for asr](#). In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 8–15. IEEE.

Knowledge Storage Ecosystem: an Open Source Tool for NLP Results Management (Documents and Semantic Information)

Julian Moreno-Schneider and Maria Gonzalez Garcia and Georg Rehm

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Alt-Moabit 91c, 10559

Berlin, Germany

julian.moreno_schneider@dfki.de

Abstract

This paper presents the Knowledge Storage Ecosystem (KSE), a tool developed for the support of storage and management of knowledge, particularly linked data. The KSE can manage not only knowledge (the semantic information that is extracted from documents using different NLP procedures), but also original documents and full text indexes, allowing full text search in an efficient way, increasing the usability of extracted knowledge in a wide variety of applications. A graphical user interface has also been developed to facilitate the usability of the KSE, allowing this tool to reach a larger audience.

1 Introduction

The development of various NLP technologies over the last decades has resulted in a wide variety of tools, services and libraries to analyze texts, thus being able to generate an enormous amount of semantic information contained in these texts. Handling all this semantic information in knowledge bases has seen a surge in popularity in recent years, because this structured way of storing information enables inference and reasoning.

The widespread use of knowledge bases has fostered the development of tools or platforms that allow the storage and management of this type of semantic information (see Section 2). The main problem we encountered is that there is no platform that allows knowledge management in addition to the original documents on which NLP processes are carried out.

In this article we present a tool that allows the management and use of knowledge as well as documents, facilitating the joint management of these two modes of conveying information. This idea is not completely new, since the World Wide Web Consortium¹ (W3C) already defined this type of

systems under the concept of the Linked Data Platform (Arwe et al., 2015). This concept only encompasses the operation rules, not stating anything about the information stored in such a system. Therefore, we go one step further by labeling the original documents as first class citizens inside our platform. The main problems that we have found in similar systems and that we are trying to solve with this platform are: (i) joint management of documents and related knowledge (especially semantic annotations); and (ii) synchronization of the stored information on CRUD (Create, Retrieve, Update and Delete) operations. In summary, the main contributions of this article are the following:

1. We have defined and implemented a platform, namely Knowledge Storage Ecosystem (KSE), that allows the joint management of knowledge, source documents and full text indexes.
2. We have designed and started the implementation of a graphical user interface that simplifies the management and usage of KSE.
3. We released the entire code of our tool (see Section 3).

2 Similar Systems

The management of semantic information (NLP annotation results) has been covered by many approaches from different perspectives. Some are more focused on the storage of linked data, platforms adhering to the Linked Data Platform standard, or combined systems including file storage or full texts. Many different tools that can be used to manage and store linked data have been developed, summarized in surveys such as those by (Zhang et al., 2021) and (Wylot et al., 2018). Platforms particularly focused on linked data are less abundant, but some alternatives exist. One example is Apache Marmota². It is composed of

¹<https://www.w3.org>

²<https://marmotta.apache.org/index.html>

several modules (for example, SPARQL module, LDP module, Reasoner module or security module among others), but apart from that, the project also develops some libraries that can be used separately such as KiWi Triple Store, LDClient or LDCache. OpenLink Virtuoso (Open-Source Edition)³ is another tool that combines Relational, Graph, and Document Data management. In many cases, Linked Data Platforms have been developed to match a specific use case or domain, such as SeCold (Keivanloo et al., 2012), an open platform for sharing software datasets; QuerioCity (Lopez et al., 2012), a platform to manage (catalog, index and query) heterogeneous information (special interest on stream integration) coming from cities; a platform that combines unstructured data from scientific literature and structured data from publicly available biological databases (Singh et al., 2020); or LinkedLab (Darari and Manurung, 2011), a Linked Data based solution for data management regarding research communities. A tool similar to ours is Trellis-LDP⁴, a platform for building linked data applications that allows storage and management of linked data and documents, but the formats of documents is rather limited, and they do not include full text search as a feature. The main issue we have with Trellis is that it does not control duplicate documents. KIM (Popov et al., 2003) provides exactly the same functionality as our system (based on GATE⁵, RDF Sesame⁶ and Lucene⁷), even integrating the information extraction. Its issues as we perceived them are that it does not store the source documents, and it is a commercial product (only freely available for research). To the best of our knowledge, there is no open-source alternative that provides the functionalities that our system is offering.

3 Knowledge Storage Ecosystem

In this article we have designed and developed a tool that allows the management of semantic information together with source documents. This tool is called Knowledge Storage Ecosystem (KSE) and its main functionality is the management of different types of information (knowledge, source documents, full text indexes) that are related and interconnected between them.

³<https://vos.openlinksw.com/owiki/wiki/VOS/>

⁴<https://www.trellisldp.org>

⁵<https://gate.ac.uk/>

⁶<https://metacpan.org/pod/RDF::Sesame::Repository>

⁷<https://lucene.apache.org/>

The architecture of KSE (shown in Figure 1) is modular and composed of four components, apart from the graphical user interface, that is considered an external extension to the KSE.

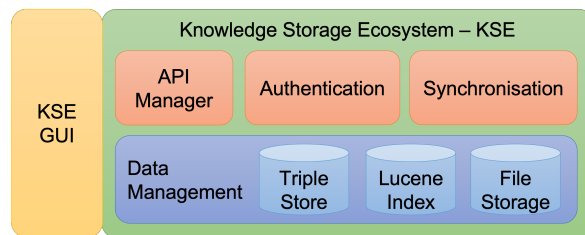


Figure 1: Architecture of the Knowledge Storage Ecosystem

With this first modular architecture of the KSE we cover the following requirements: (i) the storage of semantic information (knowledge) in a specific tool, namely triple store (see Section 3.1.2) allowing inference over the semantic information; (ii) indexing of full text using Lucene (see Section 3.1.3) to simplify search in source documents; (iii) handling of source documents (see Section 3.1.4) and linking them with semantic information through the document identifier in the triple store; and (iv) a first attempt to handle the synchronization of information inside the tool between information types (see Section 3.2).

The entire code, technical documentation and usage examples of KSE are available at <https://gitlab.com/speaker-projekt/knowledge-management/knowledge-storage-ecosystem>.

3.1 Data Management

The first and most important component of the KSE is the data management module, whose main functionality is the management (storage, recovery, modification and deletion) of information inside the system. The information stored in this system is organized in three different categories: source files or documents (PDF, DOCX, TXT, etc.), semantic information (knowledge as Linked Data) associated with the source document and full text obtained from the source document. For each category, the KSE has a specific information storage module, as described below.

3.1.1 Data Structures

The management of the information inside of KSE is made through specific data structures, that we have defined for this purpose. The

most relevant structures defined are `Collection`, `LDDocument` and `Triple`, as well as a `Converter` that allows us to convert these structures to files. `Collection` is a simple structure that has been defined to manage the set of documents that are grouped under the same collection. A collection consists of a *collection identifier*, a *name*, a *description*, and a *list of documents*. `LDDocument` is a more complex structure, because it has to group the three types of information related to a source: knowledge, source document and full text. A document is composed of the following variables: *document identifier*, *text*, a *list of triples* and *path* of the source document. `Triple` is a simple structure, and it is a set of three elements (subject, predicate and object) of a relationship or basic semantic unit. This structure has been defined to facilitate its internal management in the system. The `Converter` is responsible for (de)serializing data structures in/from files, so that they can be included in the KSE or exported from the KSE. It supports standardized semantic web formats such as RDF, TURTLE or JSON-LD.

When a document is created in the system (by uploading it via the REST API (Richardson and Ruby, 2007)), the system assigns it a unique identifier. This identifier is obtained from an encryption algorithm applied to the text of the document. The algorithm used is SHA-256 Cryptographic Hash Algorithm (Handschuh, 2011). There is a possibility that the text of the document is not provided by the user who adds it to the system, in which case an identifier is generated based on the timestamp in which the document was added. We are currently working on improving this process to use the binary content of the original document, thus being able to manage duplicates on the platform, referring to the same document and not generating a new one, as is the case with some alternatives.

3.1.2 Semantic Information Storage

The semantic information storage, or triple store, is a module that is responsible for the efficient management of knowledge (semantic information). There are many tools that are already implemented for performing this task, therefore we decided not to reinvent the wheel and use one of the available options.

We decided to use OpenLink Virtuoso (Open-Source Edition) because we already used it in several projects and the learning curve was shorter. Besides, Virtuoso offers the possibility to easily

install as an independent module and use it through socket calls, which minimizes the potential of interconnection problem within modules.

In order to perform the CRUD operations with Virtuoso, we have defined specific SPARQL queries. Due to space limitation we only show one document creation example in Listing 1.

```
sparql insert into graph <col_1> {
  docURI sp_ont:documentId
  "docId" .
  <subject> <predicate>
  <object> .
}
```

Listing 1: Example of SPARQL query for creating a document in Virtuoso.

3.1.3 Full Text Index

This module allows the search for textual information in documents in an extremely efficient way, something that is supported in triple stores, but is inefficient if text gets longer. Therefore, we are using the well-known and extensively used and tested Lucene⁸ (McCandless et al., 2010) tool. This is the basic Apache technology for full text search. Although in last years newer technologies have been developed (such as Solr or Elasticsearch), which include much more functionality, we decided to stay with the most basic technology in order to keep it simple and easy to use and integrate in our tool. Besides, the direct usage of Lucene allows us to redefine any component that we need, for example, the Document Parsers needed for the specific `LDDocument` structure.

We have defined a simple index containing three fields: *identifier* inside the Lucene index, *KSE document identifier* and *full text*. At indexing we use two different analysers to process these fields: A `Whitespace analyser` for the identifiers, and an `N-Gram Analyser` for the text. The `N-Gram analyser` converts the text in n-grams ($n = 3$) in order to index them as the minimal textual unit.

3.1.4 File Storage

This module is responsible for storing the original files within the platform. To do this, and in order to implement the module as simply as possible, we have used the file system. Original files are stored as files in a folder that is identified by the name of the collection the files belong to, for example, if we upload a file called *'Report.pdf'* and add it to the *'shared_documents'* collection, then the file system will be as shown in Listing 2.

⁸<https://lucene.apache.org>


```
kse_collections /
\--- shared_documents
    \--- Report.pdf
1 directory, 1 file
```

Listing 2: Folder structure of the file storage after including a file named 'Report.pdf' to the collection 'shared_documents'.

The main functionality of this module is to keep accessible the original documents on which the NLP analyses are performed. In this way one can reproduce experiments or display results directly on the source documents, for example, integrating entity highlights in PDFs.

3.2 Synchronization Module

The synchronization of the information is essential in our system, because when integrating other tools (Lucene, Virtuoso, etc.), it may happen that the semantic information related to a document is modified, while this document is included in the result of a textual search. Or even worse, that the document is deleted, but continues to be used in searches or statistics until it is permanently deleted from all tools. For this we have defined a synchronization mechanism that prohibits or blocks the use of a document if it is being used by some modification operation (update or delete). For this we use the synchronization mechanisms of Java (through three methods: `documentIsBlocked(docId) {...}`, `blockDocument(docId) {...}` and `unblockDocument(docId) {...}`), together with a `HashMap` that stores the identifiers of all documents stored in the system (`HashMap blockedDocuments`).

3.3 API Manager

This module is responsible for the access to the entire tool functionality, from administrative control to information management through HTTP REST API endpoints.

The administrative control of the tool is done through configuration files, which are included directly in the source code including examples (available [here](#)). Nevertheless, we have included endpoints to manage these configuration parameters, being able to create, read, modify or delete them. All the endpoints defined for administrative tasks are listed in Figure 2.

The information management is completely done through endpoints that are accessible through

management-api	
POST	/kse/mngt/addProperty
GET	/kse/mngt/listProperties
DELETE	/kse/mngt/deleteProperties
PUT	/kse/mngt/modifyProperties

Figure 2: Administration endpoints.

HTTP REST API, and are divided into two categories: endpoints for CRUD operations (Create, Retrieve, Update and Delete) of information, regarding Collections and Documents (7 endpoints), and endpoints for information search: SPARQL for knowledge and full text search for document content. In both cases, the original documents can also be retrieved. The document content must be provide manually by users, because automated PDF scraping/content extraction is still not supported.

All the endpoints defined for information management are listed in Figure 3.

crud	
GET	/kse/size
GET	/kse/listCollections
POST	/kse/createDocument
POST	/kse/addDocument
DELETE	/kse/deleteDocument
POST	/kse/retrieveDocument
POST	/kse/createCollection
search	
POST	/kse/search
POST	/kse/sparql

Figure 3: Information management endpoints.

3.4 Authentication

The authentication will not be limited to access the website, but it will be a much more detailed and resource-specific authentication policy. The basic authentication unit will be a 'user', which will be granted access to different resources: (i) websites in the graphical user interface that this user can access; (2) information resources (Collections,

Documents, Semantic annotation of documents or full text indexes) that this user can use, being able to specify if the user can read, write, etc. the resources.

Actually these roles have not been implemented, but we are planning to integrate Keycloak⁹ as independent authentication module, which we will leave to future work.

3.5 Graphical User Interface

The system that we present in this article (Knowledge Storage Ecosystem) has been designed with its integration in larger software systems in mind, hence the access to it has been predetermined through the HTTP REST API. This way of accessing the system requires users to have knowledge of programming. To ease interfacing with the system, we additionally created a graphical user interface (GUI) that allows users without programming knowledge to use KSE as well.

The graphical user interface that we present here is a Web system that has been designed for managing all the functionalities of KSE that are accessible through HTTP REST API endpoints. Its main objective is to be functional and styling the interface is added to the list of future work items. The existing pages (shown in Figure 4) in the graphical interface are: (1) Dashboard: introductory page where KSE is presented and links to the other pages are provided; (2) Management/Configuration: management of configuration parameters; (3) Users: user management; (4) Collections: management of collections, as well as being able to create new collections; (5) Collection: management of an individual collection, as well as being able to add documents to it; (6) Document: management of individual documents; (7) Text Search: KSE can be searched textually. The results are displayed in document list format; and (8) Sparql Endpoint: SPARQL queries can be made to the KSE. The results are displayed in table format.

The code and technical documentation of the graphical user interface for KSE is available at <https://gitlab.com/speaker-project/knowledge-management/kse-graphical-user-interface>.

4 Conclusions and Future Work

The joint storage of documents and semantic information associated with them is not a resolved task.

⁹<https://www.keycloak.org>

While there are solutions that have approached this problem from different angles, none of these solutions seem definitive, and there are unresolved issues. Besides, there are few existing tools that allow this functionality out-of-the-box. Therefore, we have implemented a system that performs this functionality in a simple way.

We implemented a solution that offers the user the desired functionality of CRUD operations over source documents and semantic information. The management of the information is done through HTTP REST API endpoints. To simplify that, we have also implemented and published a graphical user interface to use and manage the KSE system.

One of the important issues that we had to address in the implementation process is the synchronization of information between data storage tools.

There are several open issues that are kept for future work. The main items are:

- Integrating external Linked Data sources, such as Knowledge Bases (DBpedia, Wikidata, Yago, etc.) is foreseen. This is the first thing we plan to work on.
- Styling the interface so that aesthetics and ease of use are taken into account in the implementation.
- Implementing the authentication module by integrating Keycloak.
- Evaluating the system. The experiments to be carried out on this system are based on the evaluation of different user-related metrics that allow us to determine the usability, simplicity and performance of the system.

The link to the demonstration video is <https://youtu.be/4T6ujG6MH4>.

Acknowledgments

The work presented in this article has received funding from the German Federal Ministry for Economic Affairs and Climate Action (BMWK) through the project SPEAKER (no. 01MK19011).

References

- John Arwe, Steve Speicher, and Ashok Malhotra. 2015. Linked data platform 1.0. W3C recommendation, W3C. <https://www.w3.org/TR/2015/REC-ldp-20150226/>.

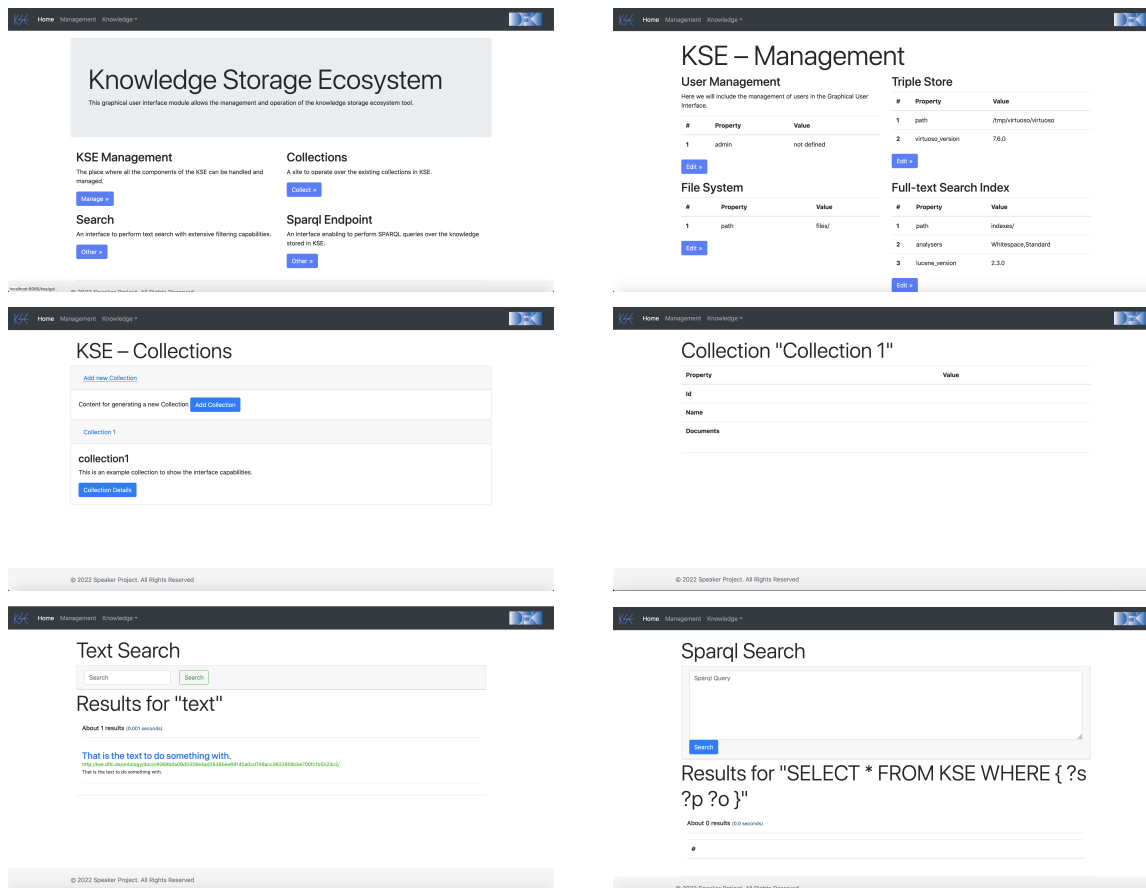


Figure 4: Screenshots of the different pages of the graphical user interface: top left is the dashboard, top right is the management and configuration, middle left is the collections management, middle right is the individual collection management, bottom left is the full text search and bottom right is the sparql endpoint.

Fariz Darari and Ruli Manurung. 2011. Linkedlab: A linked data platform for research communities. In *2011 International Conference on Advanced Computer Science and Information Systems*, pages 253–258.

Helena Handschuh. 2011. [Sha-0, sha-1, sha-2 \(secure hash algorithm\)](#). In Henk C. A. van Tilborg and Sushil Jajodia, editors, *Encyclopedia of Cryptography and Security (2nd Ed.)*, pages 1190–1193. Springer.

Iman Keivanloo, Christopher Forbes, Aseel Hmood, Mostafa Erfani, Christopher Neal, George Peristerakis, and Juergen Rilling. 2012. [A linked data platform for mining software repositories](#). In *2012 9th IEEE Working Conference on Mining Software Repositories (MSR)*, pages 32–35.

Vanessa Lopez, Spyros Kotoulas, Marco Luca Sbordio, Martin Stephenson, Aris Gkoulalas-Divanis, and Pól Mac Aonghusa. 2012. [Querocity: A linked data platform for urban information management](#). In *The Semantic Web – ISWC 2012*, pages 148–163, Berlin, Heidelberg. Springer Berlin Heidelberg.

Michael McCandless, Erik Hatcher, and Otis Gospodnetic. 2010. *Lucene in Action, Second Edition: Cov-*

ers Apache Lucene 3.0. Manning Publications Co., USA.

Borislav Popov, Atanas Kiryakov, Angel Kirilov, Dimitar Manov, Damyan Ognyanoff, and Miroslav Goranov. 2003. [Kim – semantic annotation platform](#). In *The Semantic Web - ISWC 2003*, pages 834–849, Berlin, Heidelberg. Springer Berlin Heidelberg.

Leonard Richardson and Sam Ruby. 2007. *RESTful Web Services*. O’Reilly, Beijing.

Gurnoor Singh, Arnold Kuzniar, Matthijs Brouwer, Carlos Martinez-Ortiz, Christian W. B. Bachem, Yury M. Tikunov, Arnaud G. Bovy, Richard G. F. Visser Finkers, and Richard. 2020. [Linked data platform for solanaceae species](#). *Applied Sciences*, 10(19).

Marcin Wylot, Manfred Hauswirth, Philippe Cudré-Mauroux, and Sherif Sakr. 2018. [Rdf data storage and query processing schemes: A survey](#). *ACM Comput. Surv.*, 51(4).

Fu Zhang, Qingzhe Lu, Zhenjun Du, Xu Chen, and Chunhong Cao. 2021. [A comprehensive overview of rdf for spatial and spatiotemporal data management](#). *The Knowledge Engineering Review*, 36:e10.

Towards a Conversational Web? A Benchmark for Analysing Semantic Change with Conversational Bots and Linked Open Data

Florentina Armaselu

University of Luxembourg, Luxembourg, florentina.armaselu@uni.lu

Christian Chiarcos

Goethe Universität Frankfurt, Germany, chiarcos@cs.uni-frankfurt.de

Barbara McGillivray

King's College London, United Kingdom, barbara.mcgillivray@kcl.ac.uk

Anas Fahad Khan

Istituto di Linguistica Computazionale
'A. Zampolli', Italy
fahad.khan@ilc.cnr.it

Ciprian-Octavian Truică

University Politehnica of Bucharest,
Romania
ciprian.truica@upb.ro

Giedrė Valūnaitė-Oleškevičienė

Mykolas Romeris University,
Lithuania
gvalunaite@mruni.eu

Chaya Liebeskind

Jerusalem College of Technology,
Israel
liebchaya@gmail.com

Elena-Simona Apostol

University Politehnica of Bucharest,
Romania
elena.apostol@upb.ro

Andrius Utkā

Vytautas Magnus University,
Lithuania
andrius.utka@vdu.lt

Abstract

The paper presents preliminary results from our experiments with large language models, linked data, and semantic change in multilingual diachronic contexts. It proposes the first steps towards a benchmark and aims at fostering discussion on the concept of conversational knowledge bots as emerging paradigms, and the use of linked open data in linguistic tasks.

1 Introduction

Developments in large language models (LLM) such as GPT-3, BLOOM and GPT-4 (Brown et al., 2020; Workshop BigScience, 2022; OpenAI, 2023) have drawn attention to the capabilities of deep learning technologies to support conversations between human and artificial agents using natural language. These types of conversation, spanning from question-answering to code generation, seem to indicate an emergent paradigm shift from current graphic- and keyword-based human-

computer interaction and search modes to a conversational way of interacting with machines and the World Wide Web. Although conversational agents such as ChatGPT and BLOOM have shown remarkable capabilities in generating human-like responses and ability to analyse and synthesise correct answers, the currently available versions may suffer from a few limitations, such as hallucinations, self-contradicting statements, or outdated information (Ji et al., 2023; Mündler et al., 2023).

The question that arises is, therefore, to what extent will this way of interacting affect present formalisms and concepts, in particular those related to the Semantic Web? Will the processing of large amounts of unstructured text and the availability of pre-trained language models with conversational abilities have an impact on the use of more structured forms of representing and accessing knowledge by means of vocabularies such as the Resource Description Framework (RDF), Web Ontology Language (OWL), Linked Open Data (LOD) or OntoLex? How might these two

paradigms influence each other and what possible forms of combining them might be imagined for applications in areas of research such as linguistics, data science and digital humanities?

Rather than providing direct answers to these questions, the aim of this paper is to discuss potential scenarios built on a use case that combines natural language processing (NLP) and linguistic linked open data (LLOD) to analyse semantic change in multilingual diachronic corpora. Sections 2 and 3 present related work and preliminary results from our experiments with ChatGPT (Brown et al., 2020), Bing (Mehdi, 2023), word2vec (Mikolov et al., 2013; Rehurek and Sojka, 2010), and OntoLex-FrAC (Chiarcos et al., 2022). Section 4 formulates questions based on these first-round observations and proposes a benchmark related to the concept of conversational knowledge bots and their application to linguistic tasks. Section 5 summarises our findings.

2 Related work

Research on semantic change, the phenomenon concerned with the change in the meaning of a lexical unit (word or expression) or of a concept over time, has seen significant progress in the natural language processing community in recent years (Tahmasebi et al., 2018; Tsakalidis et al., 2019; Schlechtweg et al., 2020). While the majority of these studies focus on corpus-driven embedding models covering different time intervals, some studies, e.g., Armaselu et al. (2022), have advocated for the integration of such distributional approaches with linked open data. Recent advances have also been reported in the area of linguistic linked data (Cimiano et al., 2020; Khan et al., 2021; McGillivray et al., 2023), which promotes the use of graph-based models to represent linguistic data, and in building AI-based conversational agents, such as OpenAI’s ChatGPT (Chat Generative Pre-trained Transformer), Microsoft Bing, and Google’s Bard. Studies on LLMs have drawn attention to both potential benefits and concerns (Maynez et al., 2020; Shuster et al., 2021; Talat et al., 2022; DIGHUM, 2023), to their ability to be trained on code, use external APIs (Chen et al., 2021; Schick et al., 2023) and integrate plugins.¹ However, to our knowledge, there have not been any enquiries on the opportunities and chal-

lenges of combining LLMs and LLOD in semantic change-related tasks. Given the trends in artificial intelligence (AI) possibly leading to a conversational Web paradigm, these forms of interaction and their impact should be considered within the linked data community. We will illustrate this point through examples from a use case in diachronic analysis.

3 Preliminary results

In this section, we describe preliminary results from semantic change experiments with ChatGPT and Bing, and word embedding methods.

3.1 ChatGPT and Bing

The experiments involved the use of the OpenAI platform that allows for interactions with the GPT conversational agent via a user account. At this stage, we carried out qualitative evaluations of GPT’s responses. Our first aim was to assess to what extent the agent can trace the semantic evolution of certain words, such as ‘revolution’, across several periods and languages or provide citations when explicitly asked to do so (Appendix A, Table 2).

Our second aim was to evaluate the agent’s ability to generate code based on specific relations, such as the sense of a word and its attestation, using OntoLex-Lemon (McCrae et al., 2017) and OntoLex(-FrAC) (Chiarcos et al., 2021, 2022), a model intended to complement its predecessor with a vocabulary necessary to represent information derived from corpora. The first results (Table 1) indicated that GPT seemed quite good at generating OntoLex. On the other hand, its responses related to OntoLex-FrAC, although sounding meaningful, were incorrect. This seemed to be due to a lack of training data using this formalism. This assumption was supported by GPT’s initial misreading of the acronym: *OntoLex-FrAC stands for OntoLex Lemon Extension for Frame-based Annotation of Corpora* (conversation 1); *OntoLex-FrAC (Ontology-Lexicon for Fine-Grained Annotations and Classification) is a model for representing fine-grained lexical and semantic information in natural language text* (conversation 7). Therefore, we tried to teach GPT about OntoLex-FrAC by asking it first to summarise a paper about the model (Chiarcos et al., 2022) by copying and pasting the full text. We tested what it learned by asking the ques-

¹<https://openai.com/blog/chatgpt-plugins>.

tion from Table 1. It proposed the same structure as `frac:attestation`, but with hallucinated vocabulary. In a third step, we provided the `le_falter_vi` example for attestation from the paper (p. 4020). The answer in Table 1 was obtained after asking the same question again in the fourth iteration.

User: Please express in OntoLex or OntoLex-FrAC the following sentence: The word 'révolution' in French, with the sense of 'mouvement circulaire effectué par un corps autour de son axe', was attested in 1727, according to Ortolang.

Chat GPT: Here is an example of how to express the sentence in OntoLex-FrAC:

CSS

```

:revolution_fr
  a ontolex:LexicalEntry ;
  rdfs:label "révolution"@fr ;
  ontolex:canonicalForm [
    a ontolex:Form ;
    ontolex:writtenRep "révolution"@fr
  ] ;
  ontolex:sense :revolution_fr_sense ;
  frac:attestation [
    a frac:Attestation ;
    rdf:value "révolution (sens de 'mouvement
circulaire effectué par un corps autour de
son axe') est attesté en 1727"@fr ;
    frac:attestationSource <https://www.ortolang.
fr/market/corpora/corpus-tcof>
  ] .
:revolution_fr_sense
  a ontolex:LexicalSense ;
  ontolex:definition [
    a ontolex:Definition ;
    rdfs:label "Définition de la révolution"@fr ;
    rdf:value "'Mouvement circulaire effectué par
un corps autour de son axe'"@fr
  ] .

```

Table 1: ChatGPT conversation 8 (23 March 2023). OntoLex-FrAC generation

To a certain extent, we managed to “teach” ChatGPT some OntoLex-FrAC, but corpus and date properties were missing, so ChatGPT made up something on its own. It still hallucinated a lot, influenced by the dialogue history. Remarkably, it learned that an attestation must contain an attestation string, and it made that up from the definition, so, with some more examples, it can produce valid OntoLex-FrAC. However, it failed to create a separate date property. Other GPT tests included similar conversations for words such as Lithuanian *ponas* (‘sir, lord’) and its equivalents in the other languages (conversation 10). Bing also misread the OntoLex-FrAC acronym. While correctly rendering OntoLex properties such as `ontolex:canonicalForm` and `ontolex:sense`, it included non-existing OntoLex-FrAC properties, e.g.,

`ontolexfrac:dataSource` and `ontolexfrac: dateOfAttestation` (Bing, conversation 1). Another aspect of the assessment referred to sources. For instance, when asked about the sources or methods used, the degree of detail of the GPT responses varied: from generic statements, *As an AI language model, I was trained on a large corpus of text data* (conversation 1); to recommendations, *I can suggest some resources [...]: National Library of Luxembourg [...], Corpus de Français Parlé à Bruxelles* (conversation 5); or to procedure descriptions, *In this example, we create a lexical entry [...] we include an attestation using the Frac vocabulary* (conversation 8).

3.2 Diachronic word embeddings

We compared the conversation results with the outcomes of our diachronic word embedding and LLOD modelling experiments using multilingual datasets (Appendix B, Table 3, 4). We trained standard word embedding techniques, such as word2vec (Mikolov et al., 2013; Rehurek and Sojka, 2010) and fastText (Bojanowski et al., 2017) on the datasets divided into time slices corresponding to centuries (LatinISE, Responsa) or smaller event-driven intervals (BnL Open Data). We extracted the neighbours of the target words in the different time slices via cosine similarity, following standard practice in semantic change detection. The goal was to query the models for similar terms expressing social, economic, cultural or historic facts, and compare them across several languages. We noted that whereas the time slice granularity of the order of centuries may point to meanings changing, emerging or fading out (LatinISE, SLIEKKAS, Responsa), the finer granularity seems to highlight polysemous usage in various contexts with no clear indication when a certain meaning has emerged or went out of use (BnL Open Data). In this respect, a combination of corpus- and dictionary-based knowledge may lead to richer contextual representations of semantic change.

4 Discussion

Section 3 experiments have shown that conversational agents such as GPT can provide information about the meanings of certain words or concepts and their evolution over time and across languages. However, to understand the mechanisms

that generated these changes, a deeper analysis of the sources providing evidence about them would be needed.

Metzler et al. (2021) consider that although state-of-the-art pre-trained language models are able to generate prose in response to an information need, they “do not have a true understanding of the world, they are prone to hallucinating, and crucially they are incapable of justifying their utterances by referring to supporting documents in the corpus they were trained over” (p. 2). In contrast, the models of the future should be able to leverage the “meta-information associated with documents like provenance, authorship, authoritativeness”, support “cross-lingual generalization”, integrate new data through “online” or “incremental” learning, and provide answers with a degree of detail close to those of a domain expert (pp. 2, 15, 16).

4.1 LLOD aggregation

Before considering the different types of knowledge agents that may assist our task in the future, we will get back to our example of diachronic analysis. For instance, the uses and meanings of the French word *révolution* in a certain country would need to be informed by knowledge representations combining corpora and dictionaries to study the term occurrences in time and space and compare them against existing attestation evidence. Listing 1 shows an example of lexical entry for *révolution* and its attestation that we created using elements from the OntoLex-FrAC model (Chiarcos et al., 2021, 2022).

Listing 1: OntoLex-FrAC modelling example

```

:rev-fr_le_1 a ontolx:LexicalEntry ;
  ontolx:canonicalForm [
    ontolx:writtenRep "révolution"@fr ] ;
  ontolx:sense :rev-fr_s_1 .
:rev-fr_s_1 a ontolx:LexicalSense ;
  frac:attestation [
    a frac:Attestation ;
    frac_new:dictionary [
      dc:source
        <http://example.org/ortolang/révolution>;
      dc:definition
        "Mec. Mouvement circulaire...";
      dc:date "1727"^^xsd:gYear ] ;
    frac:corpus [
      dc:source
        <http://example.org/ark:70795/dqgfr3/
          pages/17/articles/DTL612>;
      dc:date "1789"^^xsd:gYear ;
      dc:title "L'art de conduire et régler
        les pendules et les montres";
      dc:publisher "A Luxembourg, Chez la Veuve
        de J. B. Kleber, Imprimeur de Sa Majesté";
      frac:quotation "La roue ...
        fait une révolution par heure ...";
      prov:agent [
        a prov:Organization ;
        foaf:name

```

```

        "National Library of Luxembourg";
      ] ;
    ] ;
  ] ;
frac:embedding [
  a frac:FixedSizeVector ;
  dc:extent "100"^^xsd:int ;
  dc:description "word2vec";
  rdf:value "[moyene, engrennat, tige ...]";
] .

```

We propose an extension of this formalism to include attestation both from dictionaries (provisionally marked by `frac_new:dictionary`) and corpora, by specifying as well the provenance and method used to obtain the corpus-based evidence. The `dc:source` identifies the dictionary entry and the document containing the corpus citation, while the `dc:date` refers to the attestation of the sense in the dictionary and the publication date of the corpus document. Complementary information may be added, such as title, publisher, author, etymology and translation relations, degree of certainty, agent identification, etc. While not all these categories of information can be available for the processed sources (especially, those from ancient times may be less complete or certain), this type of structured aggregation may provide more context and ground for possible inferences on the circulation of knowledge and the meaning of a term and its evolution across space, time, languages and cultures.

4.2 Knowledge bots

Therefore, we imagine different forms of knowledge agents, from bots that provide outlines and connections between various themes, such as ChatGPT, to specialised agents able to focus on particular tasks and resources and return well documented responses. These responses can vary from answers to general questions, recommendations for reading or relevant resources, to dedicated search and processing of target datasets, code generation, and expert advice on a given topic. Such agents may also be taught to produce correct LLOD representations. This might lower the entry barrier for data providers, since the conversion can be automatised via GPT-like engines. For consumers, it may also lower the entry barrier, since it can help to explain turtle code in human language. In either way, it is not a substitute for having OntoLex/RDF data in the first place, but a complementary technology. LLMs lack semantic transparency and verifiability, and this is what LLOD can provide.

While transparency, interoperability, connectiv-

ity, unique identification, and ontological precision are chief assets of the Semantic Web technologies, the advances in AI-based unstructured data processing and content generation would probably imply changes in the way we create and interact with structured data on the Web. From this perspective, a series of questions should be addressed, such as: (1) What forms of knowledge agents can be foreseen to combine conversational abilities in natural language with search, processing and automatic generation of structured data in formats such as RDF, OWL and LLOD? (2) What is the role of the human agent and what types of task, interaction scenarios and potential threats can be envisaged within the human-bot interrelations? (3) How may the current Semantic Web formalisms evolve to accommodate these emerging modes of interaction and knowledge representation? (4) What new forms of collaboration between the LOD and NLP communities can be imagined to underpin the development of a conversational and more “content-aware” Web? To foster further discussion on these topics, we propose to create a shared repository of benchmarks related to combined LLM and LLOD scenarios within various use cases.

5 Conclusion and future work

We presented preliminary tests with language and linked data models in multilingual diachronic analysis. Taking into account the potential of AI-based agents, able of human-like conversations, and of an emerging conversational Web, we propose to create a benchmark repository shared within the (L)LOD community for use cases that combine conversational and linked data knowledge paradigms.

Acknowledgment

This article is based upon work from COST Action *Nexus Linguarum*, *European network for Web-centred linguistic data science*, supported by COST (European Cooperation in Science and Technology). www.cost.eu.

Authors' contribution

FA wrote the manuscript, led the ChatGPT conversations 1-6, 9, 11, and contributed to the design of the semantic change experiments for French and LLOD modelling; CC led the ChatGPT conversations 7, 8, and contributed to Sections 3.1 and 4.2;

BM contributed to the revisions of the manuscript and to the design of the semantic change experiments, and provided the analysis of the Latin words; AFK contributed to the RDF modelling of the example in section 4.1 and proofreading; COT led the Bing conversation 1, and contributed to the revisions of the manuscript and Section 2; GVO contributed to the revisions of the manuscript and Section 3, and led the ChatGPT conversation 10; CL contributed to the revisions of the manuscript and Section 3, and led the ChatGPT conversation 12; ESA contributed to the revisions of the manuscript and Section 2; AU contributed to Section 1 and the overall revision of the manuscript. All authors reviewed the final manuscript.

References

- Florentina Armasele, Elena-Simona Apostol, Anas Fahaad Khan, Chaya Liebeskind, Barbara McGillivray, Ciprian-Octavian Truică, Giedrė Valūnaitė-Oleškevičienė, and Marieke van Erp. 2022. [LL\(O\)D and NLP perspectives on semantic change for humanities research](#). *Semantic Web*, 13(6):1051–1080.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. [Enriching word vectors with subword information](#). *Transactions of the Association for Computational Linguistics*, 5:135–146.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, Vancouver, Canada.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebguss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant

- Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. [Evaluating large language models trained on code](#). (arXiv:2107.03374). ArXiv:2107.03374 [cs].
- Christian Chiarcos, Elena-Simona Apostol, Besim Kabashi, and Ciprian-Octavian Truică. 2022. [Modelling frequency, attestation, and corpus-based information with OntoLex-FrAC](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4018–4027.
- Christian Chiarcos, Thierry Declerck, and Maxim Ionov. 2021. [Embeddings for the lexicon: Modelling and representation](#). In *Workshop on Semantic Deep Learning*, page 1319.
- Philipp Cimiano, Christian Chiarcos, John P. McCrae, and Jorge Gracia. 2020. *Linguistic Linked Data. Representation, Generation and Applications*. Springer International Publishing.
- DIGHUM. 2023. [ChatGPT a catalyst for what kind of future? statement of the digital humanism initiative](#).
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. [Survey of hallucination in natural language generation](#). *ACM Computing Surveys*, 55(12):1–38.
- Anas Fahad Khan, Christian Chiarcos, Thierry Declerck, Daniela Gifu, Elena González-Blanco García, Jorge Gracia, Max Ionov, Penny Labropoulou, Francesco Mambrini, John McCrae, Émilie Pagé-Perron, Marco Passarotti, Salvador Ros, and Ciprian-Octavian Truică. 2021. [When linguistics meets web technologies. Recent advances in modelling linguistic linked open data](#).
- Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. [On faithfulness and factuality in abstractive summarization](#). In *Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919. ACL.
- John P McCrae, Julia Bosque-Gil, Jorge Gracia, and Paul Buitelaar. 2017. [The OntoLex-Lemon model: development and applications](#). In *Proceedings of eLex 2017 Conference*.
- Barbara McGillivray, Pierluigi Cassotti, Pierpaolo Basile, Davide Di Pierro, and Stefano Ferilli (in press). 2023. [Using graph databases for historical language data: Challenges and opportunities](#). In *Proceedings of the 19th Italian Research Conference on Digital Libraries, Bari, Italy, February 23-24, 2023*, CEUR Workshop Proceedings. CEUR-WS.org.
- Yusuf Mehdi. 2023. [Reinventing search with a new AI-powered microsoft bing and edge, your copilot for the web](#).
- Donald Metzler, Yi Tay, Dara Bahri, and Marc Najork. 2021. [Rethinking search: Making domain experts out of dilettantes](#). *SIGIR Forum*, 55(1):1–27.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. [Efficient estimation of word representations in vector space](#). pages 1–12.
- Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2023. [Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation](#).
- OpenAI. 2023. *GPT-4 Technical Report*.
- Radim Rehurek and Petr Sojka. 2010. [Software framework for topic modelling with large corpora](#). In *LREC 2010 Workshop New Challenges for NLP Frameworks*, pages 45–50.
- Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. [Toolformer: Language models can teach themselves to use tools](#). (arXiv:2302.04761). ArXiv:2302.04761 [cs].
- Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. [SemEval-2020 task 1: Unsupervised lexical semantic change detection](#). In *Workshop on Semantic Evaluation*, pages 1–23.
- Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. [Retrieval augmentation reduces hallucination in conversation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803.
- Nina Tahmasebi, L. Borin, and A. Jatowt. 2018. [Survey of computational approaches to lexical semantic change](#). *arXiv: Computation and Language*.
- Zeerak Talat, Aurélie Névéal, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, Shanya Sharma, Arjun Subramonian, Jaesung Tae, Samson Tan, Deepak Tunuguntla, and Oskar Van Der Wal. 2022. [You reap what you sow: On the challenges of bias evaluation under multilingual settings](#). In *BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 26–41, virtual+Dublin.
- Adam Tsakalidis, Marya Bazzi, Mihai Cucuringu, Pierpaolo Basile, and Barbara McGillivray. 2019. [Mining the UK Web Archive for semantic change detection](#). In *International Conference on Recent Advances in Natural Language Processing*, pages 1212–1221.
- Workshop BigScience. 2022. [BLOOM: A 176B-parameter open-access multilingual language model](#). (arXiv:2211.05100).

Appendix A. ChatGPT

a) User	Chat GPT
What do you know about semantic change?	Semantic change is the process by which the meaning of a word or phrase changes over time. [...]
Can you trace the evolution of the word 'révolution' from 16 century to now?	[...] 16th century: The French word "révolution" first appeared to describe the motion of celestial bodies. [...] 17th century: The term began to be used [...] to describe any significant or sudden change [...] 18th century: [...] The term "révolution" came to be associated with the idea of overthrowing a despotic government or monarch [...]
Can you trace the evolution of the word "מהפכה" (mahapecha) from its first attestation to now?	[...] The earliest known use of the term "מהפכה" appears in the Hebrew Bible [...] the term refers to a sudden and dramatic reversal of fortunes [...]. In the medieval period, the term [...] was used to describe a military defeat or a natural disaster [...] During the late 19th and early 20th centuries, [...] the term [...] became associated with the idea of a Jewish national revolution [...]
Can you trace the evolution of the word "revoliucija" in Lithuanian [...]?	The Lithuanian word "revoliucija" first appeared in written texts in the 19th century, during the period of the Lithuanian National Revival [...] During the early 20th century, "revoliucija" continued to be used to describe political upheavals [...]
What is the equivalent of the word 'revolution' in Latin?	The equivalent [...] in Latin is "revolutio". It is derived from the verb "revolvere", which means "to roll back" or "to turn over". In Latin, "revolutio" was primarily used to describe the cyclic movement of celestial bodies [...]
Is there any equivalent of the word "επανástasi" (epanástasi) in Ancient Greek?	The Ancient Greek word that is most commonly used to refer to a "revolution" or "uprising" is "στάσις" (stasis). The word [...] is used in several works of Ancient Greek literature, including Thucydides' "History of the Peloponnesian War" and Aristotle's "Politics" [...] to describe the internal conflicts and power struggles [...]
b) User	Chat GPT
[...] please give me a citation from an earlier period that uses a related term, such as "revolte" or "revolucion".	[...] "Si comme la revolucion du ciel, qui jamais ne se peut arrester ne reposer, mais tous jours et toutes heures va, porte et conduit toutes choses." - Jean de Meun, Le Roman de la Rose, 1280-1290. This passage [...] uses the term "revolucion" in the sense of a continuous and unceasing movement or cycle [...]

Table 2: a) ChatGPT conversations 2, 6, 12 (11, 19, 28 March 2023). Semantic change over time and language; b) ChatGPT conversation 4 (18 March 2023). Citations

Appendix B. Word embeddings


Dataset	Language	Time span	Size
LatinISE	Latin	2nd c. BCE - 20th c. CE	ca.10 mil. word tokens
Diorisis	Ancient Greek	8th c. BCE - 5th c. CE	10,206,421 word tokens
SLIEKKAS	Old Lithuanian	16th - 18th c.	10 texts, 350,000 words
BnL Open Data	French	1690 - 1918 (monographs)	15,966 documents, 6,514,941 word tokens
Responsa	Hebrew	11th - 21st c.	about 100 mil. word tokens


Table 3: Core datasets


LatinISE	SLIEKKAS	BnL Open Data	Responsa
450BCE-1BCE: <i>civitas</i> ('citizenship')	16th c.: <i>ponas</i> (rich person, title 'mister'; religious, 'lord', e.g., Jesus)	1690-1794: <i>révolution</i> (Mec. motion of a body around an axis)	11th-16th c.: <i>מהפכה</i> (revolution) (religious context, 'atheism', 'repentance')
1CE-450CE: <i>civitas</i> ('city')	18th c. <i>ponas</i> (rich person; independent person, 'master')	1831-1866: <i>révolution</i> (Geom. motion of a figure around an axis)	16th c.: <i>מהפכה</i> (frequency of the word declines)
451CE-900CE: <i>civitas</i> ('city')		1867-1889: <i>révolution</i> (Geol. natural phenomena)	17th-19th c.: <i>מהפכה</i> (context of war and tragedy)
		1890-1918: <i>révolution</i> (Pol. Hist. great political change)	20th c.-present: <i>מהפכה</i> (industrial, medical, ideological revolution)

Table 4: Word embedding results. Excerpts


Adopting Linguistic Linked Data Principles: Insights on Users' Experience

Verginica Barbu Mititelu  Romanian Academy, Research Univ. of Naples "L'Orientale"
Institute for Artificial Intelligence
vergi@racai.ro

Maria Pia di Buono  Univ. of Coimbra, Portugal
Napoli, Italy
mpdibuono@unior.it

Hugo Gonçalo Oliveira 
Univ. of Coimbra, Portugal
CISUC, DEI
hroliv@dei.uc.pt

Blerina Spahiu 
University of Milan-Bicocca, Italy
blerina.spahiu@unimib.it

Giedrė Valūnaitė Oleškevičienė 
Mykolas Romeris university, Lithuania
gvalunaite@mruni.eu

Abstract

Despite the advantages, Linguistic Linked Data (LLD) best practices and principles seem far from being widely adopted. Such a situation can be related to existing challenges in the creation, reusing, and exposing of LLD resources. In this paper, we present the results of a survey which examined users' perspective and experience in the use and application of LLD principles, to evaluate the impact, prospects, requirements, or challenges encountered in LLD adoption. The survey was organized in several sections to collect information about participants' background, LLD knowledge, use, development, publishing, and metadata use. The results show that some bounds have to be overstepped to ensure the penetration of LLD principles in a wider community and fully exploit their potential.

1 Introduction

Linguistic Linked Data (LLD) best practices and principles aim at describing language resources and conveying useful linguistic information about them, allowing linking among resources, interoperability across datasets and systems, as well as their federation (Chiarcos et al., 2020).

Despite their advantages, including for under-resourced languages (Bosque-Gil et al., 2022), LLD best practices and principles seem to be far from being widely adopted. Such a situation can be related to some challenges in the creation, reusing, and exposing of LLD. In this paper, we present the results of a survey, conducted within the COST Action "CA18209 - European network for Web-centred linguistic data science"¹ (Nexus Linguarum, NL CA), Working Group (WG) 1 - Task

1.2 in collaboration with Tasks 1.4 and 1.5, which investigated the users' perspective and experience in the use and application of LLD principles, in order to evaluate the impact, prospects, requirements, or challenges encountered in LLD adoption.

Such an evaluation complements another survey carried out within NL CA (Khan et al., 2022), as it offers another (i.e., the (potential) user's) perspective on the adoption of LLD and could be of interest not only to other WGs within NL CA, but also to other stakeholders, including people and categories involved in European initiatives and projects, such as the European Language Grid² and the Prêt-à-LLOD³ projects.

The paper is organized as follows: in Section 2 we report on related work; in Section 3 we describe the survey aims and structure, while in Section 4 we present the results. Section 5 is devoted to discussing some of our findings and, finally, in Section 6 we conclude and envisage future work.

2 Related Work

LLD is known to offer numerous advantages and opportunities. Lezcano et al. (2013) observed that the simple syntactic model of RDF, which allows organizing structured data into a set of simple triples, makes linguistic data suitable for carrying out tasks combining data from different sources. Also, as Linked Data (LD) is comparatively straightforward, data discovery and harvesting become an accessible task for performing without full knowledge of the data structure. While discussing their survey, Lezcano et al. (2013) pointed out that RDF requires a standardized representation

¹<https://nexuslinguarum.eu/>

²<https://live.european-language-grid.eu/>

³<https://pret-a-llod.github.io/>

of the annotation semantics. The authors identified some legal and economic issues concerning copyrighting and pricing of Language Resources (LR), that act as barriers to LR interoperability and propose that the adoption of LLD approaches to LR exchange may have a positive impact on these matters. They also identified an open issue – the development of mechanisms and knowledge to support the alignment of different features and aspects of LRs which allow for ensuring semantic and conceptual interoperability in the LOD cloud⁴. Some other areas of LLD are to be considered for improvement concerning the languages covered and types of linguistic datasets presented in the LOD cloud.

Geddes (2019) acknowledged that LLD provides the opportunity to use the data freely and connect the data to other existing data; however, the focus on the user, user's needs and capacities is of key importance in the process of sustaining a healthy data ecosystem. As LLD technologies facilitate information integration and interoperability, they require making the entities addressed in an unambiguous way, so that they could be accessed and interpreted. Also, it should be ensured that entities associated on a conceptual level are physically associated with each other as well.

The LLD applications reveal the potential of the technology in linguistics, but there is still a considerable barrier for linguists who are not advanced users of RDF and related technologies. Since the early days of the Semantic Web, the "cognitive overhead" of learning RDF and related technologies was pointed out as an obstacle to its adoption by a broader community (Marshall and Shipman, 2003). This identifies the necessity of the technology to achieve a certain level of user-friendliness suitable for its non-advanced users (Chiarcos et al., 2020).

An overview of the existing guidelines and best practices in LLD development, interlinking, publication, and validation was given by the data collection carried out as part of the survey on LLD models (Khan et al., 2022) performed as part of Task 1.1 of the NL CA. The process included the compilation of a survey of LLD-relevant projects and other relevant initiatives (i.e. W3C community groups). Khan et al. (2022) identified that the advantages of LLD and the numerous opportunities it offers as a means of publishing linguistic data require a

certain level of technical appreciation of the Semantic Web, of RDF and other formalisms as well as a number of other technologies. In order to increase the uptake of LLD amongst non-specialists, it is important to make sure the available materials are made accessible to non-specialists and provide clear instructions and ways of doing common tasks which could be ensured by Guidelines (GLs) and Best Practices (BPs). The authors provided a list of the areas for improvement for LLD GLs/BPs supported by the experience of the authors, consumers, and compilers of the documents:

- access to documents should be provided to speakers of more (ideally any) languages, not only English;
- the documents should be easily findable and freely accessible;
- the documents should be clear and self-contained;
- the documents should be designed for different levels of expertise and for covering at least the types of resources listed in the LLOD cloud and the four tasks (generation, interlinking, publication, and validation);
- the documents should refer to existing tools that can be integrated into the workflow;
- the documents should be regularly updated with the latest technology/models/tools.

The provided list of important areas helps to evaluate the already existing materials and the trends of use which we have found in the survey, as well as to suggest the directions to prioritize in the process of producing new materials.

3 Survey

With the aim of identifying potential obstacles preventing (potential) users from adopting LLD principles, we conducted a survey, whose structure is rendered in Figure 1, to collect information about participants' background, LLD knowledge/use, development, publishing, and metadata use.

The insights coming from the survey results are relevant for:

- the penetration of LLD, especially among linguists and language professionals/experts;

⁴<https://linguistic-lod.org/llod-cloud>

- the causes preventing potential contributors/users from applying/(re)using LLD principles/resources;
- the causes preventing potential developers from creating LLD resources or converting resources to LD format, as well as from publishing them;
- highlighting possible limitations of LLD resources/technologies (including current vocabularies);
- the extension/integration of vocabularies and models suitable to describe different linguistic information and language phenomena;
- the extent to which metadata are used to describe resources, as well as the user's preference with respect to their type.

The survey was open from July 2021 to February 2022, with two main calls for participation, distributed through social media, i.e., Twitter, and mailing lists, e.g., Corpora list, NL CA mailing list, and personal contacts. The total number of responses is 84, received from different participants.

4 Results

We present here the results of the survey with respect to the four major lines of interest (LLD use, development, publishing and metadata), as shown in Figure 1.

The survey reached both witting and unwitting researchers in LLD. From the former group, there were 58 participants ($\approx 69\%$) to the survey, while from the latter there were 26 participants ($\approx 31\%$). The results presented below are based on the responses provided by the 58 participants, because, as can be seen in Figure 1, the other 26 did not answer the questions related to LLD experience.

The distribution of the 58 participants according to their declared background is shown in Figure 2, where we notice that this distribution is quite balanced. Beware that no further division within each group of specialists (computer scientists, computational linguists and linguists) is made, although we admit the categories are broad.

4.1 Use

Although aware of LLD, about one third (19) of the 58 participants never used LLD.

When inspecting the reasons provided⁵ for not using LLD resources, we see that the main one is that the tools and resources they work with do not support this format ($\approx 50\%$). Two other reasons are that they did not find a useful resource ($\approx 37\%$) and they were not familiar with LLD ($\approx 32\%$). To some extent, these are all related: i.e., for someone not familiar with LLD, even if they do not assume it, it will be harder to find useful resources. This relation may also explain why no participant gave both reasons.

Reasons like the lack of documentation (2, $\approx 10\%$), and, consequently, not knowing how to access this data (1, $\approx 5\%$) were also given. Both participants that refer to the lack of documentation also answer that they did not find a useful resource that fits their needs. The lack of documentation seems to be an obstacle to the adoption of LLD resources and technologies. As recently highlighted by Khan et al. (2022), there are not enough materials available fulfilling the role of guidelines and best practices for LLD, and, moreover, a lot of what exists has not been updated for years, thus being unable to reflect the latest developments in the field.

Another relevant reason for not using LLD resources was that the dump or SPARQL endpoint of a resource they were interested in was not working ($\approx 20\%$). This is not surprising: di Buono et al. (2022) recently noted that in the metadata of the 136 linguistic datasets in the LLOD Cloud, only 41 included a SPARQL endpoint and none included the URL of their dump. This is more related to the maintenance of LLD, which can be quite complex for the creators of this kind of data. The fact that many resources listed in the LLOD Cloud and other hubs are not accessible is definitely not good advertising for LLD, and may push potentially interested users away. Together with the lack of documentation, this contributes to one last reason: not understanding the advantages of LD over other formats (2, $\approx 10\%$). Both of these participants also say that they are not familiar with LLD and SPARQL. All the above mentioned reasons for not using LLD resources are presented in Table 1.

The long discussed “cognitive overhead” of learning underlying technologies (Marshall and Shipman, 2003) plays a role here, i.e., it requires

⁵The participants could provide more than one reason for not using LLD resources.

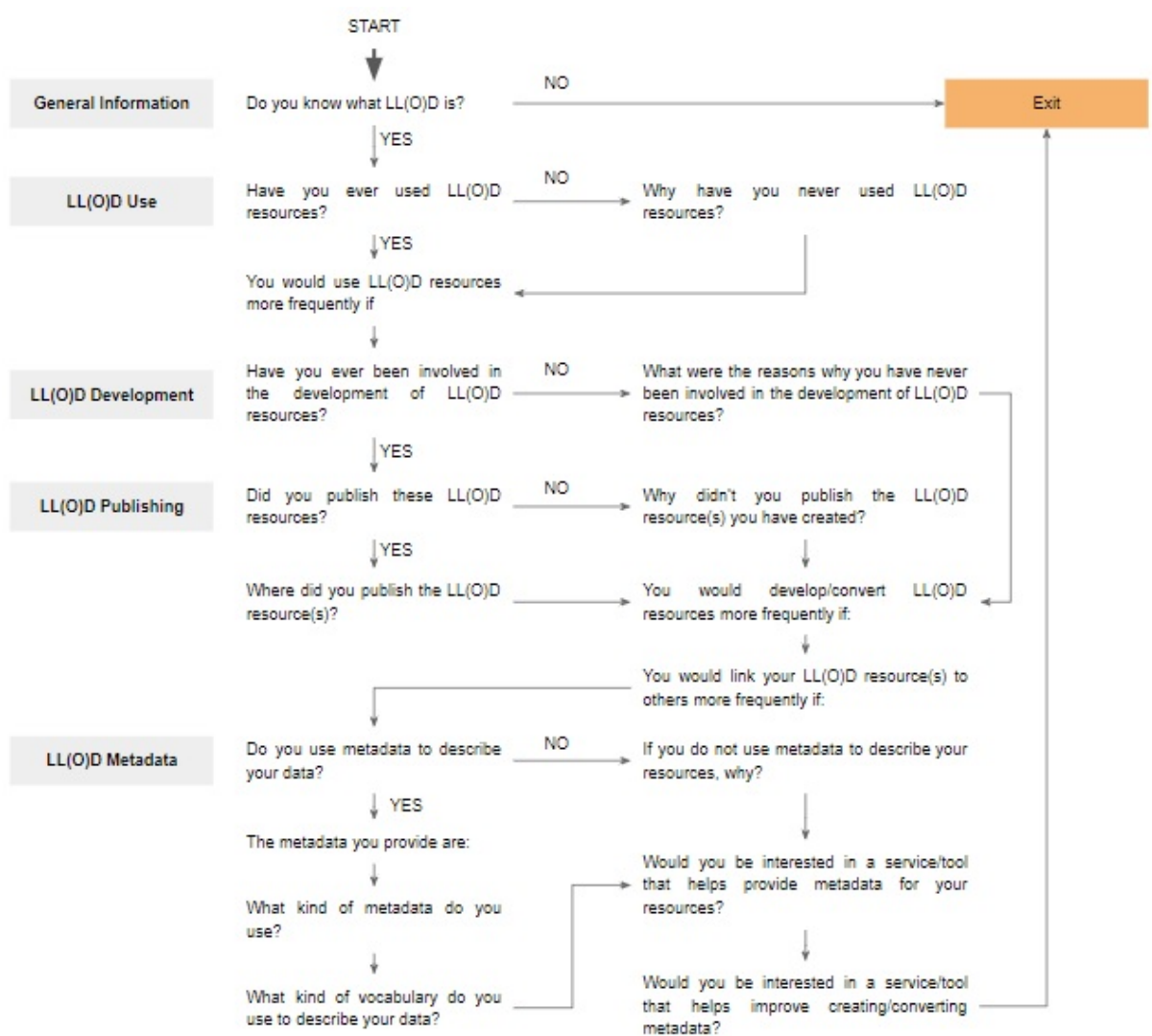


Figure 1: Diagram of survey flow. Some questions in the General Information section have been omitted given space constraints.

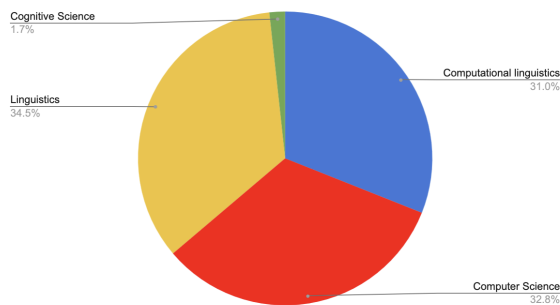


Figure 2: The distribution of the 58 LLD-aware participants to the survey according to their declared background.

time to become familiar with technologies like RDF and SPARQL.

Table 1: Reasons for never using LL(O)D resources

Reason	#
The tools/resources I work with do not support this format	9
I didn't find a useful resource that fits my needs	7
I am not familiar with the LLD models/SPARQL	6
The dump/SPARQL endpoint of the resource I was interested in was unavailable	4
I don't understand the advantages of Linguistic Linked (Open) Data resources over other formats (e.g. CoNLL-U)	2
I don't know much on how to access them	1
The documentation for the resource I was interested in was missing	2

All 58 participants were asked about the conditions (one or more) under which they would use LLD resources more frequently (see Table 2). Among the multiple choices, 30 ($\approx 52\%$) highlight the need for more documentation to help them using LD resources, and 23 ($\approx 40\%$) say that they would need more documentation about the resources they would potentially use. Moreover, 38 ($\approx 66\%$) and 29 (50%) participants, respectively, selected the availability of tools/services suitable to use and discover LLD resources.

Table 2: Conditions under which participants to the survey would use LL(O)D resources more frequently

Condition	#
You were aware of a user-friendly service/tool to help you use LD resources	38
You were aware of more informative documentation to help you use LD resources	30
You were aware of a user-friendly service/tool to help you discover LD resources	30
The resources I would potentially use had (better) documentation	23
Other	5

Table 3: Reasons for not developing LD resources. % is calculated from the total number of reasons provided.

Reason	#	%
incompatibility with other tools/resources used	14	50
lacking knowledge about adequate model/vocabulary	6	21
models not totally appropriate for representing data	5	18
unclear example or guidelines	8	29
unclear advantages that LD has over other formats	3	11
Other	4	14

4.2 Development

The shares of participants that develop resources in LLD format and of those who do not are almost equal, with a slight dominance of the former: $\approx 51\%$ of the participants are also developers of LLD resources, while $\approx 48\%$ do not develop them.

More than one reason could be provided for not developing LD resources and the answers given are summarized in Table 3: incompatibility with other tools or resources is the reason invoked by half of the respondents, 21% of all participants mentioned the lack of knowledge about the appropriate model or vocabulary for the resource under focus, while 17% of them complain about the inability of models to model data thoroughly.

4.3 Publishing

Developing LLD resources does not necessarily imply their publishing. According to the results of this survey, only 57% of these resources get published. Figure 3 shows this publication tendency per different types of resources (as classified in the LLOD Cloud): we notice that the few typological databases developed have also been published, two-thirds of the terminologies, thesauri, and databases have been so, only a little more than half of the other types of resources have been published, and less than half of the linguistic resource metadata have been published.

Participants who responded positively to the development of resources (30 respondents, i.e. $\approx 52\%$) were then asked to answer about publishing/exposing such resources and only 23 ($\approx 77\%$) of them published the resources, mainly in local repositories (15 people, i.e., $\approx 48\%$) and in the LLOD cloud (8 people, $\approx 26\%$). Considering other infrastructures/repositories for linguistic resources/language technologies, $\approx 17\%$ of the respondents (4) published their resource in CLARIN and only $\approx 9\%$ (2) in ELG. We note that none of

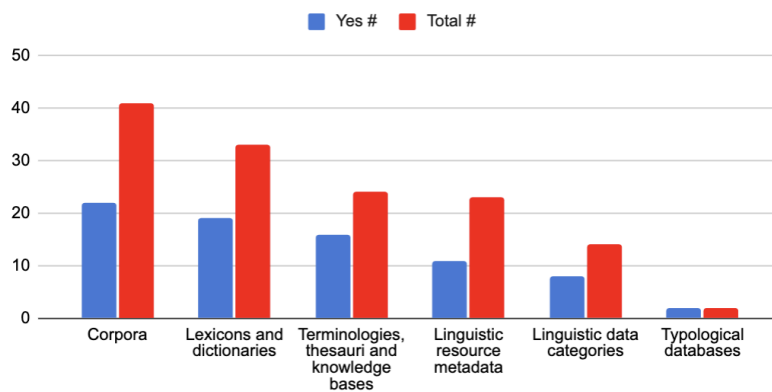


Figure 3: The proportion of users who are only developers of LD resources (in blue) and those who are both developers and publishers of LD resources (in red), for each type of LD resource.

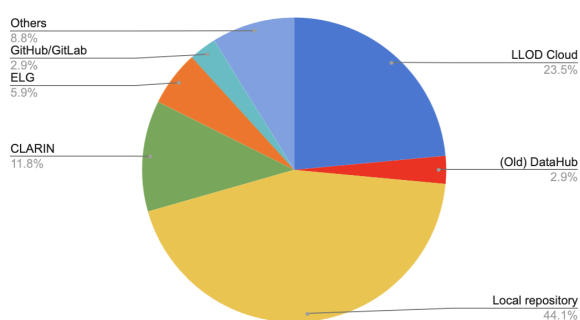


Figure 4: Repositories for publishing LD resources.

the respondents used META-SHARE to publish their resources – see Figure 4.

With reference to the reasons preventing publishing resources, copyright policies were the main one, as invoked by $\approx 57\%$ of the respondents not publishing the developed LL(O)D resources. The lack of knowledge about how/where to publish these resources, the cost/effort needed to publish/maintain the resources, and the lack of motivation have been equally given reasons ($\approx 14\%$ each).

4.4 Metadata

Metadata allows people to organize data in such a way that is meaningful to other people while making their findability easier (Zuiderwijk et al., 2012; Schmachtenberg et al., 2014). It is also a way of keeping the data consistent and enabling decisions in data handling (Spahiu et al., 2019). There are thus many advantages to producing and maintaining metadata.

In fact, 52 people ($\approx 90\%$) confirmed that they do use metadata to describe their data. On the other hand, 6 ($\approx 10\%$) participants do not use metadata.

The most shared reasons for not using metadata for describing the data are: (i) task consuming task; (ii) manual effort is required; and (iii) there is a lack of harmonization among metadata models. Only one user mentioned that the reason why they do not use metadata is that they have difficulties finding the right model.

Understanding and interpreting LLD is difficult as information about the context of the data is often missing (di Buono et al., 2022). Still, even for the available metadata, there are issues. Searching through or browsing LOD is not straightforward because the metadata is often not structured and not machine-readable (Zuiderwijk et al., 2012). However, the majority of the participants (30, $\approx 58\%$) have declared that they provide metadata in machine-readable format (see Figure 5). Participants who declared that they do not provide the metadata in a machine-readable format have the following backgrounds: two are computational linguists, and one is a linguist. Most of the participants who have declared that they provide metadata in a machine-readable format are computer scientists.

Regarding the type of metadata that participants use (Figure 6), it seems that descriptive metadata is the most used. 52 ($\approx 98\%$) participants use such metadata to describe the content of the data. Among such metadata, we can find the title, keywords, abstract, etc. Moreover, the descriptive elements that fall into this type support also the discovery, and the locating of such resources and they are also used to track the origin of the data.

Then, the types provenance (26, 50%) and technical (25, 48%) metadata were the second and the third most used types of metadata declared. Prove-

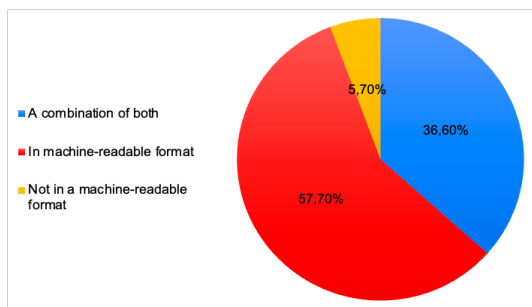


Figure 5: The distribution of the types of metadata reported as used for describing LD resources: machine-readable (red, $\approx 58\%$), not machine-readable (orange, $\approx 6\%$) and both (blue, $\approx 36\%$).

nance metadata provides information about the digital resource's history helping track its lifecycle, while technical metadata provides information related to how a system functions or metadata behaves.

Administrative metadata, which aims at providing information about managing and administering collections and information resources, is the fourth most used type with declared by 24 participants (46%). The second less-used type of metadata is the Use metadata (19, $\approx 37\%$) which provides information related to the level and type of use of collections and information resources. Finally, Preservation metadata (12, $\approx 23\%$) are the ones that provide information about the preservation management of the resource.

Vocabularies are means of sharing information and documenting definitions that should be clear, thus reducing the ambiguity of terms used in the data. In order to describe the data, data producers use existing vocabularies or ad-hoc developed ones. When creating a vocabulary, it is a common practice to use or extend pre-existing ontologies and vocabularies, which favors communication between people and computer applications. However, most of the participants (27, $\approx 52\%$) declare that they develop their own vocabulary, while 25 ($\approx 48\%$) use external vocabularies.

We asked all participants if they would be interested in a service or tool that supports them in the process of metadata creation or conversion. Figure 7 shows that 40 (69%) participants declared that they would be interested in such a service, 15 (26%) said that they might be interested, while 3 (5%) said that they have no interest. In fact, looking at the answers, 4 (66%) of the participants who did not use metadata to describe the data are in-

terested in such a service, and 2 (34%) said that they might be interested. However, all the participants that do not have an interest in such a service do provide metadata about their data. This might be related to the fact that such users have already set the process of metadata creation and have no interest in a new service.

When it comes to the improvement of the metadata creation process, Figure 8 shows that 44 (76%) participants declared that they would be interested in a service that supports them in improving the metadata creation process; 13 (22%) said that they could be interested, and only 1 (2%) does not have any interest in such a service. The latter participant further declared that they use metadata to describe the data.

Table 4 contains the list of vocabularies and the number of times they were mentioned by the participants. The most used vocabulary is DublinCore⁶, which is a set of fifteen “core” elements (properties) for describing resources. These properties are: Contributor, Coverage, Creator, Date, Description, Format, Identifier, Language, Publisher, Relation, Rights, Source, Subject, Title, and Type. In fact, all 8 participants who use DublinCore use Descriptive Metadata for their data. The second most used vocabulary is META-SHARE⁷, which is used to describe language resources (corpora, lexical/conceptual resources, models, grammar, etc., and language processing tools and services) for Language Technology needs. DCAT⁸ (Data Catalog Vocabulary) and OntoLex⁹ are the third most used vocabularies. While DCAT is used with the aim of facilitating interoperability between data catalogs published on the Web, OntoLex is used to take care of the representation of lexica relative to ontologies. The less used vocabularies are used for specific purposes and include DRMJ¹⁰, Preservica¹¹, etc.

5 Analysis and Discussion

In this section we try to correlate the responses to the different parts of the survey, with the aim of better understanding the conditions that prevent the wider adoption of LLD principles in the language resources community.

⁶<https://www.dublincore.org/>

⁷<http://www.meta-share.org/>

⁸<https://www.w3.org/TR/vocab-dcat-3/>

⁹<https://www.w3.org/2016/05/ontolex/>

¹⁰<http://drmj.eu/>

¹¹<https://preservica.com/>

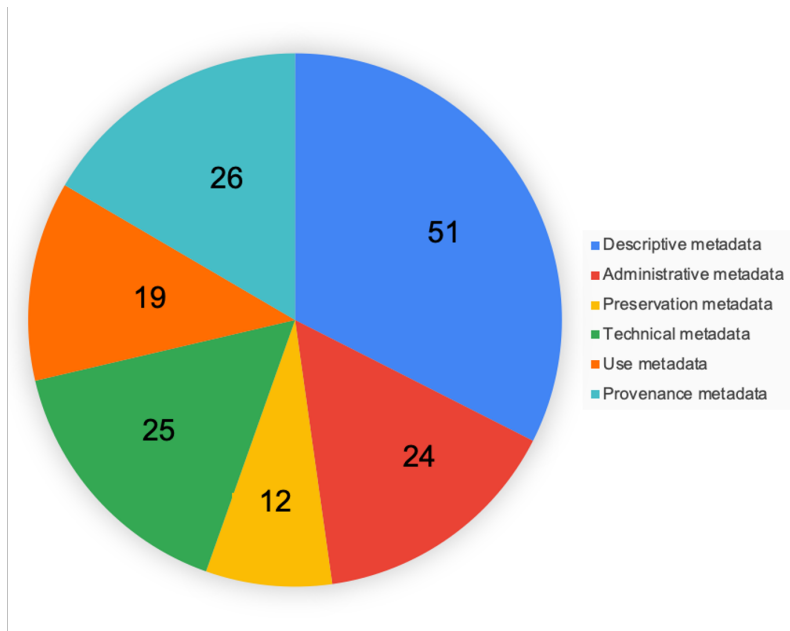


Figure 6: Kinds of metadata reported as used to describe the developed LD resources.

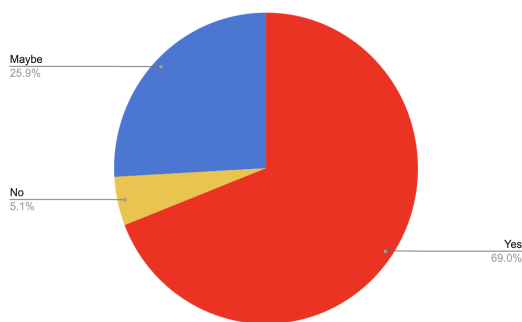


Figure 7: Distribution of participants who declared themselves interested in (red, 69%), not interested in (orange, ≈5%) and hesitant (blue, ≈69%) about a service/tool that would help provide metadata for LD resources.

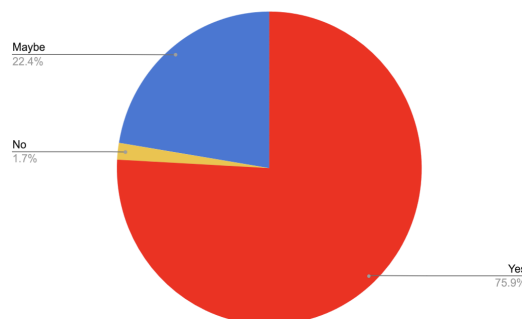


Figure 8: Distribution of participants who declared themselves interested in (red, ≈76%), not interested in (orange, ≈2%) and hesitant (blue, ≈22%) about a service/tool that would help improve creating/convertig metadata?

Table 4: List of used vocabularies and times mentioned.

Vocabulary	#
DublinCore	8
METAShare	4
DCAT	3
Ontolex	3
CLARIN	2
LIME	2
Lexinfo	1
Wiki Vocabularies	1
IMDI	1
Preservica	1
EDM	1
Eurovoc	1
Prov Ontology	1
DDML	1
DataID	1
VoID	1
DPV	1
http://drmj.eu/	1

The **use** or non-use of LLD resources is highly correlated with their declared background: as shown in Figure 9, most (95%) computer scientists, many (77%) computational linguists, but only a third (33%) of the linguists used LD resources before.

We notice the same tendency when correlating the involvement of the participants in LLD resources **development** with their background: many (74%) computer scientists, a little more than half (56%) of the computational linguists, but only almost a third (30%) of the linguists were involved in the development of LLD resources. This distribution is rendered in Figure 10.

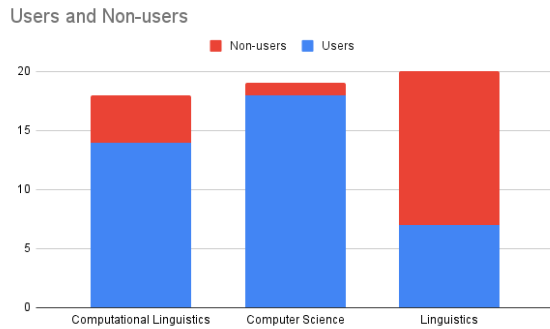


Figure 9: Participants as users of LLD resources (in blue) and non-users of LLD resources (in red), according to their declared background.

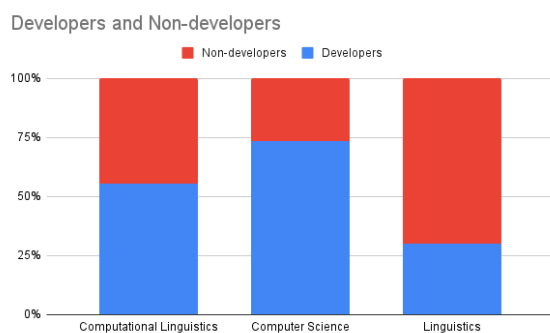


Figure 10: The involvement (in blue) and lack of involvement (in red) in the development of LD resources of participants according to their declared background.

Representing data in LLD format requires programming skills (Marshall and Shipman, 2003), which linguists rarely have. Thus, when asked under what conditions they would develop or convert LD resources more frequently, 71% of the participants mentioned the existence of user-friendly tools to help them do this. The creation of such tools, however, might come with a cost: while easing the job of those less skilled in programming, such tools may work only for some domains or contexts, given the different nature of the data to be represented in various fields (Marshall and Shipman, 2003).

Looking at the background of those who publish or do not publish resources, we notice that computational linguists tend to publish the resources they develop more than linguists, while computer scientists tend not to do so (see Figure 11).

With respect to the relation between the background of the 28 non-developers of LLD resources and the reasons for not developing such resources, we find the data in Table 5, where we show the

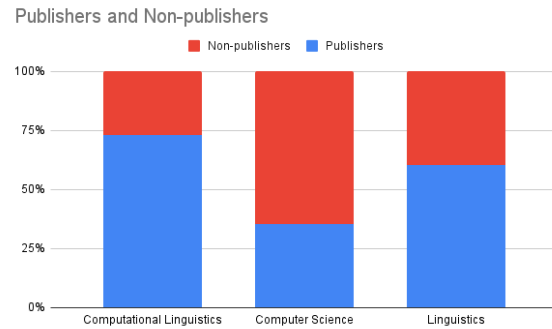


Figure 11: The correlation between the background and the tendency to publish LL(O)D resources.

distribution of participants according to their declared background¹². We can see that incompatibility between LD resources and other resources is a problem, especially for linguists (theoretical or computational), while rarely do computer scientists have it. The other reasons are invoked by members of all communities.

6 Conclusion

In this paper, we presented the results of a survey we conducted within the NL CA – WG1 to collect information useful to support the penetration of LLD, the identification of causes preventing such penetration, and possible limitations of such resources/technologies.

What emerged is that some bounds have to be overstepped in order to spread LLD principles to a wider community and fully exploit their potential. This survey results come as a confirmation of what the LLD community has already been aware of, thus reaffirming the need to take action.

We need to promote knowledge and skill transfer to support linguists in acquiring the necessary competencies for adopting LLD principles and technologies to their resources. On the other hand, the engagement of computer scientists in sharing knowledge and data as early as possible in the research process in open collaboration with all relevant knowledge actors (Von Schomberg, 2019) could contribute to support *open scholarship*¹³.

¹²One of the 14 participants mentioning incompatibility with other tools/resources as reason declared cognitive science as his/her background and this is not rendered in the table.

¹³We adopt the term *open scholarship* instead of *open science* to adhere to the European policies, directed toward “open scholarship”, as “open scholarship” reflects the inclusion of the humanities in the equation as well as emphasising the open input side to science in the form of open collaboration and active data and knowledge sharing prior to publishing and

Reason	Total #	#CS	#CL	#Ling
Incompatibility with other tools/resources used	14	1	5	7
Lacking knowledge about adequate model/vocabulary	6	–	2	4
Models not totally appropriate for representing data	5	2	3	–
Unclear example or guidelines	8	2	2	3
Unclear advantages that LD has over other formats	3	1	2	–
Other	4	2	–	2

Table 5: Reasons for not developing LD resources correlated with participants' background. CS = computer science, CL = computational linguistics, Ling = linguistics.

At the same time, easing the (re)use, the creation, and the exposure of such resources could spread the adoption of LLD. This goal can be achieved through the development of specific adaptive tools, able to support different domains and languages, as well as formats to facilitate resource exchange and integration.

Furthermore, existing resources suffer from not being easily accessible, both in terms of findability, mostly due to the lack of harmonised and full-informative metadata descriptions, and usability, as LLD documentation is reported as scarce and inadequate.

With reference to the use of metadata, the current scenario could be improved by the availability of (semi)automatic solutions to reduce the time and effort for enriching resources manually, providing useful and consistent descriptions.

The documentation limits also affect the creation of new resources, preventing the adoption of LLD vocabularies/models to formalise linguistic data. This issue could be addressed by ensuring updated and maintained guidelines, enhanced by different examples and use cases and tailored to different backgrounds and levels of expertise, to support also less expert contributors/users through the whole cycle of linguistic linked datafication of their resources.

In future work, we intend to provide our contribution to defining some of the requirements to meet in order to ensure a large adoption of LL(O)D principles and promote a collaborative evolution of such resources.

Acknowledgment

This work has been carried out within the COST Action CA 18209 European network for Web-centred linguistic data science (Nexus Linguarum). Maria Pia di Buono has been supported by Fondo FSE/REACT-EU - Progetti DM 1062

other scientific open outputs (Burgelman et al., 2019).

del 10/08/2021 "Ricercatori a Tempo Determinato di tipo A) (RTDA)". Azione IV.4 - Dottorati e contratti di ricerca su tematiche dell'innovazione/Azione IV.6 - Contratti di ricerca su tematiche Green.

The authors thank Julia Bosque Gil, Liudmila Rychkova and Max Ionov for their contribution to the survey drafting.

References

- Julia Bosque-Gil, Verginica Barbu Mititelu, Hugo Gonçalo-Oliveira, Max Ionov, Jorge Gracia, Liudmila Rychkova, Giedre Valunaite-Oleskeviciene, Christian Chiarcos, Thierry Declerck, and Milan Dojchinovski. 2022. Balancing the digital presence of languages in and for technological development. A Policy Brief on the Inclusion of Data of Under-resourced Languages into the Linked Data Cloud. DOI: 10.5281/zenodo.7142513.
- Jean-Claude Burgelman, Corina Pascu, Katarzyna Szkuta, Rene Von Schomberg, Athanasios Karalopoulos, Konstantinos Repanas, and Michel Schouppe. 2019. Open science, open data, and open scholarship: European policies to make science fit for the twenty-first century. *Frontiers in Big Data*, 2:43.
- Christian Chiarcos, Bettina Klimek, Christian Fäth, Thierry Declerck, and John Philip McCrae. 2020. On the Linguistic Linked Open Data infrastructure. In *Proceedings of the 1st International Workshop on Language Technology Platforms*, pages 8–15.
- Maria Pia di Buono, Hugo Gonçalo Oliveira, Verginica Barbu Mititelu, Blerina Spahiu, and Genaro Nolano. 2022. Paving the way for Enriched Metadata of Linguistic Linked Data. *Semantic Web Journal*, 13(6):1133–1157.
- Margaret R Geddes. 2019. Strategies to support wider adoption of Linked Open Data in smaller museums. Johns Hopkins University.
- Fahad Khan, Christian Chiarcos, Thierry Declerck, Maria Pia Di Buono, Milan Dojchinovski, Jorge Gracia, Giedre Valunaite Oleskeviciene, and Daniela Gifu. 2022. A survey of guidelines and Best Practices for the Generation, Interlinking, Publication,

- and Validation of Linguistic Linked Data. In *Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference*, pages 69–77, Marseille, France. European Language Resources Association.
- Leonardo Lezcano, Salvador Sánchez-Alonso, and Antonio J Roa-Valverde. 2013. A survey on the exchange of linguistic resources: Publishing Linguistic Linked Open Data on the Web. *Program*, 47(3).
- Catherine C. Marshall and Frank M. Shipman. 2003. Which Semantic Web? In *Proceedings of the 14th ACM conference on Hypertext and Hypermedia*, pages 57–66.
- Max Schmachtenberg, Christian Bizer, and Heiko Paulheim. 2014. Adoption of the Linked Data best practices in different topical domains. In *The Semantic Web–ISWC 2014: 13th International Semantic Web Conference, Riva del Garda, Italy, October 19–23, 2014. Proceedings, Part I 13*, pages 245–260. Springer.
- Blerina Spahiu, Andrea Maurino, and Robert Meusel. 2019. Topic profiling benchmarks in the linked open data cloud: Issues and lessons learned. *Semantic Web*, 10(2):329–348.
- Rene Von Schomberg. 2019. Why responsible innovation? In *International handbook on responsible innovation*, pages 12–32. Edward Elgar Publishing.
- Anneke Zuiderwijk, Keith Jeffery, and Marijn Janssen. 2012. The potential of metadata for Linked Open Data and its value for users and publishers. *JeDEM—Journal of eDemocracy and Open Government*, 4(2):222–244.

GPT3 as a Lexical Knowledge Base for Portuguese?

Hugo Gonalo Oliveira

University of Coimbra DEI,
CISUC Coimbra, Portugal
hroliv@dei.uc.pt

Ricardo Rodrigues

Polytechnic Institute of Coimbra, ESEC
CISUC
Coimbra, Portugal
rmanuel@dei.uc.pt

Abstract

We test the GPT3 language model in zero- and few-shot acquisition of lexico-semantic knowledge in Portuguese, with simple instruction prompts, and compare it with a BERT-based approach. Results are assessed in two test sets: TALES and the Portuguese translation of BATS. GPT3 outperforms BERT in all relations, with the few-shot approach being the best overall and for the majority of relations. Scores in both datasets further suggest that, despite their different creation approaches, they are equally suitable for this kind of evaluation.

1 Introduction

Large Language Models (LLMs) (Petroni et al., 2019) have been exploited in the acquisition of semantic relations, and as potential knowledge bases. When considering lexico-semantic relations, such models could be seen as alternatives to word-nets (Fellbaum, 1998).

BERT (Devlin et al., 2019), a bidirectional LLM pretrained on the masked language modelling task, is the most explored model in previous works, with fewer having explored GPT models (Radford et al., 2019; Brown et al., 2020). However, GPT3 is known for its adaptation to many tasks, often without requiring additional training, in zero- or few-shot approaches.

We take steps on the exploration of GPT3 for acquiring lexico-semantic knowledge in Portuguese, which contributes to better understanding this black-box model and to conclusions on its potential as a lexical knowledge base. Lexico-semantic relations are obtained through instruction-like prompts, in both zero- and few-shot learning scenarios. Performance is compared with previously used methods based on lexical patterns and masked language modelling with BERT (Gonalo Oliveira, 2023). Experiments are performed in two analogy test sets, TALES (Gonalo Oliveira et al., 2020) and a recent translation of the Bigger Analogy

Test Set (BATS) (Gladkova et al., 2016) to Portuguese (hereafter, BATS-PT). Reported experiments are the first using the latter dataset, so we also look at differences between BATS-PT, resulting from manual translation, and TALES, created automatically from lexical resources in Portuguese.

Despite the simple and direct prompts used in GPT3, the BERT-based approach was outperformed overall and for every relation, with the best performance achieved by the few-shot approach. Moreover, scores in BATS-PT and TALES were not much different, which suggests that, despite their different creation approaches, they are equally suitable for this kind of evaluation.

The remainder of the paper is structured as follows: Section 2 overviews work on relation acquisition and analogy solving; Section 3 describes the experimentation setup; Section 4 reports and discusses the results; Section 5 concludes it.

2 Related Work

Semantic relations have been obtained from pretrained word embeddings, with simple analogy solving methods, such as: the vector offset with a single example (Mikolov et al., 2013); the average offset or a classifier of related words learned from a set of examples (Gladkova et al., 2016). These were assessed in the then proposed BATS, a test set that covers several relations types, including lexico-semantic relations.

More recently, semantic relations were obtained from Transformer-based LLMs, by prompting models with handcrafted (Petroni et al., 2019; Ushio et al., 2021) or induced lexical patterns (Bouraoui et al., 2020), in some cases (Bouraoui et al., 2020; Ushio et al., 2021) also assessed in BATS.

Pretrained models are generally used, as knowledge tends to be forgotten during the fine-tuning process (Wallat et al., 2020). Much work exploits BERT, by taking advantage of masked language modelling for acquiring relations with cloze-

style prompts (e.g., Paris is the capital of [MASK]). GPT, another popular model, has not been so explored, also due to access limitations. Yet, there are examples using models of this family: an approach based in GPT2 (Radford et al., 2019) outperformed BERT and other LLMs in BATS (Ushio et al., 2021); a method (Liu et al., 2021) was proposed for searching for the best prompts when acquiring semantic relations with GPT2; and, among many tasks, GPT3 (Brown et al., 2020) was originally tested on a dataset of 374 analogies in English, in zero- and few-shot scenarios.

Lexico-semantic relations are especially challenging to acquire and to assess because, in opposition to morphological and to several encyclopedic relations (e.g., capitalOf, hasCurrency), they are not functions (e.g., a concept often has many hyponyms or parts). For Portuguese, related work has focused on these relations: word embeddings were exploited for enriching OpenWordNet-PT (Gonçalo Oliveira et al., 2021); BERT was used for the detection of hyponymy pairs (Paes, 2021), and for completing a range of lexico-semantic relations (Gonçalo Oliveira, 2023). The latter was assessed in TALES, similar to BATS, but for Portuguese. Previous work for Portuguese (Gonçalo Oliveira, 2022) has also suggested that GPT2 was not a good option for validating instances of lexico-semantic relations, and BERT would be better suited.

3 Experimentation Setup

This section describes the datasets and models used in this work, the approach for testing GPT3, and the adopted evaluation metrics.

3.1 Datasets

BATS comprises 40 files, each targeting a different linguistic relation. Each file has 50 entries, with two columns: a source word and a list of target words, related to the former by the relation specified in the filename.

Relations are organised in four groups: grammatical inflections, word-formation, lexicographic and encyclopedic relations. BATS was originally created for English, but the files of the ten lexicographic relations have recently been translated into several languages, in the scope of a use case in the NexusLinguarum COST Action¹. These files comprise: *hypernyms* (*animals*, *miscellaneous*);

hyponyms, *meronyms* (*whole-substance*, *member-group*, *whole-part*); *synonyms* (*intensity*, *exact*); *antonims* (*gradable*, *binary*). We use the Portuguese translation of BATS. Table 1 illustrates this dataset with one line for each covered relation, its original BATS identifier, and an example entry.

TALES is a similar test set, but created automatically, based on the most frequent relations and their instances in ten Portuguese lexical resources. It adopts the same format as BATS, but covering 14 lexico-semantic relations, which are not exactly the same: *has-hypernym* and *hypernym-of*, each between abstract nouns, concrete nouns, and verbs; *part-of*, *has-part*; *purpose-of*, *has-purpose*; *synonym* (nouns, verbs, and adjectives); *antonym* (adjectives).

Both BATS and TALES can be used for assessing language models in the acquisition of lexico-semantic knowledge, based on predicting the target words for a given source.

3.2 Models

Two transformer models were used for acquiring lexico-semantic relations in Portuguese. GPT3 is an auto-regressive LLM with 175B parameters, 96 attention layers and a 3.2M batch size. We have used the *text-davinci-003* engine, available through the OpenAI API². GPT3 is known to be multilingual, and may thus be prompted in Portuguese for generating text in this language. Temperature was set to 0.1, to force the model to produce the most probable sequences, and to avoid a non-deterministic behaviour. The results of GPT3 are compared to those by BERTimbau-large (Souza et al., 2020), a BERT model pretrained for Brazilian Portuguese, with 24 layers and 335M parameters, which can be seen as a baseline.

3.3 Approach

GPT3 was used in two scenarios in which it is known to perform well: zero-shot, where the model was prompted with an instruction that included the source word; and few-shot, where a similar prompt was concatenated to the same instruction instantiated for five examples of the same type, each followed by the respective list of target words. We used simple generic instructions asking for ten related words and changed the relation name accordingly (see Table 2). Since GPT3 is very flexible with its prompts, we did not put much effort on

¹<https://nexuslinguarum.eu/>

²<https://openai.com/api>

ID	Relation	Entries
L01	Hypernyms (animals)	anaconda cobra/réptil/boa/serpente/ofídio (<i>anaconda snake/reptile/boa/serpent/ophidian</i>)
L02	Hypernyms (misc)	banheira contentor/artefacto/unidade/objeto/... (<i>tub container/artefact/unit/object</i>)
L03	Hyponyms	igreja capela/abadia/basilica/catedral (<i>church chapel/abbey/basilica/cathedral</i>)
L04	Meronyms (whole-substance)	atmosfera gás/oxigénio/hidrogénio/nitrogénio/... (<i>atmosphere gas/oxygen/hydrogen/nitrogen</i>)
L05	Meronyms (member-group)	pássaro bando (<i>bird flock</i>)
L06	Meronyms (whole-part)	academia faculdade/universidade/instituto (<i>academia college/university/institute</i>)
L07	Synonyms (intensity)	choro grito/chio/guincho/berro/pranto (<i>cry scream/shriek/screech</i>)
L08	Synonyms (exact)	fazenda tecido/têxtil/pano (<i>cloth fabric/material/textile</i>)
L09	Antonyms (gradable)	capaz cobra/réptil/boa/serpente/ofídio (<i>able unable/incompetent/unequal</i>)
L10	Antonyms (binary)	anterior posterior (<i>anterior posterior</i>)

Table 1: Example entries in the Portuguese BATS files and their English translation (original).

their tuning, and leave this for future work. Still, we empirically discovered that prompts should specifically ask for Portuguese words, otherwise we would risk that, for some entries, GPT3 generates words in other languages, often Spanish. Moreover, including the number of required answers, in this case, 10, conditions the model to generate a numbered list of this size, in any case, easy to parse. Since the number of target words in the dataset is variable and it would be incoherent to give examples asking for ten but followed by a different number, we drop the 10 from the instructions in the few-shot approach.

The BERT approach followed [Gonçalo Oliveira \(2023\)](#) closely. BERT was prompted with a set of masked lexical patterns indicative of the target relations — e.g., a [MASK] é um tipo de <s> (in English, [MASK] is a type of <s>) for hyponyms. For TALES, we relied on the same patterns³, also used for relations in BATS-PT. We only had to make a few additions to the *part-of* patterns, to better cover the *whole-substance* and *member-group* sub-types.

Differently from previous work, instead of looking at individual performances for each pattern, we add a “training” step where the best patterns are selected for each relation. The final top-10 predictions result from ranking the top-10 predictions of each of the top-5 patterns, considering their overall scores, given by the model — if there were patterns

ex aequo, more than 5 patterns could be selected, which happened in some cases.

In order to select the best patterns, datasets were split into training and test. This had been done in BATS, for instance, by [Bouraoui et al. \(2020\)](#), who opted for 90%–10%, and by [Rezaee and Camacho-Collados \(2022\)](#), 50%–50%. The latter was our option: one half of the entries was assigned to the train portion, and the other to the test.⁴ A 90%–10% split was not seen as an option because testing in only five examples (10%×50) of each relation would be too narrow for any conclusions.

Splitting the dataset was not necessary for GPT3 but, for comparison over the same data, we also run GPT3 in the test portion only. Moreover, in the few-shot scenario, the five given examples were randomly selected from the training portion, which introduced some variability in the prompts.

3.4 Metrics

Accuracy (Acc) is a common metric for assessing analogy solving in datasets like BATS. It computes the proportion of source words for which the first prediction is one of the targets. Since this is too restrictive for most lexico-semantic relations, we also compute the more relaxed Accuracy@10 (Acc@10) — i.e., the proportion of source words for which one of the targets is among the top-10 predictions; and the Mean Average Precision@10 (MAP@10), which, considering that there may be

³BERTimbau patterns for TALES are available from https://github.com/NLP-CISUC/PT-LexicalSemantics/blob/master/Patterns/BERT_patterns_for_TALES_v2.txt

⁴For reproducibility, we make the TALES splits available at https://github.com/NLP-CISUC/PT-LexicalSemantics/tree/master/TALESv1.1_splits.

ID	Prompt
L01/L02	lista 10 hiperónimos, em português, da palavra <s>:
L03	lista 10 hipónimos, em português, da palavra <s>:
L04	lista 10 substâncias, em português, da palavra <s>:
L05	lista 10 conjuntos ou grupos, em português, da palavra <s>:
L06	lista 10 partes, em português, da palavra <s>:
L07/L08	lista 10 sinónimos, em português, da palavra <s>:
L09/L10	lista 10 antónimos, em português, da palavra <s>:

Table 2: Prompts used for relation acquisition from GPT3. Each translates to *list 10 <r>, in Portuguese, of the word <s>:*, where *r* is a name typically given to the related words, and *s* is the source word.

more than one correct answer in the top-10, accounts for the number of predicted target words and their ranking.

4 Results and Discussion

Tables 3 and 4 report on the scores of the three tested approaches, respectively in BATS-PT and in TALES. Scores are presented for each relation and as an average of all. In addition to zero- and few-shot with GPT3, we tested three variations of the BERT approach, with the best patterns optimised for each metric. However, since differences were minimal, we present only the scores of the patterns optimised for accuracy.

The few-shot approach is clearly the best in both datasets. In BATS-PT, it achieves the best performance in every relation in each of the three metrics, except for meronyms (member-group), with the best scores in two metrics, and for synonyms (intensity) and antonyms (gradable), with the best score in only one (in *ex aequo*). In TALES, the results are quite similar. Only in a handful of cases few-shot is outperformed by zero-shot (or has the same score), and fewer yet by BERT. Surprisingly, despite no training nor prompt tuning, zero-shot GPT3 is better than BERT for almost every relation and metric.

Performance is variable across relation types. In BATS-PT, *hypernyms (animals)* is one of the best relations for all approaches, whereas zero- and few-shot perform equally well for *antonyms (gradable)*. Lowest performances by few-shot are for *meronyms (member-group)* and *synonyms (intensity)*, the same as for the zero-shot. Specifically in the *member-group* relation, we observe some confusion with hypernymy and co-hyponymy (e.g., *parlamentar [parliamentarian]* and *legislador [legislator]* for *senador [senator]*) and, for zero-shot, answers that are groups of other things (e.g., *rebanho [herd]* or *matilha [pack]*, for *pássaro [bird]*). In few-shot,

however, shorter lists are generated, often with less or no incorrect answers.

In TALES, all approaches perform especially well for *antonyms*, and zero-shot achieves top-performance in *synonyms (verbs)*. The other synonymy relations are among the top-performing in few-shot, whereas the best performance of BERT is for *has-hypernym (abstract)*.

We highlight that the average scores in BATS-PT are not substantially different from those in TALES. Overall, few-shot performs slightly better in BATS-PT, and zero-shot in TALES. BERT is very similar in both test sets. Moreover, there is a similar trend for equivalent relations: models generally perform better for *antonymy* and *hypernymy*, and worse for *meronymy*. BATS-PT was not originally created for Portuguese, but it is the result of thorough manual translation, whereas TALES was created specifically for Portuguese, but automatically. To some extent, this validates the approach adopted for creating TALES. But it does not mean that any of the datasets cannot be improved. In fact, low scores in TALES' *has-part* and *part-of* relations can be partially explained by limitations of the dataset. TALES is based on redundancy across lexical resources and the following reasons may result in less consensual and incomplete entries: (i) to reach the 50 entries, *has-part* and *part-of* are the relations for which required redundancy was the lowest (Gonçalo Oliveira et al., 2020); (ii) there are several sub-types of meronymy, defined differently across resources.

Reference scores for TALES (Gonçalo Oliveira, 2023) use the same BERT-based approach, but in the full dataset, without combining patterns. Though not comparable, differences suggest that the combination of patterns is not always beneficial. Yet, the best patterns have to be selected from part of the data. Moreover, we should add that, with only 50 entries, the train-test split has a noticeable impact on the selection of patterns and on the

Relation	BERT			GPT3 (zero-shot)			GPT3 (five-shot)		
	Acc	Acc@10	MAP@10	Acc	Acc@10	MAP@10	Acc	Acc@10	MAP@10
L01	0.76	0.92	0.60	0.84	0.96	0.79	1.00	1.00	0.93
L02	0.42	0.88	0.35	0.29	0.71	0.43	0.42	0.96	0.62
L03	0.21	0.50	0.24	0.46	0.58	0.44	0.50	0.67	0.50
L04	0.24	0.60	0.31	0.20	0.36	0.24	0.52	0.68	0.53
L05	0.08	0.44	0.19	0.04	0.08	0.06	0.20	0.28	0.24
L06	0.00	0.20	0.04	0.20	0.40	0.22	0.32	0.64	0.38
L07	0.08	0.12	0.09	0.36	0.72	0.44	0.20	0.72	0.43
L08	0.12	0.32	0.14	0.48	0.68	0.50	0.60	0.92	0.73
L09	0.32	0.48	0.35	0.76	0.88	0.71	0.72	0.80	0.71
L10	0.48	0.78	0.48	0.57	0.61	0.57	0.74	0.83	0.73
Average	0.27	0.52	0.28	0.42	0.60	0.44	0.52	0.75	0.58

Table 3: Performance in BATS-PT.

Relation	BERT			GPT3 (zero-shot)			GPT3 (five-shot)		
	Acc	Acc@10	MAP@10	Acc	Acc@10	MAP@10	Acc	Acc@10	MAP@10
Antonyms (adjectives)	0.48	0.52	0.42	0.76	0.88	0.80	0.96	0.96	0.94
Purpose-of	0.16	0.20	0.18	0.20	0.32	0.23	0.40	0.40	0.35
Has-Purpose	0.16	0.36	0.23	0.32	0.64	0.45	0.60	0.60	0.60
Has-Hypernym (abstract)	0.44	0.80	0.35	0.44	0.80	0.50	0.80	0.80	0.45
Has-Hypernym (concrete)	0.24	0.64	0.24	0.40	0.68	0.43	0.80	0.80	0.49
Has-Hypernym (verbs)	0.08	0.48	0.20	0.60	0.88	0.62	0.92	0.92	0.57
Hypernym-Of (abstract)	0.20	0.68	0.29	0.36	0.68	0.39	0.68	0.68	0.40
Hypernym-of (concrete)	0.48	0.88	0.50	0.48	0.72	0.52	0.96	0.96	0.74
Hypernym-Of (verbs)	0.04	0.48	0.17	0.52	0.84	0.56	0.76	0.76	0.47
Has-Part	0.16	0.36	0.22	0.08	0.12	0.09	0.36	0.36	0.26
Part-Of	0.08	0.40	0.14	0.16	0.20	0.17	0.48	0.48	0.34
Synonyms (nouns)	0.24	0.72	0.33	0.60	0.92	0.65	1.00	1.00	0.74
Synonyms (verbs)	0.32	0.76	0.33	0.56	0.96	0.56	0.56	0.96	0.54
Synonyms (adjectives)	0.20	0.76	0.28	0.48	0.72	0.47	0.96	0.96	0.67
Average	0.23	0.57	0.28	0.43	0.67	0.46	0.73	0.76	0.54

Table 4: Performance in TALES.

performance achieved for some relations.

There are no reference scores for BATS-PT, but there are for the English version, where the accuracy reported by Ushio et al. (2021) is 81% with GPT2, substantially higher than few-shot’s 56% in BATS-PT. Despite GPT3 being more powerful, the lower performance is a consequence of a simpler approach, and suggests that there is room for improvement, for instance, if we invest in prompt tuning. Yet, languages are different, and BATS-PT may have resulted in a more challenging dataset, for a less-resourced language.

5 Conclusions

We have seen that, to some extent, GPT3 can be used as a lexical knowledge base for Portuguese. When compared to handcrafted knowledge bases, the coverage of GPT3 is difficult to meet. Moreover, performance is variable across relations, but this also happens for automatically created knowledge bases. GPT3 clearly outperformed a BERT-based approach, which had shown improvements against approaches based on static word embed-

dings (Gonalo Oliveira, 2023). The best performance is achieved with a few-shot approach with simple direct prompts, without previous tuning, which suggests that there is still room for improvement.

This was also the first time that BATS-PT was used as a benchmark. The fact that the scores achieved were comparable to those in TALES, despite its automatic creation, contributes to validating the utility of both datasets.

Future directions would be to test alternative prompts and to experiment with more recent LLMs, such as the recently release GPT4 (OpenAI, 2023). However, we should not forget that GPT is a black-box architecture, which prevents a deeper analysis and a direct fix of its errors. This adds to the fact that we know that GPT3 and GPT4 were trained in much data, but not exactly on which data, which may raise relevant questions for evaluation — e.g., did it learn from the test examples? While it cannot have learned from BATS-PT, because the dataset has not been released yet, we may question whether it learned from the original dataset, which, through deep inference, may help with other languages.

Acknowledgements: We would like to thank the remaining translators of BATS to Portuguese, Purificação Silvano and Sara Carvalho. This work was based upon work in COST Action CA18209 Nexus Linguarum, supported by COST (European Cooperation in Science and Technology). <http://www.cost.eu/>; and partially supported by the Portuguese Recovery and Resilience Plan (PRR) through project C645008882-00000055, Center for Responsible AI.

References

- Zied Bouraoui, Jose Camacho-Collados, and Steven Schockaert. 2020. Inducing relational knowledge from BERT. In *Proc of AAAI Conference on Artificial Intelligence*, pages 7456–7463. AAAI Press.
- Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings 2019 Conference of North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186. ACL.
- Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.
- Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuo. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Procs of NAACL 2016 Student Research Workshop*, pages 8–15. ACL.
- Hugo Gonalo Oliveira. 2023. On the acquisition of WordNet relations in Portuguese from pretrained masked language models. In *Procs of 12th Global WordNet Conference, GWC*. Global WordNet Association.
- Hugo Gonalo Oliveira, Tiago Sousa, and Ana Alves. 2020. TALES: Test set of Portuguese lexical-semantic relations for assessing word embeddings. In *Procs of ECAI 2020 Workshop on Hybrid Intelligence for Natural Language Processing Tasks (HI4NLP 2020)*, volume 2693 of *CEUR Workshop Proceedings*, pages 41–47. CEUR-WS.org.
- Hugo Gonalo Oliveira. 2022. Exploring transformers for ranking Portuguese semantic relations. In *Procs of the 13th Language Resources and Evaluation Conference, LREC 2022*, pages 2573–2582, Marseille, France. ELRA.
- Hugo Gonalo Oliveira, Fredson Silva de Souza Aguiar, and Alexandre Rademaker. 2021. On the Utility of Word Embeddings for Enriching OpenWordNet-PT. In *Procs of 3rd Conference on Language, Data and Knowledge (LDK 2021)*, volume 93 of *OASICs*, pages 21:1–21:13, Dagstuhl, Germany. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.
- Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021. GPT understands, too. *arXiv preprint arXiv:2103.10385*.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Procs of Workshop track of ICLR*.
- OpenAI. 2023. [GPT-4 Technical Report](#). ArXiv:2303.08774 [cs].
- Gabriel Escobar Paes. 2021. Deteco de hiperônimos com bert e padres de hearst. Master's thesis, Universidade Federal de Mato Grosso do Sul.
- Fabio Petroni, Tim Rocktschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proc 2019 Conf on Empirical Methods in Natural Language Processing and 9th Intl Joint Conf on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473. ACL.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.
- Kiamehr Rezaee and Jose Camacho-Collados. 2022. Probing relational knowledge in language models via word analogies. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3930–3936.
- Fbio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: Pretrained BERT models for Brazilian Portuguese. In *Procs of Brazilian Conf on Intelligent Systems (BRACIS 2020)*, volume 12319 of *LNCS*, pages 403–417. Springer.
- Asahi Ushio, Luis Espinosa Anke, Steven Schockaert, and Jose Camacho-Collados. 2021. [BERT is to NLP what AlexNet is to CV: Can pre-trained language models identify analogies?](#) In *Procs of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3609–3624, Online. Association for Computational Linguistics.
- Jonas Wallat, Jaspreet Singh, and Avishek Anand. 2020. [BERTnesia: Investigating the capture and forgetting of knowledge in BERT](#). In *Procs of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 174–183, Online. Association for Computational Linguistics.

A uniform RDF-based Representation of the Interlinking of Wordnets and Sign Language Data

Thierry Declerck¹, Sam Bigeard², Dorianne Callus³, Benjamin Matthews³,
Sussi Olsen⁴, Loran Ripard Xuereb³

¹DFKI GmbH, Multilingual Technologies, Saarland Informatics Campus, D-66123 Saarbrücken, Germany

²Institute of German Sign Language and Communication of the Deaf University of Hamburg, Germany

³Institute of Linguistics and Language Technology, University of Malta, Malta

⁴Centre for Language Technology, NorS, University of Copenhagen, Denmark

declerck@dfki.de, sam.bigeard@uni-hamburg.de, dcall01@um.edu.mt,
saolsen@hum.ku.dk, benjamin.matthews@um.edu.mt,
contact for Loran Ripard Xuereb: benjamin.matthews@um.edu.mt

Abstract

We present ongoing and incremental work dealing with a Linked Data compliant representation of approaches using wordnets and possibly other lexical data, as representative semantic resources for the description of Spoken Language (SpL), for linking multilingual Sign Language (SL) data sets. The base for our work is given by data sets produced by the European EASIER research project, which makes use of shared IDs of the Open Multilingual Wordnet (OMW) infrastructure for linking SL glosses and basic lexical information associated with three SL data sets: British, German and Greek. We transformed the EASIER data sets onto RDF and OntoLex representations. We acted similarly with a Danish data set, which links Danish SL data and the wordnet for Danish. This transformation work was extended to other Nordic wordnets, aiming at supporting cross-lingual comparisons of Nordic SLs. We started recently work on the Maltese Sign Language Dictionary, with the challenge, that no Maltese wordnet is available for linking LSM to other SLs. The final objective of our work is to include SL data sets (and their conceptual cross-linking via wordnets, but also via other SpL lexical resources) in the Linguistic Linked Open Data cloud.

1 Introduction

Our work is pursued in the context of an initiative aiming at representing and publishing Sign Language (SL) data sets in the Linguistic Linked Data (LLOD) cloud, which is a subset of the Linked Open Data (LOD) cloud.¹ We can observe that SL data are not represented in the data sets currently

¹Those clouds can be accessed respectively at <http://linguistic-lod.org/llod-cloud> and <https://lod-cloud.net/>

included in the LLOD cloud. Also the “Overview of Datasets for the Sign Languages of Europe” published by the EASIER European project (Kopf et al., 2022)² does not mention any SL data set being available in a Linked Data compliant format.

We see in this a gap that needs to be bridged, as an important type of natural language is missing from the LLOD, while the motivation behind the creation of this infrastructure is that it can ease the linking of all types of natural language resources.³

The prerequisite for publishing linguistic data in the LLOD cloud is to have it formally represented within the Resource Description Framework (RDF).⁴ And as a de facto standard for representing lexical information in RDF, the OntoLex-Lemon specifications,⁵ already exist, we investigate the re-use of those specifications in order to accommodate the description and the publication of Sign Language data sets in the LLOD. Figure 1 displays the core module of OntoLex-Lemon.

A first experiment in representing SL data within RDF and OntoLex-Lemon was building on top of an approach consisting in using wordnets for interlinking British, German and Greek SL data, as originally described in Bigeard et al. (2022).⁶ This approach makes use of shared IDs

²Available as a public deliverable at <https://www.project-easier.eu/deliverables/>

³See (Chiarcos et al., 2012) for a first description of the motivations leading to the creation of the LLOD, and (Cimiano et al., 2020) for a more recent and much more detailed description of all aspects of the LLOD infrastructure.

⁴See <https://www.w3.org/TR/rdf11-primer/> for an introduction to RDF.

⁵See <https://www.w3.org/2016/05/ontolex/> and (McCrae et al., 2017).

⁶The data set was created in the context of the European project EASIER (<https://www.project-easier.eu/>). It is available at https://www.sign-lang.uni-hamburg.de/easier/sign-wordnet/index_core_synsets.html

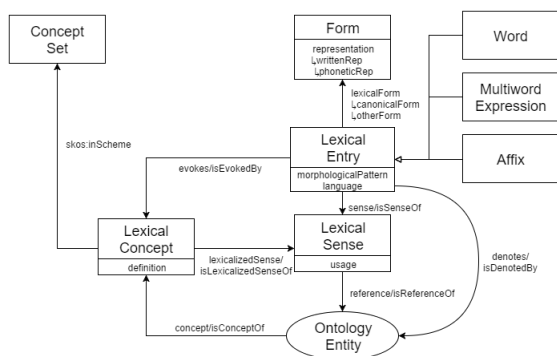


Figure 1: The core module of OntoLex-Lemon, taken from <https://www.w3.org/2016/05/ontolex/>

of the Open Multilingual Wordnet (OMW)⁷ infrastructure as a base for interlinking SL data sets.

The OntoLex-Lemon model is also therefore a good candidate for our work, as it supports the representation of WordNet data, which are encoded with the SKOS⁸ vocabulary, where the WordNet synsets are encoded as instances of the `ontolex:LexicalConcept` subclass of the `skos:Concept` class.⁹ This feature offers a good starting point for transforming into RDF and OntoLex-Lemon the EASIER data sets.

Declerck et al. (2023) presents a first RDF- and OntoLex-based representation of such interlinking of OMW and SL data. Dealing with the languages covered by EASIER, adding to it French (see Section 5) and Danish (see Section 6), while starting to work also on other Nordic Languages (Declerck and Olsen, 2023).¹⁰ We describe in this paper those stages of our incremental work, and we also introduce the most recent data set we started to work on, the Maltese Sign Language Dictionary (LSM), with a new challenge, as we cannot refer to a Maltese wordnet for cross-linking the Maltese signs to the signs of other SLs. LSM is introduced in Section 7.

⁷See (Bond and Paik, 2012) and (Bond and Foster, 2013) for more details on the Open Multilingual Wordnet and the interlinking between OMW data sets.

⁸SKOS stands for “Simple Knowledge Organization System”. see <https://www.w3.org/TR/skos-primer/> for more details.

⁹See for example (Declerck, 2019).

¹⁰A general overview of Nordic Sign Languages is given in Bergman and Engberg-Pedersen (2010) while Aldersson and McEntee-Atalianis (2008) offer a comparison of the Icelandic and the Danish Sign Languages.

2 The Open Multilingual WordNet (OMW) Infrastructure

The motivation behind the Open Multilingual Wordnet (OMW) initiative (Bond and Paik, 2012; Bond and Foster, 2013) is to ease the use of wordnets in multiple languages. OMW proposes a shared CSV-based format for supporting the interlinking of language-specific wordnets. Version 1 of OMW¹¹ offers 28 wordnets,¹² all linked to the Princeton Wordnet of English (PWN),¹³ which functions thus as a pivot wordnet for establishing links between all the other wordnets included in OMW (Version 1).

A very helpful feature of OMW Version 1 is given by its online search facility, where one can type a word and obtain all the related PWN synsets in user-selected languages.¹⁴ Searching, for example, for the word “protection” we obtain 7 synsets returned. Focusing on the synset 00817680-n, with the English lemma “protection” and the Princeton WordNet gloss “the activity of protecting someone or something”, we obtain the (linked) OWM lemmas for selected Nordic languages, as presented in Table 1.

Table 1: The Danish, Finnish, Norwegian (Nynorsk and Bokmål) and Swedish lemmas, linked to the shared synset ID “00817680-n”, as returned by the query “protection” in the OMW search engine

Danish	forsvar, forsorg, værn, beskyttelse
Finnish	suojelu
Swedish	beskydd
Nynorsk	forsvar, beskytting, vern, omsorg
Bokmål	forsvar, beskyttelse, vern, omsorg

¹¹See <https://omwn.org/omw1.html>

¹²While there are over 150 wordnets that have been processed by OMW, only those with a licence allowing free redistribution are listed in OMW Version 1.

¹³See (Fellbaum, 2010) for more details on WordNet. A queryable online version of PWN is available at <https://wordnet.princeton.edu/>

¹⁴<https://compling.upol.cz/ntumc/cgi-bin/wn-gridx.cgi?gridmode=grid>

3 Aligning several SL Resources via the Open Multilingual WordNet Infrastructure

The work reported on in this section is developed within the EASIER research project,¹⁵ which aims to ease the communication between deaf and hearing individuals with the help of MT technologies. As such, linking different SLs through semantics is a priority. We chose to use the Open Multilingual Wordnet (OMW) infrastructure (Bond and Paik, 2012; Bond et al., 2016)¹⁶ as a (semantic) pivot between SL data.

We are dealing with four languages (German, Greek, English and Dutch sign languages). The resources involved in our approach are the DGS corpus (Prillwitz et al., 2008), Noema+ GSL dictionary (Efthimiou et al., 2016), BSL signbank (Jordan et al., 2014), and the NGT global signbank (Crasborn et al., 2020). These resources contain various types of spoken language words associated with each sign. They may be keywords, equivalents, or SL glosses.¹⁷ They are used as a starting point to match with the lemmas present in the corresponding aligned language versions of OMW. Then, native signers manually validate the potential matches. By using the Open Multilingual Wordnet, we aim to identify the signs with the same (or related) senses across languages.

Each resource involved has different structures, and so, the method must be flexible enough to exploit all the data available and avoid mistakes. As an example, the DGS Corpus has a multi-level structure, where each sign can be a type, a sub-type, or a variant. Semantics are attached to the sub-type level. If a sense has been associated with a sub-type, it can be spread down to the variants associated with it, but not up to the type. The DGS Corpus also contains synonymy links that can be exploited to spread senses to other signs.

We describe in the following paragraphs elements of SLs that need to and could be semantically aligned across languages and language types.

Phonological transcriptions: While in an ideal world, those transcriptions from videos displaying

signs could be used for establishing links between SL data for different languages, different SL data sets are transcribed with different transcription systems, e.g. HamNoSys (Hanke, 2004), SignWriting (Sutton, 2014) or others, as in the case of the Swedish SL data¹⁸

Besides, even if two resources use the same transcription system, the level of accuracy or precision of the transcription is not the same for all data. In some cases the transcription can be either semi-automatically generated or produced by human transcribers with different skills and views on which phonological elements of a sign should be transcribed.¹⁹

We are aware of efforts being made toward analysing and processing the videos directly using machine learning, rather than comparing and aligning transcriptions, but those are not in the scope of our current work.

Glosses: Many projects dealing with SL use glosses to identify signs. A gloss is, typically, a spoken language word optionally followed by a sequence of numbers or letters, to allow several signs to share the same word. The word is typically related to the meaning or iconicity of the sign, in the surrounding SpL, for easier identification. But the used word is ultimately somewhat arbitrary. Two unrelated projects working on the same sign language might have different glosses for the same sign, or the same gloss for different signs. This creates an obstacle toward linking resources together.

While many SL resources use glosses for labelling their data, the low accuracy/precision of automated tagging and the low Inter-Annotator Agreement (IAA) between human annotators for such tagging made the glosses difficult to use as a potential cross-language instrument for interlinking SL data in various languages.²⁰

For linking to the IDs in OMW, we preferably use keywords and translations as a starting point to approximate the meaning of the sign, and only use glosses as a last resort. However, we use glosses as identifiers.

¹⁸See (Bergman and Björkstrand, 2015) for a detailed description, and also <https://zrajm.github.io/teckentranskription/intro.html> on recent developments on a tool to support this transcription system.

¹⁹Power et al. (2022), for example, report in their experiment that the similarity (but not the exact matching) of transcriptions by two undergraduate research assistants working in a related project was 0.69.

²⁰Forster et al. (2010) discuss, among others, best practices for gloss annotation, in order to mitigate the issues of divergent tagging results, even in one and the same corpus.

¹⁵See <https://www.project-easier.eu/> for more details.

¹⁶See also <https://omwn.org/> for more details.

¹⁷The term “gloss” in the SL community is carrying a different meaning as in the case of WordNet. On the specificity of glosses used for naming (or labelling) SL data in corpora, see (Ormel et al., 2010). See also further below in this section.

4 An Example of the Use of shared OMW IDs for interlinking SL Data

We describe in this section how the EASIER project is making use of shared OMW IDs for interlinking data in British, German and Greek Sign Languages.

[omw.00806502-v](#) approve, O.K., okay, sanction | give sanction to

- [bsl.3572](#) goodness, virtue, good, virtuous, approve, adore, well, great, all right
- [dgs.54171](#) \$GEST-KEIN-PROBLEM1^
- [dgs.13555](#) GUT1^
- [dgs.16122](#) OKAY1A^
- [dgs.93765](#) OKAY1B^
- [gsl.1000](#) εγκρίνω

Figure 2: A screenshot showing how British, German and Greek Sign Language data are interlinked via a shared OMW index, as proposed by the EASIER project. Taken from https://www.sign-lang.uni-hamburg.de/easier/sign-wordnet/index_core_synsets.html

In Figure 2, we can see that various glosses and lemmas are linked to the OMW synset [omw.00806502-v](#). Links are directing to related videos displaying corresponding signs in three languages: BSL (British Sign Language), DGS (German Sign Language) and GSL (Greek Sign Language). Clicking on, for example, the link [dgs.16122](#), the user is landing at the page containing the video displaying the sign, with some additional information, as shown in Figure 3.

This way, a DGS sign can be linked to both a BSL and a GSL sign, based on a shared OMW ID, which is much more accurate than going only via translation of glosses or lemmas. Those elements: videos, glosses, phonetic transcriptions (if available), links to OMW, are the elements we are encoding in a unified and harmonised Linked Data compliant format.

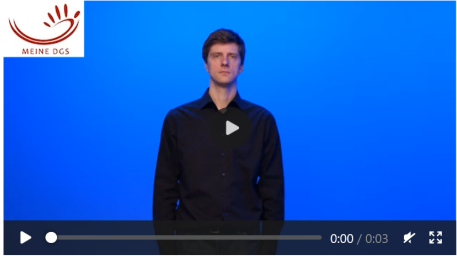
5 Extending the EASIER Approach with additional Signs

We searched for other SL resources in order to extend the approach described in [Bigéard et al. \(2022\)](#), thus linking SL data and wordnets, and then transforming those SL-wordnet combinations into RDF and OntoLex-Lemon. We found a basic lexicon of 1000 concepts associated with SL data in 4 languages, British, French, German and Greek, a result of the past Dicta-Sign project

dgs.16122 OKAY1A^

View more data about this sign in its original resource: [DOI link](#)

[direct link](#)



Synset ID and links	Synset lemmas	Synset definition	Synset examples	Type of validation	Also attested in these languages
omw.00806502-v omw link internal link	<ul style="list-style-type: none"> • approve • O.K. • okay • sanction 	give sanction to	<ul style="list-style-type: none"> • I approve of his educational policies 	Manual validation	BSL GSL

Figure 3: The video corresponding to the link ‘[dgs.16122](#)’ (see Figure 2). Taken from <https://www.sign-lang.uni-hamburg.de/easier/sign-wordnet/sign/dgs.16122.html>

(Matthes et al., 2012), which is available at the University of Hamburg.²¹ This resource is directly relevant to our purposes, as the included videos are equipped with SL glosses and HamNoSys transcriptions, as shown in Figure 4.

In Figure 4, we observe that the gloss and the HamNoSys transcription for the German video are identical with those deployed in the data used by the EASIER project for linking German SL data and wordnets, as can be seen at https://www.sign-lang.uni-hamburg.de/meinedgs/types/type13990_de.html.

This concordance of gloss and HamNoSys transcriptions²² not only allows for the association of two videos representing this German sign to one OWM ID,²³ but it also permits the addition of signs in an additional language, French, extending

²¹https://www.sign-lang.uni-hamburg.de/dicta-sign/portal/concepts/concepts_eng.html

²²But we can observe that in the one case the gloss is realised as a noun and in the second case as a verb. Signs are often ambiguous with respect to PoS, and in the future we will link the videos to both the nominal and verbal synsets, if both are available in the corresponding wordnet.

²³As the page https://www.sign-lang.uni-hamburg.de/dicta-sign/portal/concepts/cs/cs_688.html is linking to a more detailed lexical description of the sign, with the same gloss and HamNoSys transcription (see <https://www.sign-lang.uni-hamburg.de/galex/glossen/g13990.html>), with another video for the sign, we can in fact have 3 videos for this German sign associated with one OWM ID.

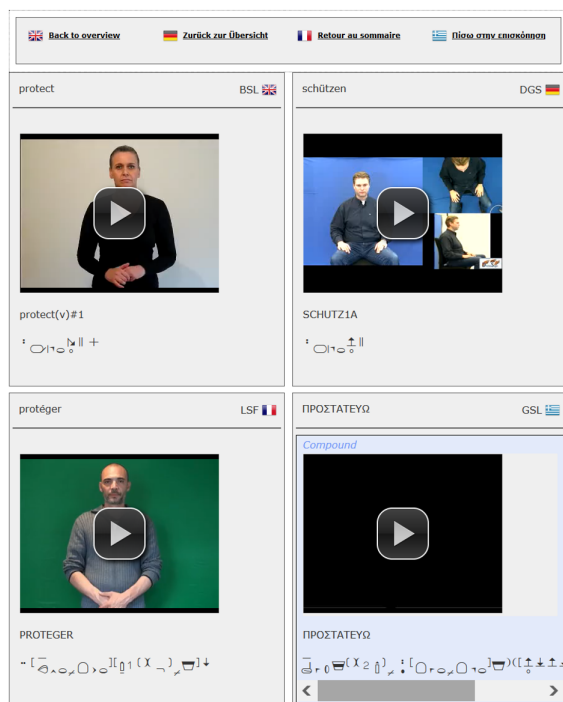


Figure 4: The concept “protect” as realised in 4 different Sign Languages. Taken from https://www.sign-lang.uni-hamburg.de/dicta-sign/portal/concepts/cs/cs_688.html

thus the multilingual coverage of the approach introduced by the EASIER project. We just need to introduce in our RDF representation new video instances (and their related glosses and transcriptions) and to link them to the same OMW ID.

Thus, the transformation of this additional data into our RDF and OntoLex-Lemon representation means organising those originally disparate and heterogeneous data sources in one harmonised formal representation, with the shared OMW IDs as the central component for the interlinking of the different data types and sources.

6 Extending our Work to Nordic Languages

We are extending our RDF representation to Nordic languages, while for now we have only for Danish a linking of SL data to its corresponding wordnet at our disposal.

Troelsgård and Kristoffersen (2018) discuss approaches for ensuring consistency between (Danish) Sign Language corpus data and the Dictionary of Danish signs. This approach aims at delivering a correspondence between the dictionary lemmas and the corpus lexicon, which consists of types introduced for lemmatising the tokens found in the

corpus annotations (glosses added to the signs). The strategy is to use words and their equivalents (also found in the dictionary) to search for signs in the corpus. In order to extend the list of potential Danish equivalents that could be used for a word-based search of signs in the corpus, Troelsgård and Kristoffersen (2018) suggest using the Danish wordnet, DanNet, which is described in Pedersen et al. (2009, 2018). This approach is thus very similar to the one described in Bigeard et al. (2022), but is monolingual. The relations between sign identifiers and lexical elements from both DanNet and other dictionary sources are encoded in a database, from which we obtained a TSV export.

In this export, we first have the signs, which correspond to entries in the Dictionary of Danish Signs (see Figure 5). A second type of data available in the export holds video links and information about the sign form (HamNoSys/SiGML).²⁴ A third type of information included in the export concerns the WordNet senses associated with the signs and their (form) variants.

Our work consisted thus in porting all those elements of the Danish data set to RDF and OntoLex-Lemon. In the OMW version of DanNet, we find for example the following information 00817680-n lemma beskyttelse, where the lemma corresponds to the OMW English wordnet 00817680-n lemma protection, thus sharing the same ID for the concept of “protection” in OMW (this holds also for French, etc.). We can therefore add the Danish sign ID (and video), which we obtained from the database, to our RDF-based infrastructure.



Figure 5: The Danish sign associated with the OMW ID “00817680-n”, corresponding to the (highlighted) lemma “beskyttelse”, here as one possible lexical realisation of the Danish SL gloss “FORSVARE” (*defend*)

Using the same strategy of deploying OMW as a pivot between concepts expressed in the

²⁴The SiGML notation is a XML transcription of the original HamNoSys code (Neves et al., 2020)

videos, we extended our approach to Icelandic and Swedish. Through OMW we can find the lemmas for Icelandic and Swedish associated with the OMW IDs “1128193-v” and “00817680-n” (corresponding to the Danish lemmas). We use these to search in the Icelandic SignWiki,²⁵ and in the Swedish Sign Language Dictionary, described in Mesch et al. (2012).²⁶ Icelandic and Swedish glosses can be easily integrated in our RDF-based representation, as can be seen for example in Listing 1, where the gloss for the Danish sign depicted in Figure 5 is augmented with glosses or lemmas from other languages.

```
dts:GLOSS_dts-722
  rdf:type sl:GLOSS ;
  rdfs:label "\FORSVARE\"@"da ;
  rdfs:label "\PROTEGER\"@"fr ;
  rdfs:label "\SCHUTZ1A^\"@"de ;
  rdfs:label "\protect(v)#1\"@"en ;
  rdfs:label "\beskydd\"@"se ;
  rdfs:label "\Vernda \"@"is ;
```

Listing 1: The RDF-based representation of the gloss “FORSVARE”, with the integration of multilingual labels from corresponding glosses

We further extended this approach to other Nordic languages, as described in Declerck and Olsen (2023). Data sets for 5 Nordic languages are included in OMW: Danish, Finnish, Norwegian (Nynorsk and Bokmål), and Swedish. Table 2 give some detailed information on the distribution of Nordic languages in OMW.

Table 2: Nordic wordnets included in OMW

Lang	Synsets	Words	Senses	Core
dan	4,476	4,468	5,859	81%
fin	116,763	129,839	189,227	100%
nno	3,671	3,387	4,762	66%
nob	4,455	4,186	5,586	81%
swe	6,796	5,824	6,904	99%

It is then straightforward to encode all the types of information on the relation between Danish SL data and DanNet into our RDF-based model. We need only to add an instance for the video displaying the sign, and its associated gloss (with language equivalents), as shown in Listing 1. The language equivalents are included, so that a Danish sign can be cross-lingually searched for, using

²⁵<https://is.signwiki.org/index.php/>

²⁶<https://teckensprakslexikon.su.se>

glosses in other languages. Then, we just need to add an `ontolex:Form` instance for the Danish sign, displayed in Listing 4, and which is linked via its corresponding lexical entry to the corresponding OMW instance, shown in Figure 5.

Listing 2 shows the encoding of the Danish video already displayed in Figure 5 above, and Listing 3 shows the RDF-based representation of the corresponding gloss.

```
<http://example.org/dts#
  SignVideos_dts-722.mp4>
  rdf:type sl:SignVideos ;
  sl:hasGLOSS dts:GLOSS_dts-722 ;
  sl:hasVideoAdresss "https://www.
  tegnsprog.dk/video/t/t_2162.mp4"^^
  rdf:HTML ;
  rdfs:label "\Video annotated with
  the gloss 'FORSVARE'\"@"en ;
```

Listing 2: The video annotated with the gloss “FORSVARE” as an instance of the RDF class “sl:SignVideos”

```
dts:GLOSS_dts-722
  rdf:type sl:GLOSS ;
  rdfs:label "\FORSVARE\"@"da ;
```

Listing 3: The RDF-based representation of the gloss “FORSVARE”

Listing 4 shows a corresponding lexical form (in this case a lemma taken from OMW) and links it to the video and to the gloss it is related to, also adding the SiGML notation, which is the XML transcription of the original HamNoSys code (Neves et al., 2020).

```
dts:Form_dts-722
  rdf:type ontolex:Form ;
  sl:hasGLOSS dts:GLOSS_dts-722 ;
  sl:hasVideo <http://example.org/dts#
  SignVideos_dts-722.mp4> ;
  sl:hasVideoAdresss "https://www.
  tegnsprog.dk/video/t/t_2162.mp4"^^
  rdf:HTML ;
  rdfs:label "\Adding transcription
  information associated with the
  video with the gloss 'FORSVARE'\"@"
  en ;
  ontolex:writtenRep "\<sigml><hns_sign
  gloss='FORSVARE'><hamnosys_manual><
  hamsymmlr/><hamfist/><hamparbegin/><
  hamextfingeru/><hampalmd/><hamplus
  /><hamextfingerr/><hampalmr/><
  hamparend/><hamparbegin/><hammoveu
  /><hamthumbside/><hamtouch/><hamplus
  /><hamnomotion/><hamparend/><
  hamrepeatfromstart/></
  hamnosys_manual></hns_sign></sigml>\
  \"@"hamnosys-sigml_;
```

```

_ontolex:writtenRep_"\beskyttelse\"
  @da_ ;

```

Listing 4: The RDF-based representation of the lexical form related to the gloss “FORSVARE” and the corresponding video

Finally, Listing 5 displays the lexical entry for which the form is a morphological realisation. The lexical entry is pointing to the OMW ID realised as a lexical concept in OntoLex-Lemon, and which itself points to the video annotated by the one gloss.

```

dts:LexicalEntry_722
  rdf:type ontolx:LexicalEntry ;
  rdfs:label "\forsvare, beskytte,
  beskyttelse\"@da ;
  ontolx:evokes wnid:omw-00817680-n ;
  ontolx:lexicalForm dts:Form_722 ;

```

Listing 5: The RDF-based representation of the lexical entry, which relates the concept and the form

```

wnid:omw-00817680-n
  rdf:type ontolx:LexicalConcept ;
  sl:hasWnLemma "\beskydd\"@se ;
  sl:hasWnLemma "\beskyttelse\"@da ;
  sl:hasWnLemma "\forsorg\"@da ;
  sl:hasWnLemma "\forsvar\"@da ;
  sl:hasWnLemma "\protection\"@en ;
  sl:hasWnLemma "\protection\"@fr ;
  sl:hasWnLemma "\vernd\"@is ;
  sl:hasWnLemma "\værn\"@da ;
  sl:hasWnLemma "\προστασία\"@el ;

  skos:definition "\the activity of
  protecting someone or something\"@en ;
  skos:definition "\παρεχόμενη φροντίδα σε
  κάποιον ώστε να προφυλάσσεται από υπαρκτούς
  ή διάφορους πιθανούς κινδύνους\"@el ;
  skos:inScheme sl:ConceptSet_OMW-DGS ;
  ontolx:isEvokedBy
  dgs:LexicalEntry_13990-2966 ;
  ontolx:isEvokedBy dts:LexicalEntry_1_2162 ;
  ontolx:isEvokedBy gsl:LexicalEntry_688 ;
  ontolx:isEvokedBy isl:LexicalEntry_vernda ;
  ontolx:isEvokedBy lsf:LexicalEntry_668 ;
  ontolx:isEvokedBy ssl:LexicalEntry_17861 ;

```

Figure 6: The encoding of the OWM ID, linking to corresponding lexical entries, which again are linked to other elements of our data set

7 The Dictionary of Maltese Sign Language (Maltese: *Lingwa tas-Sinjali Maltija*, LSM)

The Dizzjunarju tal-Lingwa tas-Sinjali Maltija (LSM, Maltese Sign Language) is an online dictionary comprising approximately 2,500 signs (as of 2023). Glosses for the LSM signs are in English and Maltese, so it is a trilingual dictionary.

Signs are transcribed using SignWriting (Sutton, 2014), and supported by photo and video illustrations. It is not currently possible to search using the SignWriting system, but words are grouped together largely by 33 semantic categories, e.g. occupations, place names, education, travel, health, etc. This means that the dictionary may also function as a glossary for people wanting to increase vocabulary in a particular field or search for semantically related terms.

This project grew out of a linguistic corpus that was begun in 1996 at the University of Malta. It has grown well beyond this, and the original research team expanded, as well as a group of collaborators representing the wider Maltese Deaf community. The dictionary has grown through sponsorship in the form of secondments of Deaf employees working in business and government posts, as well as the hard work of Deaf and Hearing volunteers.

Maltese Sign Language is a visual-gestural language of the Maltese Deaf community. There are no official statistics available on the number of people who use LSM, though the number of people in Malta who are Deaf or Hard of Hearing is estimated to be around 1500.²⁷ The current form of the language is of relatively recent origin, having its sources partly in a support/play group for deaf children, which began in the mid 1970s. Malta has been an independent country since 1964, but it has maintained strong ties to the UK, and more recently to the EU. Because of the shared history, shared use of the English language, and ongoing cultural ties between the UK and Malta, there is some influence from British Sign Language (BSL) in basic signs, though the language does not appear to be part of the BSL language family. There is also influence from other signed languages. Signing systems that were used by Deaf individuals and their families before the formation of LSM in its current form are largely undocumented. Fingerspelling, a method for borrowing words from spoken languages, uses a one-handed alphabet with 29 letters of the standard Maltese alphabet. (There is a dedicated handshape for the digraph <g> but not for <ie>.)

Today, LSM classes are offered at the University of Malta, MCAST, and community settings. Significant linguistic research and documen-

²⁷See <http://www.deafmalta.com/> accessed: 2023-06001] for more details.

tation began in the early 2000s and has carried on (Galea, 2014; Azzopardi-Alexander, 2009, 2018; Hoffmann-Dilloway, 2021; Hoffmann-Dilloway and Xerri, 2022) The first professional interpreter began working in 2001, and Deaf interpreters have presented a daily TV news bulletin since 2012. The Maltese government passed the Maltese Sign Language Recognition Act in 2016, which provides for the promotion of the use and development of Maltese Sign Language, whilst declaring that the Maltese Sign Language is to be considered an official language of Malta. This same act also set up the Sign Language Council of Malta, which is a forum for the Deaf community to be consulted on matters relating to LSM.

There exists thus a rich dictionary for the Maltese Sign Language, but we do not have a Maltese wordnet with which we can connect the videos displaying LSM sign. We are currently working on analysing alternative semantic lexical resources, including the LSM category system, for adding a combination of Maltese SpL and SL data to our RDF-based infrastructure.

8 Conclusions and Future Work

Our RDF-based encoding results in a harmonised representation of data from both spoken and sign languages that was originally stored in different formats in different locations. Taking advantage of the work proposed by Bigeard et al. (2022) and Troelsgård and Kristoffersen (2018), we can include the links between SL data and wordnets under the umbrella of RDF and by re-using elements of OntoLex-Lemon. The Open Multilingual Wordnet infrastructure plays a central role in this work, as the shared OMW IDs across various languages are at the core of the interlinking of the distinct data types and sources. The resulting unified RDF-based representation supports a dense linking of different types of information.

We are continuously extending our work to other languages. For Finnish and Norwegian we expect it to be a rather straightforward, although time consuming task, since for both these languages we have OMW entries as well as SL portals. It will be more difficult to expand to languages with fewer digital resources, as we can see while dealing with Maltese, for which we do not have a wordnet at our disposal.

The resulting data sets will be made available on Github.

Acknowledgments

The presented work is pursued in the context of the COST Action NexusLinguarum – European network for Web-centered linguistic data science (CA18209), 731015). The contributions by DFKI and University of Malta are also pursued in the context of the LT-BRIDGE project, which has received funding from the European Unions Horizon 2020 Research and Innovation Programme under Grant Agreement No 952194. Contributions by the University of Hamburg are in part supported by the EASIER (Intelligent Automatic Sign Language Translation) Project. EASIER has received funding from the European Unions Horizon 2020 research and innovation programme, grant agreement No 101016982.

References

- Russell Aldersson and Lisa McEntee-Atalianis. 2008. *A lexical comparison of signs from icelandic and danish sign languages*. *Sign Language Studies*, 9:45–87.
- Marie Azzopardi-Alexander. 2009. Iconicity and the development of maltese sign language. In Ray Fabri, editor, *Maltese Linguistics: A Snapshot in Memory of Joseph A. Cremona (1922-2003)*, pages 93–116. Brockmeyer, Bochum.
- Marie Azzopardi-Alexander. 2018. *Maltese Sign Language: Parallel interwoven journeys of the Deaf community and the researchers*. In Patrizia Paggio and Albert Gatt, editors, *Languages of Malta*. Language Science Press.
- Brita Bergman and Thomas Björkstrand. 2015. Teckentranskription. Technical Report XXV, Stockholm University, Sign Language.
- Brita Bergman and Elisabeth Engberg-Pedersen. 2010. *Transmission of sign languages in the Nordic countries*, Cambridge Language Surveys, page 7494. Cambridge University Press.
- Sam Bigeard, Marc Schulder, Maria Kopf, Thomas Hanke, Kyriaki Vasilaki, Anna Vacalopoulou, Theodore Goulas, Athanasia-Lida Dimou, Stavroula-Evita Fotinea, and Eleni Efthimiou. 2022. *Introducing sign languages to a multilingual Wordnet: Bootstrapping corpora and lexical resources of Greek Sign Language and German Sign Language*. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 9–15, Marseille, France. European Language Resources Association.

- Francis Bond and Ryan Foster. 2013. [Linking and extending an open multilingual Wordnet](#). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362, Sofia, Bulgaria. Association for Computational Linguistics.
- Francis Bond and Kyonghee Paik. 2012. A survey of wordnets and their licenses. In *Proc. of the 6th Global WordNet Conference (GWC 2012)*, Matsue. 64–71.
- Francis Bond, Piek Vossen, John P. McCrae, and Christiane Fellbaum. 2016. Cili: the collaborative interlingual index. In *Proceedings of the Global WordNet Conference*, volume 2016.
- Christian Chiacros, Sebastian Hellmann, and Sebastian Nordhoff. 2012. The Open Linguistics Working Group of the Open Knowledge Foundation. In *Linked Data in Linguistics*, pages 153–160. Springer, Heidelberg.
- Philipp Cimiano, Christian Chiacros, John P. McCrae, and Jorge Gracia. 2020. [Linguistic Linked Data - Representation, Generation and Applications](#). Springer.
- Onno Crasborn, Richard Bank, Inge Zwitterlood, Els van der Kooij, Ellen Ormel, Johan Ros, Anique Schüller, Anne de Meijer, Merel van Zuilen, Yasmine Ellen Nauta, Frouke van Winsum, and Max Vonk. 2020. [Ngt dataset in global signbank](#).
- Thierry Declerck. 2019. Ontolex as a possible bridge between wordnets and full lexical descriptions. In *Proceedings of Global WordNet Conference 2019*.
- Thierry Declerck and Sussi Olsen. 2023. Linked open data compliant representation of the interlinking of nordic wordnets and sign language data. In *Proceedings of the 2nd Workshop on Resources and Representations for Under-Resourced Languages and Domains*, pages 62–69.
- Thierry Declerck, Thomas Troelsgård, and Sussi Olsen. 2023. Towards an rdf representation of the infrastructure consisting in using wordnets as a conceptual interlingua between multilingual sign language datasets. In *GWC 2023: 12th International Global Wordnet Conference, Proceedings*. To appear.
- Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, Julie A. Hochgesang, Jette Kristoffersen, and Johanna Mesch, editors. 2016. [Proceedings of the LREC2016 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining](#). European Language Resources Association (ELRA), Portorož, Slovenia.
- Christiane Fellbaum. 2010. Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer.
- Jens Forster, Daniel Stein, Ellen Ormel, Onno Crasborn, and Hermann Ney. 2010. Best practice for sign language data collections regarding the needs of data-driven recognition and translation. In *Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (CSLT 2010)*.
- Maria Galea. 2014. [SignWriting \(SW\) of Maltese Sign Language \(LSM\) and its development into an orthography: Linguistic considerations](#). Ph.D. thesis, University of Malta.
- Thomas Hanke. 2004. [HamNoSys – representing sign language data in language resources and language processing contexts](#). In *Proceedings of the LREC2004 Workshop on the Representation and Processing of Sign Languages: From SignWriting to Image Processing. Information techniques and their implications for teaching, documentation and communication*, pages 1–6, Lisbon, Portugal. European Language Resources Association (ELRA).
- Erika Hoffmann-Dilloway. 2021. [Shadows and mirrors: Spatial and ideological perspectives on sign language competency](#). *Journal of Linguistic Anthropology*, 31(3):320–334.
- Erika Hoffmann-Dilloway and Annabelle Xerri. 2022. # deafmum: A deaf maltese activists strategies for addressing hearing parents of deaf children. *Practicing Anthropology*, 44(4):10–14.
- Fenlon Jordan, Kearsy Cormier, Ramas Rentelis, Adam Schembri, Katherine Rowley, Robert Adam, and Bencie Woll. 2014. [Bsl signbank: A lexical database of british sign language \(first edition\)](#).
- Maria Kopf, Marc Schulder, and Thomas Hanke. 2022. [D6.1 overview of datasets for the sign languages of europe](#).
- Silke Matthes, Thomas Hanke, Anja Regen, Jakob Storz, Satu Worseck, Eleni Efthimiou, Athanasia-Lida Dimou, Annelies Braffort, John Glauert, and Eva Safar. 2012. [Dicta-Sign -Building a Multilingual Sign Language Corpus](#). In *5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon. Satellite Workshop to the eighth International Conference on Language Resources and Evaluation (LREC-2012)*, Istanbul, Turkey.
- John P. McCrae, Paul Buitelaar, and Philipp Cimiano. 2017. The OntoLex-Lemon Model: development and applications. In *Proc. of the 5th Biennial Conference on Electronic Lexicography (eLex)*.
- Johanna Mesch, Lars Wallin, and Thomas Björkstrand. 2012. [Sign language resources in Sweden: Dictionary and corpus](#). In *Proceedings of the LREC2012 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon*, pages 127–130, Istanbul, Turkey. European Language Resources Association (ELRA).

- Carolina Neves, Luísa Coheur, and Hugo Nicolau. 2020. [HamNoSys2SiGML: Translating HamNoSys into SiGML](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6035–6039, Marseille, France. European Language Resources Association.
- Ellen Ormel, Onno Crasborn, Els van der Kooij, Lianne van Dijken, Ellen Yassine Nauta, Jens Forster, and Daniel Stein. 2010. [Glossing a multi-purpose sign language corpus](#). In *Proceedings of the LREC2010 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, pages 186–191, Valletta, Malta. European Language Resources Association (ELRA).
- Bolette Sandford Pedersen, Manex Aguirrezabal Zabaleta, Sanni Nimb, Sussi Olsen, and Ida Rørmann Olsen. 2018. Towards a principled approach to sense clustering a case study of wordnet and dictionary senses in danish. In *Proceedings of Global WordNet Conference 2018*. Global WordNet Association. Null ; Conference date: 08-01-2018 Through 12-01-2018.
- Bolette Sandford Pedersen, Sanni Nimb, Jørg Assmussen, Nicolai Hartvig Sørensen, Lars Trap-Jensen, and Henrik Lorentzen. 2009. DanNet — the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary. *Language Resources and Evaluation*, 43(3):269–299.
- Justin Power, David Quinto-Pozos, and Danny Law. 2022. [Signed language transcription and the creation of a cross-linguistic comparative database](#). In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 173–180, Marseille, France. European Language Resources Association.
- Siegmond Prillwitz, Thomas Hanke, Susanne König, Reiner Konrad, Gabriele Langer, and Arvid Schwarz. 2008. [DGS Corpus project – development of a corpus based electronic dictionary German Sign Language / German](#). In *Proceedings of the LREC2008 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, pages 159–164, Marrakech, Morocco. European Language Resources Association (ELRA).
- V. Sutton. 2014. *Lessons in Sign Writing: Textbook*, fourth edition. Deaf Action Committee for Sign Writing.
- Thomas Troelsgård and Jette Kristoffersen. 2018. [Improving lemmatisation consistency without a phonological description. the Danish Sign Language corpus and dictionary project](#). In *Proceedings of the LREC2018 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community*, pages 195–198, Miyazaki, Japan. European Language Resources Association (ELRA).

CURED4NLG: A Dataset for Table-to-Text Generation

Nivranshu Pasricha¹, Mihael Arcan² and Paul Buitelaar^{1,2}

¹SFI Centre for Research Training in Artificial Intelligence

²Insight SFI Research Centre for Data Analytics
Data Science Institute, University of Galway

n.pasricha1@nuigalway.ie

Abstract

We introduce CURED4NLG, a dataset for the task of table-to-text generation focusing on the public health domain. The dataset consists of 280 pairs of tables and documents extracted from weekly epidemiological reports published by the World Health Organisation (WHO). The tables report the number of cases and deaths from COVID-19, while the documents describe global and regional updates in English text. Along with the releasing the dataset, we present outputs from three different baselines for the task of table-to-text generation. The first is based on a manually defined template and the other two on end-to-end transformer-based models. Our results suggest that end-to-end models can learn a template-like structure of the reports to produce fluent sentences, but may contain many factual errors especially related to numerical values.

1 Introduction

Data-to-text generation systems aim to produce meaningful texts in a human language from non-linguistic representation of information such as tables or graphs in the input (Reiter and Dale, 2000). Traditionally, such systems have been designed using a rule-based approach relying on a modular pipeline architecture and have included applications in domains such as weather reporting (Goldberg et al., 1994), sports (Robin, 1995; Tanaka-Ishii et al., 1998) and healthcare (Binsted et al., 1995; Cawsey et al., 1997). Recently, there has been increasing interest in end-to-end approaches for data-to-text generation with neural encoder-decoder architectures. To aid further research in this direction, a number of datasets have been released in the last few years with different input data structures covering various domains. Examples include WIKIBIO (Lebret et al., 2016), ROTOWIRE (Wiseman et al., 2017), WebNLG (Gardent et al., 2017), E2E (Novikova et al., 2017), ToTTo (Parikh et al., 2020) and DART (Nan et al., 2021).

A popular strategy applied to data-to-text generation tasks is to split the problem along two fundamental axes aiming to answer the questions, *what to say?* (content determination) and *how to say it?* (microplanning and linguistic realisation). Datasets such as WebNLG, E2E and DART are only concerned with the planning and realisation aspects and do away with content selection aspect of the task. A more recent dataset, ToTTo includes content selection explicitly by highlighting relevant cells in the input table. However, the output for ToTTo is usually one or two sentences which is typically easier to generate compared to a document.

We present CURED4NLG¹ (COVID-19 Update Reports from Epidemiological Data for Natural Language Generation), a dataset for table-to-text generation, where the input data is structured in the form of a table, typically comprising of 6 to 60 rows with 7 to 9 columns (see Table 1). Each table reports the number of new cases of COVID-19 and related deaths during a week-long time period along with cumulative totals recorded since the start of the pandemic. A document corresponding to each table describes the important information contained in the table in about 200 – 300 words in English as shown in Figure 1. Hence, the goal of the table-to-text generation task is to automatically generate an output document describing the data in the input table. With CURED4NLG, we aim to enrich research in table-to-text generation with the goal of generating documents longer than one sentence in the output conditioned on structured input data while also addressing the issues related to content determination. We present outputs and results from two baseline models, based on end-to-end approaches, and compare them with a template-based system. Initial results suggest that end-to-end models are able to generate fluent outputs but can struggle to generate sentences which are faithful to the input tables.

¹<http://github.com/cured4nlg/cured4nlg>

WHO Region	New cases in last 7 days (%)	Change in new cases	Cumulative cases (%)	New deaths in last 7 days (%)	Change in new deaths	Cumulative deaths (%)
Europe	1989636 (54%)	11%	13144973 (26%)	25531 (47%)	44%	311542 (25%)
Americas	1031573 (28%)	3%	21509104 (43%)	17289 (32%)	<1%	656629 (53%)
South-East Asia	390157 (11%)	2%	9641945 (19%)	5132 (9%)	10%	149326 (12%)
Eastern Mediterranean	214072 (6%)	18%	3307411 (7%)	5675 (10%)	23%	84305 (7%)
Africa	33687 (1%)	2%	1357945 (3%)	831 (2%)	30%	30616 (2%)
Western Pacific	31370 (1%)	19%	765197 (2%)	377 (1%)	-5%	15942 (1%)
Global	3690495 (100%)	8%	49727316 (100%)	54835 (100%)	21%	1248373 (100%)

In the past week, the global number of cases of COVID-19 has increased by 8% compared to the previous week, totalling more than 3.6 million new cases, while new deaths have increased by 21% to over 54000. This brings the cumulative numbers to over 49.7 million reported cases and over 1.2 million deaths globally since the start of the pandemic. The European Region continues to account for the greatest proportion of new cases and deaths in the past 7 days, the Region reported over half (54%) of all new cases and nearly half (47%) of new deaths. Although it still accounts for only 2% of the global total number of cases and deaths, this week the Western Pacific Region showed the largest relative proportional increase in new cases (19%) compared to the previous week followed by the Eastern Mediterranean Region (18%) and the European region (11%). The three regions reporting the highest proportional increases in newly reported deaths in the past 7 days compared to the previous week are Europe (44%), Africa (30%) and the Eastern Mediterranean (23%). The Western Pacific Region was the only region to report a decrease in deaths (5%) this week compared to the previous week.

Figure 1: Example of a table and corresponding epidemiological report from the CURED4NLG dataset.

2 Related Work

Natural language generation (NLG) in the health-care domain has seen significant interest over the years (Cawsey et al., 1997; Pauws et al., 2019). Applications here usually involve generating personalised reports or medical explanations for individual patients (Binsted et al., 1995; McKeown et al., 1997; Mahamood and Reiter, 2011) and typically are not concerned with mass communication of general public health advice. However, during the COVID-19 pandemic, public dashboards (Ritchie et al., 2020; Dong et al., 2020; Wissel et al., 2020) became immensely popular for communicating information about the spread of the disease globally. These dashboards rely on visuals such as maps and charts but do not usually provide textual updates. An exception to this is a dashboard² by Microsoft and Arria NLG reporting automatically generated narratives describing the number of cases and deaths for COVID-19 along with an interactive map (Reiter and Sripada, 2020). Tangential to this, automatic generation of data-driven narratives for mass communication of news (Leppänen et al.,

²<https://www.arria.com/covid19-microsoft/>

2017) and automated journalism (Graefe, 2016) have also received significant interest over the last few years. However, since most of these systems for automatic report generation are built in-house by private organisations, the details about the underlying architecture and the actual data used are usually not publicly available (Dale, 2020). With CURED4NLG we hope to motivate research in this domain with a publicly available dataset.

In terms of the structure of the input and output data, ROTOWIRE (Wiseman et al., 2017) can be considered most similar to CURED4NLG among existing NLG datasets. ROTOWIRE consists of about 3,000 basketball box-scores paired with descriptive summaries and is one of best examples of what a real-world application of data-to-text generation might look like. However, it has some significant challenges associated with it. For instance, Wang (2019) observed that only 60% of the output textual summary content can be grounded to the boxscore data. This misalignment leads to hallucinations where a model generates a set of unconditioned random statements that are unfaithful to the input. Thomson et al. (2020) also observed data partition contamination issues where boxscore data from

	Min.	Max.	Average
Columns	7	9	7.86
Rows	7	62	33.28
Cells	49	496	265.28
Document Length	63	643	249.75

Table 1: Number of rows, columns, cells and document length (number of words) in the CURED4NLG dataset.

some games ended up in both training set as well as test/validation set. Another issue highlighted in their analysis is that random partition of the data ignores the inherent temporal dimension in the data leading to further hallucinations.

3 CURED4NLG Dataset

The CURED4NLG dataset is created from 40 epidemiological update reports published by WHO and consists of 280 pairs of tables and documents. Since August 2020, an update report has been published on the WHO website³ once a week in PDF format to provide an overview of the global and regional situation for COVID-19. Each weekly update highlights key data and trends as well as other pertinent epidemiological information concerning the pandemic. We extract the tables from Annex 1 of the PDF reports using optical character recognition (OCR) followed by a manual verification step to fix formatting and spelling errors. The resulting tabular data is saved as a file with tab-separated values, while the corresponding update reports are stored as plaintext files. Some texts include additional information about patient demographics and regional restrictions as well as references to charts and figures elsewhere in the report. Such sentences go beyond the data in the tables, hence, we filter these out and create a *cleaned* version of the CURED4NLG dataset.

The dataset is split into training, validation and test sets such that the inherent temporal aspects of the data are maintained. Data from the first 30 reports is used for training, data from the next three weeks is used for validation and the data from the five most recent weekly reports is taken to be the test set. Each update report consists of a global table along with six regional tables, hence, the training set, validation set, and the test set contain 210, 21 and 49 instances respectively (see Table 2).

³<https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports>

	train	valid	test
Examples	210	21	49
Tokens	56250	4103	9555
Types	3711	478	869
Avg. Doc. Length	267.9	195.4	195.0
<i>cleaned</i>			
Tokens	43526	4091	9506
Types	2243	476	869
Avg. Doc. Length	207.3	194.8	194.0

Table 2: Number of examples, tokens and types for text documents in the CURED4NLG dataset.

Compared to ROTOWIRE (Wiseman et al., 2017), this dataset is smaller by an order of magnitude in size. It is also much smaller than other NLG datasets which usually consist of several thousands examples. Nonetheless, CURED4NLG can be useful for data-to-text generation tasks as it is representative of a real-world application scenario for NLG and presents an opportunity to focus on the various challenges such as content selection, document planning and linguistic realisation. One limitation of this dataset might be that the sentence structure is simple in most instances and there is minimal linguistic variation in the texts. Despite that, we find state-of-the-art end-to-end NLG systems struggle to outputs with high accuracy and this dataset can be useful in studying the limitations of such systems. Since this dataset is created from weekly reports by WHO, it includes an additional challenge of working with data that contains an inherent temporal dimension which might be difficult to model using end-to-end techniques.

Since June 2021, WHO stopped publishing the tables containing detailed case statistics in the weekly epidemiological reports. The reports published since then only contain an update in the form of texts while the tables are available on the online WHO portal⁴. Hence the number of new cases and deaths reported in the tables do not always exactly match the figures reported in the text of recent weekly epidemiological reports. It is due to this reason the data in CURED4NLG is limited until May 2021. However, we plan to further extend this dataset, with data until 2023 by manually verifying the numbers reported across the tables and the texts, and aligning them correctly, where needed.

⁴<https://covid19.who.int/data>

<p>The region of <i>Europe</i> reported over <i>1.4 million</i> new cases and <i>25000</i> new deaths this week, a <i>10% decrease</i> and a <i>4% decrease</i> respectively compared to the previous week. The highest numbers of new cases were reported from <i>Turkey</i> (<i>378771</i> new cases; <i>449.1</i> new cases per 100000 population; a <i>9% decrease</i>), <i>France</i> (<i>211674</i> new cases; <i>325.5</i> new cases per 100000 population; a <i>9% decrease</i>) and <i>Germany</i> (<i>145156</i> new cases; <i>174.5</i> new cases per 100000 population; a <i>1% increase</i>). The highest numbers of new deaths were reported from <i>Poland</i> (<i>3383</i> new deaths; <i>8.9</i> new deaths per 100000 population; a <i>6% decrease</i>), <i>Russian Federation</i> (<i>2650</i> new deaths; <i>1.8</i> new deaths per 100000 population; a <i>2% increase</i>) and <i>Ukraine</i> (<i>2537</i> new deaths; <i>5.8</i> new deaths per 100000 population; a <i>8% decrease</i>).</p>	<p>In the past week, the <i>European Region</i> reported over <i>1466000</i> new cases and over <i>25000</i> new deaths, a <i>decrease</i> of <i>1%</i> and an <i>increase</i> of <i>1%</i> respectively compared to the previous week. The three countries reporting the highest numbers of new cases were <i>Kosovo</i> (<i>2662</i> new cases; <i>57</i> new cases per 100000; a <i>1%</i> decrease), <i>Turkey</i> (<i>378771</i> new cases; <i>57</i> new cases per 100000; a <i>1% decrease</i>), <i>France</i> (<i>211674</i> new cases; <i>158.8</i> new cases per 100000; a <i>7% decrease</i>). The three countries reporting the highest numbers of new deaths this week were the <i>United Kingdom</i> (<i>157</i> new deaths; <i>3.4</i> new deaths per 100000; a <i>3% decrease</i>), <i>Germany</i> (<i>1650</i> new deaths; <i>3.4</i> new deaths per 100000; a <i>3% decrease</i>), the <i>Russian Federation</i> (<i>2650</i> new deaths; <i>3.7</i> new deaths per 100000; a <i>3% decrease</i>) and the <i>Russian Federation</i> (<i>2345</i> new deaths; <i>3.4</i> new deaths per 100000; a <i>3% decrease</i>).</p>
--	---

Table 3: Example of an output epidemiological report for the European region generated by the template baseline (left) and the T5 model (right). Text in blue italics shows information filled in from the input table by the baseline template. Text in green italics shows tabular values correctly produced by the T5 model while underlined text in red shows the mistakes. Outputs from all end-to-end trained baselines for this example are presented in Appendix A.3.

4 Baselines

We present baseline results for the task of table-to-text generation with CURED4NLG using two different approaches – a templated baseline and two transformer-based encoder-decoder models. The overall task is defined as follows:

Given a set of one or more tables in the input, generate a text document in English in the output describing the tabular data.

Template baseline: We define a global and a regional template to generate an epidemiological report based on input tabular data. The template for the global report includes sentences describing new and cumulative totals of cases and deaths for COVID-19 along with changes in trends from the week prior. The template also generates sentences describing the most affected continental region as well as the five most affected countries globally. Similarly, the template for a regional report describes new numbers as well as the change in numbers from the previous week, followed by a sentence describing the three most affected countries in a specific region. The exact templates used to generate the output documents are defined in Appendix A.1.

End-to-End baselines: We use the hierarchical model (Rebuffel et al., 2020) as one of the end-to-end baseline models. It is designed for data-to-text tasks and follows a two-level encoder-decoder architecture for modeling structured data in the input. We use the state-of-the-art T5 model (Raffel et al., 2020) as another end-to-end neural baseline. It is based on the transformer architecture (Vaswani et al., 2017) and pre-trained on the “Colossal Clean Crawled Corpus” using a masked language modelling objective. Since the T5 architecture expects the input to be encoded as a sequence of text, we linearise the input table by concatenating all the rows into a single sequence. The rows in each table are arranged in decreasing order of number of new cases by default.

To assess the performance of the end-to-end baseline systems on content selection, we perform an experiment where we randomly shuffle the rows of the table to see how well the transformer-based models pay attention to the relative positioning of the rows in the input table. We perform another experiment where we include only a subset of the first ten rows in the input and evaluate the model performance. And as another experiment we train with the *cleaned* version of the CURED4NLG dataset.

	BLEU (↑)	METEOR (↑)	TER (↓)	PARENT		
				Precision (↑)	Recall (↑)	F1 (↑)
Template baseline	64.48	41.76	32.19	76.55	19.93	29.97
Hierarchical model	29.86	27.64	67.49	43.10	17.65	22.80
T5 (no pre-training)	20.31	18.47	99.55	41.07	8.38	12.24
T5 (pre-trained)	43.32	32.77	52.10	56.38	17.15	24.68
+ <i>shuffled</i>	41.16	31.67	49.89	56.07	14.75	21.97
+ <i>subset</i>	42.99	33.33	55.58	56.75	18.73	26.13
+ <i>cleaned</i>	44.57	33.37	49.85	57.07	17.35	25.05

Table 4: Results for baselines on the CURED4NLG dataset.

The details for training the end-to-end baseline models along with the chosen hyperparameter values for each model are described in Appendix A.2.

All the code for training, generating and evaluating the baseline models along with the generated outputs is available to download at <https://github.com/cured4nlg/cured4nlg>.

5 Results and Discussion

We report results on the outputs generated from the baselines using four automatic evaluation metrics, BLEU (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014), TER (Snover et al., 2006) and PARENT (Dhingra et al., 2019) as shown in Table 4. The first three are popular metrics used for measuring lexical similarity between generations and references while PARENT is a recently proposed metric specifically for table-to-text evaluation as it computes precision and recall for n -grams in generated and reference texts aligned to table data.

We find the template baseline to outperform the end-to-end models across all the automatic evaluation metrics. Earlier reports published by the World Health Organization in 2020 contained more varied text, however, reports published since March 2021 appear to follow a template-like structure. Since the validation and test sets exclusively contain data from this period, because the dataset was split in such way that the inherent temporal dimension of the data remains intact, we observe high scores across the automatic evaluation metrics with the template baseline.

We observe that the end-to-end baseline models are able to generate fluent outputs by learning the template-like sentence structure but contain many factual errors as shown in Table 3. The pre-trained T5 model performs better than the hierarchical

baseline on the metrics measuring lexical similarity as well as the precision score. However, the hierarchical model achieves a similar recall score. We further observe that shuffling the rows in the table leads to worse performance for the T5 model as it makes the task more difficult. However, we observe slight improvements in the scores with the *cleaned* version of the dataset and further notice improvements in recall and F1 scores when only a subset of the top 10 rows is considered. This suggests that the model struggles to perform content selection, especially for larger tables.

A limitation of the PARENT metric is that it cannot detect paraphrases accurately. In almost every gold-standard reference of the CURED4NLG dataset, large numbers are either written in words or rounded to nearest thousand in text while the tables contain exact numerical values. For example, in Table 1, the number of new cases reported in the input table is 3690495, while the reference text report describes this value as “*more than 3.6 million*”. To account for this and other errors related to the accuracy of the generated texts, we manually count the number of errors in the outputs of the hierarchical model and the pre-trained T5 model on a subset of 21 examples from the test set. We use the same error categories of incorrect *Number* (for numerical values), *Name* (for region names) and *Word* (for words such as increase, decrease, rise, decline, etc.) as defined by Thomson and Reiter (2020). The rest of the errors are classified in the *Other* category. We find outputs from both models contain about 20 – 25 errors on average with most of the errors being associated with numerical values as shown in Table 5. Further work is required in designing error annotation guidelines specific to the CURED4NLG dataset as well as evaluation strategies which can identify paraphrasing of numbers.

Error Category	Hierarchical		T5	
	Total	Avg.	Total	Avg.
Number	346	16.5	294	14.0
Name	63	3.0	34	1.6
Word	85	4.0	72	3.4
Other	9	0.4	6	0.3
Total	503	23.9	406	19.3

Table 5: Counts of errors in the outputs generated by end-to-end baselines on a subset of 21 examples.

6 Conclusion

We introduced CURED4NLG, a dataset for table-to-text generation which can be useful as a benchmark for data-to-text generation. Initial baseline results suggest that end-to-end text generation models can learn a template-like structure of the documents to generate fluent outputs but at the same time are prone to hallucinating and generating erroneous statements particularly related to numerical values.

Acknowledgments

This work was conducted with the financial support of the Science Foundation Ireland Centre for Research Training in Artificial Intelligence under Grant No. 18/CRT/6223 and co-supported by Science Foundation Ireland under grant number SFI/12/RC/2289 2 (Insight), co-funded by the European Regional Development Fund.

References

- Kim Binsted, Alison Cawsey, and Ray Jones. 1995. Generating personalised patient information using the medical record. In *Artificial Intelligence in Medicine*, pages 29–41, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Alison J. Cawsey, Bonnie L. Webber, and Ray B. Jones. 1997. *Natural Language Generation in Health Care*. *Journal of the American Medical Informatics Association*, 4(6):473–482.
- Robert Dale. 2020. *Natural language generation: The commercial state of the art in 2020*. *Natural Language Engineering*, 26(4):481–487.
- Michael Denkowski and Alon Lavie. 2014. *Meteor universal: Language specific translation evaluation for any target language*. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.
- Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William Cohen. 2019. *Handling divergent reference texts when evaluating table-to-text generation*. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4884–4895, Florence, Italy. Association for Computational Linguistics.
- Ensheng Dong, Hongru Du, and Lauren Gardner. 2020. An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*, 20(5):533–534.
- Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. *The WebNLG challenge: Generating text from RDF data*. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133, Santiago de Compostela, Spain. Association for Computational Linguistics.
- E. Goldberg, N. Driedger, and R.I. Kittredge. 1994. *Using natural-language processing to produce weather forecasts*. *IEEE Expert*, 9(2):45–53.
- Andreas Graefe. 2016. *Guide to automated journalism*. Technical report, Tow Center for Digital Journalism, Columbia University.
- Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213.
- Leo Leppänen, Myriam Munezero, Mark Granroth-Wilding, and Hannu Toivonen. 2017. *Data-driven news generation for automated journalism*. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 188–197, Santiago de Compostela, Spain. Association for Computational Linguistics.
- Saad Mahamood and Ehud Reiter. 2011. *Generating affective natural language for parents of neonatal infants*. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 12–21, Nancy, France. Association for Computational Linguistics.
- Kathleen R. McKeown, Desmond A. Jordan, Shimei Pan, James Shaw, and Barry A. Allen. 1997. *Language generation for multimedia healthcare briefings*. In *Fifth Conference on Applied Natural Language Processing*, pages 277–282, Washington, DC, USA. Association for Computational Linguistics.
- Linyong Nan, Dragomir Radev, Rui Zhang, Amrit Rau, Abhinand Sivaprasad, Chiachun Hsieh, Xiangu Tang, Aadit Vyas, Neha Verma, Pranav Krishna, Yangxiaokang Liu, Nadia Irwanto, Jessica Pan, Faiaz Rahman, Ahmad Zaidi, Mutethia Mutuma, Yasin Tarabar, Ankit Gupta, Tao Yu, Yi Chern Tan, Xi Victoria Lin, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. 2021. *DART: Open-domain structured data record to text generation*. In

- Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 432–447, Online. Association for Computational Linguistics.
- Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2017. [The E2E dataset: New challenges for end-to-end generation](#). In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 201–206, Saarbrücken, Germany. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Ankur P Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of EMNLP*.
- Steffan Pauws, Albert Gatt, Emiel Kraemer, and Ehud Reiter. 2019. Making effective use of healthcare data using data-to-text technology. In *Data Science for Healthcare: Methodologies and Applications*, pages 119–145. Springer.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Journal of Machine Learning Research*, 21(140):1–67.
- Clément Rebuffel, Laure Soulier, Geoffrey Scoutheeten, and Patrick Gallinari. 2020. A hierarchical model for data-to-text generation. In *Advances in Information Retrieval*, pages 65–80, Cham. Springer International Publishing.
- Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*, 1 edition. Cambridge University Press.
- Ehud Reiter and Yaji Sripada. 2020. [Using arria nlg to give visual analytics dashboards the power of language](#). Technical report, Arria NLG.
- Hannah Ritchie, Esteban Ortiz-Ospina, Diana Beltekian, Edouard Mathieu, Joe Hasell, Bobbie Macdonald, Charlie Giattino, Cameron Appel, Lucas Rodés-Guirao, and Max Roser. 2020. Coronavirus pandemic (covid-19). *Our World in Data*. <https://ourworldindata.org/coronavirus>.
- Jacques Pierre Robin. 1995. *Revision-Based Generation of Natural Language Summaries Providing Historical Background: Corpus-Based Analysis, Design, Implementation and Evaluation*. Ph.D. thesis, Columbia University, USA. UMI Order No. GAX95-33653.
- Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231.
- Kumiko Tanaka-Ishii, Koiti Hasida, and Itsuki Noda. 1998. [Reactive content selection in the generation of real-time soccer commentary](#). In *COLING 1998 Volume 2: The 17th International Conference on Computational Linguistics*.
- Craig Thomson and Ehud Reiter. 2020. [A gold standard methodology for evaluating accuracy in data-to-text systems](#). In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 158–168, Dublin, Ireland. Association for Computational Linguistics.
- Craig Thomson, Ehud Reiter, and Somayajulu Sripada. 2020. Sportsett: Basketball-a robust and maintainable dataset for natural language generation. In *IntelLanG: Intelligent Information Processing and Natural Language Generation*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.
- Hongmin Wang. 2019. [Revisiting challenges in data-to-text generation with fact grounding](#). In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 311–322, Tokyo, Japan. Association for Computational Linguistics.
- Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. [Challenges in data-to-document generation](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.
- Benjamin D Wissel, P J Van Camp, Michal Kouril, Chad Weis, Tracy A Glauser, Peter S White, Isaac S Kohane, and Judith W Dexheimer. 2020. [An interactive online dashboard for tracking COVID-19 in U.S. counties, cities, and states in real time](#). *Journal of the American Medical Informatics Association*, 27(7):1121–1125.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

The region of {REGION} reported over {NEW_CASES} new cases and {NEW_DEATHS} new deaths this week, a {NEW_CASES_CHANGE}% {INCREASE/DECREASE} and a {NEW_DEATHS_CHANGE}% {INCREASE/DECREASE} respectively compared to the previous week.

The highest numbers of new cases were reported from {MOST_AFFECTED_COUNTRIES_BY_CASES}.

The highest numbers of new deaths were reported from {MOST_AFFECTED_COUNTRIES_BY_DEATHS}.

Globally, over {NEW_CASES} new cases and {NEW_DEATHS} new deaths have been reported to WHO in the past week.

A cumulative total of {CUMULATIVE_CASES} cases and {CUMULATIVE_DEATHS} deaths have been reported since the start of the outbreak.

The number of new cases {INCREASED/DECREASED} by {NEW_CASES_CHANGE}% and the number of new deaths {INCREASED/DECREASED} by {NEW_DEATHS_CHANGE}% globally in the last 7 days.

The WHO Region of {MOST_AFFECTED_REGION} was the most affected region with {MOST_AFFECTED_NEW_CASES} new cases and {MOST_AFFECTED_NEW_DEATHS} new deaths.

This region noted {INCREASE/DECREASE} of {MOST_AFFECTED_CASES_CHANGE}% in new cases since the last week and accounts for {MOST_AFFECTED_NEW_CASES_SHARE}% of all new cases.

Figure 2: Global Template

The placeholder values inside curly braces are filled in from the input tables. The relative change in the number of new cases and new deaths reported is calculated using the data from the current week and the previous week and reported as the

The region of {REGION} reported over {NEW_CASES} new cases and {NEW_DEATHS} new deaths this week, a {NEW_CASES_CHANGE}% {INCREASE/DECREASE} and a {NEW_DEATHS_CHANGE}% {INCREASE/DECREASE} respectively compared to the previous week.

The highest numbers of new cases were reported from {MOST_AFFECTED_COUNTRIES_BY_CASES}.

The highest numbers of new deaths were reported from {MOST_AFFECTED_COUNTRIES_BY_DEATHS}.

Globally, over {NEW_CASES} new cases and {NEW_DEATHS} new deaths have been reported to WHO in the past week.

A cumulative total of {CUMULATIVE_CASES} cases and {CUMULATIVE_DEATHS} deaths have been reported since the start of the outbreak.

The number of new cases {INCREASED/DECREASED} by {NEW_CASES_CHANGE}% and the number of new deaths

A.2 End-to-End Baselines

Each end-to-end baseline model is trained on a single Nvidia GeForce GTX 1080 Ti GPU for 5,000 steps with the following set up:

- **Hierarchical Model (Rebuffel et al., 2020):** This model consists of a transformer encoder and an LSTM decoder with a hierarchical attention mechanism. We use the same set up and hyperparameter values as described in the original repository⁵, except, the number of entities in the encoder is set to 10 here instead of 24 as defined in the original paper. The maximum sequence length is set to 1000 and beam search is applied during inference with beam size equal to 10. It took approximately 8 hours to train this model on a single GPU.
- **T5 Model (Raffel et al., 2020):** We use the implementation of the T5 small model (60M parameters) from the transformers⁶ library by Hugging Face (Wolf et al., 2020). The model comprises 6 layers each in the encoder and decoder with a multi-head attention sub-layer consisting of 8 attention heads. The word embeddings are 512-dimensional and the fully-connected feed-forward sublayers are 2048-dimensional. Sequence length for input and output is set to 1024. The model is trained with the Adam optimizer with a learning rate of 5×10^{-5} . During inference, beam search is applied with a beam of size 10. All the other hyperparameter values are set to their default values. The training process took about 2 hours with a batch size of 4.

A.3 Additional Output Examples

We present outputs generated by the end-to-end baselines as well the template baselines for three tables from the test set of the CURED4NLG dataset.

Table 6 shows a truncated version of the input table for the European region along with corresponding outputs generated by the end-to-end baseline models. Similarly, Table 7 shows the table and outputs generated for an instance in the test set corresponding to the region of Eastern Mediterranean. Finally, Table 8 shows an example of a table from the test set of the CURED4NLG along with the global epidemiological reports generated by the hierarchical and the T5 baseline models.

⁵<https://github.com/KaijuML/data-to-text-hierarchical>

⁶<https://huggingface.co/transformers/>

Reporting Country/ Territory/Area	New cases in last 7 days	Cumulative cases	Cumulative cases per 100k population	New deaths in last 7 days	Cumulative deaths	Cumulative deaths per 100k population	Transmission classification
Europe	1466680	50714995	5435.3	25341	1061218	113.7	-
Turkey	378771	4591416	5444.0	2403	38011	45.1	Community transmission
France	211674	5390187	8287.6	2110	102031	156.9	Community transmission
Germany	145156	3287418	3952.8	1650	81564	98.1	Community transmission
Italy	92074	3949517	6622.1	2345	119021	199.6	Clusters of cases
Ukraine	78761	2025271	4630.9	2537	42323	96.8	Community transmission
Poland	70831	2758856	7268.2	3383	65415	172.3	Community transmission
Russian Federation	60468	4762569	3263.5	2650	108232	74.2	Clusters of cases
Netherlands	57991	1453058	8347.3	135	17038	97.9	Community transmission
Spain	32476	3456886	7303.4	214	77496	163.7	Community transmission
...

<p>The <i>European Region</i> reported over <u>1.4 million</u> new cases and over <u>25000</u> new deaths, a <u>12%</u> and a <u>5% decrease</u> respectively compared to the previous week. The trend in both new cases and deaths is decreasing in the last two weeks. The highest numbers of new cases were reported from <i>Turkey</i> (<u>378771</u> new cases; <u>449.1</u> new cases per 100000; a <u>9% decrease</u>), <i>France</i> (<u>211674</u> new cases; <u>325.5</u> new cases per 100000; a <u>9% decrease</u>), and <i>Germany</i> (<u>145156</u> new cases; <u>174.5</u> new cases per 100000; a <u>1% increase</u>). The highest numbers of new deaths were reported from <i>Poland</i> (<u>3383</u> new deaths; <u>8.9</u> new deaths per 100000; a <u>6% decrease</u>), the <i>Russian Federation</i> (<u>2650</u> new deaths; <u>1.8</u> new deaths per 100000; a <u>2% increase</u>), and <i>Ukraine</i> (<u>2537</u> new deaths; <u>5.8</u> new deaths per 100000; an <u>8% decrease</u>).</p>	<p>The <i>European Region</i> reported over <u>1.2 million</u> new cases and over <u>36000</u> new deaths, a <u>decrease of 18%</u> and <u>8%</u> respectively when compared to the previous week. The three countries reporting the highest numbers of new cases were <i>Spain</i> (<u>7666</u> new cases; <u>160.5</u> new cases per 100000, a <u>15% decrease</u>), <i>France</i> (<u>211674</u> new cases; <u>113.2</u> new cases per 100000; a <u>31% decrease</u>), and <i>France</i> (<u>3958</u> new cases; <u>2.4</u> new cases per 100000, a <u>119021% increase</u>). The highest numbers of new deaths were reported from the <i>Russian Federation</i> (<u>2650</u> new deaths; <u>0.4</u> new deaths per 100000 population; a <u>88% decrease</u>), <i>Germany</i> (<u>214</u> new deaths; <u>0.1</u> new deaths per 100000, a <u>3848% decrease</u>) and <i>France</i> (<u>2110</u> new deaths; <u>0.4</u> new deaths per 100000; a <u>88% decrease</u>).</p>
--	--

Reference	Hierarchical Model
<p>In the past week, the <i>European Region</i> reported over <u>1466000</u> new cases and over <u>25000</u> new deaths, a <u>decrease of 1%</u> and an <u>increase of 1%</u> respectively compared to the previous week. The three countries reporting the highest numbers of new cases were <i>Kosovo</i> (<u>2662</u> new cases; <u>57</u> new cases per 100000; a <u>1% decrease</u>), <i>Turkey</i> (<u>378771</u> new cases; <u>57</u> new cases per 100000; a <u>1% decrease</u>), <i>France</i> (<u>211674</u> new cases; <u>158.8</u> new cases per 100000; a <u>7% decrease</u>). The three countries reporting the highest numbers of new deaths this week were the <i>United Kingdom</i> (<u>157</u> new deaths; <u>3.4</u> new deaths per 100000; a <u>3% decrease</u>), <i>Germany</i> (<u>1650</u> new deaths; <u>3.4</u> new deaths per 100000; a <u>3% decrease</u>), the <i>Russian Federation</i> (<u>2650</u> new deaths; <u>3.7</u> new deaths per 100000; a <u>3% decrease</u>) and the <i>Russian Federation</i> (<u>2345</u> new deaths; <u>3.4</u> new deaths per 100000; a <u>3% decrease</u>).</p>	<p>The <i>European Region</i> reported over <u>1466000</u> new cases and over <u>25000</u> new deaths, a <u>decrease of 1%</u> and <u>2%</u> respectively compared to the previous week. The three countries reporting the highest numbers of new cases were <i>Kosovo</i> (<u>2662</u> new cases; <u>57</u> new cases per 100000; a <u>1% decrease</u>), <i>Turkey</i> (<u>378771</u> new cases; <u>57</u> new cases per 100000; a <u>1% decrease</u>), <i>France</i> (<u>211674</u> new cases; <u>189.8</u> new cases per 100000; a <u>1% decrease</u>), and <i>Germany</i> (<u>1650</u> new cases; <u>1.9</u> new cases per 100000; a <u>7% decrease</u>). The highest numbers of deaths were reported from the <i>Russian Federation</i> (<u>2650</u> new deaths; <u>3.7</u> new deaths per 100000; a <u>3% decrease</u>), and <i>Germany</i> (<u>1410</u> new deaths; <u>0.4</u> new deaths per 100000; a <u>3% decrease</u>).</p>

T5 (pre-trained)	T5 (pre-trained) + clean
<p>The <i>European Region</i> reported over <u>1466000</u> new cases and over <u>25000</u> new deaths, a <u>decrease of 1%</u> and <u>2%</u> respectively compared to the previous week. The three countries reporting the highest numbers of new cases were the <i>United Kingdom</i> (<u>17232</u> new cases; <u>157</u> new cases per 100000 population; a <u>1% decrease</u>), the <i>Russian Federation</i> (<u>60468</u> new cases; <u>2650</u> new cases per 100000; a <u>1% decrease</u>), and the <i>United Kingdom</i> (<u>17232</u> new cases; <u>157</u> new cases per 100000; a <u>1% decrease</u>). The highest numbers of deaths were reported from the <i>United Kingdom</i> (<u>157</u> new deaths; <u>3.7</u> new deaths per 100000; a <u>3% decrease</u>), the <i>Russian Federation</i> (<u>2650</u> new deaths; <u>3.7</u> new deaths per 100000; a <u>3% decrease</u>), the <i>Russian Federation</i> (<u>2650</u> new deaths; <u>3.7</u> new deaths per 100000; a <u>3% decrease</u>), and the <i>Russian Federation</i> (<u>2650</u> new deaths; <u>3.7</u> new deaths per 100000; a <u>3% decrease</u>).</p>	<p>In the past week, the <i>European Region</i> reported over <u>1466000</u> new cases and over <u>25000</u> new deaths, a <u>decrease of 1%</u> and <u>2%</u> respectively compared to the previous week. The three countries reporting the highest numbers of new cases were <i>Kosovo</i> (<u>2662</u> new cases; <u>57</u> new cases per 100000; a <u>1% decrease</u>), <i>Turkey</i> (<u>378771</u> new cases; <u>57</u> new cases per 100000; a <u>1% decrease</u>), <i>France</i> (<u>211674</u> new cases; <u>59.6</u> new cases per 100000; a <u>21% decrease</u>). The highest numbers of new deaths were reported from the <i>Russian Federation</i> (<u>2650</u> new deaths; <u>3.7</u> new deaths per 100000; a <u>3% decrease</u>), <i>Germany</i> (<u>1650</u> new deaths; <u>3.4</u> new deaths per 100000; a <u>3% decrease</u>), and <i>Poland</i> (<u>2537</u> new deaths; <u>0.1</u> new deaths per 100000; a <u>3% decrease</u>).</p>

T5 (pre-trained) + shuffle

T5 (pre-trained) + subset

Table 6: Sample output for an epidemiological report for the European region generated by the T5 model and the hierarchical model for a table of data in the test set of CURED4NLG. Text in blue italics shows information filled in from the input table by the baseline template. The text in green italics shows tabular values correctly produced by the end-to-end baseline models while underlined text in red shows the errors in the generated texts. Any hallucinations or repetitions generated are highlighted in purple.

Reporting Country/Territory/Area	New cases in last 7 days	Cumulative cases	Cumulative cases per 100k population	New deaths in last 7 days	Cumulative deaths	Cumulative deaths per 100k population
Eastern Mediterranean	220035	9648410	1320.2	4709	193761	26.5
Iran (Islamic Republic of)	99205	2739875	3262.0	2109	76633	91.2
Iraq	28359	1136917	2826.6	189	15930	39.6
Pakistan	20511	874751	396.0	670	19467	8.8
Bahrain	11188	199093	11700.5	59	737	43.3
United Arab Emirates	10486	544931	5509.7	19	1629	16.5
Egypt	8248	244520	238.9	424	14269	13.9
Saudi Arabia	6827	432269	1241.7	88	7147	20.5
Kuwait	6725	290801	6809.4	52	1687	39.5
Tunisia	6320	325832	2756.9	429	11779	99.7
Oman	5569	204913	4012.7	101	2184	42.8
Jordan	4112	723345	7089.4	167	9243	90.6
Lebanon	2964	535233	7841.7	125	7585	111.1
...

<p>The Eastern Mediterranean Region reported over 220000 new cases and over 4700 new deaths, a 22% and a 16% decrease respectively compared to the previous week. Case and death incidences have decreased steeply for the past four and two weeks respectively. The highest numbers of new cases were reported from the Islamic Republic of Iran (99205 new cases; 118.1 new cases per 100000; a 20% decrease), Iraq (28359 new cases; 70.5 new cases per 100000; a 26% decrease), and Pakistan (20511 new cases; 9.3 new cases per 100000; a 29% decrease). The highest numbers of new deaths were reported from the Islamic Republic of Iran (2109 new deaths; 2.5 new deaths per 100000; a 13% decrease), Pakistan (670 new deaths; 0.3 new deaths per 100000; a 20% decrease), and Tunisia (429 new deaths; 3.6 new deaths per 100000; a 21% decrease).</p>	<p>The Eastern Mediterranean Region reported over 61700 new cases the past week, a 14% decrease compared to the previous week and reported just under 800 new deaths, a 16% decrease. Of the 22 countries in the region the number of new cases rose in 6 (27%) this week while they fell in 10 of 22 (1189%). The number of new deaths increased in 1 of 22 countries (14%) and declined in 5 of 22 (23%) this week. The three countries reporting the highest numbers of new cases in the region this week were Iraq (28359 new cases; 90.2 new cases per 100000; a 15% decrease), Lebanon (2964 new cases; 4.0 new cases per 100000; a 1% decrease) and Iraq (28359 new cases; 40.4 new cases per 100000; a 13% increase). The highest numbers of new deaths were reported from Islamic Republic of Iran (19 new deaths; 5.6 new death per 100000 population, an 8% decrease), and Iraq (19 new deaths, 4.3 new deaths per 100000, a 12% decrease).</p>
--	--

Reference	Hierarchical Model
<p>In the past week, the Eastern Mediterranean Region reported over 220035 new cases, a decrease of 9% compared to last week. The region reported over 4709 new deaths, a 2% decrease. The highest numbers of new cases were reported from the Islamic Republic of Iran (99205 new cases; 2109 new cases per 100000 population; a 2% decrease), Iraq (28359 new cases; 189 new cases per 100000; a 2% decrease), Pakistan (20511 new cases; 59 new cases per 100000; a 24% increase). The highest numbers of new deaths were reported from the same countries, the Islamic Republic of Iran (2109 new deaths; 4.7 new deaths per 100000; a 2% decrease), Pakistan (670 new deaths; 4.7 new deaths per 100000; a 2% decrease) and Egypt (2109 new deaths; 0.1 new deaths per 100000; a 2% decrease).</p>	<p>In the past week, the Eastern Mediterranean Region reported over 220035 new cases, a decrease of 9% compared to last week. The region reported over 4709 new deaths, a 2% decrease. The highest numbers of new cases were reported from the Islamic Republic of Iran (99205 new cases; 2109 new cases per 100000 population; a 2% decrease), Iraq (28359 new cases; 189 new cases per 100000; a 2% decrease), Pakistan (20511 new cases; 59 new cases per 100000; a 7% decrease). The highest numbers of new deaths were reported from the same countries including the Islamic Republic of Iran (2109 new deaths; 4.7 new deaths per 100000; a 2% decrease), Pakistan (670 new deaths; 3.7 new deaths per 100000; a 2% decrease) and Egypt (29 new deaths; 0.4 new deaths per 100000; a 2% decrease).</p>

T5 (pre-trained)	T5 (pre-trained) + clean
<p>In the past week, the Eastern Mediterranean Region reported over 220035 new cases and over 4709 new deaths, a decrease of 1% and 2% respectively compared to the previous week. The three countries reporting the highest numbers of new cases this week were Islamic Republic of Iran (99205 new cases; 29.5 new cases per 100000 population; a 21% decrease), United Arab Emirates (10486 new cases; 189 new cases per 100000; a 1% decrease), and United Arab Emirates (10486 new cases; 59 new cases per 100000; a 7% decrease). The highest numbers of new deaths were reported from the Islamic Republic of Iran (2109 new deaths; 3.7 new deaths per 100000; a 3% decrease), Lebanon (429 new deaths; United Arab Emirates new deaths per 100000; a 3% decrease) and United Arab Emirates (United Arab Emirates new deaths; United Arab Emirates new deaths per 100000; a 3% decrease).</p>	<p>In the past week, the Eastern Mediterranean Region reported over 220035 new cases, a decrease of 1% compared to last week. The region reported over 47000 new deaths, a 2% decrease. The highest numbers of new cases were reported from the Islamic Republic of Iran (99205 new cases; 2109 new cases per 100000 population; a 2% decrease), Iraq (28359 new cases; 189 new cases per 100000; a 2% decrease), Pakistan (20511 new cases; 59 new cases per 100000; a 7% decrease). The highest numbers of new deaths were reported from the Islamic Republic of Iran (2109 new deaths; 0.3 new deaths per 100000; a 3% decrease), Pakistan (670 new deaths; 0.3 new deaths per 100000; a 3% decrease) and Saudi Arabia (88 new deaths; 0.4 new deaths per 100000; a 2% decrease).</p>

T5 (pre-trained) + shuffle

T5 (pre-trained) + subset

Table 7: Sample output for an epidemiological report for the region of Eastern Mediterranean generated by the T5 model and the hierarchical model for a table of data in the test set of CURED4NLG. Text in blue italics shows information filled in from the input table by the baseline template. The text in green italics shows tabular values correctly produced by the end-to-end baseline models while underlined text in red shows the errors in the generated texts. Any hallucinations or repetitions generated are highlighted in purple.

WHO Region	New cases in last 7 days (%)	Change in last 7 days	Cumulative cases (%)	New deaths in last 7 days (%)	Change in last 7 days	Cumulative deaths (%)
Americas	1272491 (23%)	-4%	63554005 (40%)	33879 (38%)	-8%	1551860 (47%)
Europe	919119 (17%)	-23%	52871662 (34%)	19056 (21%)	-18%	1104629 (34%)
South-East Asia	2877410 (52%)	6%	25552640 (16%)	28977 (32%)	15%	309197 (9%)
Eastern Mediterranean	280853 (5%)	-13%	9428375 (6%)	5605 (6%)	-13%	189052 (6%)
Africa	40656 (1%)	-5%	3357846 (2%)	1034 (1%)	3%	83904 (3%)
Western Pacific	127073 (2%)	-4%	2597134 (2%)	1691 (2%)	34%	39179 (1%)
Global	5517602 (100%)	4%	157362408 (100%)	90242 (100%)	-4%	3277834 (100%)

<p>The number of new COVID-19 cases and deaths globally decreased slightly this week, with over 5.5 million cases and over 90000 deaths (Figure 1). Case and death incidence, however, remains at the highest level since the beginning of the pandemic. New weekly cases decreased in the regions of Europe and Eastern Mediterranean, while the South-East Asia Region continued an upward trajectory for 9 weeks and reported a further 6% increase last week (Table 1). Death incidence increased in the South-East Asia and Western Pacific regions. While India continues to account for 95% of cases and 93% of deaths in the South-East Asia Region, as well as 50% of global cases and 30% of global deaths, worrying trends have been observed in neighbouring countries. In all WHO Regions there are countries which have been showing a sustained upward trend in cases and deaths over several weeks. The highest numbers of new cases were reported from India (2738957 new cases; 5% increase), Brazil (423438 new cases; similar to previous week), the United States of America (334784 new cases; 3% decrease), Turkey (166733 new cases; 35% decrease), and Argentina (140771 new cases; 8% decrease).</p>	<p>Globally, over 5.5 million new cases and 90000 new deaths have been reported to WHO in the past week. A cumulative total of 157.3 million cases and 3.2 million deaths have been reported since the start of the outbreak. The number of new cases increased by 4% and the number of new deaths decreased by 4% globally in the last 7 days. The WHO Region of South-East Asia was the most affected region with 2.8 million new cases and 28000 new deaths. This region noted an increase of 6% in new cases since the last week and accounts for 52% of all new cases. Regions reporting an increase in new cases include South-East Asia. Regions reporting a decline in new cases include Africa, Americas, Eastern Mediterranean, Europe and Western Pacific. Regions reporting an increase in new deaths include Africa, South-East Asia and Western Pacific. Regions reporting a decline in new deaths include Americas, Eastern Mediterranean and Europe. The highest numbers of new cases were reported from India (2738957 new cases; 5% increase), Brazil (423438 new cases; similar to previous week), United States of America (334784 new cases; 3% decrease), Turkey (166733 new cases; 35% decrease) and Argentina (140771 new cases; 8% decrease).</p>
--	---

Reference	Template
<p>In the past week, the number of new COVID-19 cases and deaths has continued to increase, with over 1.2 million new cases reported globally, a 4% increase compared to the previous week (Figure 1). The Region of the Americas continues to account for the greatest proportion of new cases and new deaths globally. The Region of the Americas continues to account for the greatest proportion of new cases and deaths globally. The Region of the Americas continues to report a decline in new cases and deaths. The Eastern Mediterranean Region reported a decline in new cases and deaths, with over 1.2 million new cases reported last week. The Region of the Americas continues to report a decline in new cases and deaths, while the Eastern Mediterranean Region reported a decline in new cases and deaths, with over 1.2 million new cases reported last week. The Region of the Americas continues to report a decline in new cases and deaths, with over 1.2 million new cases reported last week, a 4% increase compared to the previous week. The Region of the Americas continues to report a decline in new cases and new deaths, with over 1.2 million new cases reported last week. The Region of the Americas [...]</p>	<p>The number of global new cases reported continues to fall for the sixth consecutive week, with 2.4 million new cases and 36000 new deaths reported globally, while the number of new deaths has remained relatively stable. As of 18 October, over 40 million cases and 1.1 million deaths have been reported globally. The further acceleration in the incidence of new cases was most notable in European Region, which reported half of global new cases (over 1.7 million cases - a 22% increase from the previous week). Moreover, the region also reported a substantial rise in the number of new deaths (a 46% increase compared with the previous week), with Global new deaths in the past week. The WHO South East Asia Region showed the highest rise in new cases in the past week, with over 500,000 new cases reported. In the European Region, new cases and new deaths have continued to increase over the past seven days compared to the previous week. Along with the Region of the Americas, the percentage change in new cases in Global the week. The Eastern Mediterranean Region reported a decline in new cases and deaths, 6% and 8% respectively, compared to the previous week. The decline is mainly due to decreases in reported cases in India and Bangladesh. For the second week in a row, the Regions of the Eastern Mediterranean and the Western Pacific reported increases in cases and deaths. Overall, during the reporting period, all the Regions showed an increase in cases except the South-East Asia Region. Countries reporting the highest number of cases in the past seven days include; India, the United States of America, Brazil, the United Kingdom and France.</p>

T5 (pre-trained)

Hierarchical Model

Table 8: Sample output for a global epidemiological report generated by the T5 model and the hierarchical model for a table of data in the test set of CURED4NLG. Text in blue italics shows information filled in from the input table by the baseline template. The text in green italics shows tabular values correctly produced by the end-to-end baseline models while underlined text in red shows the errors in the generated texts. Any hallucinations or repetitions generated are highlighted in purple.

Beyond Concatenative Morphology: Applying OntoLex-Morph to Maltese

Maxim Ionov

University of Cologne, Germany
mionov@uni-koeln.de

Michael Rosner

University of Malta
mike.rosner@um.edu.mt

Abstract

OntoLex-Morph is an extension of OntoLex-lemon, (a *de facto* standard vocabulary for publishing lexical data) that is designed to accommodate the description of morphological phenomena into lexical datasets. It is intended to be universally applicable, but so far its application has been focused on the more familiar European languages. This article attempts to show that the morphology extension to OntoLex-lemon can also be applied to Maltese, and by extension, to other Semitic languages. We present our modelling, show how generation rules can be used, and offer some recommendations for changes to the module which would considerably improve the transparency of descriptions that make use of it. Finally, we conclude that if such recommendations are accepted, future discussion should attempt to better delimit the scope of the module to avoid incorporation of information that rightly belongs elsewhere.

1 Introduction

OntoLex is a formal model for representing lexical resources, such as dictionaries and thesauri, in a machine-readable format.¹ It was developed to provide a standardised framework for representing lexical entities and relationships between them, with the aim of improving interoperability and reusability of lexical data across different applications and domains.

OntoLex is an RDF model built on top of existing semantic web standards. This allows for the interoperability and integration of lexical resources with other semantic web resources, and for the querying and analysis of lexical data using RDF-based tools and applications.

The model was designed to be modular and extensible, with different modules representing different aspects of lexical information, such as lexical

senses, syntactic frames, and semantic relations. This allows for the representation of complex lexical information in a structured and flexible way, and for the customisation of the model to suit different linguistic and domain-specific needs.

One of the modules that is currently being developed is *OntoLex-Morph*, a module that allows representing rich morphological information that is often provided in lexicographic resources. In addition to representing static data such as morphemes and their grammatical information, the module provides the means to model information on how to generate wordforms given lexical entries and finite state-like rules. Despite being developed with a goal to support a wide variety of languages and language phenomena, to the best of our knowledge, it has not yet been applied to languages with nonconcatenative morphology.² Semitic languages, having a system of consonantal roots with a complex system of inflection and derivation, belong to this category. In this paper we show how OntoLex-Morph can be applied to model lexical data from one such language, Maltese. Although various computational approaches to Maltese lexical and morphological data have been proposed (e.g. [Borg and Gatt \(2017\)](#); [Ravishankar et al. \(2017\)](#); [Sagot and Walther \(2013\)](#)), this is a first time a linked-data approach has been investigated. We present a small subset of a Maltese dictionary together with a discussion of issues encountered along the way. Additionally, we provide a reference implementation for form generation, bringing the model one step closer to completion.

The rest of the paper is structured as follows: Section 2 provides an overview of the Maltese language and describes the phenomena we chose for this paper. Section 3 gives an overview of OntoLex and OntoLex-Morph vocabularies. In Section 4 we talk about modelling decisions for both static

¹<https://www.w3.org/2016/05/ontolex/>.

²At least to languages where it is the primary way of inflection and derivation.

data and generation rules and present our reference implementation for form generation. Finally, we discuss what we found along the way, whether the model as it is right now is suitable for such data (spoiler: we think so), and suggest some additions that could help the model transparency.

2 The Maltese Language

Maltese is a mixed language made up of Semitic and romance substrates, which respectively share many important characteristics of other languages in those classes. In this article we focus on the Semitic substrate which manifests itself both lexically and morpho-syntactically with respect to different syntactic categories. Thus, the Maltese words *kelb* (Eng. “dog”) and *kiteb* (Eng. “write”) not only resemble their counterparts in e.g. Arabic and Hebrew from a lexical perspective, but are susceptible to morphological processes for generating nominal and verbal paradigms similar to those operating in such languages. These processes are a superset of the affixation phenomena that characterise most European languages, primarily because word formation in Semitic languages is based on roots and templates. The formation of a word is effected in part by *interdigitation* whereby a pair of vowels called a *vocalism* is inserted into a sequence of consonants. To give a simple example, the word *kiteb* is formed by interdigitating *i-e* with *k-t-b*.

The result of such interdigitation may be a word in its own right or may, as in the case of verbs, be subjected to further processes to yield a complete conjugation paradigm. These vary greatly in complexity, from simple affixation to subtle vowel changes depending on considerations of syllabic structure and vowel harmony. Maltese has no infinitive form, so for citing lexical entries for verbs, the de facto convention is to use the third person singular masculine (3SG.M) perfective form since many other verbal forms can be derived from it relatively easily. We refer the reader to [Rosner and Borg \(2022\)](#) for further details on the Maltese language.

In this article we focus on the extent to which it is possible to generate complete paradigms using the morphological rules proposed by *OntoLex-Morph*. We are primarily concerned with the generative capacity of such rules. Subsequently we will turn to some considerations of their descriptive efficiency. We start with the easiest case of well-behaved Mal-

person/gender	perfective	imperfective
1SG	<i>ktibt</i>	<i>nikteb</i>
2SG	<i>ktibt</i>	<i>tikteb</i>
3SG.M	<i>kiteb</i>	<i>jikteb</i>
3SG.F	<i>kitbet</i>	<i>tiktbe</i>
1PL	<i>ktibna</i>	<i>niktbu</i>
2PL	<i>ktibtu</i>	<i>tiktbu</i>
3PL	<i>kitbu</i>	<i>jiktbu</i>

Figure 1: Conjugation of *kiteb*

tese strong verbs (such as *kiteb*).

2.1 Maltese Strong Verbs

Many verbs within the Semitic substrate of Maltese are trilateral i.e. built from a skeleton of three consonants. There are two aspects: perfective and imperfective.

In *perfective* aspect, suffixes *-t*, *-et*, *-na*, *-tu*, *-u* mark person, first vowel is deleted with consonant-initial suffixes; second vowel *e* before consonant initial suffix (i.e. when stressed), becomes *i*. In *imperfective* aspect, prefixes mark person, suffix *-u* marks plural. A completely regular example is *kiteb* (Eng. ‘he wrote’) which conjugates as shown in Figure 1.

There are many ways these conjugations can vary when one of the root consonants (radicals) falls into a certain category. Thus, when the *first* radical is *silent gh* or *h*, the first vowel is retained when there is a consonant-initial suffix. So for the verb *ghamel* (Eng. ‘he made’) we have *ghamilt* instead of **ghmilt* as shown in Figure 6 in the Appendix.

On the other hand, when the *second* radical is a *liquid* consonant, i.e. an *l*, *m*, *n*, *r*, an issue arises in terms of pronunciation of plural imperfective forms, so a helping (euphonic/epenthetic) vowel is required and placed between the first and second root consonants. So for the verb *telaq* (Eng. ‘he left’) we have e.g. *nitilqu* instead of **nitlqu* as shown in Figure 7 in the Appendix.

These examples are by no means exhaustive, but they clearly illustrate the need to discriminate behaviour on the basis of consonant classes.

2.2 Maltese Alphabet

The Maltese alphabet is based on the Latin one and comprises 6 vowels — *a e i o u ie* — and 24 consonants — *b ċ d f ġ g ħ h ħ j k l m n p q r s t v w x ž z*. It poses two challenges when formulating

Class	Characters
silent	<i>gh, h</i>
liquid	<i>l, m, n, r</i>
normal	<i>b c d f g g h j k p q s t v w x z z</i>

Figure 2: Character classes

replacement rules regardless of a formalism. First, it contains digraphs, *ie* and *gh*. So if there are rules that operate with concepts like a “letter”, a vowel, or a consonant it cannot be just assumed that one letter is one character. When working with regular expressions, for example, it would be incorrect to simply use `.` or `\w` to represent any letter of the alphabet. Furthermore, if we aim to minimise the number of rules and create them as universal as possible, we need means to refer to certain character classes. Based on the examples above, in order to discriminate amongst the classes of verb to which the above cases belong, we need to distinguish at least between silent, liquid and normal consonants as listed in Figure 2.

3 OntoLex and OntoLex-morph

OntoLex-lemon (McCrae et al., 2017) is the *de facto* standard for publishing lexical resources in RDF, compliant with established web standards. The central class in the core model, depicted in Figure 3 is `LexicalEntry` — a lexeme or a dictionary entry. It must have at least one (word)form (`canonicalForm`) and can have a number of other forms, a number of senses, which can then be then linked to either lexical concepts or entities in an ontology. Basic morphological information like a part of speech and grammatical categories can be provided for lexical entries and forms using elements of any suitable vocabulary, such as LexInfo.³

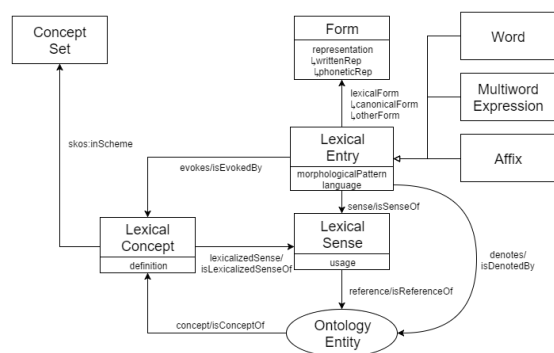


Figure 3: OntoLex-Lemon core model

³<https://lexinfo.net/>.

One thing to note is that a single lexical entry cannot have more than one part of speech, which is an important factor for our design decisions described below.

Although there is a place for including basic morphological information in the core model, it does not allow the representation of paradigmatic relationships between lexical entries and forms (inflectional morphology) or derivational relationships between lexical entries. In order to close this gap, an extension to the core module, OntoLex-Morph is being developed.⁴ The model, depicted in Figure 4 consists of three main parts: derivation (left), inflection (right), and rules for generating new forms, both for inflection and derivation (top). The central part of the module is the class `Morph`, which corresponds to a morph — a specific realisation of a morpheme. It is a subclass of `LexicalEntry`, which might be a bit counterintuitive at first, but this allows for resources where morphs are dictionary entries of their own.

Another part of OntoLex-Morph important for us is a representation of rules that can be used to generate forms from lexical entries (or, more specifically, from their forms). The mechanism behind this is the following:⁵ (i) A lexical entry can be a part of an inflectional paradigm. (ii) For each paradigm, there can be a number of rules, each of them having information on how to produce a form and grammatical meaning that should be assigned to this form; (iii) The formalism to encode a rule is a (POSIX-compatible) regular expression.

For example, a rule for forming a standard English plural form can look as following:

```
<rule_plural>
  a morph:InflectionRule ;
  morph:replacement [
    a morph:Replacement ;
    morph:source "$" ;
    morph:target "s"@en ; ] ;
  morph:involves [
    a ontolex:Affix ;
    rdfs:label "-s"@en ;
    morph:grammaticalMeaning [
      a morph:MorphologicalMeaning ;
      lexinfo:number lexinfo:plural ] ] .
```

It is, of course, possible to use instances of the `morph:Morph` class (and its subclasses) instead of blank nodes, and in most situations this will be the case. However, this will depend on the

⁴<https://www.w3.org/community/ontolex/wiki/Morphology>.

⁵Here, we focus on the rules for generating inflected forms. For the more complete description of the model refer to Chiaros et al. (2022).

lated words, and should therefore be explicitly reflected in the modelling.

This creates a choice: to model a root as a lexical entry and all the forms derived from it as forms, or to represent each lexeme as a lexical entry, with verbs having their 3SG.M form as their canonical form, additionally connecting each lexical entry to its root. There are good reasons to prefer the latter. First of all, the principle of separating lexemes into different lexical entries while preserving root information is shared by printed dictionaries of Maltese, e.g. Aquilina (1987) and other Semitic languages. The resource we are modeling, Ġabra, also shares this design. Second, lexical entries in OntoLex cannot have more than one part of speech, which makes using roots as lexical entries problematic, if not impossible. Additionally, this fits into the model's dichotomy of inflection vs. derivation, where semantically related entries (e.g. 'to write' vs. 'writer') could be distinct lexical entries connected by a derivational relationship instead of two forms, members of the same inflectional paradigm.

We therefore represent root consonants as a `lexinfo:RootMorph`, a subclass of `morph:Morph`, and each form that stems from that root cluster is connected to that morph with the property `morph:consistsOf`.

This way, for each verb we are modelling, there is a single lexical entry and a canonical form that corresponds to a 3SG.M perfective form. That form is connected to the corresponding root morph. Also, this form is connected to the lexical entry as a `morph:baseForm`, which means that its written representation will be used as a base for form generation. Furthermore, the lexical entry links to a corresponding `morph:Paradigm` to specify an inflectional paradigm for that word:

```
roots:k-t-b    a lexinfo:RootMorph ;
               rdfs:label "k-t-b" .

:1    a ontolex:Word ;
      lexinfo:partOfSpeech lexinfo:verb ;
      morph:morphologicalPattern
        <kiteb_paradigm> ;
      ontolex:canonicalForm <l_form> ;
      morph:baseForm <l_form> .

<l_form>    a ontolex:Form ;
            morph:consistsOf roots:k-t-b ;
            ontolex:writtenRep "kiteb"@mlt.
```

Instead of explicitly providing the forms, we provide rules for how the forms should be generated for each of the verbs as described in Section 2.1. As described above, the core of each rule is a map-

ping as specified by a pair of regular expressions: a source and a replacement. Unlike the example for English plural above, we need to match the whole form and replace it with a new one. Since we know the number of characters in the base form, we can simply match each of them to a capturing group. To illustrate this with respect to the perfective 3SG.M → 1SG mapping of *kiteb* we can use the following:

```
source:      (.) (.) (.) (.) (.)
replacement: \\1\\3i\\5t
```

The input specifies a sequence of 5 segments. The dot is an unrestricted wildcard matching any character. Thus the input matches any sequence of 5 characters, which become bound, in order, to numerical variables 1–5. Thus after matching *kiteb*, 1=k, 3=t, 5=b, and the output, `\\1\\3i\\5t = kiibt`

The problem with this approach comes from the fact that it assumes that each letter corresponds to one character, which is not true for Maltese alphabet. Instead, we need to provide a list of possible options for each of the positions:

```
(b|ċ|d|f|ġ|g|għ|h|h̄|j|k|l|m|n|p|q|r|s|t
|v|w|x|ż|z) (a|i|e|e|i|o|u)
(b|ċ|d|f|ġ|g|għ|h|h̄|j|k|l|m|n|p|q|r|s|t
|v|w|x|ż|z) (a|i|e|e|i|o|u)
(b|ċ|d|f|ġ|g|għ|h|h̄|j|k|l|m|n|p|q|r|s|t
|v|w|x|ż|z)
```

This can be slightly simplified by tailoring each group to symbols that can appear in a given paradigm, but even in this case, rules produced this way are clearly unwieldy. A simple yet elegant approach would be to use character classes like:

```
source:      (C) (V) (C) (V) (C)
```

where C and V respectively stand for the sets of consonants and vowels. Using this logic, it is possible to use more specific character classes, e.g. liquid consonants, to reduce the number of paradigms by creating more universal rules. However, this would again make the rules more complex and less readable. In our dataset we tried to keep the balance, creating three paradigms (and three sets of rules) for each of the cases described in Section 2.1.

4.2 Character classes and generation

An important question with regards to character classes is where and how to model them. We see three distinct possibilities: (i) externally, using a preprocessor to generate rules without character classes or generate forms directly; (ii) with a dataset-specific property; (iii) with a property specified in OntoLex-Morph. While the first two options

are less invasive and prevent the module from growing in complexity, it is worth noting that only the last option allows interoperability and reusability, not only for rules themselves, but also for any software that will use these rules. In our modelling, we propose a class `CharacterClass` that can be used in the following way:

```
gabra:V a gabra:CharacterClass ;
    rdfs:label "V" ;
    rdfs:member "a", "e", "i", "o", "u" .
```

5 Conclusion

We have verified the hypothesis that Morph can be applied to some key non-concatenative morphological phenomena in Maltese. The implication is that this generalises to other Semitic languages. We have also illustrated the need to provide facilities for incorporating definitions of character classes. The dataset, our implementation of form generation, and additional information can be found on GitHub.⁹

The main discussion point to emerge is whether such definitions should be external or internal to OntoLex-Morph. The pros of keeping character classes external is that the module remains lightweight. However there is a price to be paid. At some point, externally defined character classes will have to be replaced in each rule with lists of characters that will become exceedingly verbose and illegible. Conversely, character class definitions could become an integral part of the module. We favour the latter approach on the grounds that the benefit of legibility for producers and consumers of morphological information far outweighs the cost of slightly increased complexity in the formalism.

Of course there are limits to this line of argumentation. It would be theoretically possible to absorb morphological processing of arbitrary complexity (e.g. to include the article used with nouns, clitic pronouns, etc. all of which end up as one word on the page). However, the inclusion of this level of expressivity would contradict the intention to keep the module reasonably simple and transparent. The module aims to represent elements involved in both the decomposition and formation of lexical entries/word forms (Klimek et al., 2019, p. 579), but fine-grained description of phonological processes involved in stem or word formation on the

phoneme level is excluded.

The line between justified and unjustified refinements to OntoLex-Morph is delicate, but somewhere in between the two is an as yet unidentified cutoff point whose placement would be an apt task for imminent future discussion.

Acknowledgements

The research described in this paper was conducted in the context of the COST Action NexusLinguarum – European network for Web-centered linguistic data science (CA18209).

References

- J. Aquilina. 1987. *Maltese-English Dictionary*. Midsea Books, Valletta.
- C. Borg and A. Gatt. 2017. [Morphological analysis for the Maltese language](#). In *Proc. 3rd Arabic NLP Workshop*, pages 25–34, Valencia, Spain. Ass. Comp. Ling.
- J. Camilleri. 2013. *A Computational Grammar and Lexicon for Maltese*. MSc Thesis, Chalmers University of Technology, Gothenburg, Sweden.
- C. Chiarcos, K. Gkirtzou, F. Khan, P. Labropoulou, M. Passarotti, and M. Pellegrini. 2022. Computational morphology with OntoLex-Morph. In *Proc. 8th Workshop on Linked Data in Linguistics*, pages 78–86.
- Brother FSC Henry. 1980. *Grammatika Maltija*. De La Salle Brother Publications, Malta.
- B. Klimek, J. McCrae, J. Bosque-Gil, M. Ionov, J. Tauber, and C. Chiarcos. 2019. Challenges for the representation of morphology in ontology lexicons. *Proceedings of eLex 2019*, pages 570–591.
- John P McCrae, J. Bosque-Gil, J. Gracia, P. Buitelaar, and P. Cimiano. 2017. The OntoLex-Lemon Model: Development and Applications. In *Proceedings of eLex-2017*, pages 19–21.
- V. Ravishankar, F. Tyers, and A. Gatt. 2017. [A morphological analyser for Maltese](#). In K. Shaalan and S. El-Beltagy, editors, *Procedia Computer Science vol 117*, pages 175–182. Elsevier.
- M. Rosner and C. Borg. 2022. Report on the Maltese Language. *Deliverable D1.25, European Language Equality Project*.
- B. Sagot and G. Walther. 2013. [Implementing a formal model of inflectional morphology](#). In *3rd International W/S on Systems and Frameworks for Computational Morphology*, volume 380 of *Communications in Computer and Information Science*, pages 115–134, Berlin. Humboldt-Universität, Springer.

⁹<https://github.com/max-ionov/maltese-morph>.

A Appendix

person/gender	perfective	imperfective
1SG	<i>ktibt</i>	<i>nikteb</i>
2SG	<i>ktibt</i>	<i>tikteb</i>
3SG.M	<i>kiteb</i>	<i>jikteb</i>
3SG.F	<i>kitbet</i>	<i>tikteb</i>
1PL	<i>ktibna</i>	<i>niktbu</i>
2PL	<i>ktibtu</i>	<i>tiktbu</i>
3PL	<i>kitbu</i>	<i>jiktbu</i>

Figure 5: Conjugation of *kiteb*

person/gender	perfective	imperfective
1SG	<i>ghamilt</i>	<i>nghamel</i>
2SG	<i>ghamilt</i>	<i>tghamel</i>
3SG.M	<i>ghamel</i>	<i>jghamel</i>
3SG.F	<i>ghamlet</i>	<i>tghamel</i>
1PL	<i>ghamilna</i>	<i>nghamlu</i>
2PL	<i>ghamiltu</i>	<i>tghamlu</i>
3PL	<i>ghamlu</i>	<i>jghamlu</i>

Figure 6: Conjugation of *ghamel*

person/gender	perfective	imperfective
1SG	<i>tlaqt</i>	<i>nitlaq</i>
2SG	<i>tlaqt</i>	<i>titlaq</i>
3SG.M	<i>telaq</i>	<i>jitlaq</i>
3SG.F	<i>telqet</i>	<i>titlaq</i>
1PL	<i>tlaqna</i>	<i>nitilqu</i>
2PL	<i>tlaqtu</i>	<i>titilqu</i>
3PL	<i>telqu</i>	<i>jitilqu</i>

Figure 7: Conjugation of *telaq*

Towards Language Acquisition Through Cross-Language Etymological Links in Linguistic Linked Open Data

<p>Maxim Dužij Prague University of Economics and Business Nám. W. Churchilla 4 130 67 Praha 3, Czech Rep. maxim.duzij@hotmail.com</p>	<p>Vojtěch Svátek Prague University of Economics and Business Nám. W. Churchilla 4 130 67 Praha 3, Czech Rep. svatek@vse.cz</p>	<p>Petr Strossa Prague University of Economics and Business Nám. W. Churchilla 4 130 67 Praha 3, Czech Rep. petr.strossa@vse.cz</p>
---	--	--

Abstract

We explore the possibility of using linguistic linked open data for supporting a foreign language acquisition application through cross-language links. The links in the used LLOD resource, the Etytree knowledge graph, are primarily of etymological nature. Through a questionnaire survey we explore what interval of an edit distance measure may be suitable as guidance for offering word pairs (in an unknown and known language), connected with an etymological chain, that are too dissimilar to immediately remind of the learned word when encountering the known word but allowing to establishing a mental association between them when seeing both. A proof-of-concept application was also designed and tested for usability. While the principles of the approach look viable after this initial study, our conclusion is that large-scale enhancement of the underlying LLOD resources will be needed before tools could be delivered for real use. An edit distance measure, particularly one sensitive to cross-language character mapping, may be useful for selecting training cases with respect to the language-acquisition proficiency of the learner.

1 Introduction

One of the important aspects of linguistic linked open data (LLOD) is the consideration of cross-language links. While many efforts have been centred on semantic equivalence links, useful for tasks such as search or translation, less attention has been paid to etymological links (whether cross- or intra-language ones). A prominent recent project is Etytree (Pantaleo et al., 2017), which produced a tool for interactively exploring etymologically related words. Its target user group are the researchers and public interested in the study of etymology, who can benefit from intuitive graph-based visualization of etymological links.

We hypothesize that another beneficiary of LLOD with etymology coverage could be *foreign*

language learners. Experts generally agree that etymology is one of language aspects (together with phonology, morphology, semantics and syntax) relevant for language acquisition (Rothstein and Rothstein, 2008). However, the studies have so far been focused on classroom educational setting, and largely agnostic of support that could be provided by online databases.

Presumably, the benefits of etymology would vary across several dimensions of language learning, such as: the prior knowledge of the target (to-be-learned) and background (native or better commanded) language/s by the learner; the closeness of those languages as such; active vs. passive vocabulary acquisition setting; written vs. spoken form of the language; personal characteristics of the learner. As a promising case we want to primarily focus on is that of *passive* acquisition of (primarily) *written* form of words in the target language that has *observable* but *not strikingly obvious* etymologically justified surface similarity to words in a background language the learner knows better. Since the probability of finding such background language words increases with the number (and, perhaps, taxonomic variety) of mastered background languages, the gain might be highest for learners moderately or highly equipped with prior knowledge of languages, who at the same time experience limitations in pure memorization of words and their meanings by heart. Let us consider the following scenario:

1. The learner is exposed to a word in the target language.
2. S/he acquires the meaning of the word using a dictionary or thesaurus.
3. In the course of time, s/he encounters the word repeatedly, and has to look the meaning up again and again – until the bond between the written word and its meaning becomes firm enough.

The key question is whether showing the word together with a *personalized* etymological context, in step 2, would reduce the number of repeated look-ups in the next phase. Obviously, while showing a given word with its generic etymological context (as performed by the Etytree application) is not much different from what even paper-based etymological resources can provide, the power of LLOD knowledge graphs might nicely manifest through such dynamically generated, personalized views.

Imagine two foreign visitors to Sweden, *A* and *B*, whose mother tongue has no manifested similarity to Swedish, and none of them has any knowledge of Swedish yet. *A* only knows her/his mother tongue, while *B* knows a bit of English and German. They both come across the words¹ “Akta huvudet!” on a sign, and acquire its meaning via translation to their mother tongue, which is “Mind your head!”. As regards *A*, for the future comprehension of these or related lexemes s/he only depends on memorization. In contrast, *B* could benefit from her/his prior knowledge as follows:

- ‘huvud/et’ has a surface similarity to its English equivalent, ‘head’
- ‘akta/r’, in turn, does not have such an obvious link for English – where instead, *false friends* such as ‘acting’ pop up. However, it does have them for German, where the ‘*achten’ family of verbs and the ‘Achtung’ noun are a part of the basic vocabulary for foreign learners.

Now, the key questions are:

1. Is it likely that *B* would *fail* to *directly* see the cross-language link/s?
2. Is it likely that *B* would *understand an etymological explanation* of the link/s if it were served to him/her?
3. Would the awareness of the etymological link positively influence the *remembering* of the meaning of the words by *B*, in long term? (Would *B* on the next occasion bow her/his head instead of invoking the translation service again prior to entering the building...?)

If the answers to all these questions are positive then the example witnesses the relevance of the research line started in this paper.

¹We use an example in the form of a phrase in order to make the example more comprehensive. Admittedly, the research described later in the paper does not attempt to go from isolated words to the meaning of phrases.

In the presented preliminary research we thus aim at exploring various issues related to the prospects of using *personalized etymological context* of words, provided via *LLOD knowledge graphs*, in *foreign passive written vocabulary acquisition*. The main axes of this research are:

- Analysis of *LLOD resources* with respect to coverage of etymological links
- Study of *cross-language word pairs* returned via such links, with respect to their ‘adequate’ adoption through etymology, in terms of the first two questions above – i.e., not too trivial (which would make the etymological explanation redundant), but not too hard either (as the words may then elude adoption even with such an explanation).
- Study of actual (longer-term) learnability of word pairs, through a *prototype application*.

Those three axes roughly correspond to the next three sections of the paper.

2 Etymological Linked Data Sources and their Limitations

By a brief analysis of the available resources, it appears that LLOD sources covering etymology have been partially or fully created using an extractor from Wiktionary, since other etymological resources are typically copyright-protected.² Note however that Wiktionary itself, being one of the biggest online sources of word etymology, is essentially an unstructured source and cannot be used directly for our purposes. We identified two relevant: *Dbnary* (Sérasset, 2015) and *Etytree* (Pantaleo et al., 2017). The former is a generic approach to Wiktionary extraction, while the latter specifically focuses on etymology and employs relatively advanced NLP-based extractors. Because of our focus on etymological relations between the languages, Etytree was selected as our primary source of data for the language acquisition (micro-)study.

It is not possible to straightforwardly interlink the two sources, as they employ each its specific set of unique identifiers and are not directly interlinked. The only connection are the *seeAlso* links that lead from Etytree entities to Wiktionary pages.

²This is probably the reason why data from <https://starlingdb.org/> have not been published, although their RDF converter (Abromeit et al., 2016) exists.

	English	Latin	German	French
English	2157076	46624	3220	13910
Latin	46624	230754	4166	24700
German	3220	4166	328340	3442
French	13910	24700	3442	214958

Table 1: Number of *:etymologicallyRelatedTo* and *:etymologicallyDerivesFrom* predicate occurrences in selected languages

Prior to starting the study, we computed the number etymology links in Etytree and its proportion wrt. the number of entities, for a subset of language, in order to be able to estimate the exploitability of this resource. The result, for four major languages, is in Tab. 1. It is apparent that there the majority of etymological links hold just within a language, and only few hold between different languages.

3 Cross-language Word Pair Analysis

Our goal was to correlate the surface similarity of etymologically related words with their perceived learnability. For this purpose, we needed to express this *surface similarity* using a suitable metric. Since our target was the written vocabulary, we had preference for *edit distance* measures over pronunciation-oriented measures such as Soundex³ (which are also more language-dependent). Edit distances count the number (or sum up the costs) of operations that must be performed to transform one string into another, see e.g. an overview (Navarro, 2001). Probably the most widely used one is the Levenshtein distance, which counts the least number of single-character insertions, deletions, and replacements. Other known measures or algorithms are e.g. Hamming distance, Jaro-Winkler distance or Damerau-Levenshtein.

We eventually opted for the *Cross-Language Levenshtein Distance* (CLLD) (Medhat et al., 2015), which supports matching names across different writing scripts and uses many-to-many mapping characters. If the mapping is successful, the partial Levenshtein distance for a specific character is ignored. The intended target for this technique had indeed been the mapping between different scripts. We have however transferred the mapping-character heuristic to a somewhat different target. Namely, our intuition was that *etymologically grounded character mappings* (an example of which is, e.g., the orthographic reflection of

³<https://www.archives.gov/research/census/soundex>

the well-known High-German consonant shift) between the target and background language/s can be to some degree appropriated by the learners (even without full understanding of the etymological circumstances). Thus words differing along such mappings should have a smaller distance than those differing in other ways. Since we were unable to easily find a structured resource of cross-language character mappings, we provisionally created *ad hoc mappings* analytically, based on our speaker experience, namely, between English and two other major languages, German and French. Examples of such mappings are “th → d” or “p → f” for English vs. German. There were 22 pairs overall, of which 15 for German and 7 for French.

Next we created a *questionnaire*, aimed at general public, to which we manually selected *word pairs* such that:

- The target language word was always a *German* one and the background language word was always an *English* one.
- The words in the pair were connected by an *etymological link* in Etytree, i.e., they were chosen from the set of 3 220 linked words as indicated in Table 1.
- The *CLLD distance* of the pair varied between 1-6.

The choice of German and English was motivated by the following. English is a known language for a high number of learners. It is also the hub language of Etytree, with the highest number of cross-language links. German, in turn, features many word-level etymological links with English due to their partially shared roots. It is also an official language of several EU countries, thus many people learn it as a foreign language.

In total, seven-word pairs were manually selected, see Table 2. The questionnaire displayed for each pair⁴ the following question: “*After reviewing this etymologically related word pair, do you think a learner can later remember the meaning of the **foreign word** when seeing it in written form?*”. The answer was a choice among three options (plus the possibility to provide one’s own answer):

- *Yes, the learner will surely remember it. The words are almost the same. Upon seeing the*

⁴German nouns, except for proper nouns, were displayed as decapitalized.

German word, its English equivalent will immediately occur to the learner.

- *Unsure if the learner will remember it. The words are somewhat different. Seeing the German word might or might not “ring the bell” with reference to the English word.*
- *It is unlikely that the learner will remember it. The words are too different.*

The foreword to the questionnaire also suggested the users to always abstract from their familiarity with either word and provide feedback relative to their expectation of a learner who would know the English word but wouldn't know the German word.

The design of the study already revealed some limitations of the current setting. First and foremost, the number of etymological links was not only small with respect to the total vocabulary of both languages (less than 1% wrt. German and less than 0.15% wrt. English, see Table 1), but it was also biased towards words with *very high visual similarity*, such as #1 and #2. Finding ‘interesting’ pairs with manifestation of mapping rules, such as #5-#7, was not easy. There are also many *proper names* among the linked words (such as #3 and #4). Those might be less useful in language acquisition, first, because their translation between languages is not essential for communication, and second, because their frequency of occurrence is on average lower than that of common nouns. This also leads us to the suggestion that etymological resources should be used for suggesting word pairs in combination with a source of *word occurrence frequency* information. Finally, #3 also possibly manifests three natural deficiencies of the CLLD metric: (1) setting the contribution of the mapped characters to the CLLD to *zero* is an overshoot; (2) *very short words* exhibit low distance despite being apparently rather dissimilar; (3) CLLD also (contrary to the commonsense of word similarity perception) does not distinguish the first letter in the calculation.

In this respect it should be noted that the scope of our word pair analysis was intentionally bound to pairs that *truly originate from our LLOD resource*. This on the one hand limits the variety of cases considered, but on the other hand contributes to the assessment whether benefits to language acquisition can be obtained even for the present-day, modest, availability of etymological links in LLOD.

The questionnaire was sent to members of general public; most audience were young university

#	English	German	CLLD
1	transphenomenal	transphänomenal	1
2	heuristic	heuristisch	2
3	Vaud	Waadt	3
4	Nuremberg	Nürnberg	3
5	ravenstone	rabenstein	3
6	oversightly	übersichtlich	5
7	sharpshooter	scharfschütze	6

Table 2: Questionnaire word pairs and their CLLD

students or graduates. It returned filled by 29 respondents. Only the first three answer options (we will nick them ‘Yes’, ‘Unsure’ and ‘Unlikely’) were used overall. By the distribution of these answers, the cases (word pairs) can be relatively clearly ranged into three apparent clusters:

- #1 and #2 (CLLD ≤ 2) got ‘Yes’ from over 90% of respondents. We hypothesize that for such pairs the etymological links might help less-proficient language learners, but would be of limited value for experienced learners, since they could see the correspondence even without having been pointed to it.
- #4 got ‘Yes’ from over 60% of respondents, and ‘Unsure’ from the remaining ones. We hypothesize that for such pairs the etymological links might help the majority of language learners. Note that, however, #4 is inseparable from #3 and #5 through CLLD. Its shifted score might be influenced by the proper name nature of the word/s, which reduces the space of notions to be matched, as well as by the match at the beginning and end of the strings.
- #3, #5, #6 and #7 got ‘Yes’ from 7-20% of respondents, ‘Unsure’ from 34-52%, and ‘Unlikely’ from 34-48%. We can hypothesize that for such word pairs the etymological links might help advanced learners who would possibly either be explicitly aware of or intuitively adopt some of the mapping rules.

We also consequently prepared another questionnaire, this time addressing *linguistics/lexicography experts* (members of the Language Acquisition workgroup of the Nexus Linguarum COST Action⁵). It contained the same word pairs, but provided additional background information (e.g., about the nature and values of the CLLD measure),

⁵<https://nexuslinguarum.eu/>

prompted at entering qualitative responses on the word pairs, and also featured a set of general questions such as: “Do you think it is more beneficial to learn etymologically connected short words rather than long words?” or “Do you think it is more beneficial to learn a pair of words that have the same meaning or, rather, a pair of words that have different meanings? The meaning will be shown during the learning process. Different meanings: gift (present) (en) - Gift (poison) (de). Same meaning: house (en) - Haus (de)”.

We collected answers from four respondents. The feedback provided through the expert questionnaire largely confirmed the quantitative findings from the first (‘lay person’) questionnaire. Interesting insights were, e.g., the following:

- If the mapping rules are applied on multiple neighboring characters (as ‘w→v’, ‘aa→au’ and ‘dt→d’ in ‘Waadt vse. Vaud’), they might be more difficult to identify.
- For compound terms affected by mapping rules (#4–#7), it might be even difficult to correctly *tell* the different compounds *apart*.

Answers to the general questions also indicated that: both *long* and *short* words are worth learning via etymology; while pairs with the *same meaning* are a most suitable learning input for beginners, advanced learners will also benefit from pairs with *different meaning*; the coupling of written-form and *pronunciation* learning was also raised as a possible future agenda.

4 Experiment with a Proof-of-concept Vocabulary Acquisition Application

A proof-of-concept *web application*⁶ was developed (in .NET with a React front end), which leverages on SPARQL⁷ queries to the Etytree database for selecting word pairs from ten available languages (the mappings rules are however only used for English, German and French, as described above). Only word pairs with CLLD distance 3 or smaller are considered by the application; pairs whose strings were either identical or only differing in diacritics are also ignored. Word *meanings* are also retrieved and presented to the user; this among

⁶Source code available at <https://github.com/Duzij/LinkedLanguages>; online demo at <https://linkedlanguages.azurewebsites.net>.

⁷<https://www.w3.org/TR/sparql11-query/>

other helps identify words that are ‘false friends’ despite being etymologically related.

The users are required to create their account and to select their known and unknown languages. The *learning phase* then consists in accepting/rejecting word pair candidates for later testing, see Fig. 1. The system relies on an SQL Server Database to cache the results of the SPARQL endpoint, and this, in turn, enables a more tailored user experience. New word pairs are retrieved from the SPARQL endpoint only in case all word pairs from cache have been used. Such an architectural decision enables *collaborative filtering*: word pairs rejected by too many users are filtered out for new users. Then the user proceeds to the *testing phase*, when the previously approved word pairs are presented, but the word in the known language is left blank; the user is to complete the pair. If s/he fails to do so, the correct answer is revealed. The number of words revealed is a metric for overall test success.

During a weeklong user testing phase, 20 users used the application, and 1 725 times word pairs were either rejected or approved by users; 391 of these were either learned or revealed. Eventually, the application was formally evaluated via a *questionnaire*, which was filled by 11 users. Their responses were collected both for the common *System Usability Scale* (SUS) (Brooke et al., 1996) and for a few application-specific questions. The feedback was generally positive; the main issue reported was the fact that the application proposed ‘niche word pairs’ that were not beneficial for an average learner. This is however related to the issues with the word pair source. The average SUS score was 69.5, which corresponds to grade B – “Good”.

5 Conclusion and Further Work

The presented research is, to our knowledge, the very first study relating language acquisition to an open etymology source on the web. It revealed that the coverage of etymological links in LLOD is so far (despite the commendable efforts in DBnary and Etytree) modest, which hinders their usage in real-world language acquisition. The major take-away message is thus an encouragement to the community to push forward the (automated, as much as possible) *RDF-ization* of etymological paths that could become part of LLOD resources, whether bootstrapped from Wiktionary or also considering other, perhaps more even more rigorously collected database resources. Aside mere increase of *word*

Figure 1: Proof-of-concept application interface: the learning phase

coverage, additional information on the given pairs would be beneficial, e.g., indicating whether the etymologically related word pairs are semantically equivalent or merely related. As another resource that could be of use if available within LLOD we identified cross-language character mappings, allowing to properly shrink the distance between etymologically related words that could be quite useful for learning that from the target language. Finally, another dimension to be considered in language acquisition is the frequency of word occurrence in the given languages – both the target and background ones. Therefore, *word frequency dictionaries* might also be exploited in future etymology-driven language acquisition applications.

In parallel, however, experiments can be undertaken even with manually constructed etymological explanations independent of LLOD, in order to study the *psychology of etymology adoption* (especially in the presence of mapping rules) in more depth – though, in contrast to earlier pure-domain-driven studies by language acquisition scholars, now also with the idea of the possible computational (LLOD-based) support in mind.

By the questionnaire (albeit limited in size), the *CLLD measure* seems to be reasonably correlated with the word pair learnability. It should be however, most likely, modified in the partial distance computation. The distance of mapped characters should be *non-zero* in general, and possibly higher at the *start* (maybe also end) of the word or for *neighboring* mapped characters, since these settings likely make the learning more difficult.

The research has been supported by the Nexus Linguarum COST Action (no. CA18209). We are indebted to G. Sérasset, E. Pantaleo and T. Di Noia for their assistance regarding Dbnary and Etytree, and to G. Hrzica, G. Valunaite Oleskevicienė, O. Dontcheva-Navrátilová and others from the LA team of Nexus Linguarum for their feedback.

References

- Frank Abromeit, Christian Chiarcos, Christian Fäth, and Maxim Ionov. 2016. Linking the tower of babel: modelling a massive set of etymological dictionaries as rdf. In *LDL 2016, at LREC*, pages 11–19.
- John Brooke et al. 1996. SUS-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7.
- Doaa Medhat, Ahmed Hassan, and Cherif Salama. 2015. A hybrid cross-language name matching technique using novel modified Levenshtein Distance. In *2015 Tenth International Conference on Computer Engineering & Systems (ICCES)*, pages 204–209. IEEE.
- Gonzalo Navarro. 2001. [A guided tour to approximate string matching](#). *ACM Comput. Surv.*, 33(1):31–88.
- Ester Pantaleo, Vito Walter Anelli, Tommaso Di Noia, and Gilles Sérasset. 2017. [Etytree: A graphical and interactive etymology dictionary based on wiktionary](#). In *Proc. 26th Int'l Conf. on World Wide Web Companion, 2017*, pages 1635–1640. ACM.
- Evelyn Rothstein and Andrew S. Rothstein. 2008. *English grammar instruction that works!: developing language skills for all learners*. Corwin Press.
- Gilles Sérasset. 2015. [DBnary: Wiktionary as a Lemon-based multilingual lexical resource in RDF](#). *Semantic Web*, 6(4):355–361.

2. Workshops & Tutorials

Introduction

This volume comprises the proceedings of the workshops and tutorials held alongside the 4th Conference on Language, Data, and Knowledge (LDK 2023) in Vienna, Austria, 12–13 September, 2023. LDK is a biennial conference series dedicated to human language technology, data science, and knowledge representation. The University of Vienna, Austria, hosted the 4th edition of this conference between 12 and 15 September.

The workshops serve as a platform for discussing and exploring emerging areas of research in language data and the semantic web. These areas include data science, artificial intelligence, big data analytics, human-computer interaction, natural language processing, and information retrieval. Researchers and practitioners from both industry and academia submitted and presented papers during these workshops.

Notably, the NexusLinguarum COST Action CA18209 “European network for Web-centered linguistic data science” provided significant support for these events.

A total of 7 workshops, 3 tutorials, and 1 community day were accepted, and all the papers presented during these sessions are included in this joint volume.

Workshops:

- Deep Learning, Relation Extraction and Linguistic Data with a Case Study on BATS (DL4LD)
- Discourse studies and linguistic data science: Addressing challenges in interoperability, multilinguality and linguistic data processing (DiSLiDaS)
- International Workshop on Disinformation and Toxic Content Analysis
- Linking Lexicographic and Language Learning Resources (4LR)
- PROfiling LINGUistic KNOWledge gRaphs (ProLingKNOWER)
- Sentiment Analysis and Linguistic Linked Data (SALLD)
- Terminology in the Era of Linguistic Data Science (TermTrends)

Tutorials:

- LODification of lexical data using Wikibase
- Perspectivized Multimodal Datasets: a FrameNet approach to image-text correlations
- The DBpedia Knowledge Graph Tutorial

Community Day:

- Day of W3C Language Technology Community Groups

We would like to thank all workshop organisers, tutorial speakers and community day organisers for their engagement and cooperation along the process.

Ana Ostroški Anić and Blerina Spahiu

LDK 2023 Workshops and Tutorials Organisation

**Deep Learning, Relation
Extraction and Linguistic
Data with a Case Study on
BATS (DL4LD)**

Validation of the Bigger Analogy Test Set Translation into Croatian, Lithuanian and Slovak

Radovan Garabík

E. Štúr Institute of Linguistics, Slovak Academy of Sciences,
garabik@kassiopeia.juls.savba.sk

Ana Ostroški Anić

Institute of Croatian Language and Linguistics, aostrosk@ihjj.hr

Sigita Rackevičienė, Giedrė Valūnaitė Oleškevičienė, Linas Selmistraitis

Mykolas Romeris University, {sigita.rackeviciene,
gvalunaite, selmistraitis}@mruni.eu

Andrius Utka

Vytautas Magnus University, andrius.utka@vdu.lt

Abstract

This paper presents ongoing work focused on the analysis of translations of the English Bigger Analogy Test Set (BATS) dataset into three languages: Croatian, Lithuanian, and Slovak. We describe our automatic validation and further manual correction of the translations and analyse the main types of issues encountered in the dataset. The validation process involves checking the translations against morphological databases in order to uncover obvious mistakes or typos. Additionally, the translations are tested for the compliance to some of the formal guidelines for the Bigger Analogy Test Set translations, and for rudimentary grammatical correctness.

Each relation is represented by 10 categories, with each category containing 50 unique word pairs, e.g. *bird – feathers* and *door – threshold* for the relation of meronymy or *bicycle – bike* and *loyal – faithful* as examples representing synonymy. This layout produces 98,000 questions for the vector offset method.

The BATS bears superficial similarity to the WordNet database of semantic relations between words. While the original WordNet project (Fellbaum, 2005) covers English, numerous other WordNets and WordNet-like databases are available for many languages (Bond and Paik, 2012; Vossen et al., 2016). However, while some of the semantic relations are identical, the similarities stop there. The WordNet aims to encompass a broad range of vocabulary, ideally to cover as much of the general language as possible, and centered on the concept of sets of semantically equivalent words (*synsets*). The BATS is a specialized dataset including a pre-selected set of words and a comprehensive range of terms related to them by the given relation, incorporating highly specialized and rare lexical items. Moreover, the majority of the WordNets include only basic vocabulary or exhibit other major gaps in lexica. Nevertheless, individual language WordNets are a valuable source to consult when translating the BATS dataset.

The current study stemmed from one of the targets of the COST action *NexusLinguarum* of the creative utilization of pre-trained neural language models in order to acquire RDF relations, which form a foundation of the Linguistic Linked Open Data (LLOD) and which in turn can be used as a valuable source of curated data for Deep Learning methods. This task requires a multilingual

1 Introduction

1.1 Description of the the Bigger Analogy Test Set

Word embeddings are widely used in various Natural Language Processing tasks and toolkits. One of the features of the embeddings is that the vector space captures relations between the words and maps them to relations between the vectors, which leads to the word analogy based on vector arithmetic (commonly cited example is *king – man + woman = queen*) (Mikolov et al., 2013). The Bigger Analogy Test Set (BATS) was developed as a balanced analogy test set with 40 morphological and semantic relations (which yielded total 99,200 questions according to (Gladkova et al., 2016)) to draw the attention of the NLP community to word embeddings and analogical reasoning algorithms in the context of lexicographic and derivational relations (Gladkova et al., 2016). BATS includes inflectional and derivational morphology, and it also covers lexicographic and encyclopedic semantics.

evaluation set of lexico-semantic relations to allow testing various potential methods for relation acquisition from neural language models across languages. Thus, the COST action started the initiative to create such a dataset by manually translating the existing English BATS dataset to as many languages as possible, by initially focusing on translating the lexico-semantic portion of the dataset. Since BATS has so far been adapted to Japanese (Karpinska et al., 2018) and Icelandic (Friðriksdóttir et al., 2022), this is indeed a large-scale initiative.

This paper presents an automated validation process developed for the purpose of assessing the translated datasets' compliance with certain formal requirements, such as spell check, basic grammar and syntax verification. It also discusses the results of validation, focusing on true and false positive results, which often indicate errors in the initial dataset or reflect deliberate decisions regarding translation equivalents.

1.2 Analysed Languages

The Slovak language belongs to the West Slavic group of Slavic languages. It is the official and main language in Slovakia, spoken by about 5 million native speakers (conservative estimate based on the 2011 census data). It can be characterized as a medium-level inflected, subject-verb-object language with three grammatical genders, seven cases¹, two grammatical numbers, three tenses and two verbal aspects. Adjectives are inflected for gender, number and case and agree with the noun in these categories. These features are shared with most Slavic languages.

Being in the group of the Western South-Slavic languages, Croatian is typologically very similar to Slovak, with which it shares many grammatical features, e.g. the level of inflectional complexity, three grammatical genders, two grammatical numbers, and agreement between nouns and adjectives. It also has seven cases, three simple and three compound tenses, three moods, and four participles (Tadić, 2007). Its standardized variety is the official language of the Republic of Croatia, and is spoken by about 7 million native speakers around the world (Eberhard et al., 2023).

The Lithuanian language is one of two liv-

ing languages of the Baltic branch of the Indo-European language family (the other living Baltic language is Latvian). It is the official state language of the Republic of Lithuania and has about 2.67 million speakers in Lithuania and about 0.6 million speakers abroad (VLE, 2023). Lithuanian is a highly inflected language. Notional parts of speech are inflected by cases (nouns, pronouns, adjectives, participles, numerals), by person (verbs) or are uninflected (adverbs). The parts of speech inflected by cases have two or three grammatical genders (nouns have two, while the other parts of speech have three), two grammatical numbers (some pronouns have, in addition, the dual number), and the declension system comprised of case paradigms, the number of which varies across the parts of speech. Nouns and adjectives agree in gender, number and case. Verbs have three grammatical persons, two grammatical numbers, four tenses, four moods and two voices. The only uninflected notional part of speech is adverb, but many adverbs still have the morphological category of degrees of comparison (Ambrazas et al., 2006).

Slovak, Croatian and Lithuanian thus share several grammatical features that make them quite compatible for the cross-linguistic comparison and this analysis. All three languages are synthetic, SVO with a relatively free word order, with medium to high level inflection, and in general they have two grammatical numbers and three genders. All have noun-adjective agreement in gender, number and case, and – not less relevant – all three have adverbs as the only uninflected part of speech that appears in the lexico-semantic part of the BATS dataset.

The remainder of the paper is structured as follows: the guidelines for translating the BATS dataset are briefly presented in the next section. In section 3, the morphological databases of Croatian, Lithuanian and Slovak are described, which were used for the validation process, explained in section 4. The results of validation are discussed in detail in section 5, from the point of view of each language.

2 Description of the BATS Translation Process

We begin by introducing several expressions that will be used throughout the article. We use the term *source word* to indicate the word from which

¹The number of cases and genders depends on the level of abstraction of morphological analysis and on inclusion of marginal features; thus sometimes we encounter six cases and four genders

the semantic relation originates. Conversely, we refer to the word related by the given semantic relation (i.e. the second member of the related pair of words), as the *target word*. The term *word* encompasses both single words and multi-word expressions in this context. It is important to note that these terms are not related to the notion of the ‘source’ or ‘target’ language. If we take meronyms as an example, in the English original dataset *roof* is the source word, while *shingles*, *tiles*, *wood*, *metal* are the target words in the meronymic relation. Similarly, in the Slovak translation, *strecha* is the source word, while *škri-dle*, *dlaždice*, *drevo*, *kov* are the target words.

By *entry*, we understand one source word, accompanied by all the target words, and all corresponding translations in the given language. We call a single source word with the corresponding translation (or multiple translations) an *item*. An *entry* is thus composed of the list of *items*.

Detailed translation guidelines to be used as internal for the *Use Case 4.1.3 – Acquiring RDF Relations with Neural Language Models* were drafted by the task coordinator specifically for the task of translating the BATS dataset into 19 European languages. However, translation processes did not all start at the same time, and they are currently at various stages. The guidelines prescribed manual translation as they were intended to focus on possible issues in finding equivalents for the original English examples strictly. In particular, machine translation and post-editing is strictly prohibited. Apart from the expected common semantic phenomena, such as polysemy and synonymy, English examples contained a large number of culturally specific words, which were deemed as potentially too language specific, and for which finding appropriate equivalents proved to be challenging. For this reason, as well as in order to achieve a high level of validation, all translations were to be carried out manually. For each English word, the most common or the most frequent equivalent in the target language was chosen. Translation equivalents could be tested with a quick Google search to compare frequencies or by consulting dictionaries, word embeddings, online resources, etc., and choosing the most relevant translation. There was a possibility to add other equivalents commonly used on the line below the final target word, not aligned with a specific target word. In order to identify duplicates, i.e. two or

more words in the target language that are used for one word in the original dataset, the label `DUPLICATE` was to be used. Similarly, in cases where there was no appropriate equivalent word in the translation, the label `NO_TRANSLATION` was used. In order to allow for replicability and comparison of the English data and the translated files, the guidelines strictly forbade changing anything in the original English dataset, including obvious errors and the duplication of words in certain pairs.

In the Slovak translation of the dataset, we decided to keep the translations blank in such instances, as it was frequently impossible to find an adequate number of valid and distinct target words. This approach differs from the use of the `NO_TRANSLATION` keyword. In the latter case, it indicates the existence of either a genuine lexical lacuna or a situation where the target word’s concept is too regional and does not have a direct (loanword) equivalent in the target language.

In Table 1 we summarize the categories, identified by prefixes of the individual files. We will use these identifiers to refer to the categories and their translations.

category ID	relation
L01	hypernyms – animals
L02	hypernyms – misc
L03	hyponyms – misc
L04	meronyms – substance
L05	meronyms – member
L06	meronyms – part
L07	synonyms – intensity
L08	synonyms – exact
L09	antonyms – gradable
L10	antonyms – binary

Table 1: List of lexical categories

3 Morphological Databases

In the validation, we use morphological databases, i.e. triplets of *lemma*, *word*, *morphosyntactic description (MSD) tag* for some validation steps. We briefly describe the databases for our analysed languages.

3.1 Croatian

The Inflectional lexicon hrLex 1.3 (Ljubešić, 2019) is an inflectional lexicon of the Croatian language in which each entry consists of a word form, lemma, MSD, MSD features, UPOS, morphological features, frequency, and per-million frequency. The wordform, lemma, and MSD frequencies are

calculated on the hrWaC v2.2 corpus. The process of compiling the initial lexicon is described in (Ljubešić et al., 2016). The database met all the validation requirements, but minor issues in initial lemmatization (e.g. that participles are lemmatized as verbs) led to creating false positives in the validation process.

3.2 Lithuanian

The Lithuanian Morphological Database was specially designed for the validation of Lithuanian BATS translation. The database contains all types and lemmas for nouns, adjectives, verbs, and conjunctions extracted from the Joint Corpora of Lithuanian, as well as their morphological analyses. The wordlist of types, which is the base of the Lithuanian Morphological Database, is freely accessible from the CLARIN-LT repository (Dadurkevičius, 2020). The database includes more than 1.43 million unique word forms (types). Since the database includes only 4 parts of speech, our validation generated errors for translation including the missing parts of speech, i.e. numerals, adverbs, prepositions, and pronouns.

3.3 Slovak

The Slovak Morphological Database is a database of lemmas and their inflected word forms. The database includes 114,634 lemmas, selected from various Slovak dictionaries and supplemented with the most frequent words from the Slovak National Corpus. Each lemma is provided with a full paradigm along with morphological tags representing grammatical information. The database currently holds about 1.3 million unique word forms, for a total of 3.8 million entries (including homonyms). The database is used for automatic lemmatization and tagging of texts in the Slovak National Corpus and other Slovak corpora (Garabík and Mitana, 2022).

4 Validation Description

4.1 Validation Levels

The automated validation process assesses the translated dataset compliance with formal requirements, which encompasses the syntax of the files, spell-check, and a simple grammar check of multiword terms. During this validation, we recognize three degrees of significance:

- ERR is a hard error, either a formatting error, or a duplicate translation. Issues labeled as

ERR have high probability of being true positives

- WARN is a less serious issue, including spelling mistakes or unusual characters in the terms. These issues are quite often false positives.
- NOTE is just a notice. This is used to indicate missing translations.

4.2 Validation Steps

The first step involves the initial validation of the formal format following the BATS translation guidelines. This step focuses on a limited set of checks to allow for progress to the subsequent validation stages. The syntactical checks, in the sense of the formal syntax of the entries, include the following criteria: the translation must not be empty, multiword expressions should use the underscore character as the word separator instead of spaces, and all-capitals entries longer than one character should only consist of the strings `DUPLICATE` or `NO_TRANSLATION` as their values.

The second step involves validating the orthography and grammar of the entries. We compare the entries against a morphological database that includes lemmas and inflected words. Since we assume single-word translations to be lemmas, the validation fails if a translation is not present in the list of lemmas from the morphological database.

In the case of two-word translations, where the first word is an adjective or a participle and the second word is a noun, the second word must be included in the list of lemmas (specifically, nominative singular in almost all cases²) to pass the validation, and the first word has to agree with the noun in gender, case and number – or to be more precise, since the intra-lexeme homonymy is significant in all the three languages, at least one of the possible triplets of *gender*, *case*, *number* should agree with the noun.

If the translation consists of more than two words, or two words that are not an adjective (or a participle) and a noun, the validation passes if all the words are present in the list of possible word forms, and they do not need to be in the basic form. These multiword translations are mostly noun phrases, and as such they usually consist of variously inflected words: nouns, adjectives and

²With the exception of pluralia tantum and some defective nouns lacking the nominative.

prepositions. However, a small portion of multi-word units are also verb phrases.

These validation steps ensure basic correctness of the translations. However, many of the original English words are in plural (for various reasons, mostly due to usage or the common perception of concepts, e.g. *claws*, *pebbles*, *whiskers*), and the translations follow them rather faithfully. Although we could have easily added the plurals to the list of lemmas, we decided to include such translations in the list of warnings, lest we overlook easily visible errors.

The third step checks for duplicate translations (identically translated target words) within one entry. We consider the duplicates in the English original to be errors of the original dataset, and ignore them in this step. Overall, there are 154 duplicates in the original English dataset out of 5866 target words, comprising about 2.6% of the data.

5 Validation Results

category	en	first run			final run		
		hr	lt	sk	hr	lt	sk
L01	828	825	967	821	835	965	821
L02	876	838	845	796	848	844	796
L03	1507	1474	1799	1700	1474	1786	1685
L04	198	199	251	199	203	250	199
L05	113	119	152	125	119	151	125
L06	834	835	852	914	835	852	909
L07	254	263	303	287	263	303	287
L08	186	211	272	213	211	273	213
L09	881	869	865	1004	869	865	994
L10	190	203	207	192	203	205	192

Table 2: Translated target words per language and category. Note that there can be more translations than the original items in the English dataset (denoted by *en* in the table)

In the following Tables 3 and 4, the originally translated data (before validation) is called the *initial run*; data where the issues identified by the validation are fixed is called the *final run*. In Table 3, we show the number of issues found in the first version of the translations, per language and per category. Note that the issues with the NOTE level (i.e. untranslated words) are not comparable between languages – the Slovak dataset often leaves the translation empty by design; the Croatian dataset has not been completely translated by the time of writing this article. Table 4 shows the results after manual corrections. The last row shows the amount of corrected issues as a percentage of the difference from Table 3. Al-

though the percentage appears to be small in some cases, the remaining issues are (confirmed by further proofreading) predominantly false positives, thus these corrections eliminated practically all the mistakes of these types. Notably, we eliminated all the ERRs and significantly reduced other issues (mostly related to typos and spelling mistakes). The increase of Slovak NOTES is caused by deleting some of the duplicates, thus moving those ERRs into NOTES.

	hr			lt			sk		
	N	W	E	N	W	E	N	W	E
L01	41	120	8	0	240	42	7	128	10
L02	85	8	7	0	97	22	1	23	3
L03	1226	20	0	1	226	32	162	293	45
L04	0	39	4	0	44	4	0	34	1
L05	0	0	0	0	6	1	3	1	0
L06	695	19	2	6	84	85	88	136	35
L07	97	20	0	0	97	10	14	17	0
L08	0	27	1	0	31	5	74	21	0
L09	597	29	2	0	162	26	226	90	16
L10	3	16	3	1	67	4	69	6	1
Σ	2744	298	27	8	1054	231	644	749	111

Table 3: Number of NOTES (N), WARNs (W) and ERRs (E) per language and category, initial run.

	hr			lt			sk		
	N	W	E	N	W	E	N	W	E
L01	4	115	0	0	152	0	10	109	0
L02	0	11	0	0	46	0	1	21	0
L03	1226	20	0	0	200	0	165	281	0
L04	0	39	0	0	49	0	0	34	0
L05	0	0	0	0	4	0	3	1	0
L06	695	18	0	0	79	0	89	134	0
L07	95	19	0	0	76	0	14	16	0
L08	0	26	0	0	28	0	74	18	0
L09	597	29	0	0	156	0	226	82	0
L10	0	9	0	0	64	0	69	3	0
Σ	2617	286	0	0	854	0	651	699	0
$-\Delta\Sigma/\Sigma$ [%]	4.6	4.0	100	100	20.0	100	-1.1	6.7	100

Table 4: Number of NOTES (N), WARNs (W) and ERRs (E) per language and category, final run.

	hr			lt			sk		
	s	d	t	s	d	t	s	d	t
L01	1	7	0	36	6	0	0	8	2
L02	0	7	0	16	6	0	1	2	0
L03	0	0	0	10	15	7	3	42	0
L04	2	2	0	3	1	0	1	0	0
L05	0	0	0	0	1	0	0	0	0
L06	0	2	0	43	30	12	7	28	0
L07	0	0	0	3	7	0	0	0	0
L08	0	1	0	5	0	0	0	0	0
L09	0	2	0	11	13	2	0	16	0
L10	0	3	0	1	3	0	0	1	0
Σ	3	24	0	128	82	21	12	97	2

Table 5: Number of ERR types, initial run.

In Table 5, we analyse the types of the errors (is-

sues with the ERR severity). We use these codes:

- *s* means there is a space in the translated item, instead of the correct underscore
- *d* means the item is a duplicate of an already existing translation within one entry
- *t* stands for a typo in the value that should have been DUPLICATE (e.g. DULICATE, DUPLICATE etc.) or NO_TRANSLATION (however, there were no misspelled NO_TRANSLATION items found)

6 Discussion of False Positive Warnings

The warnings produced by the automated validation process are of three different types: agreement, spelling, capitalisation. They include false positive cases, the number of which depends on the design of each morphological database used for validation.

6.1 False Positive Warnings in Slovak

Slovak stands out with very few false positive warnings. Somewhat surprisingly, the adjective+noun orthographic/grammar check resulted in only two warnings in the Slovak translations, in L09 *cobwebby* → *pokrytý_pavučinami* (covered-NOM-MS-CG cobwebs-INS-FEM-PL, i.e. ‘covered by cobwebbs’) and *doddering* → *upadajúci_vekcom* (declining-NOM-MS-CG age-INS-MS-CG, i.e. ‘declining because of age’), both false positives.

6.2 False Positive Warnings in Croatian

There were no agreement warnings for the Croatian data. False positives in the Croatian data mostly referred to participles, which are lemmatized in the inflectional lexicon as verbs. Common warnings referred to adjectives when they had been translated in their definite form, instead of using a canonical indefinite form commonly appearing in traditional dictionaries of Croatian, e.g. *besmrtni*, *uzlazni*, *završni* instead of the indefinite forms *besmrtan*, *uzlazan*, *završan*, ‘immortal, rising, final’. However, this also depends on the type of an adjective, e.g. relational adjectives are always used in their definite form, while possessive adjectives always appear in the indefinite form.

Other false positives in the Croatian data related to spelling include adjectives in the form of participles, e.g. *natopljen* ‘saturated’, *pobjesnio* ‘outraged’, *prestrašen* ‘scared’, *ukočen* ‘stiff’,

uspaničen ‘panicky’, *zarobljen* ‘trapped’, *zaspao* ‘asleep’ and a small number of proper adjectives correctly spelled, e.g. *koščat* ‘bony’, *majušan* ‘tiny’. Adverbs were another category triggering warnings, e.g. *isprijed* ‘ahead’, *napolju* ‘outside’, and *postrani* ‘aside’ as well as colloquial words probably not found in the morphological database, e.g. *bajk* ‘wheel’, *bajs* ‘cycle’, *klinac* ‘kid’, *deran* ‘tike’, and *lupež* ‘rascal’. As expected, plural forms were also not recognized, as previously mentioned *šape* ‘paws’, *oči* ‘eyes’, *zubi* ‘teeth’, and *jaja* ‘eggs’, as well as specialized terms such as *cementit* ‘cementite’, *lubanjac* ‘craniate’, *patkarica* ‘anseriform bird’, *plodvaš* ‘placental’, and *svitkovac* ‘chordate’, most of which have a place in the animal taxonomy in the category L01 hypernyms-animals.

6.3 False Positive Warnings in Lithuanian

In the Lithuanian data, 24 false positive adjective+noun agreement warnings have been produced. This is due to the limits of the Lithuanian Morphological Database, which does not include inter-lexeme homonyms, e.g. the word form of the definite adjective *baltosios* ‘white’ may be used as singular genitive or as plural nominative; the word forms of the adjective *lengva* ‘light, not heavy’ and the noun *kamera* ‘camera’ may be used as singular nominative or singular instrumental; however, in all these and similar cases, the database includes only one of the word forms and occasionally the included word form does not coincide with the one which has to be in the translation. E.g., in the translation, the adjective *žydra* ‘bluish’ has to be in singular nominative (as it agrees with the noun in singular nominative), but the database includes only the word form *žydra* tagged as singular instrumental; therefore, such a case produced an adjective+noun agreement warning.

In addition, in the Lithuanian data, many false positive spelling warnings were produced. They were of two major types: the ones related to lemmatisation and the ones related to the limits of the Lithuanian Morphological Database.

The false positive warnings related to lemmatisation were produced in the cases where the provided single-word translations were included in the database, but did not match with the lemmata in the database. The following categories of translations produced the false positive warnings of this type:

1) single-word translations which are definitive adjectives as they are lemmatised as indefinite adjectives in the database, e.g. *aukštesnysis* ‘euthertian’ is lemmatised as *aukštas*;

2) single-word translations which are participles as they are lemmatised as infinitives in the database, e.g. *svyruojantis* ‘hesitant’, *dvejojantis* ‘inconclusive’ – lemmas *svyruoti*, *dvejoti*;

3) single-word translations which are nouns in plural nominative as they do not coincide with lemma-forms in the database, e.g. *plėviasparniai* ‘hymenopteron’, *papuošalai* ‘jewellery’ – lemmas *plėviasparnis*, *papuošalas*;

4) single-word translations which are nouns in singular genitive or plural genitive as they do not coincide with lemma-forms in the database, e.g. *placentos* ‘placental’, *kaukolės* ‘cranial’, *šunų* ‘canine’, *žinduolių* ‘mammalian’ – lemmas *placenta*, *kaukolė*, *šuo*, *žinduolis*;

The false positive warnings related to the limits of the Lithuanian Morphological Database were produced in the cases where the provided translations were words or comprised words that were not included in the database. The following categories of translations produced the false positive warnings of this type:

1) specialised single-word terms such as *aspidas* ‘elapid’, *liugeris* ‘lugger’ or multi-word terms that include highly specialised words such as *katinių šeimos gyvūnas* ‘felid’;

2) single-words which do not comply to the language norms, but were used for translation because they are frequent in the daily speech, such as *hamburgeris* ‘hamburger’, *fišburgeris* ‘fishburger’;

3) single-words of parts of speech that were not included in the database or multi-words which comprise parts of speech that were not included in the database (pronouns, adverbs, prepositions, etc.), e.g. *kažkas* ‘somebody’, *aukštyn* ‘up’, *žemyn* ‘down’, *virš* ‘above’, *po* ‘under’, *liūdnas ir kartu malonus* ‘bittersweet’, *dirbinys iš vielos* ‘wire-work’, *išvesti iš proto* ‘madden’.

7 Conclusions

The validation process proved valuable, particularly in identifying duplicate translations and highlighting spelling mistakes.

Numerous false errors and warnings (false positives) have various causes. Some stem from incomplete morphological databases used for validation, indicating insufficient coverage in certain

languages like Lithuanian. Others arise from errors and decisions made during the creation of the original dataset or reveal language-specific variations in lemmatization (e.g., indefinite vs. definite adjectives or participles lemmatized as verbs). Additionally, there may be missing highly specialized terms in domains such as biological taxonomy or nautical terminology. Given that we could not modify the original dataset, we had to find appropriate equivalents that accurately reflect the relationships found in the original. These often involved using lemmas in the plural form, colloquial or culturally specific words, etc.

However, the warnings and notices generated during validation also served as additional checks in cases where there was no existing translation. This could occur due to oversight during the translation process or the absence of a suitable equivalent. In such cases, the validation process provided an opportunity to compare these translation gaps with equivalents in other languages and potentially find effective solutions. While this paper primarily focuses on the formal aspect of translating BATS into different languages, it is worth noting that there were numerous lexical gaps specific to English-speaking regions of the world, as well as many domain-specific words or terms requiring verification in terminological resources. These translations had few or no occurrences even in very large corpora, especially within the meronym categories.

The analysis reveals that the accuracy of the initial translations varied among the languages, primarily due to differences in the effort invested in the translations, the approaches taken to the guidelines, and the resolution of problematic entries in the original dataset, rather than inherent differences between the languages.

Acknowledgements

This study is based upon work from the COST Action NexusLinguarum - European network for Web-centered linguistic data science (CA18209), supported by COST (European Cooperation in Science and Technology, <https://www.cost.eu/>).

References

Vytautas Ambrazas, Emma Geniušienė, Aleksas Girdenis, Nijolė Slizienė, Dalija Tekorienė, Adelė

- Valeckienė, and Elena Valiulytė. 2006. *Lithuanian Grammar*. Baltos lankos.
- Francis Bond and Kyonghee Paik. 2012. A survey of wordnets and their licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, pages 64–71, Matsue.
- Virginijus Dadurkevičius. 2020. *Assessment Data of the Dictionary of Modern Lithuanian versus Joint Corpora*. CLARIN-LT digital library in the Republic of Lithuania.
- David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2023. *Ethnologue: Languages of the World*. SIL International, Dallas, Texas. Twenty-sixth edition.
- Christiane Fellbaum. 2005. WordNet and wordnets. In Keith Brown and et al., editors, *Encyclopedia of Language and Linguistics*, second edition edition, pages 665–670. Elsevier, Oxford.
- Steinunn Rut Friðriksdóttir, Hjalti Daníelsson, Steinþór Steingrímsson, and Einar Sigurdsson. 2022. *IceBATS: An Icelandic Adaptation of the Bigger Analogy Test Set*. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4227–4234, Marseille, France. European Language Resources Association.
- Radovan Garabík and Denis Mitana. 2022. Accuracy of Slovak Language Lemmatization and MSD Tagging – MorphoDiTa and SpaCy. In *LLOD Approaches for Language Data Research And Management, Abstract Book*, pages 93–95, Vilnius. Mykolo Romerio universitetas.
- Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuo. 2016. *Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't*. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15, San Diego, California. Association for Computational Linguistics.
- Marzena Karpinska, Bofang Li, Anna Rogers, and Aleksandr Drozd. 2018. Subcharacter Information in Japanese Embeddings: When Is It Worth It? In *Proceedings of the Workshop on the Relevance of Linguistic Structure in Neural Architectures for NLP*, pages 28–37, Melbourne, Australia. Association for Computational Linguistics.
- Nikola Ljubešić. 2019. *Inflectional lexicon hrLex 1.3*. Slovenian language resource repository CLARIN.SI.
- Nikola Ljubešić, Filip Klubička, Željko Agić, and Ivo-Pavao Jazbec. 2016. *New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian*. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4264–4270, Portorož, Slovenia. European Language Resources Association (ELRA).
- Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. *Linguistic Regularities in Continuous Space Word Representations*. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.
- Marko Tadić. 2007. Building the Croatian Dependency Treebank: the Initial Stages. *Suvremena lingvistika*, pages 85–92. 63/1.
- VLE. 2023. *Visuotinė lietuvių enciklopedija (General Lithuanian Encyclopedia)*. <https://www.vle.lt/straipsnis/lietuviu-kalba/>. LNB Mokslo ir enciklopedijų leidybos centras.
- Piek Vossen, Francis Bond, and John McCrae. 2016. Toward a truly multilingual Global WordNet Grid. In *Proceedings of the 8th Global WordNet Conference (GWC2016)*, pages 419–426, Bucharest.

Workflow Reversal and Data Wrangling in Multilingual Diachronic Analysis and Linguistic Linked Open Data Modelling

Florentina Armaselu

University of Luxembourg, Luxembourg, florentina.armaselu@uni.lu

Barbara McGillivray

King's College London
United Kingdom

barbara.mcgillivray@kcl.ac.uk

Chaya Liebeskind

Jerusalem College of Technology
Israel

liebchaya@gmail.com

Giedrė Valūnaitė Oleškevičienė

Mykolas Romeris University, Lithuania
gvalunaite@mruni.eu

Andrius Utka

Magnus University, Lithuania
andrius.utka@vdu.lt

Daniela Gifu

Romanian Academy - Iasi Branch, Romania
daniela.gifu@iit.academiaromana-is.ro

Anas Fahad Khan

Cnr-Istituto di Linguistica Computazionale "Antonio Zampolli", Italy, fahad.khan@ilc.cnr.it

Elena-Simona Apostol

University Politehnica of Bucharest
Romania

elena.apostol@upb.ro

Ciprian-Octavian Truică

University Politehnica of Bucharest
Romania

ciprian.truica@upb.ro

Abstract

The article deals with data wrangling in a multilingual collection intended for diachronic analysis and linguistic linked open data modelling for tracing concept change over time. Two types of static word embeddings are used: word2vec (French and Hebrew data sets), and fastText (Latin and Lithuanian data sets). We model examples from these embeddings via the OntoLex-FrAC formalism. To address the challenge of heterogeneity, we use a minimalist workflow design allowing for both convergence and flexibility in attaining the project goals.

The data wrangling phase described in this proposal is intended to prepare the data for tracing the evolution of concepts in different languages and historical periods through NLP and LLOD approaches. The main challenges of this type of task consist in the heterogeneity of the data sets to be considered for analysis, the need for harmonisation among the different teams involved, and the lack of an established methodology for dealing with the process of data preparation within a multilingual, multi-format, and multi-team context.

Although reported as taking 80% of the data scientist's time (Paton, 2019), data wrangling seems to be less studied so far in digital humanities (DH), and especially in areas that combine natural language processing (NLP), such as diachronic word embeddings, and LLOD representations including spatio-temporal dimensions. Our proposal addresses the question of how to optimise *collaboration* within a DH use case that requires multilingual multi-format corpora (pre-)processing and LLOD modelling by several teams. We approached this question through an adaptation of a method origi-

1 Introduction

In data wrangling, the "data required by an application is identified, extracted, cleaned and integrated, to yield a data set that is suitable for exploration and analysis" (Furche et al., 2016, p. 473). The tasks often referred to in this process pertain to data organisation, including data integration and transformation, and data quality, including missing data or anomaly identification (Nazabal et al., 2020). These tasks have also raised questions about the possibilities of automating them (Paton, 2019).

nated in the domain of engineering, called *workflow reversal* (Chen et al., 2019). It implies an inverse uncertainty propagation and workflow reversal with input-output variable swap to deal with the issue of “handling pre-defined uncertainty associated with design objectives (targets) or constraints (requirements)” (p. 1). We applied the idea in a more general, abstract way, by considering that some requirements and targets can be precisely specified in the workflow, while others can remain under-specified and allow a certain degree of design and implementation flexibility to the different teams.

2 Method

In this section, we present the methodology and the current status of our solution. The main problem was that our data sets varied in many aspects: language, format (TXT, XML; vertical, PoS-tagged, lemmatised), number of files (single, multiple), folder structure (flat, hierarchised), time coverage (ancient, medieval, modern) and genre (Appendix A, Table 1). Although initially we considered unifying all the data formats for the downstream tasks, we realised that this will involve non-trivial preparation and harmonisation work. Finally, for the exploratory design phase, we decided that a certain degree of format variability and independence among the teams can be afforded, provided that a number of common conditions are met at specific points in the processing flow. Therefore, despite the differences in the intermediary steps for our data sets and teams, we were able to define convergence points, through common requirements and outputs in the workflow, that had to be fulfilled for all the involved parts. The main tasks of the workflow were: 1) generate a set of terms and their neighbours resulting from word embedding (word2vec or fastText) and cosine similarity measures; 2) model via OntoLex-FrAC the word embedding results and possibly combine them with dictionary evidence, to represent the evolution of a set of parallel or related concepts in the studied languages.

Figure 1 illustrates the minimal requirements (brace callout) that are demanded by each module or target (rectangular blocks) from the previous modules to accomplish its objectives. Hence, the reversed sense of the arrows, with a left-to-right reading for targets and their needs, and right-to-left, for the actual order of the processing operations. While the types of data wrangling, target tasks, and constraints are specific to our project, we assume

that the general method of workflow reversal, understood as a way of identifying the minimal set of specifications and common targets viewed from the reversed perspective of what is needed or intended to be achieved, can be applied to other projects that deal with issues such as the heterogeneity of data and approaches, and multi-team collaboration.

3 Results

Currently, we are in the phase of LLOD modelling, intended to use the OntoLex-FrAC formalism for RDF-based machine-readable dictionaries combined with corpus observables and observations (Chiarcos et al., 2022). The data wrangling and diachronic word embedding tasks included so far experiments with the French, Latin, Hebrew, and Lithuanian data sets. Partial findings from these experiments are expected to be applied to the other corpora from the collection. The data preparation involved different strategies depending on the format characteristics of each data set.

The Lithuanian data set comprised three layers. The representation layer used the original spelling which was transliterated into modern Lithuanian on the next layer, followed by linguistic and morphological annotations. The text was lemmatised and English translations were provided. The decision was to work with the transliteration into the modern Lithuanian layer. Then, the procedures involved extracting text and metadata from XML files and organising the resulting text files by time slice, to prepare them for diachronic word embedding. It was chosen to use FastText, as it is acknowledged to work better for word embeddings in morphologically rich languages, with experimentally proven results in the Lithuanian language (Petkevicius and Vitkute-Adzgauskiene, 2021). The corpus was split into three time periods: 16th, 17th and 18th century. FastText embeddings were generated for each subcorpus for further analysis.

For the Latin corpus, we extracted the publication dates from the metadata available in the corpus file, and normalised the dates so that they were all in a numeric format. This required converting centuries in years or assigning the midpoint between the two extremes in the case of a data range. The input to the embedding training was the lemmatised version of the corpus. We split the corpus into three time intervals: from 450 BCE to 1BCE, from 1CE to 450 CE, and from 451CE to 900 CE. We generated FastText embeddings for each subcor-

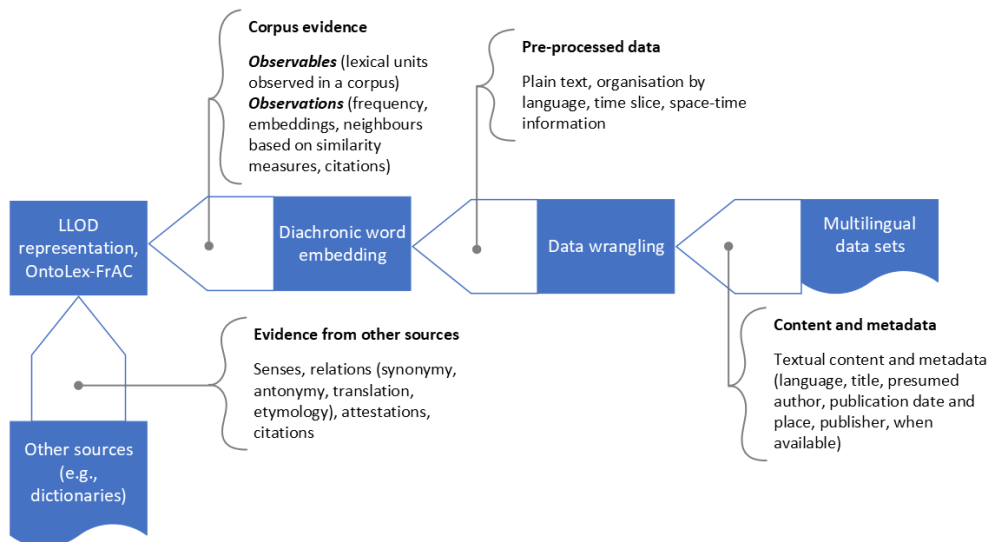


Figure 1: Workflow reversal for multilingual diachronic analysis and LLOD representation

pus, with 100 dimensions, a context window of 5 words to the left and to the right of the target word, and a minimum frequency threshold of 50. In order to make the semantic spaces comparable, we aligned the semantic spaces using the Procrustes Alignment algorithm (Schönemann, 1966).

Minimal pre-processing was performed on the Hebrew Responsa data set before the word embedding (word2vec) phase. Considering the poor performance of a state-of-the-art modern Hebrew POS taggers on the Responsa (Liebeskind et al., 2012), this pre-processing consisted only of white space tokenisation. We split the Responsa into four time intervals: the 11th century until the end of the 15th century, the 16th century, the 17th through the 19th centuries, and the 20th century until today (Liebeskind and Liebeskind, 2020).

The preparation of the Romanian data set included operations such as: acquisition of primary textual data, clearing of copyrights, OCR in some cases, interpretative transliterations in some others, storing, cleaning of data, and metadata completion. From the input DOC and PDF files, raw text was extracted and lists of words were generated. The extracted text was passed to the PoS-tagger that outputs XML files with unknown words marked as NotInDict (Not In Dictionary), i.e., words whose lemmas were not found in the DEXonline lexical database, but also numbers, including years, and proper names. The PoS-tagger included sentence segmentation, tokenisation, and lemmatisation. To create the word embeddings, Radim Rehurek’s gensim package, for instance, could be used.

For the BnL Open Data, containing thousands of XML files in a hierarchy of folders and sub-folders, an automatic pre-processing was necessary. Figure 3 (Appendix B) illustrates the preparation of the monograph subset (the arrows indicate the input-output direction). The pipeline was produced with KNIME, a software for creating data science workflows. It extracted text and metadata from the BnL hierarchy of folders and XML files, selected only French documents and generated new file names, plain text files, and a new folder structure. The longest horizontal branch (ReadXML to CSV Writer) extracted the textual content from the XML files, and created a flat folder with all the resulting TXT files for French. To the original file names, a prefix was added (language code and publication date from the XML file) to be used in the second KNIME workflow. The three other branches (ending with CSV Writer) produced files for metadata (language, publication date, publisher, persistent ARK identifier), statistics (word and document count by language), and issues (lists of files missing language information). A second KNIME workflow organised the text files by time slice,¹ taking into account elements from the history of Luxembourg, e.g., military and political events, royal decrees and school laws. Other platforms were also tested (OpenRefine and Karma). KNIME was selected since it was open source and dealt well with XML and folder hierarchy processing, and missing data and inconsistency detection.

¹BnL monographs, time slices: 1690 – 1794; 1795 – 1814; 1815 – 1830; 1831 – 1866; 1867 – 1889; 1890 – 1918.

4 Discussion

For our experiments, we used static word embeddings and `gensim word2vec` (Rehurek and Sojka, 2010) for French and Hebrew, and `fastText` (Bojanowski et al., 2017) for Latin and Lithuanian. This required tokenised text, with and without lemmas and PoS, and sentence segmentation. The corpora were structured by time slice (year, decade, century) to determine semantic changes. For each language, we trained our own word embeddings, and we intend to compare the results across language and time period. For example, we were able to qualitatively assess the Latin diachronic embeddings against known instances of lexical semantic change. To mention one such case, the neighbours of the embeddings for the Latin word *pontifex* display evidence of the shift from the domain of the traditional Roman religion (e.g. *sacerdos* ‘priest’ and *aedes* ‘temple’ towards terms related to Christianity, such as *missa* ‘mass’ and *beatus* ‘blessed’).

Qualitative assessment was performed for the French data set, after having applied `word2vec` (5 word window, 100 dimension vectors) by time slice. We compared the list of neighbours resulting from word embedding with dictionary attestations, and found corpus evidence of emerging polysemy within the time period of the data set. For example, we aligned the embedding results of the term *révolution* (*revolution*) with different senses attested by Ortolang, such as: ‘motion of a body around an axis’, ‘motion of a figure around an axis’, ‘natural phenomena’, and ‘political change’. While the attestations always referred to earlier dates than the time intervals of the embeddings, the analysis provided a snapshot of the senses on a timeline and their dictionary-corpus contextualisation.

The word *מהפכה* (*revolution*) has appeared in numerous contexts throughout the Responsa (as evidenced by its top neighboring terms). The majority of references to revolution in the first era are made in a religious context (*כפירה* (*atheism*), *תשובה* (*repentance*)). In the second era, the word is used less frequently. However, it occurs in the context of war and tragedy (*אונס* (*rape*), *הרג* (*killing*), *מיחה* (*death*)) in the third era, which corresponds to the eras in the French corpus, as a consequence of the pogroms that Jews faced during this time. Industrial (*מכונות* (*machines*), *אנרגיה* (*energy*)) and medical (*החיאה* (*resuscitation*), *אנאטומיה* (*anatomy*)) revolutions, and revolutionary ideological movement (*רפורמים* (*Judaism Reform*)), *הילוניות* (*secu-*

larism)) pertain to the fourth period.

A qualitative assessment performed on the Lithuanian data set by comparing word embeddings to the dictionary entries revealed that, for example, for the word *ponas* (*mister; lord*) the polysemy identified in the data set could be attested by the Lithuanian language dictionary².

These first results served for exploratory analysis and estimation of the possible outcomes obtained from our data sets, which led us to consider a combination of corpus and lexicographic resources for the subsequent LLOD modelling task. The OntoLex-FrAC model seemed appropriate to it.

No generally agreed upon way of representing diachronic constructs in linked data exists, despite of several proposals within the OntoLex-Lemon framework (see (Armaselu et al., 2022) for a discussion). Currently, we experiment with the Frequency, Attestation, and Corpus information (FrAC) extension of the OntoLex module (McCrae et al., 2017) to represent word embeddings and the relationship between lexical entries and the relevant corpora (Chiarcos et al., 2022), also considering previous work in modelling etymological information in lexical linked data resources (Khan, 2018).

Figure 2 provides a generic example of OntoLex-FrAC combining corpus and dictionary-based attestation for a lexical entry in language *ll*. This may be connected to other senses, lexical concepts, and entries in other languages through etymological and translation relations. We propose to add a new property and class (`new:dictionary`, `new:Dictionary`) to indicate a dictionary attestation, and a property (`new:neighList`) to store the neighbours in a structured form such as a list. Each neighbour can be represented as an instance of one of the subclasses of `frac:Observable` (lexical entry, lexical sense, form, lexical concept). This type of resource may be used for queries and inferences about semantic change, or enrichment.

The interplay between semantics and pragmatics (e.g., determined by historical, socio-cultural, communication-related factors) should also be considered in representing semantic change and its context. This may involve knowledge- and language-oriented theoretical frameworks, and properties such as `ontolex:usage` for modelling usage and pragmatic nuances of word meaning (Armaselu et al., 2022), or other forms of encoding linguistic

²Lietuvių kalbos žodynas (Lithuanian language dictionary, electronic edition). 2017. Vilnius: Lietuvių kalbos institutas. <http://www.lkz.lt> (accessed 10 January 2022).

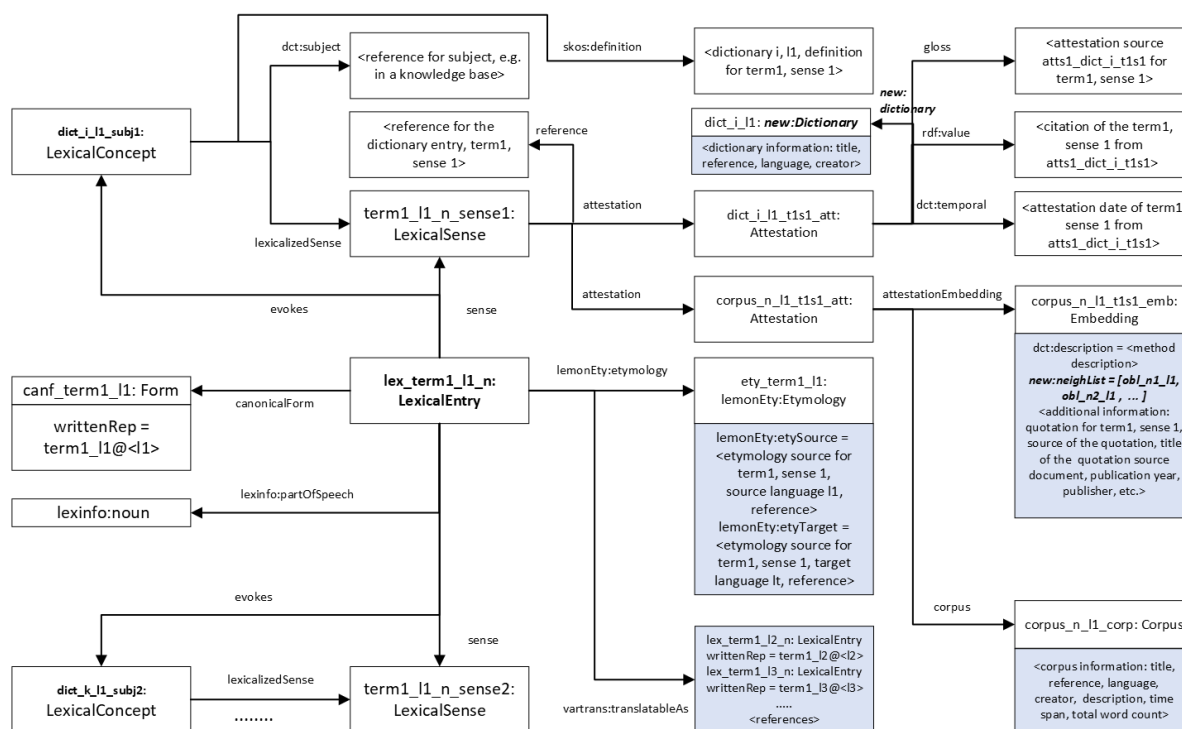


Figure 2: OntoLex-FrAC example combining corpus and dictionary-based attestation (angle brackets: single-item free descriptions; blue-shaded cells: aggregated descriptions)

content as LLOD still under investigation (Bosque-Gil et al., 2018; Gromann et al., 2022).

5 Conclusion

The proposal focuses on data wrangling in multi-language data sets with various sizes, formats, time spans, and downstream tasks. We argue that a combination of NLP methods and LLOD formalisms, such as diachronic word embedding and OntoLex-FrAC, as well as corpus- and lexicographic-based evidence, can serve in creating inter-operable and more context-rich LLOD resources for detecting and representing semantic change.

We applied the concept of workflow reversal as a general framework for devising a common yet flexible roadmap for our data preparation phase. We defined a minimal set of functional blocks and requirements necessary to accomplish the intended tasks and allowed a certain degree of freedom in their implementation, according to the specificity of each data set, language, and team. The main challenge in applying this type of method may consist in finding a balance between the under-specified and the well-defined parts of the workflow, and avoiding downstream divergence that can impede the project goals. We will use this exploratory design phase to

refine and apply the implementation requirements to each language, with the aim of building a multi-lingual sample of interconnected LLOD diachronic ontologies. Since some of the data sets were rather limited in time coverage, it may be envisaged to complement them, for instance by using multilingual corpora available online via repositories such as Wikimedia Downloads.

Acknowledgment

This article is based upon work from COST Action *Nexus Linguarum*, *European network for Web-centred linguistic data science*, supported by COST (European Cooperation in Science and Technology). www.cost.eu.

Authors' contribution

F.A. wrote the manuscript and contributed to the methodological design, data processing and analysis for French and LLOD modelling; BMcG conducted the processing and analysis of the Latin data, contributed to the methodological design and wrote the parts of sections 3, 4 and Table 1 relative to Latin; C.L. conducted the processing and analysis of the Hebrew data and contributed to sections 3, 4 and Table 1 relative to Hebrew; G.V.O.

and A.U. conducted the processing and analysis of the Lithuanian data and contributed to sections 3, 4 and Table 1 relative to Lithuanian; D.G. conducted the processing of the Romanian data and contributed to sections 1, 3 and Table 1 relative to Romanian; A.F.K. contributed to discussions on the modelling of the results as LLOD in section 4; E.S.A. contributed to sections 1, 3 and 5; C.O.T. contributed to sections 2 and 3. All authors reviewed the final manuscript.

References

- Florentina Armaselu, Elena-Simona Apostol, Anas Fahad Khan, Chaya Liebeskind, Barbara McGillivray, Ciprian-Octavian Truică, Giedrė Valūnaitė Oleškevičienė, and Marieke van Erp. 2022. **LL(O)D and NLP perspectives on semantic change for humanities research**. *Semantic Web*, 13(6):1051–1080.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. **Enriching word vectors with subword information**. *Transactions of the Association for Computational Linguistics*, 5:135–146.
- Julia Bosque-Gil, Jorge Gracia, Elena Montiel-Ponsoda, and Adrián Gómez-Pérez. 2018. **Models to represent linguistic linked data**. *Natural Language Engineering*, 24(6):811–859.
- Xin Chen, Arturo Molina-Cristóbal, Marin D. Guenov, Varun C. Datta, and Atif Riaz. 2019. **A Novel Method for Inverse Uncertainty Propagation**, volume 48 of *Computational Methods in Applied Sciences*, pages 353–370.
- Christian Chiarcos, Elena-Simona Apostol, Besim Kabashi, and Ciprian-Octavian Truică. 2022. **Modelling frequency, attestation, and corpus-based information with OntoLex-FrAC**. In *International Conference on Computational Linguistics*, pages 4018–4027.
- Tim Furché, Georg Gottlob, and Leonid Libkin. 2016. **Data wrangling for big data: Challenges and opportunities**. In *International Conference on Extending Database Technology*, pages 473–478.
- Jolanta Gelumbeckaite, Mindaugas Šinkunas, and Vytautas Zinkevicius. 2012. **Old Lithuanian reference corpus (SLIEKKAS and automated grammatical annotation)**. *Journal for Language Technology and Computational Linguistics*, 27(2):83–96.
- Daniela Gifu. 2016. *Lexical Semantics in Text Processing. Contrastive Diachronic Studies on Romanian Language*. PhD thesis, "Alexandru Ioan Cuza" University of Iași, Romania.
- Dagmar Gromann, Elena-Simona Apostol, Christian Chiarcos, Marco Cremaschi, Jorge Gracia, Katerina Gkirtzou, Chaya Liebeskind, Verginica Mititelu, Liudmila Mockiene, Michael Rosner, Ineke Schuurman, Gilles Sérasset, Purificação Silvano, Blerina Spahiu, Ciprian-Octavian Truică, Andrius Utka, and Giedrė Valūnaitė Oleskeviciene. 2022. **Multilinguality and LLOD: A survey across linguistic description levels**. *Semantic Web 1 (0) 1–39*, IOS Press (currently under review).
- Anas Khan. 2018. **Towards the representation of etymological data on the Semantic Web**. *Information*, 9(12):304.
- Chaya Liebeskind, Ido Dagan, and Jonathan Schler. 2012. **Statistical thesaurus construction for a morphologically rich language**. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics—Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 59–64.
- Chaya Liebeskind and Shmuel Liebeskind. 2020. **Deep learning for period classification of historical Hebrew texts**. *Journal of Data Mining & Digital Humanities*, 2020:5864.
- John P McCrae, Julia Bosque-Gil, Jorge Gracia, and Paul Buitelaar. 2017. **The OntoLex-Lemon model: development and applications**. In *eLex 2017 Conference*, pages 587–597.
- Barbara McGillivray and Adam Kilgarriff. 2013. **Tools for historical corpus research, and a corpus of Latin**. *New Methods in Historical Corpus Linguistics, Tübingen: Narr*, pages 10.
- Alfredo Nazabal, Christopher K. I. Williams, Giovanni Colavizza, Camila Rangel Smith, and Angus Williams. 2020. **Data engineering for data analytics: A classification of the issues, and case studies**. *arXiv:2004.12929 [cs]*. ArXiv: 2004.12929.
- Norman W. Paton. 2019. **Automating data preparation: Can we? Should we? Must we?** In *International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data*, pages 1–5.
- Mindaugas Petkevicius and Daiva Vitkute-Adzgauskiene. 2021. **Intrinsic word embedding model evaluation for Lithuanian language using adapted similarity and relatedness benchmark datasets**. In *IVUS*, pages 122–131.
- Radim Rehurek and Petr Sojka. 2010. **Software framework for topic modelling with large corpora**. In *LREC 2010 Workshop New Challenges for NLP Frameworks*, pages 45–50.
- Peter H. Schönemann. 1966. **A generalized solution of the orthogonal procrustes problem**. *Psychometrika*, 31:1–10.
- Alessandro Vatri and Barbara McGillivray. 2018. **The Diorisis Ancient Greek corpus: Linguistics and literature**. *Research Data Journal for the Humanities and Social Sciences*, 3(1):55–65.

Appendix A. Datasets³

Data set	Language	Time span	Size	Format	Genre
LatinISE	Latin	2nd c. BCE - 20th c. CE	10 mil. word tokens	TXT, vertical format, lemmatised, PoS-tagged	Literature, philosophy, law, religion, technical writings, letters
Diorisis	Ancient Greek	8th c. BCE - 5th c. CE	10,206,421 word tokens	TXT, enriched with morphological information, lemmatised, PoS-tagged	Literature, philosophy, historiography, scriptures, technical writings, letters
RODICA	Romanian	19th c. (second decade)	over 5 mil. lexical tokens	TXT, XML, PoS-tagged, lemmatised	Newspapers from Moldavia, Wallachia, Transylvania and Bessarabia
SLIEKKAS	Old Lithuanian	16th - 18th c.	10 texts, 350,000 words	TXT, representation layer (old alphabet); transliterated layer (modern Lithuanian alphabet); linguistic and morphological annotations; lemmatised; English translations	Prose and poetry, religious texts (prayers, catechisms, hymnals and sermons)
BnL Open Data	French, German, Luxembourgish	1690-1918 (monographs); 1841-1878 (newspapers)	23,663 newspaper issues, 510,505 articles; 504 monographs, 33,477 chapters	XML, Dublin Core	Monographs: literature, history, philosophy, geography, religion; newspapers
Responsa	Hebrew	11th -21st c.	76,710 articles, about 100 mil. word tokens	TXT	Questions and rabbinic answers on daily issues (law, health, commerce, marriage, education, Jewish customs)

Table 1: Description of the data sets

Appendix B. KNIME workflow

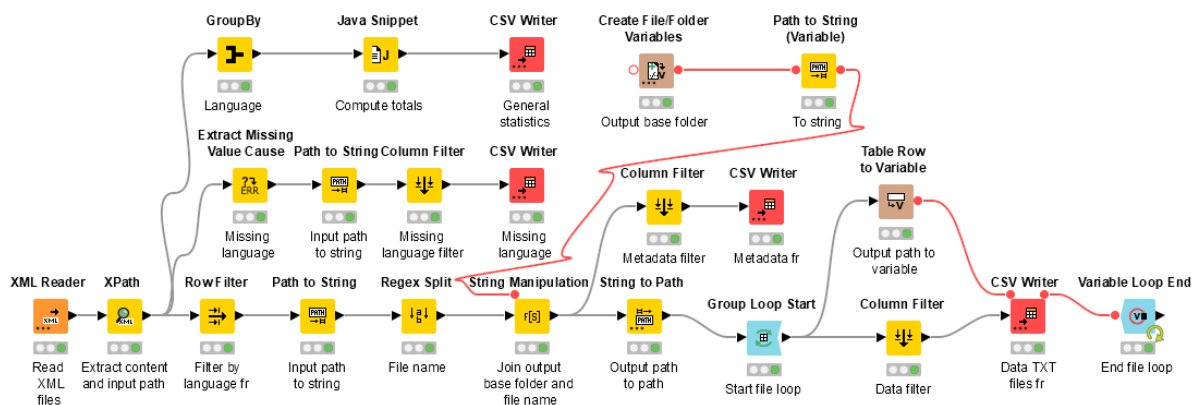


Figure 3: KNIME workflow for the preparation of the BnL Open Data set (French monographs)

³LatinISE (McGillivray and Kilgarriff, 2013); Diorisis (Vatri and McGillivray, 2018); RODICA (ROmanian DIachonic Corpus with Annotations) (Gifu, 2016); SLIEKKAS (Gelumbeckaite et al., 2012); Bibliothèque nationale du Luxembourg, BnL Open Data; Responsa (Liebeskind and Liebeskind, 2020).

DBnary2Vec: Preliminary Study on Lexical Embeddings for Downstream NLP Tasks

Nakanyseth Vuth and Gilles Sérrasset

Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG
38000 Grenoble
France

first.last@univ-grenoble-alpes.fr

Abstract

In this preliminary study, we experiment with the use of DBnary, a big lexical knowledge graph, to create word embeddings that could be used in NLP downstream tasks. Our gamble is that word embeddings created from lexical data (instead of language corpora) may exhibit less biases while still being usable as the first layer of deep learning approaches to NLP tasks.

We tried very basic method of embedding creation from lexical graph and evaluate (1) the intrinsic performance of the created embeddings on word similarity and word analogy test sets and their extrinsic quality in POS tagging and NER downstream tasks, along with (2) the biases they may exhibit. Such embeddings show promising performances outperforming word2vec on few specific tasks, while still not on par on most others, but we confirm that they exhibit less bias overall.

1 Introduction

Most NLP tasks now use word or sub-word embeddings as their first ingredient. Such embeddings are created based on the proximity of words with others in a corpus. These embeddings have proven to be a valid approach in many practical systems, but they do suffer from biases, leading to research to de-bias through better selection of the training corpus or ad-hoc debiasing techniques on the embeddings themselves.

At the same time, there exists several huge lexical datasets that provide curated information on the words, word forms and senses of different natural languages. With growing size, such datasets are largely disregarded in current deep learning approaches to NLP tasks.

In this paper, we would like to know if training word embeddings from a lexical dataset could be an alternative to corpus based embeddings computation. This work is a preliminary attempt to answer 2 research questions: (1) *is it possible to create*

embeddings solely from a lexical graph that could be an alternative to corpus based embeddings for downstream tasks? and (2) *do embeddings learned from lexical graphs suffer from the main biases identified in the corpus-based embeddings literature?*

For this first attempt, we will use the DBnary dataset that we present in section 2. Then, we discuss the evaluation of the adequacy of such embeddings in downstream tasks and of their potential biases in sections 4 and 5. Section 6 presents and discusses the experiments performed to address the research questions at hand.

2 DBnary, a multilingual lexical graph

DBnary (Sérrasset, 2015)¹ is a lexical resource extracted from 23 language editions of Wiktionary.² This dataset is structured in RDF (Resource Description Framework), a W3C standard for modelling and exchanging metadata about web resources where information is given about resources using triples that consist of subject-predicate-object statements.³

DBnary data can be downloaded or queried online using the SPARQL language⁴, accessed interactively through a faceted browser⁵ or accessed by dereferencing any of the resource URI it defines,⁶

¹See <http://kaiko.getalp.org/about-dbnary/> for the current state of development of DBnary.

²See <https://www.wiktionary.org/>.

³See <https://www.w3.org/TR/rdf11-primer/> for more details

⁴The *SPARQL Protocol and RDF Query Language* is the “standard query language and protocol for Linked Open Data on the web or for RDF triplestores”, quoted from <https://www.ontotext.com/knowledgehub/fundamentals/what-is-sparql/>.

The SPARQL endpoint of DBnary can be accessed at <http://kaiko.getalp.org/sparql>

⁵The browser can be accessed at <http://kaiko.getalp.org/fct/>

⁶Each DBnary resource has a URI that can be queried using any web browser or any programmable HTTP client.

making it fully compliant with the guidelines of Linguistic Linked Open Data (LLOD) framework (Declerck et al., 2020).⁷

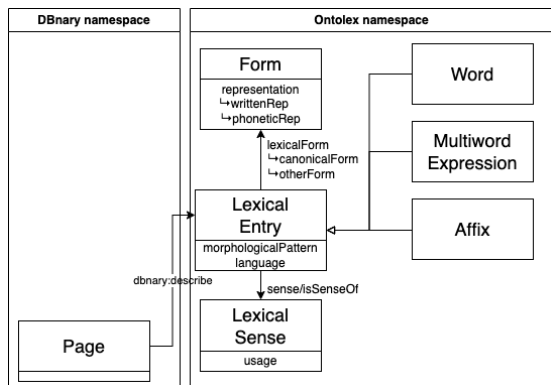


Figure 1: The OntoLex-Lemon core module excerpt (taken from <https://www.w3.org/2016/05/ontolex/#core>) that is used by DBnary, along with the additional `dbnary:Page` class that is used to represent a Wiktionary page describing several lexical entries.

The data consists of a huge multilingual graph where nodes (resources) are lexical objects (pages, lexical entries, forms, word senses, etc.), and edges (properties) are structural properties or lexical relations (translation, synonym, antonym, etc.). DBnary uses the core vocabulary of the OntoLex-Lemon model (McCrae et al., 2017) which was developed and which is further extended in the context of the W3C Community Group “Ontology Lexica”.⁸ As depicted in figure 1, an additional `dbnary:Page` class has been added to account for the fact that Wiktionary data is organised mainly as a set of pages, where each page describes several lexical entries (possibly in several languages). Other properties and classes are present in the dataset but are not currently used in this work.

The DBnary dataset has steadily grown since its first description (Sérasset, 2012, 2015) and, at the time of writing, contains more than 414M triples describing 6.7M lexical entries in 23 languages.

Figure 2 shows a (simplified) excerpt of the DBnary graph for `dbnary:Page` “cat”. In this preliminary study, we only used the DBnary English subgraph.

E.g. http://kaiko.getalp.org/dbnary/bass_noun_1 represents one of the Lexical Entries described at page *bass* in the English edition of the Wiktionary project.

⁷See also <http://www.linguistic-lod.org/>.

⁸See <https://www.w3.org/community/ontolex/> for more details.

3 Building embeddings from graphs

Current node embedding methods, which create embeddings for nodes in a graph, do not take into account most of the information available in the DBnary graph (namely, typing of the nodes or labelling of the relation). Hence, we have to create graphs suitable for embedding computation from DBnary.

For all our experiments, we use the same general modelling for graphs, but propose two graph topologies.

3.1 Graph Modeling

Formally, we model the graph as follows. Let $G = (V, E)$ denote the graph, where V denotes the set of nodes and E denotes the set of edges. In this graph, each node $x_w \in V$ represents a node in DBnary, such as a page, lexical entry, or word sense. Thus, we have:

$$V = \{x_w : w \in DBnary\} \quad (1)$$

and each edge $e_{u,v} \in E$ represents a relationship between two words u and v of weight $w_{x_u, x_v} \in \mathbb{R}$. The weight reflects the strength or relevance of the relationship between u and v . Graph G can be (un)directed or (un)weighted, depending on the type of graph being modeled.

3.1.1 DBnary topology

The first graph topology, we experiment with, directly uses the relational topology present in DBnary. We extracted all the pages, lexical entries, word sense, and their relations between them from the database and used this information to construct the graph. Each of them is represented as a node in the graph, while each relation between nodes is represented as an edge connecting the corresponding nodes.

Based on the topology, an edge e is formulated as:

$$e_{u,v} = \{(x_u, x_v, w(rel_{x_u, x_v})) : u, v \in V\} \quad (2)$$

For example, consider the node x_{cat} in Figure 2, which represents a page in DBnary. It is connected to another page node x_{kitty} through a $synonym_{x_{cat}, x_{kitty}}$ relationship. Additionally, it has a *describes* relationship with its lexical entries, namely $x_{cat_Adjective_1}$ and $x_{cat_Noun_1}$. Each of these lexical entries is also linked to its corresponding word sense. The weights of the edges are defined based on the relation property. For instance,

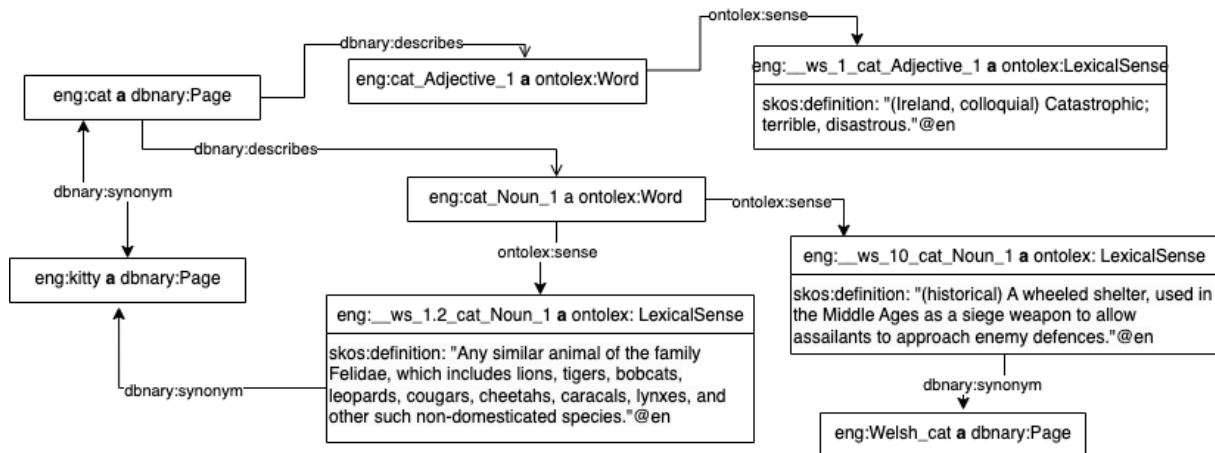


Figure 2: Excerpt of DBnary graph depicting page "cat", along with 2 of its lexical entries and some word senses, with their definition. DBnary graph also contains lexico-semantic relations (synonymy, antonymy, hyponymy...) between pages, lexical entries and/or word senses.

synonym has a higher weight than *antonym*, and so on. This allows us to capture the strength of the relationship between different nodes. Furthermore, we can use the "nyms" relationships (e.g., synonym, antonym, hypernym, hyponym) to establish connections between lexical entries, word senses, and other nodes.

Using the DBnary topology, we construct the graph as a list of edges consisting of two nodes and a weight value based on their relationship. Specifically, an edge between nodes x_u and x_v with a weight of w_{x_u, x_v} is represented as:

$$\langle x_u \rangle \langle x_v \rangle \langle w_{x_u, x_v} \rangle \quad (3)$$

For instance, the relationship between the nodes "cat" and "kitty" with a weight of 10 can be denoted as $\langle cat \rangle \langle kitty \rangle \langle 10 \rangle$ in this format, where *cat* and *kitty* correspond to the two nodes and the weight value of 10 indicates the strength of the edge. This format will be used in the graph embedding models, which will be described further in Section 3.2.

3.1.2 Text to Graph

The second graph topology involves utilizing the definitions of each word sense node to create a training corpus and representing the relationship between words in the corpus as edges in the graph. Specifically, we implemented a method that converts sentences into a graph by considering each word as a node and connecting them based on bi-grams co-occurrence. The weight of each edge is based on the co-occurrence frequency of the bi-gram in the entire corpus.

$$w_{(t_i, t_{i+1})} = count_occur(t_i, t_{i+1}) \quad (4)$$

where t_i and t_{i+1} are the two words in the bi-gram and *count_occur* is a function that returns the number of times the bi-gram appears in the corpus. The resulting edge can be represented as:

$$e = \{(v_{t_i}, v_{t_{i+1}}, w_{(t_i, t_{i+1})}) : t_i, t_{i+1} \in S\} \quad (5)$$

where S is the set of all unique words in the corpus, v_{t_i} and $v_{t_{i+1}}$ are the corresponding nodes in the graph, and $w_{(t_i, t_{i+1})}$ is the weight assigned to the edge between these nodes.

3.2 Embedding methods

In the context of our preliminary studies into graph embedding techniques, we have opted to examine three widely recognized algorithms for producing graph embeddings, namely DeepWalk (Perozzi et al., 2014), LINE (Tang et al., 2015), and node2vec (Grover and Leskovec, 2016). These techniques have been demonstrated to be effective in a variety of applications and have attained state-of-the-art performance in numerous benchmarks. In addition, we have incorporated the prevalent Skip-Gram technique (Mikolov et al., 2013a), word2vec, as a fundamental model for comparative analysis.

3.2.1 SGNS (word2vec)

The Skip-Gram with Negative Sampling is a well-known embedding method that aims to learn a dense, continuous vector representation for each

word in a given corpus. SNGS model predicts the surrounding context words given a center word. It focuses on maximizing probabilities of context words given a specific center word, which can be written as

$$P(w_{i-c}, w_{i-c+1}, \dots, w_{i-1}, w_{i+1}, \dots, w_{i+c-1}, w_{i+c} | w_i) \quad (6)$$

3.2.2 DeepWalk

DeepWalk is an unsupervised learning method for generating node embeddings by utilizing random walks on the graph. The objective of DeepWalk is to learn a representation for each node in the graph, which captures its structural context in the graph. The method starts by generating random walks on the graph, where each walk starts from a randomly selected node and traverses the graph by following its edges. The walks are then treated as sentences, and the Skip-gram model from word2vec is used to learn node embeddings by predicting the context nodes for each target node in the walk.

3.2.3 LINE

LINE on the other hand, aims to learn node embeddings by considering the global structure of the graph. The method uses a first-order proximity and a second-order proximity objective to capture the local and global structure of the graph, respectively. The first-order proximity objective is to maximize the probability of observing a context node given a target node in a random walk, similar to DeepWalk. The second-order proximity objective, on the other hand, is to maximize the probability of observing a node u being the second-order neighbor of node v .

3.2.4 node2vec

node2vec is another method for learning node embeddings by utilizing random walks on the graph. Similar to DeepWalk, the objective of node2vec is to learn a representation for each node in the graph that captures its structural context in the graph. node2vec improves upon DeepWalk by introducing a biased random walk strategy that allows for the generation of walks that balance the exploration and exploitation of the graph structure which in turn leads to representations obeying a spectrum of equivalences from homophily to structural equivalence. Specifically, node2vec uses a two-parameter family of random walks, where the parameters control the trade-off between depth-first and breadth-first search. It uses second-order biased random walks to generate sequences of nodes

or “sentences” from a given graph. Once the sequences of nodes are generated, they are used as input to the SGNS model to learn embeddings for nodes.

4 Evaluating embeddings

As outlined in (Bakarov, 2018), the field of word embedding evaluation has developed two primary classes of methods for assessing the quality of embedding models: intrinsic and extrinsic evaluators. Intrinsic evaluators assess the quality of embedding models through specific tasks that are independent of downstream NLP applications. Extrinsic evaluators, on the other hand, use the vector representations of the embedding models in downstream NLP tasks, such as part-of-speech tagging and named entity recognition. These evaluations measure the effectiveness of embedding models in improving the performance of NLP tasks. It is important to note that both intrinsic and extrinsic evaluations have their limitations. Intrinsic evaluations may not necessarily correlate with the performance of embedding models in real-world NLP applications, while extrinsic evaluations may be affected by other factors such as the quality of the downstream NLP task. Therefore, it is better to use both intrinsic and extrinsic evaluations to get a comprehensive understanding of the quality of embedding models.

4.1 Intrinsic evaluator

Intrinsic evaluation is a method for assessing the quality of word embeddings by testing their ability to capture certain linguistic properties and relationships. The primary objective of intrinsic evaluation is to determine how well an embedding model captures semantic and syntactic information. This approach involves assessing the embedding quality through specific tasks that are independent of downstream NLP applications. Two commonly used intrinsic evaluation methods are word similarity and word analogy tasks. Intrinsic evaluation is an important step in assessing the quality of word embeddings, as it provides insight into the model’s ability to capture linguistic properties and relationships.

4.1.1 Word similarity

Word similarity tasks are designed to measure the degree of similarity between pairs of words. These tasks typically involve a list of word pairs along with human judgments of the degree of similarity between the pairs. The model’s performance

is evaluated based on its ability to produce similarity scores that match human judgments using cosine similarity. It measures the cosine of the angle between the two vectors and ranges from -1 to 1, where 1 represents identical vectors, 0 represents independent orthogonal vectors, and -1 represents opposite vectors. The cosine similarity between vectors a and b is calculated as follows:

$$\cos(w_a, w_b) = \frac{w_a \cdot w_b}{\|w_a\| \|w_b\|} \quad (7)$$

where \cdot represents the dot product of two vectors, and $\|w_a\|$ and $\|w_b\|$ denote the Euclidean norms of vectors w_a and w_b , respectively.

4.1.2 Word analogy

Word analogy tasks, on the other hand, assess the model's ability to capture the relationships between words, such as analogies. In these tasks, a set of word pairs is provided, and the model is required to complete an analogy by finding a fourth word that is related to the third word in the same way as the second word is related to the first word. For example, given the pair "man:woman," the model should find the word "queen" when presented with the pair "king:?". This task is calculated using the 3CosAdd method (Mikolov et al., 2013b). Given a pair of words a and a^* and a third word b , the analogy between a and a^* can be used to determine the word b^* that corresponds to b . It is mathematically expressed as:

$$a : a^* :: b : _ \quad (8)$$

It solves for b^* using the following formula:

$$b^* = \underset{b'}{\operatorname{argmax}} (\cos(b', b + a^* - a)) \quad (9)$$

This method normalizes the vector length using cosine similarity. Alternatively, there is a refined method called 3CosMul (Levy and Goldberg, 2014) which is defined as:

$$b^* = \underset{b'}{\operatorname{argmax}} \frac{\cos(b', b) \cos(b', a^*)}{\cos(b', a^*) + \epsilon} \quad (10)$$

where $\epsilon = 0.001$ is used for preventing zero division.

4.2 Extrinsic evaluator

Extrinsic evaluators are NLP downstream tasks that directly use embedding models to improve the performance of the task at hand. By using

the embeddings as input features for these tasks, we can evaluate the effectiveness of the embedding model in contributing to the downstream task performance. In our preliminary study, we have chosen two specific tasks, Part-of-Speech (POS) tagging and Named Entity Recognition (NER), as extrinsic evaluators for our embedding models.

4.2.1 Part-of-speech tagging

Part-of-Speech (POS) tagging is a fundamental task in NLP that involves the identification of the grammatical category of words in a sentence. The goal of POS tagging is to automatically assign a specific part-of-speech tag (such as noun, verb, adjective, etc.) to each word in a sentence, based on its context and the grammatical rules of the language. POS tagging is an essential preprocessing step for many NLP applications, such as text classification, information retrieval, and machine translation. It is a challenging task, as words often have multiple possible tags, and the same word can have different meanings and functions in different contexts.

4.2.2 Named entity recognition

Named Entity Recognition (NER) is a task in NLP that involves identifying and extracting named entities from unstructured text. Named entities refer to specific objects, people, places, or concepts that have a unique name or identity. The goal of NER is to automatically identify and classify named entities in text, and assign them a pre-defined label such as PERSON, ORGANIZATION, LOCATION, etc. The task is crucial for a wide range of NLP applications, such as information extraction, document retrieval, and machine translation, and it is a challenging task due to the variability and complexity of named entities in text.

5 Biases in embeddings

Word embeddings have proven to be valuable tools for natural language processing tasks, but they are not immune to biases. Biases in embeddings arise from the underlying biases present in the training data, leading to certain groups or concepts being over-represented or under-represented in the embedding space (Garg et al., 2018). These biases can manifest in various forms, including gender, race, ethnicity, religion, and more. Recognizing and addressing these biases is crucial to ensure fairness, equity, and non-discrimination in NLP applications. Studies have highlighted the

presence of biases in word embeddings, revealing how societal biases can seep into the learned representations. For example, Bolukbasi et al. (Bolukbasi et al., 2016) demonstrated the existence of gender bias in word embeddings through the analogy "man:programmer::woman:homemaker", where the embedding model associated men with the profession of programmer and women with the role of homemaker. This finding illustrates how gender biases present in the training data can be reflected in the learned embeddings.

The consequences of biases in embeddings can be far-reaching and detrimental. Biased embeddings can perpetuate and reinforce harmful stereotypes, leading to discriminatory outcomes in downstream NLP applications. For instance, automated hiring systems that utilize biased embeddings may unfairly discriminate against certain demographic groups, resulting in an inequitable hiring process (Dastin, 2022). Search engines that rely on biased embeddings can produce biased search results, reinforcing existing societal biases and limiting access to diverse perspectives and information (Kay et al., 2015; Caliskan et al., 2017). Furthermore, automated hate speech detection models trained on biased corpora can inadvertently exhibit racial bias, potentially amplifying harm inflicted upon marginalized communities (Sap et al., 2019). Because of this, it is essential to gain an understanding of the biases that are present in word embeddings and to work to eliminate them in order to stop the negative effects they have on society.

6 Experiments

The following section presents the experimental setup used to evaluate the embedding models, as well as the evaluation results that highlight both performance and bias along with evaluation datasets used in this study.

6.1 Experimental Setup

We selected four models for our study, comprising three graph embedding models, DeepWalk, LINE, and node2vec, as well as a traditional word embedding model, SGNS. To obtain a comprehensive analysis, we trained the graph embedding models using two approaches described in Section 3.1, resulting in a total of six graph embedding models. We trained text-to-graph based models and SGNS using DBnary definition nodes that contained 945,525 definitions/sentences. Table 1 il-

Graph	# Edges	# Nodes	# Vocab
DBnary topology	2396346	3284911	1120225
Text to graph	1772040	276619	276617

Table 1: Graph’s properties

lustrates the properties of the graph used in the graph embedding models. For the node2vec approach, we used the official implementation⁹. We used Graphvite (Zhu et al., 2019) to train DeepWalk and LINE. Finally, we trained the SGNS model using Gensim (Rehurek and Sojka, 2011) word2vec library. To ensure consistency in our results, we used the same default settings for all the graph embedding models, including walk length $l = 40$, number of walks per node $r = 100$, and ($p = 1$, $q = 1$) specifically for node2vec method, and window size $w = 10$ for SGNS. We chose to use 256 dimensions for all the embedding models in our study.

6.1.1 Intrinsic

The embedding models were evaluated intrinsically through word similarity and word analogy tasks. In this study, we have selected a total of eight benchmark datasets for the purpose of evaluating word similarity. These datasets are presented in Table 2. The Google analogy test set (Mikolov et al., 2013a) and the Bigger Analogy test set (BATS) (Gladkova et al., 2016) were selected to serve as the benchmark datasets for the word analogy test. Both of these tasks were evaluated using GluonNLP¹⁰

6.1.2 Extrinsic

Extrinsic evaluation was performed using two different NLP downstream tasks: 1) part-of-speech tagging, and 2) named-entity recognition. We trained each task with the same architecture, which consisted of running a vanilla RNN on the Keras library (Chollet et al., 2015) for 25 epochs with 64 hidden dimensions, and a batch size of 128. The CoNLL-2000 (Tjong Kim Sang and Buchholz, 2000) from NLTK (Bird et al., 2009) and the CoNLL-2003 (Tjong Kim Sang and De Meulder, 2003) from HuggingFace¹¹ were used for part-of-speech tagging and named-entity recognition tasks, respectively. The data split for both tasks is presented in Table 3.

⁹<https://github.com/eliorc/node2vec>

¹⁰<https://nlp.gluon.ai/index.html>

¹¹<https://huggingface.co/datasets/conll2003>

Dataset	Pairs
WordSim-353 (Finkelstein et al., 2001)	353
WordSim-353-SIM (Agirre et al., 2009)	203
WordSim-353-REL (Agirre et al., 2009)	252
MEN (Bruni et al., 2014)	3000
RadinskyMTurk-287 (Radinsky et al., 2011)	287
RareWords (Luong et al., 2013)	2034
SimLex-999 (Hill et al., 2014)	999
SimVerb-3500 (Gerz et al., 2016)	3500
YangPowerVerb-130 (Yang and Powers, 2006)	130
SemEval17Task2(Camacho-Collados et al., 2017)	518

Table 2: Word Similarity benchmark datasets. The MEN dataset has been partitioned into a dev set consisting of 2000 pairs and a test set consisting of 1000 pairs. The SemEval17Task2 dataset is divided into two distinct subsets, comprising 18 pairs for the trial set and 500 pairs for the test set.

Table 3: Dataset splits for extrinsic tasks

Dataset	Train	Validation	Test
CoNLL-2000	7909	1396	1643
CoNLL-2003	14041	3250	3453

6.2 Bias experiment

To evaluate the presence of bias in our embedding models, we utilized the code ¹² which replicates the paper of (Badilla et al., 2020). Following this paper, we used four metrics to measure biases: 1) the Word Embedding Association Test (WEAT) (Caliskan et al., 2017), 2) the WEAT effect size, 3) the Relative Norm Distance (RND) (Garg et al., 2018), and 4) the Relative Negative Sentiment Bias (RNSB) (Sweeney and Najafian, 2019). Details on the queries utilized in our study can be found in Table 4. Due to the size of the corpus used for training our text-to-graph models and SGNS model, we were only able to measure biases in Gender and Religion, as many of the embeddings for Ethnicity queries were not present in our models.

6.3 Embeddings evaluation and biases

This section presents the evaluation results of our embedding models in terms of their performance using intrinsic and extrinsic evaluators, as well as their biases.

Intrinsic - Word similarity results: We evaluated the performance of all our models on

¹²https://github.com/dccuchile/wefe/blob/master/examples/WEFE_rankings.ipynb

13 different datasets, and the results are presented in Table 5. Our experimental findings reveal that the node2vec topology-based model outperforms the other models in capturing the similarity and relationship of word pairs, as evidenced by its superior performance in datasets such as SimLex999, SimVerb3500, and YangPowerVerb-130. These datasets were designed to focus more on measuring a range of semantic relationships. On the other hand, the SGNS model generally outperforms all other models in most datasets, except the ones that specifically focus on capturing semantic relationships. However, our node2vec text-to-graph model also shows promising results, coming in second after SGNS and outperforming the node2vec topology-based method in most cases. It is important to note that not all models were able to cover all pairs in the evaluation datasets, as shown by the percentage of out-of-vocabulary pairs in Table 6 word similarity.

Intrinsic - Word analogy results: The results obtained using 3CosAdd and 3CosMul methods for two datasets are presented in Table 7. We observe that the topology-based models perform the worst, with SGNS model achieving the highest scores in both datasets and methods. These findings suggest that while topology-based models may excel at capturing similarity and semantic relationships between word pairs, they do not perform as well in word analogy tasks. This could be attributed to the fact that topology-based models rely heavily on the graph structure, which may not always capture the full extent of the semantic relationships between words. Furthermore, the results also reveal some interesting insights into how the models perform on specific word analogy tasks. For instance, for the pair "man:king::women:?", our model predicted "face-sit" with a score of 0.70, and "queen" with a score of 0.68. This could be explained by the fact that in DBnary, the node "face-sit" shares an edge connection through a synonym relation to one of "queen"'s word senses, which leads to this result. Another example is the pair "Athens:Greece::Bangkok:?", where our model predicted "Krung_Thep" instead of "Thailand". This occurred because in DBnary, "Krung_Thep" is synonymous with "Bangkok" and the node "Bangkok" does not have an edge connecting to the node "Thailand" at all.

Table 4: Bias experiment queries

	Target set	Attribute sets
Gender query	{Male terms, Female terms}	{Career, Family}, {Math, Arts}, {Science, Arts}, {Intelligence, Appearance}, {Intelligence, Sensitive}, {Pleasant, Unpleasant}, {Positive words, Negative words}, {Man Roles, Women Roles}
Religion query	{Christianity terms, Islam terms}	{Pleasant, Unpleasant}, {Conservative, Terrorism}, {Positive words, Negative Words}
	{Christianity terms, Judaism terms}	{Pleasant, Unpleasant}, {Conservative, Greed}, {Positive Words, Negative Words}
	{Islam terms, Judaism terms}	{Pleasant, Unpleasant}, {Terrorism, Greed}, {Positive Words, Negative Words}

Table 5: Word similarity evaluation results

	Word Similarity Datasets												
	WS-all	WS-sim	WS-rel	MEN-full	MEN-dev	MEN-test	MTurk	RW	SimLex	SimVerb	YP	SEval-trail	SEval-test
node2vec	0.3664	0.6350	0.1140	0.4284	0.4420	0.4022	0.2717	0.2289	0.4630	0.4269	0.6672	0.4757	0.4062
deepwalk	0.2900	0.4911	0.0955	0.2163	0.2128	0.2240	-0.0132	0.1238	0.2092	0.2378	0.3702	0.1889	0.2267
line	0.2501	0.4566	0.0103	0.2302	0.2325	0.2248	-0.0135	0.1163	0.1922	0.2476	0.3369	0.0918	0.2236
node2vec_t2g	0.5080	0.6174	0.4354	0.5745	0.5703	0.5839	0.5099	0.1778	0.2225	0.1393	0.1951	0.7523	0.3986
deepwalk_t2g	0.2877	0.4141	0.2368	0.4433	0.4545	0.4190	0.3322	0.1444	0.2031	0.1878	0.2695	0.6491	0.3390
line_t2g	0.2873	0.4132	0.2378	0.4417	0.4524	0.4180	0.3389	0.1459	0.2070	0.1858	0.2638	0.6347	0.3388
SGNS	0.5511	0.6278	0.4555	0.6282	0.6283	0.6279	0.4635	0.3562	0.3427	0.2661	0.3438	0.7957	0.5268

Table 6: Word similarity out-of-vocabulary percentage

	WS-all	WS-sim	WS-rel	MEN-full	MEN-dev	MEN-test	MTurk	RW	SimLex	SimVerb	YP	SEval-trail	SEval-test
Topology	0.00%	0.00%	0.00%	0.23%	0.35%	0.00%	18.82%	1.13%	0.00%	0.00%	0.00%	0.00%	5.00%
Text to graph	0.00%	0.00%	0.00%	0.43%	0.50%	0.30%	21.25%	20.94%	0.20%	0.06%	0.00%	0.00%	3.80%
SGNS	0.00%	0.00%	0.00%	0.43%	0.50%	0.30%	21.25%	20.85%	0.20%	0.06%	0.00%	0.00%	3.80%

Extrinsic Evaluation: Our embedding models were evaluated on two extrinsic tasks: part-of-speech tagging and named entity recognition using the F1 score as the performance metric. The experiment was run thrice, and the average F1 score was taken to obtain the final results, which are presented in Table 8. We observe that the text-to-graph based models outperform the topology-based and SGNS models in both tasks, with DeepWalk performing the best in named-entity recognition, and node2vec in part-of-speech tagging. This is an indication that our text-to-graph models have captured more contextual and semantic information and are able to better understand the relationship between words in a sentence.

Bias Evaluation: To evaluate the presence of bias in our experiment, we measured the similarity between the target sets (T1, T2) and attribute sets (A1, A2) for each bias query. For instance, in

the case of Gender bias, we used Male Terms and Female Terms as target sets, and Intelligence and Appearance as attribute sets. Our bias evaluation results, presented in Table 9, demonstrate that the DeepWalk topology-based model exhibits the lowest bias in Gender queries, while the node2vec topology-based and SGNS models display the highest bias. Interestingly, for Religion bias, we found that the LINE topology-based model has the least bias, while the SGNS model shows the highest bias, with DeepWalk text-to-graph ranking second. We have also calculated the overall cumulative ranking for each model on both queries, and we present the results in Table 10. Our findings demonstrate that the traditional SGNS embedding method exhibits the most bias compared to the Lexical embedding methods.

Table 7: Word Analogy evaluation results

	Word Analogy Datasets					
	GoogleAnalogyTestSet			BiggerAnalogyTestSet		
	3CosAdd	3CosMul	% OOV pair	3CosAdd	3CosMul	% OOV pair
node2vec	0.0063	0.0073	0.00%	0.0161	0.0157	0.98%
deepwalk	0.0105	0.0092	0.00%	0.0135	0.0106	0.98%
line	0.0097	0.0086	0.00%	0.0131	0.0106	0.98%
node2vec_t2g	0.0578	0.0627	10.55%	0.0418	0.0422	9.65%
deepwalk_t2g	0.0495	0.0483	10.55%	0.0427	0.0373	9.65%
line_t2g	0.0511	0.0497	10.55%	0.0424	0.0378	9.65%
SGNS	0.1425	0.1452	10.55%	0.0873	0.0851	9.65%

Table 8: Extrinsic evaluation results

	POS		NER	
	Macro F1	Weighted F1	Macro F1	Weighted F1
node2vec	0.8089	0.8686	0.3729	0.9694
deepwalk	0.7831	0.8356	0.3651	0.9691
line	0.7809	0.8351	0.3520	0.9685
node2vec_t2g	0.8345	0.9141	0.4782	0.9786
deepwalk_t2g	0.8317	0.9115	0.5002	0.9790
line_t2g	0.8321	0.9121	0.4996	0.9790
SGNS	0.8274	0.9054	0.4767	0.9784

Model	Rank
line	1
deepwalk	2
line_t2g	3
node2vec_t2g	4
node2vec	5
deepwalk_t2g	6
word2vec	7

Table 10: Bias Ranking. Sorting by the best to the worst model.

7 Conclusion and future works

In our preliminary study, we proposed methods to create lexical embeddings for downstream NLP tasks using the DBnary Lexical Database. We conducted comprehensive evaluations and bias analysis of graph-based embeddings and compared them with the traditional SGNS corpus-based embedding model. Our results indicate that graph-based embeddings generated from the relational topology of the lexical graph outperform SGNS embeddings in capturing semantic relationships between words. However, further research is needed to explore methods for assigning edge weights automatically instead of relying on manual assignments.

We observed that text-to-graph-based models perform better than topology-based models in most datasets except for those that focus on semantic relationships, where text-to-graph-based models rank second after SGNS. To improve the performance of text-to-graph-based models, better weight assignment methods need to be developed, for instance, using word probability. Moreover, the quality of the DBnary graph needs to be assessed to address missing and irrelevant nodes.

In addition to performance evaluations, we conducted a bias analysis of the embeddings. Our results demonstrated that SGNS embeddings exhibited higher levels of bias compared to lexical graph embeddings. This highlights the importance of considering bias in word embeddings and underlines the potential benefits of using lexical graphs to mitigate bias. However, a more comprehensive study is needed to gain a deeper understanding of the underlying factors contributing to bias, such as the characteristics of the training data and the embedding methods. Future research should also explore debiasing techniques to mitigate biases in the models. Furthermore, as our experiments utilized default parameters, future work will focus on hyperparameter tuning to optimize the performance of the lexical graph embedding models. Additionally, an interesting path for future exploration lies in leveraging the DBnary graph topology to employ Knowledge Graph Embedding methods for computing vector representations. By comparing the performance and characteristics of our baseline methods with a more specialized knowledge graph embedding technique we can gain insights into the advantages and limitations of different approaches.

Beyond improving current results, however, we acknowledge that this experiment is very preliminary and contains many limitations that should

	Gender				Religion			
	WEAT	WEAT ES	RND	RNSB	WEAT	WEAT ES	RND	RNSB
node2vec	6 (0.117)	7 (0.263)	7 (0.116)	6 (0.061)	2 (0.027)	1 (0.258)	4 (0.101)	6 (0.074)
deepwalk	2 (0.057)	5 (0.206)	1 (0.029)	1 (0.019)	5 (0.043)	5 (0.439)	2 (0.078)	1 (0.019)
line	1 (0.056)	3 (0.182)	4 (0.049)	2 (0.02)	1 (0.018)	3 (0.387)	1 (0.074)	3 (0.02)
node2vec_t2g	4 (0.099)	4 (0.2)	5 (0.065)	3 (0.023)	4 (0.041)	4 (0.408)	3 (0.09)	2 (0.019)
deepwalk_t2g	5 (0.103)	2 (0.175)	3 (0.043)	5 (0.042)	6 (0.048)	6 (0.478)	6 (0.112)	5 (0.04)
line_t2g	3 (0.091)	1 (0.138)	2 (0.043)	4 (0.039)	3 (0.037)	2 (0.37)	5 (0.112)	4 (0.04)
word2vec	7 (0.181)	6 (0.245)	6 (0.092)	7 (0.16)	7 (0.062)	7 (0.878)	7 (0.3)	7 (0.109)

Table 9: Bias evaluation results for Gender and Religion queries. Lower scores indicate lower bias w.r.t to a metric.

be handled if we want to provide alternatives to current first layer initialization steps in deep learning based models. We decided for the moment to focus on word embeddings as words represent a token granularity shared with lexical datasets, however, current approaches are now using so called subwords as tokens bringing better results and handling of out of vocabulary terms. In the near future, we will address such approaches using lexical data. Moreover, many tokenizer/embedders are now multilingual, hence we will also experiment with other languages available in DBnary, either in a monolingual setting or in multilingual setting.

References

- Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Pasca, and Aitor Soroa. 2009. A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27, Boulder, Colorado. Association for Computational Linguistics.
- Pablo Badilla, Felipe Bravo-Marquez, and Jorge Pérez. 2020. **WEFE: The Word Embeddings Fairness Evaluation Framework**. In *Twenty-Ninth International Joint Conference on Artificial Intelligence*, volume 1, pages 430–436.
- Amir Bakarov. 2018. **A Survey of Word Embeddings Evaluation Methods**.
- Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc."
- Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16*, pages 4356–4364, Red Hook, NY, USA. Curran Associates Inc.
- E. Bruni, N. K. Tran, and M. Baroni. 2014. **Multimodal Distributional Semantics**. *Journal of Artificial Intelligence Research*, 49:1–47.
- Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. **Semantics derived automatically from language corpora contain human-like biases**. *Science*, 356(6334):183–186.
- Jose Camacho-Collados, Mohammad Taher Pilehvar, Nigel Collier, and Roberto Navigli. 2017. **SemEval-2017 Task 2: Multilingual and Cross-lingual Semantic Word Similarity**. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 15–26, Vancouver, Canada. Association for Computational Linguistics.
- François Chollet et al. 2015. Keras. <https://keras.io>.
- Jeffrey Dastin. 2022. **Amazon Scraps Secret AI Recruiting Tool that Showed Bias against Women ***.
- Thierry Declerck, John Philip McCrae, Matthias Hartung, Jorge Gracia, Christian Chiarcos, Elena Montiel-Ponsoda, Philipp Cimiano, Artem Revenko, Roser Saurí, Deirdre Lee, Stefania Racioppa, Jamal Abdul Nasir, Matthias Orlikowski, Marta Lanau-Coronas, Christian Fäth, Mariano Rico, Mohammad Fazleh Elahi, Maria Khvalchik, Meritxell Gonzalez, and Katharine Cooney. 2020. **Recent developments for the linguistic linked open data infrastructure**. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5660–5667, Marseille, France. European Language Resources Association.
- Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. **Placing search in context: The concept revisited**. In *Proceedings of the 10th International Conference on World Wide Web, WWW '01*, pages 406–414, New York, NY, USA. Association for Computing Machinery.
- Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. **Word embeddings quantify 100 years of gender and ethnic stereotypes**. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

- Daniela Gerz, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. 2016. [SimVerb-3500: A Large-Scale Evaluation Set of Verb Similarity](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2173–2182, Austin, Texas. Association for Computational Linguistics.
- Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuo. 2016. [Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't](#). In *Proceedings of the NAACL Student Research Workshop*, pages 8–15, San Diego, California. Association for Computational Linguistics.
- Aditya Grover and Jure Leskovec. 2016. [Node2vec: Scalable Feature Learning for Networks](#).
- Felix Hill, Roi Reichart, and Anna Korhonen. 2014. [SimLex-999: Evaluating Semantic Models with \(Genuine\) Similarity Estimation](#).
- Matthew Kay, Cynthia Matuszek, and Sean Munson. 2015. [Unequal representation and gender stereotypes in image search results for occupations](#).
- Omer Levy and Yoav Goldberg. 2014. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning (CoNLL-2014)*, pages 171–180.
- Thang Luong, Richard Socher, and Christopher Manning. 2013. Better Word Representations with Recursive Neural Networks for Morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104–113, Sofia, Bulgaria. Association for Computational Linguistics.
- John P. McCrae, Paul Buitelaar, and Philipp Cimiano. 2017. The OntoLex-Lemon Model: development and applications. In *Proc. of the 5th Biennial Conference on Electronic Lexicography (eLex)*.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. [Efficient Estimation of Word Representations in Vector Space](#).
- Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751.
- Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. [DeepWalk: Online Learning of Social Representations](#). In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 701–710.
- Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. 2011. [A Word at a Time: Computing Word Relatedness using Temporal Semantic Analysis](#). In *Proceedings of the 20th International Conference on World Wide Web, WWW 2011*, pages 337–346.
- Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).
- Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. [The Risk of Racial Bias in Hate Speech Detection](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.
- Gilles Sérasset. 2012. [Dbnary: Wiktionary as a LMF based Multilingual RDF network](#). In *Language Resources and Evaluation Conference, LREC 2012*, Istanbul, Turkey. Nicoletta Calzolari and Khalid Choukri and Thierry Declerck and Mehmet Uğur Doğan and Bente Maegaard and Joseph Mariani and Jan Odijk and Stelios Piperidis.
- Gilles Sérasset. 2015. [DBnary: Wiktionary as a lemon-based multilingual lexical resource in RDF](#). *Semantic Web*, 6(4):355–361.
- Chris Sweeney and Maryam Najafian. 2019. [A Transparent Framework for Evaluating Unintended Demographic Bias in Word Embeddings](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1662–1667, Florence, Italy. Association for Computational Linguistics.
- Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. [LINE: Large-scale Information Network Embedding](#). In *Proceedings of the 24th International Conference on World Wide Web*, pages 1067–1077.
- Erik F. Tjong Kim Sang and Sabine Buchholz. 2000. Introduction to the CoNLL-2000 Shared Task Chunking. In *Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*.
- Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.
- Dongqiang (东强) Yang and David Powers. 2006. Word similarity on the taxonomy of WordNet.
- Zhaocheng Zhu, Shizhen Xu, Meng Qu, and Jian Tang. 2019. [GraphVite: A High-Performance CPU-GPU Hybrid System for Node Embedding](#). In *The World Wide Web Conference*, pages 2494–2504.

Information Extraction of Political Statements at the Passage Level

Juan-Francisco Reyes

Institute of Computer Science

Brandenburgische Technische Universität Cottbus-Senftenberg

`jf.reyes@b-tu.de`

Abstract

This research addresses the challenge of accurately identifying, extracting, and publishing political statements from the web. The thesis proposes a broader definition of a political statement and presents a novel information extraction system. The system leverages natural language processing techniques, a web crawler, a taxonomy of political issues, a Political Discourse Analyzer, machine learning as a service, and a content management system. The goal is to develop a theoretical model for the efficient extraction of political statements, reducing the need for human.

This research aims to use NLP techniques to accurately and automatically identify, extract, and publish the most politically relevant passage-level statements with minimal human intervention. By ensuring these statements are readily accessible through a suitable website, the proposed research seeks to empower political discourse stakeholders with a streamlined and efficient means of accessing valuable political and linguistic analysis. This problem-driven research will contribute to the advancements of NLP techniques and support a more informed and transparent democratic practice in the digital age.

1 Introduction

The WWW abounds with discursive content authored by politicians but dispersed in several web resources, such as governmental websites, news websites, social media platforms, or APIs. However, finding, selecting, extracting, and making the most relevant political statements properly accessible to political scientists, journalists, linguists, citizens, and others interested in political discourse is time-consuming and involves intense human effort.

How can we adapt and enhance existing natural language processing (NLP) techniques to accurately and automatically identify, extract, and publish the most politically relevant passage-level statements with minimal human intervention while ensuring their accessibility through a suitable website?

2 Existing solutions

Before reviewing the existing solutions, we must first answer the apparently simple and naive question, "What is a political statement?". In Linguistics, a statement is a "declarative sentence" that 1. expresses a fact, idea or opinion, 2. consists of one sentence, and 3. includes a clause composed of a doing verb and a subject, the thing or person it refers to.

In linguistics, a political statement is "a declarative sentence with political content". Nevertheless, in political science, a political statement is a verbal or non-verbal form of communication with political content that 1. expresses an intention to influence the recipient's decision, attitude or action, and 2. (in verbal form) consists of one or many sentences (a passage).

The literature review does not give a consensual definition of a political statement. Hence, rather

than being normative, we make a fundamental assumption by choosing a broader and more pragmatic way to define a political statement in this research a political statement is a coherent passage, sentence, or clause of political discourse with political content that conveys a political intention.

No existing solutions address the very same problem, but we found multiple solutions that partially solve the problem divided across the domains of computer science, political science and linguistics.

- **In computer science:** Multiple research aims to identify and extract relevant multi-sentence text (passage extraction) from an extensive collection of documents in response to a user's query (information retrieval) or in response to a user's question (question-answering) or for presenting a summary that better captures the critical information and ideas (text summarization) (Kenter et al., 2018; Xu et al., 2011).
- **In political science/Linguistics:** Multiple research aims to identify text genre/profile by analyzing linguistic features of text based on genre theory (analyzing generic constructs and the contexts in which such genres are produced, interpreted, and used), linguistic profiling by extracting lexical, grammatical and semantic features that characterize language variation, political discourse analysis (PDA) by analyzing discourse in political forums (such as debates, speeches, and hearings) and computational sociolinguistics by studying the relation between language and society from a computational perspective (Dunmire, 2012).

Most of the research in information extraction (IE) of political statements extract the embedded knowledge in a single-sentence text (not at the passage level) to populate a knowledge graph, using machine learning (ML) methods to automate the extraction task without digging much into the nuances of the political language (Bamman &

Smith, 2015). Another research area in IE studies one specific aspect of the political language, such as sentiment (Bonikowski & Zhang, 2023), stance (Gambini et al., 2022), or election forecasting (Jérôme et al., 2022). More recent research analyses specific political rhetoric traits in political discourse to extract argumentation (Lapesa et al., 2020).

3 Research questions

How can we design and implement an IE system that can accurately and automatically identify and extract the most relevant political statements from the WWW by analyzing domain-specific discourse and linguistic markers while minimizing the need for human intervention?

This broad research question can be broken down into more specific sub-questions, which encompass both theoretical and practical implications for examining linguistic complexity and its computational processing, including:

1. What are the most effective computational methods for analyzing morpho-syntactic structures and patterns to automatically identify and extract coherent and cohesive statements at the passage level?
2. How can NLP techniques be adapted or combined to accurately identify and extract factually correct and politically relevant statements while minimizing the reliance on manually annotated data or human intervention?
3. What are the critical political discourse markers and linguistic features that, once systematically detected and analyzed using NLP techniques (operationalized), predict more effectively politically relevant statements?

4 Solution approach

Contrary to traditional information extraction's scope of extracting relations, entities, and facts,

extracting political statements at the passage level should identify and consolidate information from various text parts to create a more comprehensive and coherent single text. Also, a relevant political statement possesses specific linguistic markers that should be computationally analyzed before proposing statements as candidates.

Thus, we propose developing an IE system that leverages NLP techniques to automatically process texts to extract and assess passages based on their political discourse markers (via syntactic and semantic features) to propose them as relevant statements. The general approach prefers rule-based methods over ML methods to study and describe the linguistic challenges thoroughly; however, ML methods are used whenever more efficiency is required. Our solution will incorporate the following components:

4.1 Web crawler

Implement a web crawler to automatically retrieve fresh political discourse texts from the US political scene from different web resources in the WWW, such as web archives, social media outlets, news websites, Etc. The crawler recognizes political discourse content on crawled pages and classifies texts in monologic (speeches, remarks) and dialogic (interviews, conferences, debates).

4.2 Taxonomy of political issues

Implement a taxonomy of political issues that classifies all political issues (persons, organizations, places, concepts, Etc.) linked to their respective representation in a knowledge base (Wikidata). Each entity has a lexicon with various ways to refer to itself ("aliases").

4.3 NLP pipeline

Implement an NLP pipeline using spaCy with the following components/tasks:

1. Named-entities recognition (NER), rule-and-lexicon-based and linked to Wikidata.

2. Named-entity disambiguation and linking (NED/NEL): used in case of ambiguous concepts or entities (i.e., Columbus [PERSON] and Columbus [PLACE]). ML model trained using automatically retrieved-context sentences from the WWW.
3. Coreference resolution: Identifying and linking different textual mentions that refer to the same entity or concept within a given text to improve the understanding of relationships between words, phrases, and sentences and to provide a more coherent representation of the text's meaning.
4. Relation extraction (RE): Using an open relation extraction (ORE) approach, which extracts relations and their arguments without a predefined schema (ClauseIE). More meaning may be inferred while extracting more relations.
5. Triple extraction: Knowledge in the form of triples is extracted using dependency parsing and matching algorithms to ensure a correct representation of facts in the real world.
6. Political Discourse Analyzer (PDA): Using multiple algorithms and matching rules, assessing political discourse markers in the statements (via syntactic and semantic features) classifies them as valid as a relevant candidate.

4.4 Machine Learning as a Service (MLaaS)

Implement ML models deployed on a cloud computing service (Google Cloud), accessed by the IE system via APIs.

4.5 Content Management System (CMS)

Implement a CMS to allow editors to promote candidate statements to be published on an observatory website.

4.6 ObPolDis –Observatory of Political Discourse

Implement a CMS to allow editors to promote candidate statements to be published on an observatory website, <https://obpoldis.netlify.app/>.

5 Evaluation methods

This research aims to comprehend the linguistic intricacies of addressing the problem and its computational implementation utilizing NLP techniques. As a result, the primary endeavor involves systematically exploring novel insights related to the studied artifacts' linguistic principles, methodologies, and performance. Substantial advancements in IE can be realized by understanding the NLP pipeline components or techniques employed to tackle the issue.

This research is problem-oriented, meaning that the research problem itself is on focus rather than the methods and tools to solve it. By identifying the existing knowledge base and its gaps through literature reviews and conducting preliminary experiments on an NLP pipeline prototype within the IE system, the core nature of the problem is determined. Once linguistic features that contribute to the relevance of political statements are defined (i.e., cohesion, coherence, correctness, accuracy, readability, complexity, and other syntactic and semantic features found in political discourse), fundamental experiments can be performed to establish relationships between variables along the NLP pipeline components.

Throughout the research, if existing theories cannot explain a phenomenon, multiple experiments are conducted to verify the accuracy of the proposed model. The focus is on achieving a strong qualitative correlation (through observation) rather than quantitative agreement. If verification fails, the model must be refined, and new observations may be required. Upon achieving a verified model, large-scale extractions can be conducted to gather information about the IE system's characteristics and performance.

Experiments involve manipulating variables along the NLP pipeline to improve the prediction of relevant political statements and evaluating individual components through experiments. For example, an essential aspect of the study is testing the Political Discourse Analyzer (PDA) component ("algorithm") with a dataset of political and non-political statements to determine which political discourse markers (or "index") better predict relevant political statements. After numerous iterations, the model's efficiency could be improved by defining linguistic attributes and political discourse markers that predict relevant political statements. The newly acquired knowledge can be framed as design considerations, which can be incorporated by modifying the initial product or developing a new design. When implemented effectively, the new product addresses the original problem.

6 Research objectives

The overall purpose of this work is to achieve a fundamental understanding of how a passage-level political statement can be extracted automatically from the WWW. This thesis aims to develop a theoretical model that describes the most efficient way to extract political statements automatically in mathematical terms.

References

- Bamman, D., & Smith, N. A. (2015). Open Extraction of Fine-Grained Political Statements. *Empirical Methods in Natural Language Processing*. <https://doi.org/10.18653/v1/d15-1008>
- Bonikowski, B., & Zhang, Y. (2023). Populism as Dog-Whistle Politics: Anti-Elite Discourse and Sentiments Toward Minority Groups. *Social Forces*. <https://doi.org/10.1093/sf/soac147>
- Dunmire, P. (2012). Political Discourse Analysis: Exploring the Language of Politics and the Politics of Language. *Language & Linguistics Compass*. <https://doi.org/10.1002/lnc3.365>
- Gambini, M., Fagni, T., Senette C. & Tesconi, M. (2022). Tweets2Stance: Users stance detection

- exploiting Zero-Shot Learning Algorithms on Tweets. arXiv.
<https://doi.org/10.48550/arXiv.2204.10710>
- Jérôme, B., Mongrain, P., & Nadeau, R. (2022). Forecasting the 2022 French Presidential Election: From a Left–Right Logic to the Quadripolarization of Politics. *PS Political Science & Politics*.
<https://doi.org/10.1017/s1049096522000488>
- Kenter, T., Borisov, A., Van Gysel, C., Dehghani, M., de Rijke, M., & Mitra, B. (2018). Neural Networks for Information Retrieval. arXiv preprint arXiv:1801.021782 .
- Lapesa, G., Blessing, A., Blokker, N., Dayanik, E., Haunss, S., Kuhn, J., & Pado, S. (2022). Analysis of Political Debates through Newspaper Reports: Methods and Outcomes. *Datenbank-Spektrum*.
<https://doi.org/10.1007/s13222-020-00344-w>
- Xu, W., Grishman, R., & Zhao, L. (2011). Passage Retrieval for Information Extraction using Distant Supervision. In H.
- Wang, & D. Yarowsky (Eds.), *IJCNLP 2011 - Proceedings of the 5th International Joint Conference on Natural Language Processing* (pp. 1046-1054). Association for Computational Linguistics (ACL)

**Discourse studies and
linguistic data science:
Addressing challenges in
interoperability,
multilinguality and
linguistic data processing
(DiSLiDaS)**

Validation of Language Agnostic Models for Discourse Marker Detection

Mariana Damova*

*Mozaika, Ltd.

*mariana.damova@mozajka.co

Giedrė Valūnaitė Oleškevičienė[⊙]

[⊙]Mykolas Romeris University

[⊙]gvalunaite@mruni.eu

Purificação Silvano[†]

[†] Centre of Linguistics of the University of
Porto

[†]msilvano@letras.up.pt

Ciprian-Octavian Truică^{§,‡}

[§]Uppsala University

[§]ciprian-octavian.truica@it.uu.se

Christian Chiarcos^{§,‡}

Goethe-Universität

chiarcos@cs.uni-frankfurt.de

Kostadin Mishev[°]

[°]Ss. Cyril and Methodius University

[°]kostadin.mishev@finki.ukim.mk

Chaya Liebeskind[◇]

[◇]Jerusalem College of Technology

[◇]liebchaya@gmail.com

Dimitar Trajanov[°]

[°]Ss. Cyril and Methodius University

[°]dimitar.trajanov@finki.ukim.mk

Elena-Simona Apostol^{§,‡}

[‡]University Politehnica of Bucharest

[‡]elena.apostol@upb.ro

Anna Bączkowska[×]

[×]University of Gdansk

[×]anna.baczkowska@ug.edu.pl

Abstract

Using language models to detect or predict the presence of language phenomena in the text has become a mainstream research topic. With the rise of generative models, experiments using deep learning and transformer models trigger intense interest. Aspects like precision of predictions, portability to other languages or phenomena, scale have been central to the research community. Discourse markers, as language phenomena, perform important functions, such as signposting, signalling, and rephrasing, by facilitating discourse organization. Our paper is about discourse markers detection, a complex task as it pertains to a language phenomenon manifested by expressions that can occur as content words in some contexts and as discourse markers in others. We have adopted language agnostic model trained in English to predict the discourse marker presence in texts in 8 other unseen by the model languages with the goal to evaluate how well the model per-

forms in different structure and lexical properties languages. We report on the process of evaluation and validation of the model's performance across European Portuguese, Hebrew, German, Polish, Romanian, Bulgarian, Macedonian, and Lithuanian and about the results of this validation. This research is a key step towards multilingual language processing.

1 Introduction

Using language models to detect or predict the presence of language phenomena in the text has become a mainstream research topic. The performance of these models heavily depends on the quantity and on the quality of the data used for training them. Producing datasets of training data is a very time-consuming and expensive process, requiring human expertise. Deep learning models have been so far built by training single languages one by one. This requires the availability of training data in each language of interest, and makes obtaining language

models for multiple languages complicated, expensive and virtually impossible for smaller or rare languages. That is why research efforts have been focusing on removing the need for manual preparation of training data by developing deep learning architectures able to produce language models for languages without training on them - language agnostic models. Language agnostic models build models based on training data in one language, and then extrapolate them to other unknown for the model languages. It is important to know how well they perform and whether the quality of the prediction results in unseen languages is good enough to adopt and further develop these approaches and architectures. This paper presents experiments with a language-agnostic model in 8 languages, trained on data in English, to detect the presence and absence of discourse markers in unseen text and discusses the process and the results of validating their performance, demonstrating the good performance and the viability of the model. In our case, the model targets discourse markers, essential pointers for the communicational setting and the speaker's attitudes. They have particular roles in facilitating discourse organization and providing text coherence and cohesion between discourse segments.

The structure of the paper is as follows: Section 2 presents related work; Section 3 describes the language-agnostic machine learning method that has been adopted for the experiment; Section 4 gives an overview of the multilingual corpus used in the experiment; Section 5 describes the experiment, discusses the validation process and the performance of the language-agnostic model; Section 6 concludes the paper.

2 Related work

Regarding NLP tasks, there have been advancements in identifying and classifying discourse markers. For instance, Zufferey (2004) describes an experiment where discourse markers are detected and assigned inferential semantic functions. For the improvement of automatic methods for discourse markers detection and classification, shared tasks such as DISRPT 2019 and 2021 editions (Zeldes et al., 2019, 2021) and Discourse Relation Classification across RST (Mann and Thompson, 1988), SDRT (Asher et al., 2003), and PDTB (Prasad et al., 2008) have played a significant role. Following CoNLL 2015 setting, Kurfali (2020) developed an experiment to determine the efficacy of

pre-trained language models in the task of shallow discourse parsing (SDP) used to identify explicit local discourse relations without resorting to tree/graphs structures. The BERT-based model and the Hugging's face Transformer library were employed with the maximum sequence length 400 for the first approach and 250 for the second. For the test set, the author used PDTB. The model evaluation was performed on top of the official results of CoNLL 2015 (Xue et al., 2015) and 2016 (Xue et al., 2016) shared tasks, and of (Knaebel et al., 2019). Regarding connective identification, the model accomplished an F1-score of 95.76%, similar to previous experiments. In the 2021 edition of the DISRPT Shared Task, the system with the best results was DisCoDisCo (Gessler et al., 2021) with a Transformer-based neural classifier. This model outperformed state-of-the-art scores from the 2019 DISRPT concerning connective detection with an F1-score of 91.22%.

3 Language agnostic methods

Language-agnostic models have been developed to allow cross-language analysis and language phenomena detection without the need to process training data in each language manually. Such model is La-BSE, which we have adopted for our experiment, based on the amount of languages it is able to cover and on its modeling architecture.

The Google's language-agnostic BERT sentence embedding (La-BSE) model supports 109 languages (Feng et al., 2020). The multilingual architecture of BERT is adapted to produce language-agnostic sentence embeddings for 109 languages. La-BSE combines the masked-language model (MLM) and translation language model (TLM) pre-training with a translation ranking task using bi-directional dual encoders. This method improves the average bi-text retrieval accuracy and establishes new state-of-the-art on the bi-text retrieval.

4 Datasets

The multilingual datasets that have been part of the experiment contain examples from nine languages English, Lithuanian, Bulgarian, German, Macedonian, Romanian, Hebrew, Polish and European Portuguese, compiled from the publicly available TED Talk transcripts. It is an ongoing expansion of TED-EHL parallel corpus LINDAT/CLARIN-LT repository¹. In addition, we have produced a list

¹<http://hdl.handle.net/20.500.11821/34>

of multiword expressions (MWE) that can occur as discourse markers in specific contexts and as content expressions in others, where ambiguity is tricky to capture. For example, the expression *you know* in examples 1 and 3 below describes the content, whereas in example 2 it describes a discourse marker.

1. By the way, just so *you know*
2. But *you know*, they have, after all, evolved in a country without telephones,
3. *you know* what I mean.

Expressions of this nature are also *I remember, I mean, I think, you see*, etc.

Other MWEs from the established discourse markers list are lexicalized discourse markers that are interpreted as such in any context. Such MWEs are *of course, for example, above all, in addition* and the like.

We have produced eight bilingual datasets with aligned parallel texts in English and another language, based on the occurrence of MWE potentially describing a discourse marker in the sentence context. The structure of English part of the aligned bilingual corpus is shown in table 1.

In the bilingual parallel corpus, another four columns to the right of the last column of the data for English contain the translations of the English examples in the given language from the eight we cover. So, we end up with a corpus of eight bilingual parallel aligned corpora with an overall size presented in table 2.

5 Experiment

The English dataset was used as a baseline. It is composed of the union of all unique sentence contexts from all language pairs, and counts 44,209 sentence contexts. From them 4777 have been manually annotated, and 1019 turned to be with a discourse marker present (1) whereas 3758 - without a discourse marker present (0). The English dataset was split 80% for training and 20% for testing. The training set is used to fine-tune the XLM-RoBERTa Large model for the classification. The test set is used to evaluate the performance on unseen samples to predict the presence or absence of discourse markers in the training dataset.

The same training dataset was used to train with the La-BSE language-agnostic method to generate a model that has been consequently run through

all languages from the bilingual parallel corpus (cf. table 2 described above). As a result, prediction for the presence or absence of discourse markers in each context for each language has been generated and output in the table structure shown in table 3. Note that the English example does not have a value for presence or absence of a discourse marker in the context (9) in table 3. This indicates that the trained model in English has been run through unseen examples in the other languages.

6 Validation

The validation of the results has had two stages. In the first stage, the prediction results have been verified against the manual annotations. Table 4 shows the evaluation for Bulgarian and Lithuanian with considerably better prediction results for Lithuanian - 0.94 precision than for Bulgarian - 0.74 precision.

As a second step, human experts manually validated the predictions of the language-agnostic model. To provide the most accurate possible outlook, we took the first 100 lines of each bilingual file, ensuring that all selected examples differ.

Then, human experts had to evaluate whether the prediction of the model was correct or not. The validation has shown that the La-BSE method, trained on English text, performs very well on unseen languages regardless of their family and on diverse unseen texts. The results are shown in table 5 below with an average of 12 wrongly predicted occurrences and 88% precision.

The reasons for the discrepancies in the correct prediction rate are still to be analyzed. We predict that they may be related to the texts themselves, the human analysts' expert judgement, and the structure of the language compared to the structure of English.

7 Conclusion

This paper presented an experiment of applying a language-agnostic machine learning method to a multilingual corpus of 9 languages to verify how well it would perform detecting discourse markers when trained in English. The two validation methods with testing corpus and with human expert assessment showed only a little discrepancy in the analysis of the results. The human expert analysis performed better than the automatic evaluation of the testing corpus. The reasons for these discrepancies are to be investigated in detail in our future

Table 1: Structure of the English part of the corpus

MWE	Sentence chunk	Context	Discourse Marker Presence
I remember	And I remembered that the old and drunken guy destroying my statistical significance of the test. So I looked carefully at this guy. He was 20-some years older than anybody else in the sample.	And I remembered that the old and drunken guy came one day to the lab wanting to make some easy cash	0
You know	But you know, these stories, because he would have pulled the mean of the group lower, giving us even stronger statistical results than we could. So we decided not to throw the guy out and to rerun the experiment.	But you know, these stories, and lots of other experiments that we've done on conflicts of interest, basically kind of bring two points	1

Table 2: Constituted multilingual datasets

language	aligned sentences with MWE
English	43600
Macedonian-English	2846
German-English	15852
Lithuanian-English	4112
Bulgarian-English	19209
Portuguese-English	4398
Polish-English	17408
Romanian-English	18946
Hebrew-English	23566

Table 3: Example of model output

DM EN	S Chunk EN	DM Presence EN	text LANG	LA-BSE prediction
in fact	In fact, she had aged a lotThe woman who as a child had skipped with him through fields and broken his heart	9	Всъщност, доста беше остаряла Жената, която като дете бе подскачала с него през полята и бе разбила сърцето му	1

work. This experiment proved that the language-agnostic models' performance is not affected significantly by the structure of the language or other lexical or grammatical peculiarities of the single

languages and gives a good prediction for the presence of discourse markers in texts in unseen by the model languages.

Table 4: Language-Agnostic Methods Results

Model	Accuracy	Precision	Recall	Specificity	F1-Score	MCC
La-BSE (BG)	0.7273	0.7403	0.7090	0.7459	0.7243	0.4551
La-BSE (LT)	0.8338	0.9412	0.8758	0.2877	0.9073	0.1228

Table 5: Human validation results

Language	Number of Wrong Predictions	Total Number of Examples	Precision ratio
BG	10	100	0,90
MK	19	100	0,81
EN	16	100	0,84
HE	5	100	0,95
PT	20	100	0,80
DE	17	100	0,83
PL	10	100	0,90
LT	12	100	0,88
RO	1	100	0,99

Acknowledgements

This work has been done within the COST Action CA18209 - NexusLinguarum.

References

- Nicholas Asher, Nicholas Michael Asher, and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.
- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.
- Luke Gessler, Shabnam Behzad, Yang Janet Liu, Siyao Peng, Yilun Zhu, and Amir Zeldes. 2021. [Dis-CoDisCo at the DISRPT2021 shared task: A system for discourse segmentation, classification, and connective detection](#). In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 51–62, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- René Knaebel, Manfred Stede, and Sebastian Stober. 2019. [Window-based neural tagging for shallow discourse argument labeling](#). In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 768–777, Hong Kong, China. Association for Computational Linguistics.
- Murathan Kurfali. 2020. Labeling explicit discourse relations using pre-trained language models. *ArXiv*, abs/2006.11852.
- William Mann and Sandra Thompson. 1988. [Rhetorical Structure Theory: Toward a functional theory of text organization](#). *Text*, 8:243–281.
- Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltasakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. [The Penn Discourse TreeBank 2.0](#). In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).
- Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, and Attapol Rutherford. 2015. [The CoNLL-2015 shared task on shallow discourse parsing](#). In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, pages 1–16, Beijing, China. Association for Computational Linguistics.
- Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Attapol Rutherford, Bonnie Webber, Chuan Wang, and Hongmin Wang. 2016. [CoNLL 2016 shared task on multilingual shallow discourse parsing](#). In *Proceedings of the CoNLL-16 shared task*, pages 1–19, Berlin, Germany. Association for Computational Linguistics.
- Amir Zeldes, Debopam Das, Erick Galani Maziero, Julian Antonio, and Mikel Iruskieta. 2019. [Introduction to discourse relation parsing and treebanking \(DISRPT\): 7th workshop on Rhetorical Structure Theory and related formalisms](#). In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 1–6, Minneapolis, MN. Association for Computational Linguistics.
- Amir Zeldes, Yang Janet Liu, Mikel Iruskieta, Philippe Muller, Chloé Braud, and Sonia Badene. 2021. [The DISRPT 2021 shared task on elementary discourse](#)

unit segmentation, connective detection, and relation classification. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 1–12, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sandrine Zufferey. 2004. Une analyse des connecteurs pragmatiques fondée sur la théorie de la pertinence et son application au TALN. *Nouveaux Cahiers de Linguistique Française*, 25:257–272.

ISO-DR-core Plugs into ISO-dialogue Acts for a Cross-linguistic Taxonomy of Discourse Markers

Purificação Silvano
University of Porto and CLUP
Via Panorâmica, s/n
4150-564 Porto, Portugal
msilvano@letras.up.pt

Mariana Damova
Mozaika Ltd
Solunska 52
Sofia 1000, Bulgaria
mariana.damova@mozajka.co

Abstract

The present paper¹ proposes an interoperable taxonomy to represent the meaning of discourse markers based on ISO DR-core (ISO 24617-8) but with a plug-in to ISO-dialogue acts (ISO 24617-2). The proposed taxonomy encompasses two dimensions: the semantic, with values regarding the discourse relations signalled by discourse markers, and the pragmatic, with values concerning the communicative function realized by discourse markers. We present a proof of concept for this two-dimensional taxonomy in a multilingual parallel dataset in three languages, English, European Portuguese and Bulgarian, comprising 165 textual segments with multiword discourse makers obtained from publicly available TED Talk transcripts. We show that the two-dimensional taxonomy can successfully annotate cross-linguistically the meaning of discourse markers and discuss linguistic evidence where extension of the proposed taxonomy can be relevant.

1 Background and Motivation

Discourse markers have been largely studied in different languages (e.g. [Schiffrin \(1987\)](#); [Fraser \(1996\)](#); [Knott and Dale \(1993\)](#); [Silvano \(2010\)](#); [Taboada \(2006\)](#); [Das \(2014\)](#); [Mendes et al. \(2018\)](#); [Stede et al. \(2019\)](#), among others) due to their relevance in discourse interpretation and, simultaneously, to their complexity regarding their multifunctional nature. Some of these studies have

¹This work was presented in the 1st Workshop on Discourse Studies and Linguistic Data Science-DiSLiDaS 2022 in Jerusalem, 24th May 2022 (https://dislidas.mozajka.co/?page_id=211)

rendered several taxonomies within different theoretical frameworks, some language independent, others - language specific, many associated to discourse relations taxonomies (eg. [Mann and Thompson \(1988\)](#); [Sanders et al. \(1992\)](#); [Asher et al. \(2003\)](#); [Prasad et al. \(2008\)](#); [Zeyrek et al. \(2018\)](#)), and most directed to written discourse (cf. eg. for spoken discourse [González \(2005\)](#); [Maschler and Schiffrin \(2015\)](#); [Crible \(2014\)](#)).

Bearing in mind, on the one hand, the diversity of frameworks described and, on the other hand, the usefulness of establishing comparisons between annotated data in the same language and across languages, there have been some efforts to reconcile different taxonomies, such [Benamara and Taboada \(2015\)](#) and [Sanders et al. \(2021\)](#). One of those unifying proposals has resulted in the *Semantic annotation framework (SemAF) — Part 8: Semantic relations in discourse, core annotation schema (DR-core) – ISO 24617-8* ([Bunt and Prasad, 2016](#); [Prasad and Bunt, 2015](#)). ISO 24617-8 (ISO, b) stipulates an interoperable core-annotation scheme for low-level discourse relations, i.e., local dependencies. Although the aforementioned aggregating schemes are designed for annotating discourse relations, since these can be explicitly marked by discourse markers that act as cue words/ expressions to infer the proper relation of meaning, it is assumed that they can also be used to represent discourse markers semantics/pragmatics. There are, however, research that design discourse markers-oriented taxonomies experimenting in more than one language, as is the case of [Crible and Zufferey \(2015\)](#).

Regardless of the theoretical approach, the uni-

fyling taxonomies lack a wide-range application to corpora across languages, genres and types of discourse to test their reliability and comprehensiveness. Regarding multilinguality, ISO (b) states that “a future part of ISO 24617 is envisaged that will complement this document by providing a complete interoperable annotation scheme for DRels (discourse relations), while also addressing the multilingual dimension of the standard”, but it has not been published so far. In what concerns written and oral discourse, Crible and Degand (2019), for example, observe that “these interoperable schemes either target written corpora or the relational meanings of spoken DMs, while specific (non-relational) spoken functions still lack a similar unifying approach to date”.

The taxonomy of discourse markers put forward in this paper addresses these two types of shortage. On the one hand, by combining ISO DR-core (ISO 24617-8) with ISO-dialogue acts (ISO 24617-2), we can represent not only the semantic meaning of discourse markers (or their relational meanings, as described by Crible and Degand (2019)) with the values of discourse relations but also their pragmatic meaning (or non-relational meaning, as proposed by Crible and Degand (2019)), making use of communicative functions. On the other hand, by applying to a multilingual dataset, which will eventually be published, we demonstrate to what extent the taxonomy is truly interoperable.

2 Related Work

One can opt for narrower and broader notions regarding discourse markers. For instance, Schiffrin (1987) presents “a definition which encompasses both “connectives” (e.g. *and*, *but*, *because*, *actually*) and pragmatic particles more specific to speech (e.g., *well*, *I mean*, *you know*). As the author puts it, this is intentionally a vague definition, not to limit the set of discourse markers. Schiffrin (1987) assigns to discourse markers a bracketing role, which Crible and Degand (2019) consider too restricting.

Schiffrin (1987) describes the multifunctionality of discourse markers distinguishing between (1) ideational structure, with relations between propositions, e.g. a cohesion relation, a topic relation or a functional relation; (2) action structure, which describes the organisation and constraints on the use of speech acts; (3) exchange structure, which is “the outcome of decision procedures by which speakers

alternate sequential roles and define those alternations in relation to each other” (Schiffrin, 1987). The author argues that discourse markers may simultaneously have roles within these three structures. Other authors have discussed the multifunctionality of discourse markers. Hovy (1995) considers that discourse markers convey rhetorical structure, interpersonal/ intentional structure, semantic structure, stylistic variants and guidance information. Additionally, CribleDegand+2019+71+99 put forward an annotation taxonomy of discourse markers in spoken language featuring two independent layers of semantic-pragmatic information, domains and functions. The four domains are the following: ideational, rhetorical, sequential or interpersonal. The model includes 15 functions (eg. addition, contrast), some based on Prasad et al. (2007)). They have tried the model in different languages (French, English, Polish, Spanish) and modalities (spoken, written, signed), attesting to their reliability and suitability for cross-lingual analysis.

Petukhova and Bunt (2009) also prove with corpus analysis that discourse markers can have multiple meanings concurrently because one dialogue act can serve several goals simultaneously. These authors adopt an empirically-based and formal approach to the semantic functions of discourse markers in dialogue capable of capturing their multifunctional nature. Within the semantic framework of Dynamic Interpretation Theory (Bunt et al., 2020), they propose a multilayered and multidimensional taxonomy with a set of communicative functions, which was the precursor of the *Semantic annotation framework (SemAF) — Part 2: Dialogue acts*, ISO 24617-2 (ISO, a), an interoperable dialogue act annotation framework with dimensions, communicative functions and qualifiers to annotate dialogue acts.

Besides the part that deals with dialogue acts, ISO 24617 comprises part 8 (ISO, b), which stipulates an interoperable core-annotation scheme for low-level discourse relations, i.e., local dependencies, according to the meaning of the relation’s arguments. Despite having been designed to annotate discourse relations, ISO 24617-8 has, nevertheless, been used to develop discourse markers lexicon such as PDTB (Prasad et al., 2008), LexConn (Roze et al., 2010), LDM-PT (Mendes et al., 2018), but always taken as triggers of discourse relations.

To sum up, in the face of the diversity of frame-

works described, on the one hand, and, on the other hand, the usefulness of establishing comparisons between annotated data in the same language and across languages, there have been some efforts to reconcile different taxonomies, and at the same time, there have been some proposals to develop an overarching model for discourse markers annotation. Some of those taxonomies can be used to annotate the meaning of discourse markers, but only a few are specifically designed for that purpose. Moreover, none attempts to use ISO standards that can capture both the semantic and pragmatic meaning of discourse markers. Furthermore, most discourse markers-oriented taxonomies lack a wide-range application to corpora across languages, genres and types of discourse to test their reliability and comprehensiveness.

Considering what has already been done and what could be done to contribute to a better understanding of discourse markers, we propose a comprehensive interoperable discourse markers taxonomy able to represent not only the semantic meaning of discourse markers but also their pragmatic meaning, and we determine its reliability by applying it to a sample of a multilingual dataset.

3 The ISO-based Unifying Taxonomy

In our proposal, we assume that discourse markers subsume words or expressions that link utterances and play different pragmatic functions (Schiffrin, 1987; Fraser, 2009; Crible, 2014). Thus, we include in this group - connectives (*as a consequence, on the one hand*) and pragmatic particles (*you know, I mean*). As is well established in the literature, we assume discourse markers to be multifunctional in the sense that they can have, in some contexts, different semantic and pragmatic meanings and also that they can have multiple meanings simultaneously (Petukhova and Bunt, 2009).

We propose an ISO-based unifying taxonomy of discourse markers to annotate both written and spoken discourse cross-linguistically. We adopt the set of core discourse relations provided by ISO 24617-8 (ISO, b), which was defined on the grounds of different theoretical approaches and annotation endeavours. According to this framework, the discourse relations are of two types: symmetric, in which case the two arguments assume relation-specific semantic role, and asymmetric, when the arguments take the same semantic role. The discourse relations are used to ascertain the semantic

meaning of discourse markers such as "as a result of" (Cause) (cf. ex.(1)), "for example" (Exemplification) (cf. ex. (2)).

- (1) It turns out that rarely do we practice under the types of conditions we're actually going to perform under, and **as a result**, when all eyes are on us, we sometimes flub our performance.
- (2) Ah, earth's oceans. They are beautiful, inspiring, life-sustaining. They are also, as you're probably quite aware, more or less screwed. In the Seychelles, **for example**, human activities and climate change have left corals bleached. Overfishing has caused fish stocks to plummet.

Notwithstanding, not all discourse markers convey a relational meaning, and instead play an interactional function, not accounted for by ISO 24617-8. It should be noted that this part of the SemAF admits pragmatic variants of discourse relations (Bunt and Prasad, 2016), that is, for each discourse relation, there is the possibility of one or both arguments expressing an implicit belief or a dialogue act. In those instances, the relevant arguments, and not the discourse relations, are annotated with that information because, according to the authors, the inference of a belief or a dialogue act depends on the arguments, and not on the discourse relation. This distinction is not, however, relevant for our taxonomy, since we aim at a typology which encodes the meaning of the discourse marker and not the nature of the discourse relation. To properly represent the interactional (or pragmatic) meaning of some discourse markers, we deemed it best to add an annotation plug-in to Semantic annotation framework (SemAF) — Part 2: Dialogue acts (ISO, a), (Bunt, 2019), (Bunt et al., 2020). This mechanism is introduced by Bunt (2019) and Bunt et al. (2020) with the inverse direction, from ISO 24617-2 to ISO 24617-8, to solve the problem of annotating semantic content of dialogue acts. In our taxonomy, we utilize the plug-in to overcome the limitations of the discourse relations set in ISO DR-core, enabling the encoding of the pragmatic meaning of discourse markers such as *you know*, which can convey the communicative function *Opening* (cf. ex.(3)), and *of course*, which expresses certainty, hence the qualifier *Certain* (cf. ex.(4)). Although the meta-model designed for ISO 24617-2 involves dimensions, communicative functions and qualifiers, for our taxonomy the last two suffice.

- (3) (Applause) Lakshmi Pratury: Just stay for a second. Just stay here for a second. (Applause) **You know**, when I heard Simon's – please sit down; I just want to talk to him for a second –
- (4) You've dissolved the barrier between you and other human beings. And this, **of course**, is the basis of much of Eastern philosophy

Table 1 summarizes the different values for each dimension.

Accordingly, there are discourse markers with a semantic dimension that receive one of the values from the first column. The discourse markers with a pragmatic dimension can be assigned a general communicative function (first column from the pragmatic dimension) or a more specific communicative function (second column from the pragmatic dimension), as discussed in example (3) above. Their interpretation may require an additional value related to notions of certainty, conditionality, and sentiment, like in examples (5), where the discourse marker plays a communication function *confirm*, in addition to carrying a value represented by the qualifier *Certain*. The multifunctional nature of discourse markers is evidenced by example (6), where the discourse marker *of course* has, concurrently, a semantic and pragmatic value, signalling the discourse relation *Expansion* and having the communication function *Confirm* and the qualifier *Certain*.

- (5) And that is, there is a sudden emergence and rapid spread of a number of skills that are unique to human beings like tool use, the use of fire, the use of shelters, and, **of course**, language, and the ability to read somebody else's mind and interpret that person's behavior.
- (6) Instead, so far, the measurements coming from the LHC show no signs of new particles or unexpected phenomena. **Of course**, the verdict is not definitive. In 2015, the LHC will almost double the energy of the colliding protons,

We acknowledge that both the semantic and pragmatic dimensions of the annotation scheme we propose can be in themselves multi-dimensional². However, although a text span can convey more

²This observation was made by one reviewer, to whom we thank.

than one communicative function and/ or be linked to another by more than one discourse relation, the same is not as frequent with discourse markers. In other words, the same discourse marker can be assigned different communicative functions and discourse relations in different contexts, but, as we will demonstrate in the next section, the concurrence of two semantic meanings or two communicative functions in the same discourse marker in the same context is rarely observed in our annotation framework and data.

4 The Proof of Concept

With the goal of determining the reliability and coverage of the proposed taxonomy, we devised a short experiment with a dataset of 165 multiword discourse makers occurrences in three languages, English, European Portuguese and Bulgarian. We selected multiword expressions because we have also been working on cross-lingual and language-agnostic methods for discourse markers prediction, and multiword discourse markers pose relevant problems when dealing with automatic detection. The data for this experiment were extracted from publicly available TED Talk transcripts. They represent a subset from a larger parallel multilingual corpus covering English, European Portuguese, Lithuanian, Bulgarian, German, Macedonian, Hebrew, Romanian, Italian and Polish, where English has been established as a pivot language for all language pairs of the dataset. A baseline annotation was performed by a linguist for the English data. Whenever necessary, annotation decisions were discussed in the working group. After establishing the gold standard, an annotation manual was created. While all languages have been annotated, we present evidence from three of them in this paper. Table 2 illustrates the result of applying the taxonomy to the three datasets.

Table 2 reveals that ISO 24617-8 adequately represents the meaning of most of the discourse markers found in the three datasets. However, the plug-in to ISO 24617-2 enables a more suitable classification of a group of discourse markers, even if they are few. A very small number of discourse markers can be classified using both dimensions (*of course*, *de facto*, разбира се)

In the set of 165 multiword discourse makers occurrences in three languages, English, European Portuguese and Bulgarian, we observed that the majority of the discourse markers convey a seman-

Table 1: Taxonomy of discourse markers.

Semantic dimension	Pragmatic dimension		
Cause	CheckQuestion	AutoNegative	conditional
Expansion	Inform	AlloPositive	unconditional
Asynchrony	Agreement	AlloNegative	certain
Concession	Disagreement	FeedbackElicitation	uncertain
Elaboration	Correction	Stalling	positive
Exemplification	Answer	Pausing	negative
Manner	Confirm	InteractionStructuring	
Condition	Disconfirm	Opening	
Negative Condition	Offer	TopicShift	
Purpose	Promise	SelfError	
Exception	AddressRequest	Retraction	
Substitution	AcceptRequest	SelfCorrection	
Conjunction	DeclineRequest	InitGreeting	
Contrast	AddressSuggest	InitSelfIntroduction	
Synchrony	DeclineSuggest	Apology	
Similarity	Request	Thanking	
Disjunction	Instruct	InitGoodbye	
Restatement	Suggest	Compliment	
	AddressOffer	Congratulation	
	AcceptOffer	SympathyExpression	
	DeclineOffer	ContactCheck	

tic meaning represented by nine different discourse relations, which are *Exemplification*, *Elaboration*, *Synchrony*, *Contrast*, *Concession*, *Conjunction*, *Restatement*, *Cause* and *Expansion*. The values of *Restatement* - inferred when the discourse marker links two arguments that represent the same situation but from different perspectives (ISO, b) -, and *Expansion* - assigned when the second argument is a situation involving some entity/entities present in the first argument, expanding the narrative or expanding on the setting relevant for interpreting the first argument (ISO, b) -, are, in our dataset, expressed by more multiword discourse markers, at least for English and European Portuguese. Although, in the case of *Restatement*, the discourse markers are variants or have very similar meanings (eg. in Portuguese, *por outras palavras*, *noutras palavras*), looking at the discourse markers that carry the value of *Expansion*, we can observe, for English and European Portuguese, more lexical variety (eg. *in fact*, *that is*, *of course*). In fact, regarding the set of discourse relations, it is not surprising that more specific ones would permit a more fine-grained distinction of the discourse markers semantic value. ISO 24617-8 already assumes that this applies to *Expansion*. It also postulates

that *Elaboration* subsumes the discourse relation *Summary* proposed by Mann and Thompson (1988). However, discourse marker *sum up* encodes a different meaning when compared to *in particular*, for instance. Other discourse markers such as *in fact*, *de facto*, *всѣщност* would be better represented with a more informative discourse relation, like, for instance, *Affirmation*.

In what concerns the pragmatic dimension, despite the extensive list of communicative functions (cf. Table 1), the sample of discourse markers subject to this experiment displays little variety, only four, to be precise. The communicative functions that the discourse markers fulfill are the following: *CcheckQuestion*, used to determine, from the addressee, whether a proposition, which forms the semantic content, is true (ISO, a); *Confirm*, utilized to inform the addressee that the proposition which constitutes the semantic content is true (ISO, a); *Opening*, to show to the addressee that the sender is ready to start the dialogue (ISO, a); and *AlloPositive*, employed to inform the addressee that the sender believes that the addressee is processing what is being said (ISO, a). The fact that the same discourse marker can signal different communicative functions, as is the case of *you know* and its

equivalents in the other three languages, or discourse relations, like *on the other hand* with a *Contrast* and *Concession*), or even simultaneously a discourse relation and a communicative function, like *in fact*, *de facto*, attests the polyfunctionality of discourse markers. Furthermore, the same discourse marker can carry a communicative function and an additional value, represented in our proposal by qualifiers, which are predicates that can "narrow down the meaning of a communicative function, called restrictive qualifiers, and those that add something to the meaning of a communicative function, called additive qualifiers" (Bunt et al., 2012). In our dataset, we only came across one discourse, *of course*, *claro*, *разбира се* to which a certainty qualifier (restrictive) was assigned.

Table 2 includes all the cases where the discourse markers translated from English to European Portuguese and Bulgarian have the same semantic and/or pragmatic in the three languages. However, on close inspection, the cross-lingual analysis of the dataset reveals that one and the same English expression gets translated with different expressions conveying distinct meanings. In Bulgarian in different contexts, for example, we encounter *правилно*, Bulgarian for the English words (*right*, *correct*), conveying a value of *CheckQuestion* (cf. ex.(7), (8)), and not *всъщност* in a context where in English *in fact* with the meaning of *Expansion* is used.

- (7) и рожденият ден на Лейди Гага. Не ви ли звучат невероятно? Но повечето хора не са съгласни. Правилно, защото техните умове не се вписват, в това което обществото смята за нормално, често биват избягвани и неразбрани.
- (8) and Lady Gaga's birthday. Don't they sound incredible? But most people don't agree. And **in fact**, because their minds don't fit into society's version of normal, they're often bypassed and misunderstood.

This leads to considerations that the different translations of the same expression can signal different meanings or communication functions and to the assumption that the thorough cross-lingual analysis can provide insight into the application and the further enrichment of the proposed taxonomy. Further, observation points to the interdependence between some conjunctions with discourse markers. It is not rare to see *in fact* preceded by *and*,

for example preceded by *so*, and much more. Although out of the scope of the present work, these phenomena present interesting evidence related to the classification and identification of the roles of discourse markers in discourse and their representation.

5 Final Remarks

In conclusion, when compared to other proposals, our taxonomy has the following strengths: a) it was specifically designed to codify the meaning of discourse markers; b) the two dimensions, semantic and pragmatic, are featured by values that are specific to those dimensions (and not generic); c) the dimensions-oriented values properly account for the role or roles each discourse marker can play in discourse; d) being the values extracted from parts of ISO 24617, tried out in different genres and text modalities and languages, grants our proposal more reliability and allows for interoperability.

Nonetheless, we still have some work to do. First, we will stabilize the taxonomy by adding more discourse relations to account for pertinent distinctions of meaning, by applying the taxonomy to a larger dataset both composed of monologues and dialogues and by defining a smaller set of relevant communicative functions taking into consideration their occurrence on the corpora. Then we will proceed to large-scale annotation, which means the annotation of the complete corpus using inter-annotator agreement. Finally, we will develop an empirical-based multilingual lexicon of discourse markers to be used as LLOD.

Acknowledgements

This research has been funded by the NexusLinguarum COST Action CA18209 - European network for Web-centred linguistic data science.

References

- a. ISO 24617-2. 2020. language resource management-semantic annotation framework (SemAF) - part 2 - dialogue acts. Standard, Geneva, CH.
 - b. ISO 24617-8. 2016. language resource management, part 8: Semantic relations in discourse (DR-Core). Standard, Geneva, CH.
- Nicholas Asher, Nicholas Michael Asher, and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.

- Farah Benamara and Maite Taboada. 2015. [Mapping different rhetorical relation annotations: A proposal](#). In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 147–152, Denver, Colorado. Association for Computational Linguistics.
- Harry Bunt. 2019. Plug-ins for content annotation of dialogue acts.
- Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis, and David Traum. 2012. [ISO 24617-2: A semantically-based standard for dialogue annotation](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 430–437, Istanbul, Turkey. European Language Resources Association (ELRA).
- Harry Bunt, Volha Petukhova, Emer Gilmartin, Catherine Pelachaud, Alex Fang, Simon Keizer, and Laurent Prevot. 2020. The ISO standard for dialogue act annotation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 549–558.
- Harry Bunt and Rashmi Prasad. 2016. ISO DR-Core (ISO 24617-8): Core concepts for the annotation of discourse relations. In *Proceedings 12th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-12)*, pages 45–54.
- Ludivine Crible. 2014. Identifying and describing discourse markers in spoken corpora. annotation protocol v.8. Technical report.
- Ludivine Crible and Liesbeth Degand. 2019. [Reliability vs. granularity in discourse annotation: What is the trade-off?](#) *Corpus Linguistics and Linguistic Theory*, 15(1):71–99.
- Ludivine Crible and Sandrine Zufferey. 2015. Using a unified taxonomy to annotate discourse markers in speech and writing.
- Debopam Das. 2014. *Signalling of coherence relations in discourse*. Ph.D. thesis, Arts & Social Sciences: Department of Linguistics.
- Bruce Fraser. 1996. Pragmatic markers. *Pragmatics*, 6:167–190.
- Bruce Fraser. 2009. An account of discourse markers. *International Review of Pragmatics*, 1(2):293–320.
- Montserrat González. 2005. [Pragmatic markers and discourse coherence relations in english and catalan oral narrative](#). *Discourse Studies*, 7(1):53–86.
- Eduard Hovy. 1995. The multifunctionality of discourse markers. In *Proceedings of the Workshop on Discourse Markers*. Egmond-aan-Zee, The Netherlands.
- Alistair Knott and Robert Dale. 1993. Using linguistic phenomena to motivate a set of rhetorical relations. *Human Communication Research Centre, University of Edinburgh*.
- William Mann and Sandra Thompson. 1988. [Rhetorical structure theory: Toward a functional theory of text organization](#). *Text*, 8:243–281.
- Yael Maschler and Deborah Schiffrin. 2015. Discourse markers: Language, meaning, and context. *The handbook of discourse analysis*, 1:189–221.
- Amália Mendes, Iria del Río Gayo, Manfred Stede, and Felix Dombek. 2018. A lexicon of discourse markers for portuguese - ldm-pt. In *LREC*.
- Volha Petukhova and Harry Bunt. 2009. [Towards a multidimensional semantics of discourse markers in spoken dialogue](#). In *Proceedings of the Eight International Conference on Computational Semantics*, pages 157–168, Tilburg, The Netherlands. Association for Computational Linguistics.
- R. Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind K. Joshi, Livio Robaldo, and Bonnie Lynn Webber. 2007. The penn discourse treebank 2.0 annotation manual.
- Rashmi Prasad and Harry Bunt. 2015. [Semantic relations in discourse: The current state of ISO 24617-8](#). In *Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-11)*, London, UK. Association for Computational Linguistics.
- Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. [The Penn Discourse TreeBank 2.0](#). In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).
- Charlotte Roze, Danlos Laurence, and Philippe Muller. 2010. [LEXCONN: a French Lexicon of Discourse Connectives](#). In *MAD 2010 - 8th Workshop Multidisciplinary Approaches to Discourse*, Proceedings of the 8th Workshop Multidisciplinary Approaches to Discourse (MAD 2010), pages 114–125, Moissac, France.
- Ted Sanders, Wilbert Spooren, and Leo Noordman. 1992. [Toward a taxonomy of coherence relations](#). *Discourse Processes - DISCOURSE PROCESS*, 15:1–35.
- Ted J.M. Sanders, Vera Demberg, Jet Hoek, Merel C.J. Scholman, Fatemeh Torabi Asr, Sandrine Zufferey, and Jacqueline Evers-Vermeul. 2021. [Unifying dimensions in coherence relations: How various annotation frameworks are related](#). *Corpus Linguistics and Linguistic Theory*, 17(1):1–71.
- Deborah Schiffrin. 1987. *Discourse Markers*. Cambridge University Press, Cambridge.
- Purificação Silvano. 2010. *Temporal and rhetorical relations: the semantics of sentences with adverbial subordination in European Portuguese*. Ph.D. thesis.

Manfred Stede, Tatjana Scheffler, and Amália Mendes. 2019. Connective-lex: A web-based multilingual lexical resource for connectives. *Discours*.

Maite Taboada. 2006. Discourse markers as signals (or not) of rhetorical relations. *Journal of Pragmatics*, 38(4):567–592. Focus-on Issue: The Pragmatics of Discourse Management.

Deniz Zeyrek, Amália Mendes, and Murathan Kurfalı. 2018. Multilingual extension of PDTB-style annotation: The case of TED multilingual discourse bank. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Table 2: The annotation of discourse markers - illustration.

Discourse markers meaning	English DM	Portuguese DM	Bulgarian DM
Discourse relations ISO 24617-8			
Exemplification	for example, for instance	por exemplo	например
Elaboration	in particular, to sum up	em suma	особено, в частност
Synchrony	so far	até agora	до сега
Contrast	on the one hand	por um lado	от една страна
Concession	on the other hand	por outro lado	от друга страна
Conjunction	on the other hand	por outro lado	от друга страна
Restatement	in other words, I mean	por outras palavras, noutras palavras, isto é	с други думи
Cause	as a result	como resultado, como consequência	в резултат
Expansion	in fact, this is, that is, of course	de facto, na verdade, ou seja, claro	всъщност
Communicative functions and qualifiers ISO 24617-2			
CheckQuestion	you know		знаеш ли, знаете ли
Confirm	of course, in fact	claro, de facto, na verdade	разбира се
Opening	You know	sabem	знаеш ли, знаете ли
AlloPositive	you see		виждаш ли, видите ли
Certain	of course	claro	разбира се

Testing the Continuity Hypothesis: A decompositional approach

Debopam Das

Åbo Akademi University
Tehtaankatu 2
20500 Turku, Finland
debopam.das@abo.fi

Markus Egg

Humboldt-Universität zu Berlin
Unter den Linden 6
10099 Berlin, Germany
markus.egg@hu-berlin.de

Abstract

The Continuity Hypothesis (CH) predicts that discontinuous discourse relations are harder to process and therefore more marked than continuous ones. To investigate this hypothesis, we annotated a corpus of discourse relations for Givón's (1993) seven continuity dimensions and also for discourse signalling, widening the perspective to discourse signals in general. Our results show that discourse relations often are simultaneously continuous and discontinuous on different continuity dimensions, and that continuity dimensions behave very differently with respect to discourse marking: Only the temporal dimension (partially) confirms the CH while the perspective dimension provides counter-evidence to the CH. Also, contrary to Givón's expectation, local discontinuity introduces more marking than global discontinuity.

1 Introduction

The signalling of discourse relations varies in kind and degree (Das, 2014; Crible, 2020). Different relation types employ different kinds of signalling; e.g., in English, CONDITION relations are mostly signalled by subordinating conjunctions like *if* or *when*, while PURPOSE relations are predominantly marked by the syntactic signal *infinitival clause*. Also, some relations are more marked than others; e.g., CONCESSION relations in comparison to HYPOTHETICAL relations.

The variation in relation signalling is often explained in terms of the Continuity Hypothesis (CH) (Murray, 1997). The CH presumes that discourse comprehension is greatly shaped by expectation, i.e., language users, while processing a text, have default assumptions about the upcoming discourse segment¹. In particular, readers have a preference

¹Comprehension based on the notion of expectedness is also accounted for by the 'causality-by-default' hypothesis (Sanders, 2005) and the Uniform Information Density (UID) hypothesis (Frank and Jaeger, 2008). For an overview of these hypotheses, see Asr and Demberg (2012).

for interpreting sequences of sentences in a continuous manner. Continuity ensues when the sentences maintain deictic dimensions such as time, reference, or perspective. Discontinuity, in contrast, arises when inter-sentential transitions are marked by deictic shifts along these dimensions. The CH predicts that discontinuous transitions between sentences are harder to process than continuous ones, and such transitions are therefore explicated more often in terms of suitable markers than continuous ones; e.g., the CONCESSION relations in (1) and (2) both convey discontinuity, but (1) is easier to understand than (2) due to the connective *even though* (examples from Zufferey and Gygax 2016, p. 533).

- (1) Peter married Jane even though he didn't love her.
- (2) Peter married Jane. He didn't love her.

Evidence for the CH mainly comes from psycholinguistic studies. Segal et al. (1991) observe that readers, when given a task to identify the relation types between successive sentences, most often chose causal or additive relations instead of contrastive relations. Murray (1997) shows that signals of discontinuity (i.e., adversative connectives like *but*) have a greater impact on on-line processing than signals of continuity. Further support for the CH comes from corpus data: Asr and Demberg (2012) observe that discontinuous relations display more explicitness than continuous ones.

In this paper, we argue that discourse relations can be simultaneously continuous and discontinuous on different continuity dimensions (*time, reference, or perspective*). We accordingly examine the CH directly on those dimensions, rather than on relation types as being categorically continuous or discontinuous. Also, unlike previous studies, we focus not only on discourse connectives (DCs), but also on non-DC signals such as lexical relations (e.g., antonymy) and syntactic structures (e.g., parallel syntactic constructions). We examine a corpus

of about 1,000 relations from five major relation types (CAUSAL, CONDITIONAL, CONTRASTIVE, ELABORATION, and TEMPORAL) that we first annotate with respect to Givón’s (1993) seven continuity dimensions (*time, space, reference, action, perspective, modality, and speech act*). We then test the CH, examining the signalling of those relations for individual continuity dimensions.

This paper is structured as follows: Section 2 outlines previous work on continuity (dimensions) in discourse relations. In Section 3, we describe the methodology adapted for the CH analysis. Section 4 presents the results and discussion. We conclude the paper with an outlook on the future work.

2 Background

2.1 Continuity and discourse relations

Previous studies on the CH generally consider continuity as a binary feature, classifying discourse relations categorically as either continuous or discontinuous. For instance, Murray (1997) considers CAUSAL relations continuous, and Zufferey and Gygax (2016) regard CONTRASTIVE relations as discontinuous. Asr and Demberg (2012) group the PDTB relations (Prasad et al., 2008) like RESULT, INSTANTIATION, and LIST as continuous and relations like PRAGMATIC CONTRAST, CONTRA-EXPECTATION, or TEMPORAL relations as discontinuous, whereas they leave CONDITIONAL relations underspecified with respect to continuity.

However, corpus evidence shows that discourse relations can be continuous on some continuity dimensions but at the same time discontinuous on other dimensions. For instance, CAUSAL relations, generally deemed continuous, can simultaneously exhibit continuity for the temporal dimension, but discontinuity for the reference dimension, as in (3).

- (3) [As some securities mature and the proceeds are reinvested,] [the problems ought to ease.]

Similarly, CONTRAST relations, usually regarded as discontinuous, can show the same configuration (continuity for time, discontinuity for reference):

- (4) [The gasoline picture may improve this quarter,] [but chemicals are likely to remain weak.]

Having noted these incongruities, we first set out to re-examine the relationship between continuity and discourse relations. To do so, we adopted a fine-grained approach, decomposing continuity into dif-

ferent continuity dimensions, following Givón’s framework (1993), as outlined below.

2.2 Givón’s continuity dimensions

Givón defines continuity in terms of thematic coherence, which distinguishes seven continuity dimensions or ‘coherence strands’. Maintaining or shifting deictic centres on these dimensions between discourse segments determines the extent of thematic coherence (continuity) or disruption (discontinuity). The seven dimensions are *time, space, reference, action, perspective, modality, and speech act*. The first four are more concrete and local, the others, more abstract and global:

local	time
	space
	reference
	action
global	perspective
	modality
	speech act

Table 1: Givón’s coherence strands

The grouping of dimensions is based on effect; consider (5)–(6) from Givón (1993, p. 319, 321). In (5), a change in the temporal continuity across the two clauses causes a local break, but does not necessarily terminate a larger coherent sequence of clauses in the text. In contrast, a change in one of the global dimensions amounts to a stronger break, which can terminate such a sequence of clauses. There is such a break in (6), because it exhibits discontinuity in perspective between the two sentences (viewpoint of the author vs. the one of the protagonist).

- (5) She flew in at midnight and left the next day.
 (6) She came in and sat on the bed. She was tired, she thought.

2.3 Operationalisation of dimensions

We operationalised Givón’s seven continuity dimensions in terms of distinctive features. As an example, consider the operationalisation of the *perspective* dimension². We distinguish three types of perspective (Pander Maat, 1998): *objective*, *author* (in the form of comments), and *other* (quotations). We consider a discourse relation continuous on the perspective dimension if its segments share the same perspective, as in (7), otherwise, as discontinuous, as in (8) (both are CONTRAST relations):

²The operationalisation of the seven dimensions is documented in detail in our previous work (Das and Egg, 2023).

- (7) [“Climate varies drastically due to natural causes,” said Mr. Thompson.] [But he said ice samples from Peru, Greenland and Antarctica all show substantial signs of warming.]
- (8) [“The earnings were fine and above expectations,” said Michael W. Blumstein, an analyst at First Boston Corp.] [Nevertheless, Salomon’s stock fell \$1.125 yesterday to close at \$23.25 a share in New York Stock Exchange composite trading.]

2.4 Continuity annotation on relations

In order to investigate how continuity interacts with discourse relations, we annotated over 1,000 tokens of discourse relations with respect to all seven continuity dimensions. The relations constitute a subset of the RST Discourse Treebank (Carlson et al., 2002), representing five major relation types: CAUSAL, CONTRASTIVE, CONDITIONAL, ELABORATION, and TEMPORAL. This selection is motivated by previous classifications, which categorise, e.g., CAUSAL and ELABORATION relations as continuous (Murray, 1997), CONTRASTIVE relations as discontinuous (Zufferey and Gyax, 2016), TEMPORAL relations as one or the other (Hopper, 1979), and CONDITIONAL relations as underspecified with respect to continuity (Asr and Demberg, 2012).

relation type	predicted continuity
CAUSAL	continuous
CONTRASTIVE	discontinuous
CONDITIONAL	not specified
ELABORATION	continuous
TEMPORAL	(dis)continuous

Table 2: Relation types and their features

We examined 1,009 relations with 276 CAUSAL, 156 CONTRASTIVE, 172 CONDITIONAL, 179 ELABORATION, and 226 TEMPORAL relations. Each relation was independently annotated by two annotators (the authors) for the seven continuity dimensions. We tested the inter-annotator agreement on 240 additional relations. Agreement was substantial according to Cohen’s kappa (Landis and Koch, 1977) for the four dimensions *time*, *reference*, *perspective*, and *modality*, as shown in Table 3. For the remaining dimensions, we also agreed, rather overwhelmingly, and no meaningful κ -values could be computed due to prevalence³.

³The agreement scores for these dimensions were 97.07% for *space*, 95.82% for *action*, and 98.74% for *speech act*.

time	reference	perspective	modality
0.72	0.69	0.70	0.76

Table 3: Inter-annotator agreement on four dimensions

3 Testing CH on continuity dimensions

3.1 Results on continuity and relations

The results from our corpus analysis, as summarised in Table 4⁴, show that continuity dimensions interact with discourse relations in varying ways. In particular, some continuity dimensions show uniformity across relation types. All relation types are found to be overwhelmingly continuous (> 98%) for the dimensions *space* and *speech act*, and almost never continuous (< 2%) for *action*. In contrast, the dimensions *time*, *reference*, *perspective*, and *modality* yield considerable differences amongst the relation types. For these dimensions, the types are not homogeneously continuous or discontinuous, but they can be simultaneously more continuous for some dimensions but less continuous or even predominantly discontinuous for other dimensions. For example, CONTRASTIVE relations are the least continuous for *reference* and *perspective*, but highly continuous for *time*. Furthermore, continuity is not found to be uniform even for a single dimension of one of these relations; e.g., only 82.61% (and not 100%) of the CAUSAL relations are continuous for *time*.

We measured the significance of the results statistically with a chi-square test, for interdependence between relation types and continuity along a specific dimension. We found that continuity correlates with relation types very significantly for *time*, *perspective*, and *modality* ($p < 0.00001$). The correlation is significant for *reference* ($p < 0.05$) and *action* ($p < 0.001$), too; but for *action*, low counts (< 5) reduce the validity of the test. No significant correlation was found between relation types and *space* or *speech act*. These findings imply that continuity and discontinuity systematically coexist in relations on the *time*, *reference*, *perspective*, and *modality* dimensions; consequently, relations are not fully continuous or discontinuous, neither on the level of the entire relation nor for any of these particular dimensions.

Since every relation type exhibits continuity and discontinuity in different continuity dimensions simultaneously, it seems incongruous to test the CH on the level of relation types. Therefore, we test

⁴The highest/lowest scores for a dimension are in bold font.

relation type	time	reference	perspective	modality	space	action	speech act
CAUSAL	82.61%	30.79%	85.87%	80.79%	97.46%	2.54%	99.64%
CONDITIONAL	81.98%	35.47%	93.61%	61.63%	98.84%	5.81%	98.26%
CONTRASTIVE	91.67%	23.72%	67.31%	77.56%	98.08%	0.00%	100%
ELABORATION	93.85%	34.64%	78.21%	85.47%	100%	0.56%	99.44%
TEMPORAL	74.34%	38.50%	90.27%	92.92%	97.35%	0.88%	98.67%
mean	84.04%	32.90%	83.94%	80.57%	98.23%	1.98%	99.21%

Table 4: Continuity scores across relation types

the validity of the CH on the level of individual continuity dimensions, that is, we examine the signalling of a relation type when it is continuous for a particular dimension as opposed to when it is discontinuous for that dimension. In our analysis, we focus only on the four dimensions, *time*, *reference*, *perspective*, and *modality*, which were distinctive for continuity and discontinuity on relations.⁵

We use the RST Signalling Corpus (RST-SC, Das et al., 2015) to examine the signals of the relations chosen for our continuity analysis. The RST-SC provides the signalling information for the discourse relations in the RST Discourse Treebank (Carlson et al., 2002), where our 1,009 relations come from. The relational signals in the RST-SC include different textual devices such as reference, lexical, syntactic, semantic, and graphical features, in addition to discourse connectives (DCs). Example (9) illustrates an RST-SC signalling annotation:

- (9) [Since Mexican President Carlos Salinas de Gortari took office last December,] [special agents have arrested more than 6,000 federal employees on charges ranging from extortion to tax evasion.]

The CIRCUMSTANCE relation is marked by the connective *since* as well as by the change of tense between two clauses (from simple past to present perfect), and also by the indicative phrase *last December*. We examine both DCs and all other signals in our examination of the CH.

4 Results and discussion

We gauge the impact of the four distinctive continuity dimensions (*time*, *reference*, *perspective*, and *modality*) on signalling in three ways. First, we compare the signalling of continuous and discontinuous tokens for each relation type for every continuity dimension. I.e., we examine how frequently a relation type is signalled (by a DC or/and by a non-DC signal) when it is continuous and

⁵For *space*, *action*, and *speech act*, relation types are found to be either almost continuous or discontinuous as a whole.

when it is discontinuous for a particular continuity dimension. The results are summarised in Table 5.

The data show that, along the *time* dimension, relation types on average and a majority of the individual subtypes (except CONTRASTIVE and ELABORATION) are marked more frequently in the absence of temporal continuity than in its presence. For *reference*, the average signalling scores do not vary much between the continuous and discontinuous relations (89.76% vs. 90.39%); still, marking in the absence of referential continuity is higher for CAUSAL and ELABORATION relations but lower for CONTRASTIVE relations. These results are not statistically significant, however.⁶

A different picture emerges for *perspective* and *modality*: Relations, when discontinuous on these dimensions, are less marked on average than the continuous ones (92.09% vs. 80.25% and 90.99% vs. 86.87%), and so are most individual relation subtypes (except CONTRASTIVE for perspective and TEMPORAL for modality continuity). In particular, the results for perspective continuity (except for CONTRASTIVE relations) provide counter-evidence against the CH. The numbers are significant here for the average ($p < .0001$) as well as for CAUSAL and CONDITIONAL relations ($p < .01$ and $p < .0001$, respectively).

We also conducted a similar analysis for DCs only, following the spirit of previous work on the CH. The results (in Table 6) for the overall distribution of the DC-only signalling were in line with the previous analysis on general signalling (in Table 5): Again, discontinuous relations tend to be more marked for *time*, but this time the positive evidence of the temporal dimension for the CH was more pronounced (significant for the average at $p < .0001$ and for CAUSAL and CONDITIONAL relations at $p < .05$ and $p < .0001$). The *reference* dimension once again does not offer evidence

⁶Lack of significance in Table 5 sometimes results from data sparsity (e.g., there is only one referentially continuous unsignalled CONTRASTIVE relation or only two relations for CONDITIONAL and ELABORATION that are temporally discontinuous and unsignalled).

relation type	time		reference		perspective		modality	
	cont	discont	cont	discont	cont	discont	cont	discont
CAUSAL	89.04%	89.58%	85.88%	90.58%	91.56%	74.36%	89.19%	88.89%
CONDITIONAL	85.11%	93.55%	87.30%	86.24%	91.30%	18.18%	89.62%	81.82%
CONTRASTIVE	90.14%	85.71%	97.37%	87.29%	89.52%	90.19%	90.91%	85.71%
ELABORATION	95.21%	83.33%	91.53%	95.83%	96.43%	87.18%	95.39%	88.89%
TEMPORAL	88.69%	98.28%	90.80%	91.37%	91.67%	86.36%	90.48%	100%
mean	89.71%	92.64%	89.76%	90.39%	92.09%	80.25%	90.99%	86.87%

Table 5: Distribution of marked relations for continuity dimensions

relation type	time		reference		perspective		modality	
	cont	discont	cont	discont	cont	discont	cont	discont
CAUSAL	49.56%	68.75%	57.65%	50.78%	55.69%	35.89%	50.90%	61.11%
CONDITIONAL	78.01%	87.50%	80.95%	79.82%	84.47%	18.18%	83.02%	75.76%
CONTRASTIVE	80.28%	78.57%	89.47%	77.12%	81.90%	76.47%	81.82%	74.29%
ELABORATION	8.38%	8.33%	5.08%	10.00%	10.00%	2.56%	7.89%	11.11%
TEMPORAL	70.24%	77.59%	70.11%	73.38%	74.02%	54.55%	71.43%	81.25%
mean	55.44%	72.39%	59.64%	57.46%	61.28%	41.98%	56.97%	63.13%

Table 6: Distribution of relations with DCs for continuity dimensions

relation type	discontinuous for	
	local	global
CAUSAL	94.44%	83.33%
CONDITIONAL	90.48%	22.22%
CONTRASTIVE	80.00%	87.50%
ELABORATION	90.00%	81.81%
TEMPORAL	97.62%	100%
mean	93.28%	75.00%

Table 7: Signalling for local and global discontinuity

for or against the CH, and the *perspective* dimension clearly goes against the predictions of the CH (significant for the average at $p < .0001$ and for CAUSAL and CONDITIONAL relations at $p < .05$ and $p < .0001$). For *modality*, unlike what we found for general signalling (Table 5), discontinuous relations are marked more frequently by DCs than continuous relations.

Next, we compared relations that are discontinuous on the local dimensions (*time* and *reference*) to those discontinuous on the global dimensions (*perspective* and *modality*). The results (in Table 7) indicate that the first group on average shows more marking than the second one. As a break in global coherence has more impact in Givón's theory, one would have expected a higher need for signalling for the second group, i.e., the reverse result.

As a third measure for the impact of continuity on marking, we attempted to gauge the effect of continuity in general (i.e., irrespective of a particular dimension) on marking. To this end, we examined the distributions of signalled and unsignalled relations for relations that are continuous on 0-4 of the four relevant dimensions. The results (in Table 8) show that, contrary to what one would expect in

the light of the CH, more continuous dimensions actually lead to an increase in marking.

We then compared the distributions of marked and unmarked signals across the five groups in terms of relative entropy $S(q, p)$ (also known as Kullback-Leibler divergence), where both p and q are distributions over signalled relations which differ in the number of continuity dimensions. In our case, $S(q, p)$ measures the influence of an additional continuous dimension on the distribution of signalled signals.

dimensions	0 vs. 1	1 vs. 2	2 vs. 3	3 vs. 4
entropy	.01285	.00210	.00001	.00005

Table 9: Relative entropy and continuous dimensions

As shown in Table 9, the impact of additional continuous dimension tends to be greater for smaller numbers of dimensions. This result once again suggests that the degree of continuity for a relation is correlated positively with discourse marking, because it can be interpreted in terms of diminishing marginal utility, e.g., the difference in marking between relations with three and four continuous dimensions is smaller than the one between relations with one and two.

5 Conclusions and outlook

We have argued that continuity functions as a multi-dimensional phenomenon in discourse relations. We have supported the claim by validating a decompositional approach of annotating relations with respect to different continuity dimensions. We have applied this decompositional approach for testing

relation type	zero dim.	one dim.	two dim.	three dim.	four dim.
CAUSAL	0%	82.35%	88.10%	90.24%	90.38%
CONDITIONAL	100%	78.57%	83.02%	90.14%	90.91%
CONTRASTIVE	66.67%	84.21%	88.89%	88.41%	100%
ELABORATION	100%	80.00%	94.59%	95.92%	94.44%
TEMPORAL	66.67%	90.91%	98.31%	87.91%	89.06%
mean	66.67%	83.33%	90.28%	90.71%	91.79%

Table 8: Scores for marked relations for different numbers of continuous dimensions

the Continuity Hypothesis for all relational signals including discourse connectives.

The results from our corpus provided no conclusive evidence for or against the CH on the level of individual continuity dimensions: Temporal continuity is found to (partially) corroborate the CH while continuity along perspective contradicts it. Furthermore, contrary to Givón's line of reasoning, global discontinuity is found to decrease the amount of discourse marking. Finally, continuity, when the specificity of its dimensions is not taken into account, correlates with discourse signalling positively, hence going counter to the CH.

We would, however, like to point out that our results on continuity and the CH are based on the newspaper genre of the corpus (RST-DT). Continuity might function differently in other genres, e.g., fiction (as in Givón's framework), and also across languages, as shown by Mendes et al. (2023).

In future work, we will incorporate more data (in terms of additional relation types and also corpus size) in the evaluation of the CH. We will also investigate whether relation types and their marking are differently susceptible to the impact of continuity. Furthermore, our results motivate searching for other potential factors for the data to explain why they do not fit in with the predictions of the CH.

References

- Fatemeh Torabi Asr and Vera Demberg. 2012. Implicitness of Discourse Relations. In *Proceedings of COLING 2012*, page 2669–2684.
- Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2002. *RST Discourse Treebank, LDC2002T07*.
- Ludivine Crible. 2020. Weak and strong discourse markers in speech, chat and writing: Do signals compensate for ambiguity in explicit relations? *Discourse Processes*, 57:793–807.
- Debopam Das. 2014. *Signalling of Coherence Relations in Discourse*. PhD dissertation, Simon Fraser University, Canada.
- Debopam Das and Markus Egg. 2023. Continuity in discourse relations. *Functions of Language*, 30:41–66.
- Debopam Das, Maite Taboada, and Paul McFetridge. 2015. *RST Signalling Corpus, LDC2015T10*.
- Austin Frank and Florian Jaeger. 2008. Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*, page 933–938.
- Talmy Givón. 1993. *English Grammar: A function-based introduction*, volume 2. John Benjamins.
- Paul Hopper. 1979. Aspect and foregrounding in discourse. In Talmy Givón, editor, *Syntax and semantics, Vol. 12. Discourse and syntax*, page 213–241. Academic Press, New York.
- Richard Landis and Gary Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.
- Amália Mendes, Deniz Zeyrek, and Giedrė Oleškevičienė. 2023. Explicitness and implicitness of discourse relations in a multilingual discourse bank. *Functions of Language*, 30(1):67–91.
- John Murray. 1997. Connectives and narrative text: The role of continuity. *Memory and Cognition*, 25:227–236.
- Henk Pander Maat. 1998. Classifying negative coherence relations on the basis of linguistic evidence. *Journal of Pragmatics*, 30:177–204.
- Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltasakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. In *Proceedings of LREC 2008*, pages 2961–2968.
- Ted Sanders. 2005. Coherence, causality and cognitive complexity in discourse. In *Proceedings/Actes SEM-05, First International Symposium on the Exploration and Modelling of Meaning*, page 105–114.
- Erwin Segal, Judith Duchan, and Paula Scott. 1991. The role of interclausal connectives in narrative structuring. *Discourse Processes*, 14:27–54.
- Sandrine Zufferey and Pascal Gygax. 2016. The role of perspective shifts for processing and translating discourse relations. *Discourse Processes*, 53:532–555.

Lexical Retrieval Hypothesis in Multimodal Context

Po-Ya Angela Wang, Pin-Er Chen, Hsin-Yu Chou, Yu-Hsiang Tseng, Shu-Kai Hsieh

Graduate Institute of Linguistics, National Taiwan University
 differe94nt@gmail.com, cckk2913@gmail.com,
 r10142008@ntu.edu.tw,
 seantyh@gmail.com, shukaihsieh@ntu.edu.tw

Abstract

Multimodal corpora have become an essential language resource for language science and grounded natural language processing (NLP) systems due to the growing need to understand and interpret human communication across various channels. This paper presents our efforts in building the first Multimodal Corpus for Languages in Taiwan (MultiMoco). Based on the corpus, we conduct a case study investigating the Lexical Retrieval Hypothesis (LRH), specifically examining whether the hand gestures co-occurring with *speech constants* facilitate lexical retrieval or serve other discourse functions. With detailed annotations on eight parliamentary interpellations in Taiwan Mandarin, we explore the co-occurrence between *speech constants* and non-verbal features (i.e., *head movement*, *facial movement*, *hand gesture*, and *function of hand gesture*). Our findings suggest that while hand gestures do serve as facilitators for lexical retrieval in some cases, they also serve the purpose of information emphasis. This study highlights the potential of the MultiMoco Corpus to provide an important resource for in-depth analysis and further research in multimodal communication studies.

1 Introduction

Over the past decades, there has been a growing interest in multimodal corpus linguistic research (Paquot and Gries, 2021), which focuses on the analysis and comprehension of information from diverse modalities, including speech, image, and gesture. To facilitate research in this field and other interdisciplinary studies, the creation of multimodal corpora, or collections of data from various modalities, has become more crucial.

We thereby introduce the Multimodal Corpus for Languages in Taiwan (the MultiMoco Corpus), a newly released multimodal corpus that includes audio, video, gestural, and textual data involving various languages and discourse contexts. The

MultiMoco Corpus is comprised of recordings of realistic interactions taken in news and interpellation in parliament, where interviews and spontaneous speech take place. The synchronization of the audio, video clips, and gesture segments enables researchers to study the link between the communication modes. These data assist researchers in annotating information on the speakers, their actions, and the communication contexts. This corpus is designed for human communication and interaction-related research, such as conversation analysis, multimodal machine learning, and natural language processing.

To demonstrate the feasibility of the MultiMoco Corpus, we conduct a case study based on the parliamentary interpellation clips in Taiwan Mandarin, aiming to validate the widely discussed Lexical Retrieval Hypothesis (hereafter, LRH) (Dittmann and Llewellyn, 1969; Ekman and Friesen, 1972; Butterworth and Beattie, 1978; Rauscher et al., 1996), which suggests that gesture and verbal disfluency tend to co-occur in spontaneous speech.

More specifically, we take *speech constants*, based on the framework of Voghera (2001), as indicators of potential verbal disfluency. We annotate one verbal feature (*speech constants*) as well as four non-verbal features, including three forms of non-verbal expressions (*head movement*, *face movement*, *hand gesture*) and *functions of hand gesture*. With careful annotation, we attempt to answer research questions as follows: (1) Could we observe co-occurrences between *speech constants* and gestures in the context of interpellation? (2) If there are co-occurrences with *speech constants*, do hand gestures mainly play the role of priming lexical items? And (3) Do the hand gestures serve other functions regarding interlocutors and the entire discourse context?

To provide guidance on utilizing the MultiMoco Corpus to address multimodal research problems, we first review studies on the multimodal corpus,

the multimodal annotation framework, and the LRH (Section 2). Following this, we outline the data collection and annotation framework for the case study in Section 3.2 and Section 3.3. Next, we analyze if the non-verbal features co-occur with *speech constants* (Section 4.1). The LRH mechanism is examined by identifying the co-occurrences between *speech constants* and LRH-related/ non-LRH-related functions of hand gesture (Section 4.2), along with the individual performances discussed in Section 4.3. Section 5 concludes the paper.

2 Related Works

2.1 Multimodal corpus

Communication, by nature, is multimodal (Carter and Adolphs, 2008), and thereby constructing multimodal corpora affords researchers the opportunity to get a comprehensive understanding of the cognitive mechanisms underlying communication. "Multimodal corpus" can be defined at varying degrees depending on its architecture (Allwood, 2008). Generally speaking, it refers to an online repository of language and communication-related content that contains several modalities. In a narrower sense, it can be specified with audiovisual materials accompanied by annotations and transcriptions.

Most earlier multimodal corpora are for specific purposes. For example, the Mission Survival Corpus (McCowan et al., 2003), the Multimodal Meeting (MM4) Corpus (McCowan et al., 2005), and the VACE corpus (Chen et al., 2006) are all built on conversations in meeting. Others are task-oriented corpora elicited in lab settings, such as the Fruit Carts corpus (Gallo et al., 2006), Culture-adaptive BEhavior Generation for interactions with embodied conversational agents (CUBE-G) (Rehm et al., 2009), and the spatial task-based dialogue corpus, SaGA (Lücking et al., 2010). Still, others include dyadic conversation in academic discourse: the Nottingham Multi-Modal Corpus (NMMC) (Knight et al., 2008) and the Pisa Audiovisual Corpus project (Camiciottoli and Bonsignori, 2015), providing domain-specific multimedia materials for English for Specific Purposes (ESP) learners in higher education.

Recent corpora attempt to be less specific and purpose-oriented. Mlakar et al. (2017) select 4 recordings of multiparty conversation in a talk show, with more spontaneous discourses and more

topics. The NTHU-NTUA Chinese interactive multimodal emotion corpus (NNIME) (Chou et al., 2017) constructed a dataset with 44 subjects majoring in drama to record performed scenes for affective behaviors. In addition, the Communicative Alignment of Brain and Behaviour (CABB) (Eijk et al., 2022) builds a dataset on recordings of 71 pairs of participants discussing innovative, unconventional objects¹ (Barry et al., 2014), which provides pre-and-post behavioral and fMRI measurement information. Nevertheless, these corpora have their limitations. Certain datasets are built on less amount of data, some are restricted to conversations revolving around narrow topics, and others are collected for particular experiments.

The MultiMoco Corpus presented in this study incorporates video and audio recordings from ten public news channels and interpellation videos, which encompass a broader spectrum of languages and communication genres.² This renders it a more balanced resource for investigating multilingual and multimodal communication in everyday conversations, with the capacity to accommodate multidimensional annotations.

2.2 Multimodal annotation framework

Various annotation frameworks have been proposed to encode labels for gesture forms and corresponding functions (Bavelas et al., 1992; McClave, 2000; Kendon, 2004; Müller, 2004; Allwood et al., 2005; Bressemer et al., 2013). According to Debras (2021)'s proposal, "articulator" (e.g., hand or head), and "configuration of articulator" (e.g., head nod, wave, or turn) should be formally annotated. Functional annotation is to indicate co-verbal intentions of gestures. The Facial Action Coding System (FACS; Ekman and Rosenberg, 1997; Clark et al., 2020), for facial expression annotations, and the Linguistic Annotation System for Gestures (LASG; Bressemer et al., 2013), for hand annotations, are both well-designed but complicated annotation systems. Annotation frameworks such as these can be time-consuming and challenging to achieve annotation agreement. Debras (2021) suggests that coarse-grained annotations can benefit the onset of the research.

We here review the annotation frameworks that will be adopted in the case study. Firstly, *speech constants* will be annotated to examine the LRH

¹"Fribbles"

²The collection and characteristics of MultiMoco Corpus data are described in Section 3.1.

evaluated by Trotta and Guarasci (2021), given that gestures tend to co-occur with verbal disfluency. Referring to the guidelines in Voghera (2001), four types of *speech constants* (i.e., *pause*, *repetition*, *truncation*, and *semi-lexical*) are taken as the annotation targets. Secondly, the non-verbal target features comprise forms and functions, namely *head movement*, *face movement* (*eyebrows and mouth*), *hand gesture*, and *functions of hand gesture*. Considering Debras (2021)'s suggestions for coarse-grained annotations, this study follows the concise annotation framework adopted by Camiciottoli and Bonsignori (2015), incorporating gesture form abbreviations by Julián (2011) and the gesture functions by Kendon (2004) and Weinberg et al. (2013). In Camiciottoli and Bonsignori (2015)'s framework, *head movement* include *head-nodding/tilting/jerking/moving* together with multiple directions and repetition; *face movement* involve the movement of eyebrows and mouth; *hand gesture* mark the movements of fingers, palm, and the whole hand. The comprehensive labels and definitions for each feature will be explained in Section 3.3.

2.3 Lexical Retrieval Hypothesis

As reviewed in Özer and Göksun (2020), multi-modal interaction in speech production and comprehension regarding individuals' cognitive tendencies has been heatedly discussed. When a speaker cannot clarify intended thoughts, gestures are incorporated during hesitation pauses or the lexical pre-planning stage (Dittmann and Llewellyn, 1969; Butterworth and Beattie, 1978). The link between verbal, non-verbal, and conceptual aspects can be addressed by the "growth point," the smallest thought unit, comprising both utterances and gestures (McNeill, 1992). Krauss (1998) has considered the relationship between thoughts, utterances, and gestures from another perspective, specifying three parts in speech production: conceptualizing, grammatical encoding, and phonological encoding. Among these three parts, phonological encoding, the retrieval of lexical form, is the part where gestures affect the verbal modality, and limited gestures reduce speech fluency when a speaker discusses spatial information (Krauss, 1998). Later, Krauss and Hadar (1999) have further proposed that concepts in the mind are stored in various forms, so activating one idea in one modality may also activate concepts in other modalities. Thus,

concepts can be fully comprehended when information from different modalities is all presented, and representations from one modality can be converted into another modality. Following the line of this discussion, the gestural modality can assist lexical retrieval in the verbal modality because of such cross-modal priming. This is termed the "Lexical Retrieval Hypothesis" (Gillespie et al., 2014; Trotta and Guarasci, 2021). Namely, LRH refers to the process that the triggered idea's lexical gestures³ (i.e., gestures that can iconically represent meanings) can semantically prime the phonological encoding of the related words, reviewed in Gillespie et al. (2014). Gillespie et al. (2014) also specify that LRH is less applicable if the speaker can resort to alternative tactics to avoid lexical access challenges, which occur in improvisational speech production.

The Lexical Retrieval Hypothesis is tested in several tasks and contexts. Hostetter and Alibali (2007) distinguish the phonemic fluency from the semantic fluency⁴, suggesting lexical access efficiency may be related to different types of gestures. Additionally, Smithson and Nicoladis (2013) have proposed that the negative association between verbal working memory and iconic gesture production in bilinguals designates gesture production's assistance in the retention and utilization of language information. Trotta and Guarasci (2021) calculate the weighted mutual information (WMI) between the hand movements and the concurrent speech disfluency features involving five kinds of *speech constants*⁵. The result concurs with the LRH since hand gestures are more related to semi-lexical features and pauses in interview contexts. It is noted that in Trotta and Guarasci (2021), *speech constants* are considered disfluency features to assess the LRH, whereas hesitation pauses may signal lexical retrieval difficulties.

As most of the studies mentioned have examined the LRH with laboratory tasks or free-form inter-

³Krauss (1998) refers to these lexical retrieval supporting gestures as "lexical gestures."

⁴As defined in Hostetter and Alibali (2007), movements that transmit information relevant to the content of the vocal communication are representational gestures. Beat gestures are short, rhythmic motions that accentuate terms without demonstrating what they mean. "Phonemic fluency" indicates thought-organizing skills associated with representational gesture rates, whereas "semantic fluency" is less correlated with representational gesture rates but has a significant correlation with beat gestures.

⁵Five kinds of *speech constants*: pause, repetition, truncation, and semi-lexical, as specified by Voghera (2001)

views, we aim to assess the LRH in formal speaking contexts (i.e., political interpellation) as well as its applicability in less colloquial speech. Meanwhile, given that investigations in multiple modalities can provide us with more comprehensive perspectives on cross-modal interaction, we also aim to extend the hypothesis testing scope by exploring how disfluency co-occurs with more gestures: face, head, and hand. Among them, different functions of hand gesture co-occurring with *speech constants* are investigated to ascertain whether or not gestures assist in lexical retrieval. This case study conjectures that gestures co-occurring with *speech constants* are not just for facilitating lexical retrieval.

3 Methodology

Our study of the lexical retrieval hypothesis is based on the multimodal data made available from MultiMoco. We first introduce the construction and contents of the MultiMoco Corpus (Section 3.1). Then, the data collection for our case study on the LRH is illustrated (Section 3.2), followed by the annotation framework for the target features (Section 3.3). The annotation results and analyses will be discussed in the subsequent sections.

3.1 MultiMoco Corpus

The MultiMoco Corpus is built on recorded videos and audios from 10 public television channels⁶ in Taiwan, including news in multiple languages (i.e., Taiwan Mandarin, Taiwan Southern Min, Hokkien, Hakka, and Formosan languages) and the interpellation of the Taiwan Legislative Yuan (the parliament of Taiwan). While the TV news is recorded by wireless television receivers, the interpellation video clips with transcriptions in Taiwan Mandarin are retrieved directly from the Internet Multimedia Video-on-Demand System for Rebroadcasting Legislative Yuan Proceedings⁷.

Figure 1 displays the data processing workflow of the MultiMoco Corpus. With 223 video clips from Taiwan public television channels and the interpellation from Taiwan Legislative Yuan, the MultiMoco Corpus provides 5,854 minutes of dialogue, accompanied by 1,485,297 characters of captions transcribed via Whisper (Radford et al.,

⁶The target channels are as follows: CTV News PTS News, PTS Taigi, Hakka TV, Taiwan Indigenous TV, TTV News, CTS News, Congress Channel I, Congress Channel II, and FTV News.

⁷<https://ivod.ly.gov.tw/Demand>

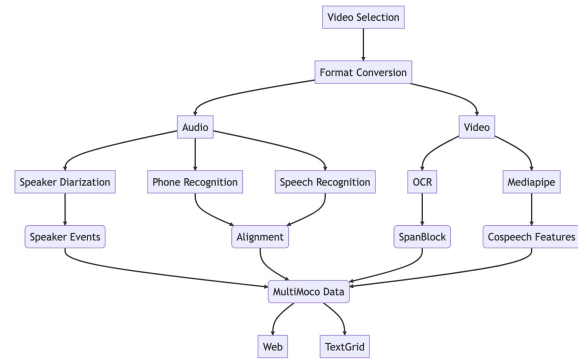


Figure 1: Establishment workflow of the MultiMoco Corpus

2022) model. In addition, 22,805 gestures identified via MediaPipe (Lugaresi et al., 2019) are also included in the corpus. The multimodal nature of the corpus allows researchers to conduct cross-modality analyses, thereby broadening the understanding of the communicative potential of various modalities beyond spoken texts. That is, the MultiMoco Corpus provides us with the potential to extend communication studies to diverse linguistic and multimodal contexts.

3.2 Data collection

Our lexical retrieval analysis data are extracted from MultiMoco Corpus, specifically focusing on spontaneous speech during interpellation involving interactions between legislators and officers. To control the gender, speech delivery performance, and speech topics of the selected data, we chose two biological females and two biological males, along with a balanced selection of speech topics. The interpellation topics are detailed in Table 1. As to speech delivery performance, we have selected interpellation clips based on the evaluation scores of 103 legislators from Citizen Congress Watch (CCW) in the 10th session of Congress⁸. we have selected interpellation clips based on the evaluation scores of 103 legislators from the Citizen Congress Watch (CCW) in the 10th session of Congress. Figure 2 shows the distribution of individually-averaged evaluation scores, with an average score of approximately 16, a minimum of 11.25, and a maximum of 17.998. After considering the evaluation score, interpellation topics, and

⁸Using the Legislative Yuan’s Internet multimedia Video on Demand System, civil jurors can evaluate the performance of parliamentarians in sessions and fill out questionnaires. Then, the evaluation score of each legislator is calculated through this procedure.

political parties, we choose four legislators (two with higher evaluation scores and two with lower evaluation scores) for subsequent multimodal analyses. In the end, we collect eight interpellation clips, each lasting between 8 and 12 minutes and featuring a male and a female legislator in each pair.

Legislator	Topic of Interpellation Clips	
high_A	Social welfare	Education and culture
high_B	Finance	Communications
low_C	Finance	Judiciary and organic laws
low_D	Social welfare	Education and culture

Table 1: Topics of the interpellation clips. The prefixes (*high* or *low*) in the Legislator column are used for identifying the evaluation scores for the legislators (i.e., A, B, C, and D).

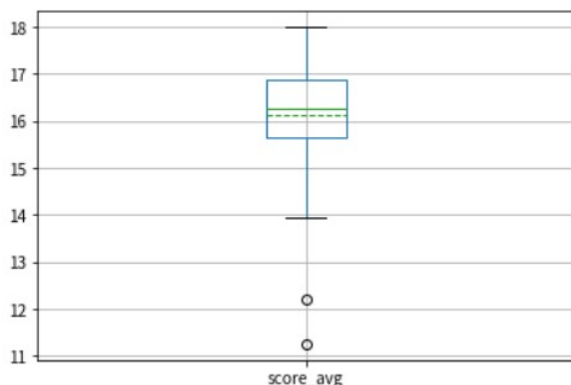


Figure 2: Descriptive statistics of citizen evaluation score

3.3 Data annotation

We investigate the functions of non-verbal features and their co-occurrence with disfluency in spontaneous speech. Three non-verbal forms (i.e., *head movement*, *face movement*, and *hand gesture*), one non-verbal function (i.e., *functions of hand gesture*), and one verbal feature (i.e., *speech constants*) are selected as our annotation targets; the latter is used to identify disfluency in speech.

Considering the specificity of each feature and the consensus in prior studies, we adopt different annotation frameworks for corresponding features. The *speech constants* are annotated based on the framework in Voghera (2001), as shown in Table 2;

Label	Definition
Pause	This marks a pause either between or within utterances.
Non-lexical item	This marks interjections (e.g., eh and ehm), or more general words that convey the meaning of an entire sentence, constituting a complete linguistic act demonstrated by their paraphrasability.
Repetition	This marks cases of repetition of utterances in order to give coherence and cohesion to the speech or self-repetition as a control mechanism of the speech programming.
Truncation	This indicates the deletion of a phoneme or a syllable in the final part of a word.

Table 2: Labels for speech constants. It is noted that the original label “semi-lexical” in Trotta and Guarasci (2021) is renamed “non-lexical item” in our study.

functions of hand gesture were annotated via Camiciottoli and Bonsignori’s framework, as presented in Table 3. The three non-verbal forms (i.e., *head movements*, *face movements*, and *hand gestures*) are classified based on Camiciottoli and Bonsignori (2015)’s framework, as illustrated in Table 4. It is noted that the labels in the table are generalized to a more coarse-grained scale regarding the entailment of the original labels.

Five native speakers annotate the five verbal and non-verbal features (i.e., *head movement*, *face movement*, *hand gesture*, *function of hand gesture*⁹, and *speech constants*) via ELAN (Sloetjes and Wittenburg, 2008)¹⁰, an open-source software appropriate for multimodal annotations and linguistic analysis. Take *speech constants* for instance, the two annotators separately mark the time periods and corresponding labels of *speech constants* that occur in all eight clips. Then, the annotated pair of tiers (made by the two annotators) for each clip are segmented into units of 100 milliseconds and aligned with each other.

For annotation consistency, the annotators are asked to annotate different features from clip segments and decide on an agreed-upon criterion for disagreed annotations. For instance, the function, *Parsing*, marks situations in which a speaker intends to initiate a new discourse turn, recur the same gesture as if beating, or make some trivial

⁹For clarity, we use the *italic* form when referring to the five targets, and we use the `typewriter` font when referring to the labels under each target.

movements that have no clear reference. In terms of our Inner Annotator Agreement (IAA), we calculate the ratio of intersecting annotation segments and the agreement ratio of the intersecting segments to measure the agreement between the annotators. As shown in Table 5, *hand gesture* (.76) and *function of hand gesture* (.81) acquire a higher ratio of intersecting segments, in which the annotators are able to identify more overlapping time periods of hand movements. Conversely, the ratio of intersecting segments for the *head movement* (.26) and *face movement* (.37) is relatively low. We suggest that the lower number of intersecting segments may relate to the different scales of movements perceptualized by the annotators. Although we generalized certain categories of the labels, we found it hard to define the degree of the speakers' movements. While one annotator perceived and marked some subtle tilting periods, the other annotator may have missed the same units. The subjectivity in continuum segmentation poses a challenge for multimodal annotation, yet since the annotators have discussed their inconsistencies and reached a consensus, the annotation results of the subsequent discussion are reliable.

As we focus on the co-occurrence and association between non-verbal features and disfluency, we will not inspect the details of the annotation results within each non-verbal feature but rather discuss the general co-occurrence with *speech constants* in the following sections.

Label	Definition
Social	social (emphasizing a message)
Repres	representational (representing object/idea)
Index	indexical (indicating a referent)
Parsing	parsing (distinguishing units of speech)
Perform	performative (illustrating speech act)
Modal	modal (expressing certainty/uncertainty)

Table 3: Labels for functions for hand gesture.¹¹ The functions of 'beat' and 'representational' in Hostetter and Alibali (2007) are represented as `Parsing` and `Representational` in this study.

4 Results & Discussions

We first examine the non-verbal features' co-occurrence with *speech constants*, which indicate verbal disfluency (Section 4.1). Then, the potential

¹⁰ELAN (<https://archive.mpi.nl/tla/elan>); Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands.

¹¹The functions of hand gestures are mutually exclusive.

Type	Label
Face	Frowning eyebrows Raising eyebrows Smile Other
Head	Nod Jerk Move Forward/Backward Tilt Side-turn Shake (repeated) Other
Hand	Finger pointing towards audience Hands sweeping sideways Hands rotating at center of body Hands wide apart moving down Hands clasped together in front of body Other

Table 4: Labels for co-speech gestures: face, head, and hand.

Target	Ratio	Agreement Rate
Head	.26	.78
Face	.37	.99
Hand gesture	.76	.70
Function of hand gesture	.81	.41
Speech constant	.49	.89

Table 5: Inter-annotator agreement on five targets. "Ratio" refers to "Ratio of Intersecting Segments." Intersecting segments are those existing on both annotation tiers (of the two annotators) after aligned to the timeline of each clip. "Agreement Rate" refers to the "Agreement Rate on Labels of the Intersecting Segments."

discourse *functions of hand gesture* will be analyzed (Section 4.2). Finally, we will discuss more comprehensive gesture functions independent of verbal disfluency but related to interlocutors and the entire discourse context in Section 4.3.

4.1 Co-occurrence overview

As we target one verbal feature (*speech constants*) and three forms of non-verbal features (head, hand, and face)¹², we calculate the co-occurrences¹³ of the six patterns by modality. Figure 3 shows that *head movement* and *speech constants* co-occur most frequently, followed by *hand* and

¹²It should be noted that one non-verbal related feature, i.e., the *functions of hand gesture*, are annotated based on the occurrence of hand gesture; thus, calculating the co-occurrences (i.e., overlapping segments) between *functions of hand gesture* and the other features would be meaningless, as it would be the same as hand gesture.

¹³The co-occurrence of one pair of features is defined as the summed number of overlapping segments; one segment is a unit of 100 milliseconds.

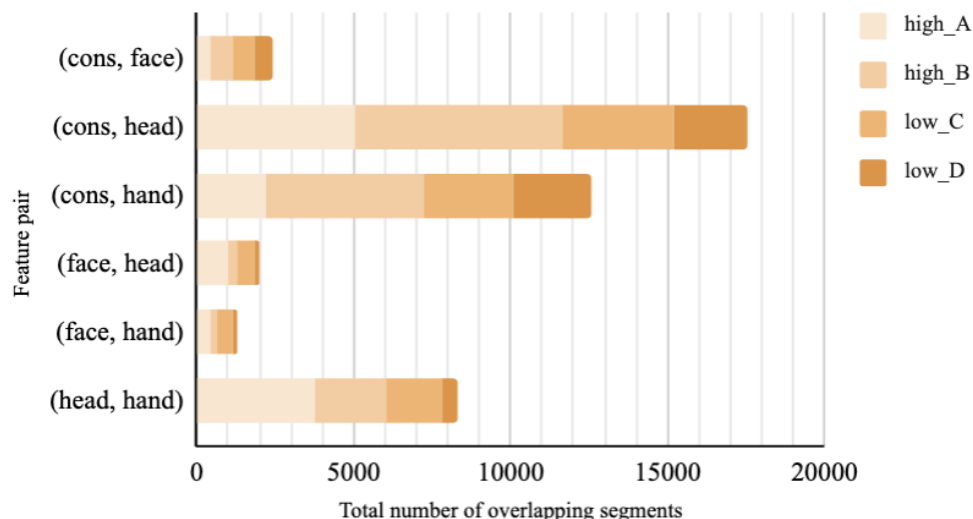


Figure 3: Co-occurrences of different feature pairs. The y-axis represent the number of overlapping segments between different pairs of annotated features.

speech constants. Face movement shows fewer co-occurrences with the other features (i.e., face & head, face & hand, and face & *speech constants*), which may relate to the few occurrences of face movement in all clips. In addition to mask-wearing situations, these few occurrences of facial movement are the result of the face movements being so frequent and inconsequential that the annotators reach an accord to only record the apparent ones, as some trivial ones may be the result of habitual movements. This annotation procedure illuminates considerations for future annotation frameworks. While the non-verbal features tend to co-occur with one another, the frequencies are far lower than their respective co-occurrence with *speech constants*. This may correspond to the LRH that when *speech constants* appear, i.e., during hesitation pauses or the lexical pre-planning stage, non-verbal gestures are possibly employed by the speaker as well (Dittmann and Llewellyn, 1969; Butterworth and Beattie, 1978). To sum up, the distribution illustrates that non-verbal characteristics are more likely to co-occur with disfluent situations than with other types of non-verbal movements. Furthermore, it demonstrates the significance of both the head and the hand in the research of verbal disfluency.

4.2 Co-occurring functions of hand gestures

As significant as the respective gesture co-occurrence with *speech constants* is, could we claim that the identified *speech constants* require gestures to facilitate lexical retrieval? To further

understand the purposes of the hand gestures co-occurring with *speech constants*, Table 6 below presents the overall frequencies of each type of *speech constants* co-occurring with different *functions of hand gesture*. *Speech constants*, especially non-lexical items and pauses, are taken as verbal disfluency traits in the LRH evaluation (Trotta and Guarasci, 2021). We would like to argue that the intentions of performing *speech constants* are various, so the functions resulting from the interplay between verbal and non-verbal modalities are complicated. Thus, in addition to using *speech constants* as markers of the possible presence of verbal disfluency, we study the functions of co-occurring hand gestures in order to realize whether the co-occurring hand gestures are lexical retrieval facilitators or carry out other functions in speech contexts.

First, we examine the distributions of *speech constants* and their co-occurring *functions of hand gesture*. Regarding *speech constants*, pause is the most frequently observed category with 345 frequencies, accounting for 72.2% co-occurrences among all. Repetition and non-lexical item both rank second. Truncation sporadically occurs in the collected dataset. As for *functions of hand gesture*, Social (i.e., to emphasize a message) is the most frequent function for the *speech constants* as a whole. The rest of the ranking goes as follows: Parsing > Indexical > Representational > Performative >

(SC / FH)	Indexical	Parsing	Performative	Representational	Social	Total
Non-lexical item	10	24	2	16	9	61
Pause	59	87	20	46	133	345
Repetition	6	16	0	7	32	61
Truncation	8	0	0	0	3	11
Total	83	127	22	69	177	478

Table 6: Contingency table of *speech constants* and *functions of hand gesture*. SC represents *speech constants*, and FH represents *functions of hand gesture*.

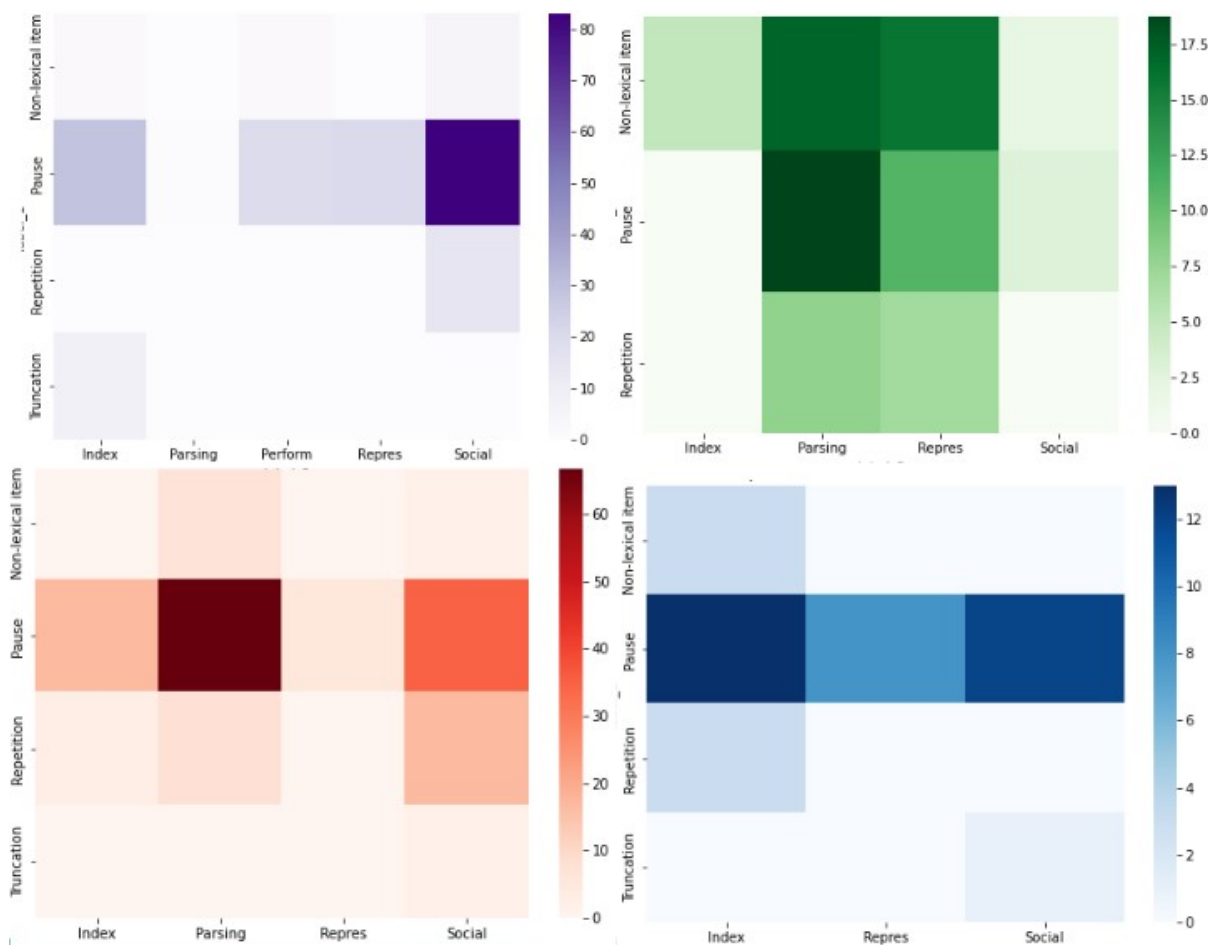


Figure 4: Heat maps of co-occurrence of *speech constants* and *functions of hand gesture* (by-legislator). The upper left image belongs to high_C, the upper right image belongs to high_B, the lower left image belongs to high_D, and the lower right image belongs to high_A.

Modal¹⁴.

Trotta and Guarasci (2021) claim that more hand gestures go with semi-lexical items (“non-lexical item” in our study) and that pauses can confirm LRH. In this way, if we take *speech constants* as the speech disfluency indicators, then pauses and non-lexical items seem to be the focused indicator to evaluate the LRH. In the following

¹⁴As there is no co-occurrence between Modal and *speech constants*, this label is not displayed in Table 6.

analysis, we focus on *function of hand gesture* co-occurring with pause and non-lexical item. These functions of concurrent hand gesture can be subcategorized into LRH-related functions (Parsing and Representational) and non-LRH-related functions (Social, Indexical, and Performative), for beat and representational gestures receptively correlate with different types of fluency (Hostetter and Alibali, 2007).

Starting from the LRH-related functions of

hand gesture, Table 6 shows that *functions of hand gesture* co-occurring with pause and non-lexical item account for 42.6%. Parsing is the second-highest intended *function of hand gesture* co-occurring with pause; this is noticeably consistent with the obvious correlation between semantic fluency and beat gestures (Hostetter and Alibali, 2007). Although pauses co-occur with hand gestures for Representational rank fourth, it still comprises 13.3% of total occurrences. In the case of non-lexical items, hand gestures for Parsing and Representational functions show higher frequencies for appearing with non-lexical item (65.5%), suggesting that hand gestures co-occurring with non-lexical item are more likely to facilitate verbal delivery in formal speech. From the discussion above, it can be concluded that pauses and non-lexical items are often accompanied by hand gestures for Parsing and Representational, which appears to correspond with the findings of how gestures prime lexical retrieval reviewed in Gillespie et al. (2014).

When it comes to non-LRH-related functions of concurrent hand gestures, the pause is highly associated with hand gestures for Social function. This indicates that pauses seem not primarily to represent hesitation pauses but rather to emphasize the primary topic of the speech in interpellation. Subsequently, Indexical is the ranked third *function of hand gestures* synchronizing with pause, implying that speakers prefer to depict the referent with visual-motion modality. Performative function is the least frequent one, but its occurrence is still significant compared to other *speech constants*. Indexical function in non-lexical item case is subtly higher than Social and Performative. As shown in Figure 4, it can be inferred that synchronous hand gestures of pause and non-lexical item also carry out information emphasis and referent depiction functions.

To sum up, in formal speech hand gestures co-occurring with *speech constants* related to speech disfluency are not just used to iconically represent the unspoken thoughts but also serve the function of reinforcing the verbal information.

4.3 Co-occurrence of individual legislators

This research takes formal speech as a research target to reexamine the applicability of LRH in individual performance since Gillespie et al. (2014) specify that LRH is less applicable if the speaker can use alternate strategies to circumvent lexical access difficulties that arise during improvised speech. Trotta and Guarasci (2021) illustrate that LRH does not confirm in all interviewers' performances, whereas the applicability of LRH in formal speech stays unclear. Accordingly, the purpose of this section is to highlight the functions adopted by all speakers and their implications related to LRH.

According to individual speaker behaviors in Figure 4, Social, Indexical, and Representational are the functions employed by all of the speakers. This exemplifies that information accentuation and referent portrayal are primary functions of synchronous hand gestures despite possible variations in individual style preferences. Notably, all speakers adopt the concurrent hand gestures for the Representational function when pausing, indicating the widespread use of nonverbal modalities to compensate for verbal delivery difficulties in improvised speech situations. This offers a new perspective to extend the suggestions presented by Gillespie et al. (2014), highlighting the general applicability of hand gestures to serve the lexical retrieval purpose in formal spontaneous speech contexts.

5 Conclusion

In conclusion, this paper highlights the creation of a multimodal corpus of Taiwanese languages and evaluates its research potential by investigating the lexical retrieval hypothesis in gestures and speech.

The case study using the MultiMoco dataset presented in this paper examines the application of multimodal corpora in the investigation of the lexical retrieval hypothesis, indicating that hand gestures often accompany *speech constants* such as pauses and non-lexical items, priming the function of lexical retrieval. By leveraging the corpus, our finding suggests that hand gestures are not solely for retrieval struggles but can also serve as means of emphasizing information. Additionally, the outcome of individual speech performances signifies the general applicability of hand gestures for the lexical retrieval purpose.

In the subsequent investigation, our emphasis will be on examining the potential correlation be-

tween hand movements and the content of regular speech (excluding non-speech elements). Following the current study, our objective is to conduct a thorough comparison of how various gesture functions are distributed in both disfluent and fluent speech contexts. We can also investigate the issue from neurolinguistic perspectives (Weisberg et al., 2017), with active learning in annotation expansion (Gal et al., 2017), or for Multimodal Learning Analytics (MMLA) applications in education disciplines (Chen et al., 2014). We believe that the continued development and utilization of the MultiMoco Corpus will pave the way for enhancing our understanding of the intricate interplay between verbal and non-verbal communication channels.

References

- Jens Allwood. 2008. Multimodal corpora. In *Corpus Linguistics. An International Handbook*, pages 207–225. Berlin: Mouton de Gruyter.
- Jens Allwood, Loredana Cerrato, Kristiina Jokinen, Costanza Navarretta, and Patrizia Paggio. 2005. The mumin annotation scheme for feedback, turn management and sequencing. In *Proceedings from the Second Nordic conference on Multimodal Communication. Gothenburg Papers in Theoretical*.
- Tom J Barry, James W Griffith, Stephanie De Rossi, and Dirk Hermans. 2014. Meet the fribbles: novel stimuli for use within behavioural research. *Frontiers in Psychology*, 5:103.
- Janet Beavin Bavelas, Nicole Chovil, Douglas A Lawrie, and Allan Wade. 1992. Interactive gestures. *Discourse processes*, 15(4):469–489.
- Jana Bressemer, Silva H Ladewig, and Cornelia Müller. 2013. 71. linguistic annotation system for gestures. In *Volume 1*, pages 1098–1124. De Gruyter Mouton.
- Brian Butterworth and Geoffrey Beattie. 1978. Gesture and silence as indicators of planning in speech. *Recent advances in the psychology of language: Formal and experimental approaches*, pages 347–360.
- Belinda Crawford Camiciottoli and Veronica Bon-signori. 2015. The pisa audiovisual corpus project: a multimodal approach to esp research and teaching. *ESP Today*, 3(2):139–159.
- Ronald Carter and Svenja Adolphs. 2008. Linking the verbal and visual: new directions for corpus linguistics. In *Language, People, Numbers*, pages 275–291. Brill.
- Lei Chen, Chee Wee Leong, Gary Feng, and Chong Min Lee. 2014. Using multimodal cues to analyze mla’14 oral presentation quality corpus: Presentation delivery and slides quality. In *Proceedings of the 2014 ACM workshop on multimodal learning analytics workshop and grand challenge*, pages 45–52.
- Lei Chen, R Travis Rose, Ying Qiao, Irene Kimbara, Fey Parrill, Haleema Welji, Tony Xu Han, Jilin Tu, Zhongqiang Huang, Mary Harper, et al. 2006. Vace multimodal meeting corpus. In *Machine Learning for Multimodal Interaction: Second International Workshop, MLMI 2005, Edinburgh, UK, July 11–13, 2005, Revised Selected Papers 2*, pages 40–51. Springer.
- Huang-Cheng Chou, Wei-Cheng Lin, Lien-Chiang Chang, Chyi-Chang Li, Hsi-Pin Ma, and Chi-Chun Lee. 2017. Nnime: The nthu-ntua chinese interactive multimodal emotion corpus. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 292–298. IEEE.
- Elizabeth A Clark, J’Nai Kessinger, Susan E Duncan, Martha Ann Bell, Jacob Lahne, Daniel L Gallagher, and Sean F O’Keefe. 2020. The facial action coding system for characterization of human affective response to consumer product-based stimuli: a systematic review. *Frontiers in psychology*, 11:920.
- Camille Debras. 2021. How to prepare the video component of the diachronic corpus of political speeches for multimodal analysis. *Research in Corpus Linguistics*, 9(1):132–151.
- Allen T Dittmann and Lynn G Llewellyn. 1969. Body movement and speech rhythm in social conversation. *Journal of personality and social psychology*, 11(2):98.
- Lotte Eijk, Marlou Rasenberg, Flavia Arnese, Mark Blokpoel, Mark Dingemanse, Christian F Doeller, Mirjam Ernestus, Judith Holler, Branka Milivojevic, Asli Özyürek, et al. 2022. The cabb dataset: A multimodal corpus of communicative interactions for behavioural and neural analyses. *NeuroImage*, 264:119734.
- Paul Ekman and Wallace V Friesen. 1972. Hand movements. *Journal of communication*, 22(4):353–374.
- Paul Ekman and Erika L Rosenberg. 1997. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press.
- Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep bayesian active learning with image data. In *International conference on machine learning*, pages 1183–1192. PMLR.
- Carlos A Gomez Gallo, Scott Stoness, Mary Swift, and Michael Tanenhaus. 2006. Software architectures for incremental understanding of human speech.
- Maureen Gillespie, Ariel N James, Kara D Federmeier, and Duane G Watson. 2014. Verbal working memory predicts co-speech gesture: Evidence from individual differences. *Cognition*, 132(2):174–180.

- Autumn B Hostetter and Martha W Alibali. 2007. Raise your hand if you're spatial: Relations between verbal and spatial skills and gesture production. *Gesture*, 7(1):73–95.
- Mercedes Querol Julián. 2011. *Evaluation in discussion sessions of conference paper presentations*. LAP LAMBERT Academic Publishing.
- Adam Kendon. 2004. *Gesture: Visible action as utterance*. Cambridge University Press.
- Dawn Knight, Svenja Adolphs, Paul Tennent, and Ronald Carter. 2008. The nottingham multi-modal corpus: A demonstration.
- Robert M Krauss. 1998. Why do we gesture when we speak? *Current directions in psychological science*, 7(2):54–54.
- Robert M Krauss and Uri Hadar. 1999. The role of speech-related arm/hand gestures in word retrieval. *Gesture, speech, and sign*, 93.
- Andy Lücking, Kirsten Bergmann, Florian Hahn, Stefan Kopp, and Hannes Rieser. 2010. The bielefeld speech and gesture alignment corpus (saga). In *LREC 2010 workshop: Multimodal corpora—advances in capturing, coding and analyzing multimodality*.
- Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. [Mediapipe: A framework for building perception pipelines](#).
- Evelyn Z McClave. 2000. Linguistic functions of head movements in the context of speech. *Journal of pragmatics*, 32(7):855–878.
- Iain McCowan, Samy Bengio, Daniel Gatica-Perez, Guillaume Lathoud, Florent Monay, Darren Moore, Pierre Wellner, and Hervé Bourlard. 2003. Modeling human interaction in meetings. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03)*, volume 4, pages IV–748. IEEE.
- Iain McCowan, Daniel Gatica-Perez, Samy Bengio, Darren Moore, and Hervé Bourlard. 2005. Towards computer understanding of human interactions. In *Machine Learning for Multimodal Interaction: First International Workshop, MLMI 2004, Martigny, Switzerland, June 21-23, 2004, Revised Selected Papers 1*, pages 56–75. Springer.
- David McNeill. 1992. Hand and mind1. *Advances in Visual Semiotics*, 351.
- IZIDOR Mlakar, ZDRAVKO Kačič, and MATEJ Rojc. 2017. A corpus for investigating the multimodal nature of multispeaker spontaneous conversations—eva corpus. *WSEAS transactions on information science and applications*, 14:213–226.
- Cornelia Müller. 2004. Forms and uses of the palm up open hand: A case of a gesture family. *The semantics and pragmatics of everyday gestures*, 9:233–256.
- Demet Özer and Tilbe Göksun. 2020. Gesture use and processing: A review on individual differences in cognitive resources. *Frontiers in Psychology*, 11:573555.
- Magali Paquot and Stefan Th Gries. 2021. *A practical handbook of corpus linguistics*. Springer Nature.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. [Robust speech recognition via large-scale weak supervision](#).
- Frances H. Rauscher, Robert M. Krauss, and Yihsiu Chen. 1996. [Gesture, speech, and lexical access: The role of lexical movements in speech production](#). *Psychological Science*, 7(4):226–231.
- Matthias Rehm, Elisabeth André, Nikolaus Bee, Birgit Endrass, Michael Wissner, Yukiko I Nakano, Afia Akhter Lipi, Toyoaki Nishida, and Hung-Hsuan Huang. 2009. Creating standardized video recordings of multimodal interactions across cultures. *Multimodal corpora*, (5509):138–159.
- Han Sloetjes and Peter Wittenburg. 2008. Annotation by category - elan and iso dcr. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*.
- Lisa Smithson and Elena Nicoladis. 2013. Verbal memory resources predict iconic gesture use among monolinguals and bilinguals. *Bilingualism: Language and Cognition*, 16(4):934–944.
- Daniela Trotta and Raffaele Guarasci. 2021. How are gestures used by politicians? a multimodal co-gesture analysis. *IJCoL. Italian Journal of Computational Linguistics*, 7(7-1, 2):45–66.
- Miriam Voghera. 2001. Teorie linguistiche e dati di parlato. *Dati Empirici E Teorie Linguistiche*, pages 75–96.
- Aaron Weinberg, Tim Fukawa-Connelly, and Emilie Wiesner. 2013. Instructor gestures in proof-based mathematics lectures. In *Proceedings of the 35th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*, volume 1119.
- Jill Weisberg, Amy Lynn Hubbard, and Karen Emmorey. 2017. Multimodal integration of spontaneously produced representational co-speech gestures: an fmri study. *Language, cognition and neuroscience*, 32(2):158–174.

Multi-word Expressions as Discourse Markers in Multilingual TED-ELH Parallel Corpus

Giedrė Valūnaitė Oleškevičienė

Institute of Humanities
Mykolas Romeris university
Ateities 20, LT-08303
Vilnius, Lietuva
gvalunaite@mruni.eu

Chaya Liebeskind

Department of Computer Science
Jerusalem College of Technology
21 Havaad Haleumi st., 9116001
Jerusalem, Israel
liebchaya@gmail.com

Abstract

In this paper, we present the outcome of the research inspired by the Nexus Linguarum network. As a theoretical basis, we discuss the multi-word word expressions as a part of the formulaic language used as discourse markers for organizing discourse. We also identify that parallel research in multiple languages may provide inter-lingual insights. We created a parallel multilingual corpus TED-ELH for our research and applied a parallel corpus alignment algorithm to extract multi-word discourse markers and their translations in Lithuanian and Hebrew. The analysis of the translations of multi-word discourse markers allowed us to identify that they demonstrate certain variability and either remain multi-word expressions or turn into one-word translations due to the linguistic characteristics of the target languages.

1 Introduction

One of natural language processing (NLP) research trends focuses on textual coherence including the relatedness of dialogical speech and also discourse relations between sentences and bigger pieces of text. Discourse relations both explicit and implicit facilitate a better understanding of the underlying relations among ideas in spoken or written texts. While implicit discourse relations could be inferred relying on the surrounding context, explicit discourse relations are realized through explicit discourse markers that belong to a number of linguistic classes including multi-word expressions. Currently, the researchers are working on both monolingual and multilingual resources. Monolingual studies and the development of the resources of discourse makers (Prasad et al., 2014; Webber et al., 2016) gave rise to multilingual studies creating multilingual corpora and comparing the use of discourse markers in various languages (Stede et al., 2016; Zufferey, 2016; Oleskeviciene et al., 2018; Zeyrek et al., 2019).

The purpose of the current study is extending the available resources working towards low-resource languages and providing linguistic processing for several languages by creating a multilingual parallel corpus (including English Lithuanian and Hebrew) based on social media texts and working on multi-word expressions in social media texts by exploring how multi-word expressions are used as discourse markers and if they remain multi-word expressions in the languages of the TED-ELH Parallel Corpus.

2 Related research

The rise of corpus linguistics and NLP brought the understanding that formulaic language plays an important role and that language users have memorized sequences which enable language generation process (Biber et al., 1999). In fact formulaic language is used as an umbrella term which covers collocations, idioms, lexical bundles or multi-word expressions and etc. Lexical bundles or multi-word expressions often perform discourse organizing functions (Biber et al., 2004) so in such cases they operate as discourse markers. As discourse markers signal discourse relations and organization researchers expect that obtaining parallel findings in different languages may serve as substantial evidence of discourse marker discourse organizing role (Zufferey, 2016). This generated research focusing on cross-linguistic mapping of discourse markers (Nedoluzhko and Lapshinova-Koltunski, 2018; Meyer and Poláková, 2013). The insights in semantic provided by Noel (Noël, 2003) stress the importance of cross-linguistic and translation studies of discourse markers as such approach may give light on contextual dimensions of the researched discourse markers. Evers-Vermeul et al. (Evers-Vermeul et al., 2011) identify that translation correspondence of discourse markers may provide the information on the pragmatic content because usually certain translator choices are guided by certain

meanings which guide the translator while looking for the equivalents or making the corresponding choices in the target context.

The research on coherence relations also stimulated research on multi-word expressions used as discourse markers (Dobrovoljc, 2017). Initially, only secondary status was given to multi-word expressions serving as discourse markers and performing pragmatic functions in corpus linguistics research. However, Wray (Wray, 2013) pointed out that multi-word discourse markers require empirical research and reconsideration. Corpus-driven research on formulaic language led to understanding that certain multi-word expressions perform discourse signaling and organizing function (Cso- may, 2013; Schnur, 2014).

3 Methodology

First, the parallel texts in English, Lithuanian, and Hebrew were extracted from TED talks by using the transcripts, and then the sentences were aligned to make a parallel corpus for further research. The corpus contains 87230 aligned sentences (published in LINDAT/CLARIN-LT repository). Then further, we focused on multi-word expressions and narrowed our research focusing on multi-word expressions which are used as discourse markers to ensure textual cohesion and according to Fraser (Fraser, 2009) relate separate discourse messages, for example, such phrases as *you know, I mean, of course*, etc. which are characteristic of spoken language (Furkó and Abuczki, 2014; Huang, 2011). Thus, 3314 aligned sentences containing the earlier mentioned multi-word expressions were extracted and then manually annotated spotting the cases when the expressions are used as discourse markers, for example in case (1) the multi-word expression *you know* is used to introduce a new discourse message, while in case (2) they are content words fully integrated into the sentence.

1. You know, I'm not even ashamed of that.
2. You know the little plastic drawers you can get at Target.

After that, the variations of the translations of discourse markers into Lithuanian and Hebrew were extracted for comparative study spotting out the variations in translation.

4 Research findings

At the initial stage of the research the manual annotation revealed the distribution of multi-word expressions used as discourse markers and content words (see Figure 1). The research revealed that some multi-word expressions are used as discourse markers more often while other multi-word expressions have a tendency to remain content words in the research corpus. The most frequent multiword expressions used as discourse markers appear to be *I think* and *you know*. It is visible in Figure 1 that such multi-word expressions as *that is* or *you see* are seldom used as discourse markers in the researched corpus, instead they are mostly content words.

Also it was identified that English multi-word expressions used as discourse markers demonstrate variability in Lithuanian and Hebrew translations: they are either translated into multi-word expressions or in one inflected word in the target languages or are omitted at all. For example, in Lithuanian multi-word expression discourse marker *you know* splits into a number of multi-word expressions and also one-word translations. Multi-word expressions could be classified into cases representing pronoun-verb phrase *jūs žinote* (you know), *jūs suprantate* (you understand), *jūs įsivaizduojate* (you imagine), *jūs esate girdėję* (you have heard) or particle-verb phrase: *(na (well), juk (after all), ir (and)) žinote* (you know), *suprantate* (you understand), or connective-verb phrase (*kaip (how), kad (that)) žinote* (you know), *matote* (you see) where connective could be used in a pre- or post- position to the verb.

One-word translations mainly include verbs, for example, *žinote* (you know), *suprantate* (you understand), *įsivaizduojate* (you imagine), and etc., which due to Lithuanian being a highly inflected language (Zinkevičius et al., 2005) fully represent the verb-pronoun cases. It should be noted that Lithuanian translations of pronoun-verb multi-word expressions and one-word verb cases could be considered as almost word for word translations. It could be said that more interesting cases which represent translator choices of particle-verb or connective-verb multi-word expressions which due to the use of particles and conjunctions also carry out certain rhetorical discourse meaning which needs to be researched further.

In Hebrew multi-word discourse marker translations demonstrate the tendency to remain multi-

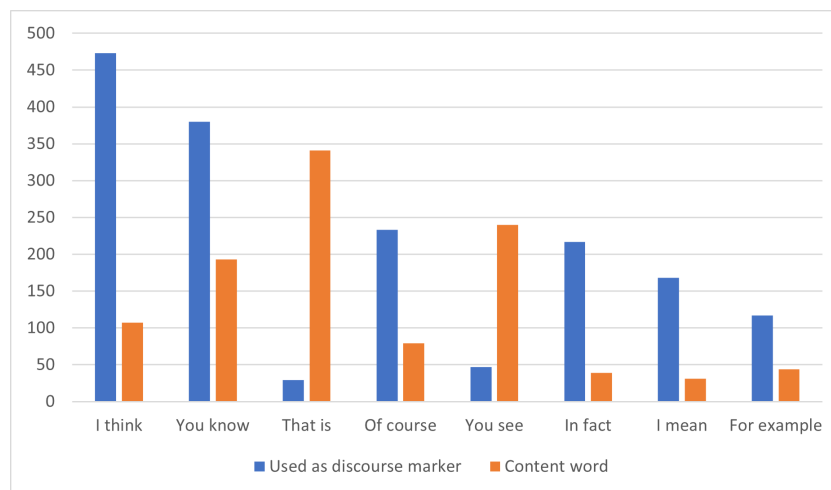


Figure 1: Multiword expressions used as discourse markers and content words

word discourse markers with a little number of one word translations. The distinctive pattern in Hebrew is the prevalence of male gender in discourse marker translations, for example, the translations of the discourse marker *you know* are mostly expressed using male gender in plural **אתם יודעים** and in singular **אתה יודע** which reveals that the translators demonstrate preference for male gender in their translation choices. Similarly to Lithuanian there are cases in Hebrew translation when a connective is added to the multi-word expression for example, **ואנו יודעים** (and we know) which also relate to the rhetorical discourse nature so further research is required to investigate the cases of additional particles and connectives used in the translation.

5 Conclusions

In conclusion, the analysis of multi-word expressions used as discourse markers identifies that there is a certain distribution of multi-word expressions used as discourse markers in the researched corpus. The analyzed multi-word expressions fall into two groups: the multi-word expression with the tendency of being used as discourse markers in the researched corpus and the multi-word expressions with the tendency of being used as content words in the researched corpus.

The initial research also reveals that in Ted talks translated transcripts English multi-word discourse markers may be translated into one-word expression probably due to the rich in inflections target languages of the research. The analysis of the translations of the multi-word expressions used as discourse markers in Lithuanian and Hebrew reveals

that there is a tendency in Lithuanian to turn them into one word discourse markers due to translator preferences to use inflected verb forms. While in Hebrew the tendency is to keep the multi-word form of discourse markers just mainly choosing the male gender both in singular and plural forms of the discourse marker translations which could be socio-culturally guided translator choice.

There are also cases of additional particles and connectives used in the translation of multi-word expressions both in Lithuanian and Hebrew. Such translator choices could be guided by the contextual pragmatic features; however, further research is needed to investigate the cases further. The mentioned cases are interesting for the research as they require insights and specific annotation to investigate which contextual pragmatic factors guided the translator choices.

The corpus building method and the extraction method of the multi-word expressions used as discourse markers tested on social media texts such as TED talks scripts can be applied to other languages. Also, it relates to expanding resources by working towards low-resource languages as the parallel corpus embracing English, Lithuanian, and Hebrew was build and it could be used as a resource for multiple scientific research.

Acknowledgements

This study is based upon work from COST Action NexusLinguarum - European network for Web-centered linguistic data science (CA18209), supported by COST (European Cooperation in Science and Technology, <https://www.cost.eu/>).

References

- Douglas Biber, Susan Conrad, and Viviana Cortes. 2004. If you look at...: Lexical bundles in university teaching and textbooks. *Applied linguistics*, 25(3):371–405. Publisher: Oxford University Press.
- Douglas Biber, Stig Johansson, Geoffrey Leech, S. Conrad, Eclward Finegan, and Randolph Quirk. 1999. Longman. *Grammar of spoken and written english*.
- Eniko Csomay. 2013. Lexical bundles in discourse structure: A corpus-based study of classroom discourse. *Applied linguistics*, 34(3):369–388.
- Kaja Dobrovoljc. 2017. Multi-word discourse markers and their corpus-driven identification: The case of mwdm extraction from the reference corpus of spoken slovene. *International journal of corpus linguistics*, 22(4):551–582.
- Jacqueline Evers-Vermeul, Liesbeth Degand, Benjamin Fagard, and Liesbeth Mortier. 2011. Historical and comparative perspectives on subjectification: A corpus-based analysis of dutch and french causal connectives. *Linguistics*, 49(2):445–478.
- Bruce Fraser. 2009. An account of discourse markers. *International review of Pragmatics*, 1(2):293–320. Publisher: Brill.
- Péter Furkó and Ágnes Abuczki. 2014. English discourse markers in mediatised political interviews. *Brno Studies in English*, 40(1).
- Lan Fen Huang. 2011. *Discourse markers in spoken English: A corpus study of native speakers and Chinese non-native speakers*. PhD Thesis, University of Birmingham.
- Thomas Meyer and Lucie Poláková. 2013. Machine translation with many manually labeled discourse connectives. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 43–50.
- A Nedoluzhko and E Lapshinova-Koltunski. 2018. Pronominal adverbs in german and their equivalents in english, czech and russian: Evidence from the parallel corpus. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual conference "Dialogue" (Moscow)*, pages 522–532.
- Dirk Noël. 2003. Translations as evidence for semantics: an illustration. *Linguistics*, 41(4):757–785.
- Giedre Valunaite Oleskeviciene, Deniz Zeyrek, Viktorija Mazeikiene, and Murathan Kurfalı. 2018. Observations on the annotation of discourse relational devices in TED talk transcripts in Lithuanian. In *Proceedings of the workshop on annotation in digital humanities co-located with ESSLLI*, volume 2155, pages 53–58.
- Rashmi Prasad, Bonnie Webber, and Aravind Joshi. 2014. Reflections on the Penn Discourse Treebank, comparable corpora, and complementary annotation. *Computational Linguistics*, 40(4):921–950. Publisher: MIT Press.
- Erin Schnur. 2014. Phraseological signaling of discourse organization in academic lectures: A comparison of lexical bundles in authentic lectures and eap listening materials. *Yearbook of Phraseology*, 5(1):95–122.
- Manfred Stede, Stergos Afantenos, Andreas Peldzus, Nicholas Asher, and Jérémy Perret. 2016. Parallel discourse annotations on a corpus of short texts. In *10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1051–1058.
- Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2016. A discourse-annotated corpus of conjoined vps. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 22–31.
- Alison Wray. 2013. Formulaic language. *Language Teaching*, 46(3):316–334.
- Deniz Zeyrek, Amália Mendes, Yulia Grishina, Murathan Kurfalı, Samuel Gibbon, and Maciej Ogródniczuk. 2019. Ted multilingual discourse bank (tedmdb): A parallel corpus annotated in the pdtb style. *Language Resources and Evaluation*, pages 1–27.
- Vytautas Zinkevičius, Vidas Daudaravičius, and Erika Rimkutė. 2005. The Morphologically annotated Lithuanian Corpus. In *Proceedings of The Second Baltic Conference on Human Language Technologies*, pages 365–370.
- Sandrine Zufferey. 2016. Discourse connectives across languages: factors influencing their explicit or implicit translation. *Languages in Contrast*, 16(2):264–279.

DRIPPS: a Corpus with Discourse Relations in Perfect Participial Sentences

Purificação Silvano

University of Porto and CLUP
Via Panorâmica, s/n
4150-564 Porto, Portugal
msilvano@letras.up.pt

João Cordeiro

University of Beira Interior
Rua Marquês d'Ávila e Bolama,
6201-001 Covilhã, Portugal
INESC TEC, Porto, Portugal
jpcc@ubi.pt

António Leal

University of Porto and CLUP
Via Panorâmica, s/n
4150-564 Porto, Portugal
jleal@letras.up.pt

Sebastião Pais

University of Beira Interior and HULTIG
Rua Marquês d'Ávila e Bolama,
6201-001 Covilhã, Portugal
sebastiao@ubi.pt

Abstract

The main objective of this paper is to introduce a new language resource for some varieties of Portuguese - European, Brazilian, Mozambican, and Angolan - and for British English, called DRIPPS (Discourse Relations In Perfect Participial Sentences). The corpus DRIPPS comprises, at the moment, 993 adverbial perfect participial sentences annotated with Discourse Relations and with the following Discourse Relational Devices: connectors, ordering of the clauses, temporal relations, tenses, and aspectual types. Additionally, an application with a *Graphical User Interface* (GUI) has been developed not only to browse and manipulate the corpus but also to allow the activation of specific Discourse Relation constraints, thereby selecting specific cases from the data set that can be analyzed separately. Besides calculating simple counts and percentages, insightful statistical graphs can be generated and visualized on the fly from the combination of the user-selected constraints and the loaded corpora. The application is pre-loaded with Portuguese and English cases and allows to import/load further cases from different languages/varieties.

1 Introduction

Discourse Relations (DRel) are meaning relations used to describe textual coherence by establishing connections between the different textual segments through meaning functions, crucial to analyze discourse structure and explain linguistic problems. For that reason, there has been a propagation of small or medium size annotated corpora of different genres (instructive, expository, descriptive, argumentative, narrative; oral, written), and in various

languages (individual or parallel): e.g. *Penn Discourse Treebank* (PDTB) (Prasad et al., 2008), *RST Spanish Treebank* (RST-ST) (da Cunha et al., 2011), *SDRT Annodis French corpus* (Afantenos et al., 2012), and *Prague Discourse Treebank* (Rysová et al., 2016). The increasing interest in annotated corpora with DRel stems from the valuable contribution that those may offer to the development of Natural Language Processing (NLP) applications, such as automatic summarization and translation, information retrieval, sentiment analysis, and opinion mining (see Webber et al. (2012) for a review of these applications).

For European Portuguese, the only existing corpora annotated with DRel are the following: a relatively small corpus of spoken discourse (TEDPT) (Zeyrek et al., 2018, 2020; Mendes et al., 2023), and CRPC-DB, a Discourse Bank for Portuguese annotated according to the Penn Discourse Treebank (PDTB) scheme (Mendes and Lejeune, 2022). Regarding other varieties, the closest is CST-news with cross-document annotated relations established between sentences aimed at summarization for Brazilian Portuguese (Cardoso et al., 2011). Aleixo and Pardo (2008) describe the annotation process of this corpus of 3534 sentences extracted from news and annotated according to *Cross-document Structure Theory*. Collovini et al. (2007) annotated a corpus of 50 news texts also in Brazilian Portuguese using *Rhetorical Structure Theory* (Mann and Thompson, 1988). Angolan and Mozambican varieties lack any annotated corpora with DRel.

Currently, the annotation of DRel in many corpora relies on a lexically grounded approach –

mostly on information conveyed by discourse connectors (conjunctions or connectives, like ‘although’, ‘because’, ‘as a result of’) – which implies leaving some discourse segments without annotation or annotated with implicit relations. Some, nonetheless, adopt a ‘complete discourse coverage’ (Benamara and Taboada, 2015) taking other information sources into account, like PDTB (Prasad et al., 2008), the American English corpus (Carlson et al., 2001, 2003) annotated with the framework of Rhetorical Structure Theory (Mann and Thompson, 1988) and the Potsdam Commentary Corpus (Stede, 2004), a corpus of German newspaper commentaries also annotated with Rhetorical Structure Theory (Mann and Thompson, 1988), using RST-Tool¹. For an exhaustive annotation of DRel, it is essential, in addition to discourse connectives, to consider other Discourse Relational Devices² (DRD) (e.g. semantic and syntactic) that are pivotal when inferring DRel. The consideration and study of these DRD lead to improved annotation and a more comprehensive and grounded explanation of discourse organization.

Structures without connectives abound in texts, and some have specific syntactic and semantic properties, which may determine the DRel. One such construction is the one with an adverbial perfect participial clause (APC). This type of sentence results from combining two complete propositions, and it can convey inter-propositional values of different types (Móia and Viotti, 2004; Leão, 2018), which can be represented by DRel. Das and Taboada (2018) consider that participial clauses, both with present and past participles, are syntactic signals of certain DRel, that is, they are themselves DRD. However, our study reveals that, although they may signal the existence of a DRel, they allow for a wide array of DRel partly because this construction is mostly devoid of discourse markers. Therefore, the speakers must rely on other sources of information to infer the relevant DRel, such as the tense of the main clause, temporal relations, aspectual type of the situations involved, position of the adverbial perfect participial clause relative to the main clause and the temporal value of the participle. Identifying these sources (or DRD) is essential to better understand how we infer DRel in APC. Moreover, this research can give essential clues to identifying the relevant sources of informa-

tion in other constructions where discourse markers are also absent. In addition to this, the results of this investigation can also benefit the automatic extraction of DRel.

The primary purpose of this paper is to present a new language resource, DRIPPS, an annotated corpus of discourse relations in sentences with perfect participial clauses in some varieties of Portuguese (European (EP), Brazilian (BP), Angolan (AP) and Mozambican (MP)) and British English (BE), which is the outcome of research that the authors have been developing (Leal, 2011; Silvano et al., 2019, 2021). The option for the aforementioned Portuguese varieties is motivated by the fact that MP and AP lack not only annotated corpora but also stabilized norms, so it is of utmost importance to uncover the differences and similarities between these Portuguese varieties and the ones that have been studied and analyzed in more depth (EP and BP). Besides, contrary to EP and BP, MP and AP are most likely impacted by other African languages typologically different from Portuguese, such as Bantu languages (e.g. Carvalho and Lucchesi (2016)), so the description of these African Portuguese varieties will contribute to bringing to light their particularities regarding both EP and BP. The inclusion of BE in the corpus is motivated by two types of reasons. From a theoretical linguistic point of view, it is essential to compare languages, especially from different branches/families. From a computational point of view, since English is a well-studied language for which many computational tools have already been developed, a corpus that contrasts the same construction in English and Portuguese can aid in adapting tools designed for English to the specificities of Portuguese.

The following two sections provide a more detailed description of DRIPPS and of an application interface for browsing the corpus. Section 2.1 is dedicated to a brief semantic and syntactic characterization of the data, i.e., sentences with adverbial perfect participial sentences in both languages (Portuguese and British English); Section 2.2 details the process of building the corpus; Section 2.3 lays out the annotation framework; and Section 2.4 presents results of the corpus analysis. Section 3 explains the interface designed to access and work with the corpus. Finally, some concluding remarks and plans for future work are provided in Section 4.

¹<http://www.wagsoft.com/RSTTool/>

²Term used by TextLink (www.textlink.ii.metu.edu.tr/).

2 DRIPPS corpus

This section describes the Discourse Relations In Perfect Participial Sentences Corpus (DRIPPS), its creation, and the annotation framework. This first version of DRIPPS gathers 993 sentences with adverbial perfect participial clauses in varieties of Portuguese (EP, BP, MP, AP) and British English (BE) annotated with discourse relations (DRel) according to ISO 24617-2:8 (ISO) and relevant discourse relational devices (DRD). More data will gradually be added in the subsequent versions.

2.1 The Data: Adverbial Perfect Participial Sentences

Adverbial perfect participial sentences (APC) (in the Portuguese grammatical tradition, *adverbial gerundive clauses with compound gerund*) are instances of subordinated clauses that, in Portuguese, have the auxiliary verb "ter" in the gerund ("tendo"), or, in English, the auxiliary verb "to have" in the -ing form ("having"), followed by the past participle of the main verb (cf. (1) and (2)).

- (1) *No passado dia 13 de novembro, o antigo avançado brasileiro já tinha sido submetido a uma intervenção cirúrgica aos rins, tendo recebido alta dois dias depois.* (from the EP dataset)
On November 13, the former Brazilian striker had already undergone kidney surgery, having been discharged two days later.
- (2) *Having served his country, he became a great believer in the need for change and to stop unnecessary wars.* (from BE dataset)

APC have been the object of much research both in Portuguese (mainly for the EP variant, e.g. Leal (2002); Lobo (2003); Mória and Viotti (2004); but also for BP, e.g. Mória and Viotti (2004); (Leão, 2018)), and English (e.g. Quirk et al. (1985); Stump (1985); Kortmann (1995); König (1995)). Overall, APC are described as being introduced, or not, by connectors (subordinating conjunctions or prepositions that function as subordinating conjunctions) and as being able to be placed in an initial and final position regarding their main clause. They are normally featured as conveying temporal interpretations of anteriority or posteriority. Additionally, some studies about the DRel that they may establish indicate that the most frequent are Narration (cf. example (1)), Explanation (cf. example (2)),

Result, Background, Elaboration and Concession (Mória and Viotti, 2004; Leal, 2011; Silvano et al., 2019).

Typologically, for European Portuguese, Lobo (2003) divides APC into peripheral clauses, which occur by default in an initial position (with a pause before the main clause) with a temporal meaning of anteriority, and coordinate clauses, which occur only in final position with a temporal meaning of posteriority. However, this proposal is not without problems, as proved by Silvano et al. (2021). The DRIPPS-based analysis carried out by Silvano et al. (2021) reveals that this distinction cannot account for the corpus data since, on the one hand, APC can be positioned initially, finally, and also medially, and, on the other hand, there is not a direct association between the position and the temporal interpretation.

2.2 Corpus Creation through Web Crawling

The corpus of sentences potentially containing APC was entirely constructed with data collected from the *World Wide Web* (Web), applying a crawling method specifically designed for that purpose. A number of well-known newspaper websites were targeted for each language and variety, and relevant sentences were extracted from online news stories. These are well-formed sentences that satisfy specific predefined linguistic patterns provided by the user. We were especially interested in selecting sentences with *adverbial perfect participial* clauses, as described in Section 2.1.

An existing common challenge in the process of selecting well-formed text from web pages is the presence of many "spurious textual segments", like in advertisements, web page structural elements (e.g., menus, sidebars, etc.), and even for news websites. These segments are absolutely unrelated to the news story, with no interest in our study. Another common characteristic of these spurious segments is the lack of an acceptable syntactical structure, even in terms of punctuation marks. Therefore, our text selection method considers these characteristics (more details in Appendix A), selecting only relevant sentences.

The corpus DRIPPS automatically extracted from public online news sources was then manually analyzed, with each sentence classified and annotated by experts from linguistics, as described in Section 2.3. The annotation process adds eight features of information to each selected sentence

related to the DRel, ending up in a data structure as shown in Table 3, as well as in the application interface shown in Figures 3 and 4. Our corpus of 993 adverbial perfect participial sentences, annotated with DRel is stored in conventional and simple CSV format, with one file for each language/variety. These files are directly loaded into the application described in Section 3 and are freely available to the community for research purposes.

Regarding legal issues, it is essential to emphasize that we are not storing whole news texts but only small portions, always keeping the reference to the original source (newspaper URL). The dataset was gathered from publicly available news sources, annotated, and kept only for language research. The decision to resort to online newspapers and not to existing corpora also derives from our intention of studying this structure in comparable, contemporary data.

At the moment, DRIPPS comprises a total of 993 adverbial perfect participial sentences annotated, 793 from four Portuguese varieties and 200 from British English. For Portuguese, DRIPPS has a total 29373 words, representing an average of 37.04 words per sentence. Details on each variety can be observed in Table 1. For the 200 British English sentences, we have a total of 5715 words, giving an average of 28.58 words per sentence.

2.3 Annotation Process

DRel integrate different semantic and pragmatic theories such as *Theory of Discourse Coherence* (Hobbs, 1985), *Rhetorical Structure Theory* (RST) (Mann and Thompson, 1987), or *Segmented Discourse Representation Theory* (SDRT) (Asher and Lascarides, 2003), which differ along several aspects, namely DRel designations, definitions, nature, number, and type of arguments. Bearing in mind, on the one hand, the diversity of these frameworks and, on the other hand, the usefulness of establishing comparisons between annotated corpora from different genres in the same language but also across languages, there have been some efforts to reconcile different proposals of annotation, which have resulted in *Semantic annotation framework (SemAF) – Part 8: Semantic relations in discourse, core annotation schema (DR-core) - ISO 24617-2:8* (ISO) (see also (Bunt and Prasad, 2016)). ISO 24617-2:8 stipulates an interoperable core-annotation scheme for low-level DRel, i.e., local dependencies. The reasons behind the

choice of ISO 24617-2:8 for our annotation scheme are two. The first reason concerns interoperability, which is fundamental (Ide and Pustejovsky, 2010) with the rapid expansion of the Semantic Web and Linguistic Linked Data (Chiarcos et al., 2020). It should be noted that, contrary to what Sanders et al. (2021) claim, ISO 24617-2:8 shows that a complete mapping between different sets of DRel proposed within various frameworks is possible. The second set of reasons derives from the first and is related to the requirements of interoperable semantic annotation (Bunt, 2015): it is language independent, general enough to be able to account for specific instances (although in some cases, more granularity is warranted³) and it has a well-defined semantics, which can be machine-interpretable.

ISO 24617-2:8 provides a set of core DRel of two types, symmetric and asymmetric: while, in the former, the arguments play the same semantic role, in the latter, Arg1 and Arg2 bear relation-specific semantic roles. Figure 1 provides the definitions of the DRel found in our corpus.

Regarding the process of DRel inference, it is widely accepted that the primary sources of information are of two types: linguistic sources (lexicon and compositional semantics) and non-linguistic sources (world knowledge and the cognitive state of the participants) (e.g. Asher and Lascarides (2003)). Although DRel may be implicit, not signalled linguistically, many are explicit, i.e. there is some linguistic marker, be it a word, lexical expression, tense or syntactic structure. These Discourse Relational Devices (DRD) are significant DRel triggers and are studied in many languages (e.g. Das (2014)). In the case of APC, in the absence of a cue phrase to signal the appropriate DRel, the process of inference must depend on other linguistic sources, namely the semantic value of the perfect participle, tense, aspect, mood and modality of the main clause, the presence of negation, or even the mere relative order of both clauses, among other factors. The study of these factors and their relative weight in the overall interpretation of APC has been pursued both for Portuguese and English (for English, e.g. Quirk et al. (1985); Stump (1985); Kortmann (1995), a.o.; for EP, e.g. Leal (2011); Lobo (2003); Silvano et al. (2021); and, for BP,

³Despite the fact that “a future part of ISO 24617 is envisaged that will complement this document by providing a complete interoperable annotation scheme for DRel, while also addressing the multilingual dimension of the standard” (ISO), it has not been published so far.

Language/Variety	#Sentcs	#Words	Words/Sentc
Angolan Portuguese	200	7772	38.86
Brazilian Portuguese	193	6734	34.89
European Portuguese	200	7605	38.03
Mozambican Portuguese	200	7262	36.31
British English	200	5715	28.58

Table 1: Corpus statistics.

	DR-core relations	Definition	Semantic Role	
			Arg1	Arg2
Asymmetric	Cause	Arg2 is an explanation for Arg1.	result	reason
	Expansion	Arg2 is a situation involving some entity/entities in Arg1, expanding the narrative of which Arg1 is a part, or expanding on the setting relevant for interpreting Arg1. The Arg1 and Arg2 situations are distinct.	narrative	expander
	Asynchrony	Arg1 temporally precedes Arg2.	before	after
	Concession	An expected causal relation between Arg1 and \neg Arg2 is cancelled or denied by Arg2.	expectation raiser	expectation-denier
	Elaboration	Arg1 and Arg2 are the same situation, but Arg2 provides more detail.	broad	specific
	Exemplification	Arg1 is a set of situations; Arg2 is an element of that set.	set	instance
	Manner	Arg2 specifies how Arg1 comes about or occurs.	achievement	means
Symmetric	Conjunction	Arg1 and Arg2 bear the same relation to some situation evoked in the discourse, explicitly or implicitly. Their conjunction indicates that they both hold with respect to that situation.		
	Contrast	One or more differences between Arg1 and Arg2 are highlighted with respect to what each predicates as a whole or to some entities they mention.		
	Synchrony	Some degree of temporal overlap exists between Arg1 and Arg2. All forms of overlap are included.		

Figure 1: Definitions of DRel-ISO 24617-2:8 (ISO; Bunt and Prasad, 2016).

Móia and Viotti (2004); Leão (2018). Our annotation scheme includes the most relevant parameters to infer DRel according to the literature. Figure 2 summarizes the framework utilized in annotating DRIPPS.

After designing the annotation scheme, two trained linguists (both EP native speakers with a good command of English) manually annotated a dataset to ensure that the guidelines were well understood. Afterwards, each annotator was assigned a different dataset to be annotated in an Excel spreadsheet. Each line had one example with only one APC. Sentences with two or more APC were duplicated, and each line was dedicated to the analysis of one and only one APC. Regarding the DRel, the annotator had to choose the most prominent DRel whenever there were two possible interpretations. Although sometimes two readings

arose, it is a fact that when the writer wrote the sentence, he/she had a specific communicative goal in mind. Whenever the interpretation was not possible due to the lack of a larger context), the example was discharged.

The inter-rater reliability between the annotators was measured with respect to DRel⁴, for each variety/language, through *Cohen's Kappa* (Cohen, 1960). Generally, the agreement obtained was significant, as shown in Table 2.

Thus, according to the Landis and Koch (1977) criteria, we can see that we have obtained three *perfect* agreements, one *moderate*, and one *substantial* agreement, shown in the third column from Table 2. The varieties where there was initially some uncertainty among the annotators were Portuguese

⁴The inter-annotator agreement regarding the DRD was not performed because their classification is clear-cut.

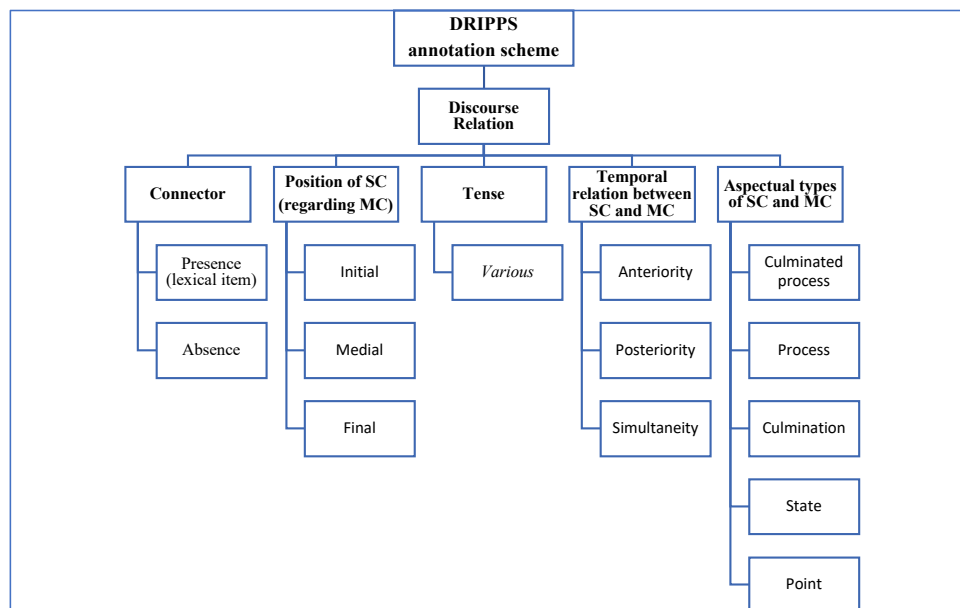


Figure 2: DRIPPS annotation scheme.

Language	<i>Kappa</i>	Agreement
PT-Brazil	0.89953	<i>perfect</i>
PT-Angola	0.55065	<i>moderate</i>
PT-Mozambique	0.67589	<i>substantial</i>
PT-Europe	0.95932	<i>perfect</i>
EN-Europe	0.88088	<i>perfect</i>

Table 2: Annotator agreement measures using *Cohen's Kappa* (Cohen, 1960).

from Angola and Portuguese from Mozambique.

The subsequent step was reaching a consensus regarding the examples of disagreement. The most relevant disagreements between the annotators involved Asynchrony and Conjunction (in AP), Asynchrony and Expansion (in MP) and Expansion (in BP). The annotators discussed the examples and agreed on the accepted DRel.

Table 3 exemplifies the result of manual annotation.

2.4 Some main results of the corpus analysis

From the complete corpus with 4222 sentences in Portuguese and 2635 in English, 993 have already been annotated (EP, AP, MP and BE – 200 sentences each; BP – 193 sentences). This first annotation has already enabled a comprehensive study of the main features annotated in DRIPPS. [Silvano et al. \(2021\)](#) demonstrate that there is crosslinguistic and intralinguistic variation. Since the main objective of this paper is not to present an in-depth

contrastive semantic analysis of the data presented in DRIPPS, we refer the reader to [Silvano et al. \(2021\)](#) and present only the main results from the research.

[Silvano et al. \(2021\)](#) conclude from the corpus analysis that, in interpreting temporal relations involving APC without connector in English, the most critical parameter is the temporo-aspectual information given by the perfect participle. In contrast, the key factors in Portuguese are the relative position of both main and subordinated clauses and their aspectual classes. Although there are no absolute restrictions regarding telicity and durativity, aspectual classes of predications are closely intertwined with temporal interpretation as anteriority and posteriority readings tend to be related to telic situations in main and subordinated clauses, whereas simultaneity readings lean on the presence of durative situations in both clauses. In English, by contrast, the combination of aspectual types in both clauses was not a relevant factor, as the anteriority reading is recurrent, irrespective of the aspectual types of both clauses. This is in line with the literature on these structures in English, which points out the anterior orientation of APC.

As for intralinguistic variation, the study also reveals that AP and MP APC are more alike EP APC and that BP is clearly different from other Portuguese varieties in what concerns the main aspects of APC. This finding goes against the idea of an Afro-Brazilian continuum of Portuguese (cf.

Sentence	Pos	TR	Tense MC	ATMC	ATSC	CNT	RR	SR-SC
A PSP do Seixal, no distrito de Setúbal, anunciou nesta terça-feira a detenção de sete pessoas por suspeita de tráfico de droga, tendo sido apreendidas mais de quatro mil doses de droga e 16 mil euros em dinheiro. (EP dataset).	Final	Ant	PP	Culm	Culm		asynchrony	before
Até à chegada da troika a Portugal, as despesas com pessoal consumiam sistematicamente 13 % a 14 % do PIB, tendo mesmo atingido o pico de 14,5 % em 2005. (EP dataset)	Final	Simul	PIMP-Ind	St	Culm		expansion	expand
As declarações estão a criar ondas de choque no meio judicial, entre magistrados e advogados, tendo levado o Conselho Superior de Magistratura (CSM) a abrir um inquérito para tirar as insinuações a limpo. (EP dataset)	Final	Post	PresPro	St	Culm		cause	result
Vários profissionais do cinema, inclusive o Exército dos Estados Unidos da América, reagiram a morte de Lee, tendo agradecido o serviço que prestou. (AO dataset)	Final	Simul	PP	Pro	Culm		elaboration	specific
No PSL, que dobrou bancada (de 1 para 2), quem fica fora é Sargento Pereira Júnior, mesmo tendo aumentado sua votação de forma considerável: de 1.267 para 1.530 votos. (BP dataset)	Final	Ant	Pres-Ind	St	CP	mesmo	concession	e-raiser
Segundo o biólogo, a invasão em Moçambique compreende duas vagas: a primeira ocorreu nos fins da década de 60 e início da década de 70, tendo afetado a Ilha da Inhaca. (MZ dataset)	Final	Simul	PP	Pro	Pro		conjunction	
If she was failing, she deserved, after having achieved so much, to be allowed to fail at the polls. (BE dataset)	Medial	Ant	Pst	St	Culm	after	cause	reason

Table 3: Sample of the annotation.

Petter (2009)).

3 The Corpus Interface Application

This section briefly presents the DRIPPS corpus interface application, focusing on the main features implemented so far. The application allows one to load corpora, Portuguese varieties, and British English, in our case, and apply a set of selection constraints to obtain different views and statistics of the data, enabling a whole range of specific corpora analyses and studies. Figure 3 presents the application’s main view, where the dataset of annotated sentences from different varieties/languages might be loaded into the main table, the main component of this view. The table presents one sentence per line with its corresponding annotations: *Discourse Relation* (DR), *Semantic Role* (SR), etc. The last column contains the sentences, which are not entirely visible. However, each table’s selected sentence is totally visible below in a specific box for that purpose (light yellow colour). The set of buttons above the table, on the right-hand side, allows one to select the varieties/languages’ examples to be shown. Each one of these buttons can be independently activated and deactivated, meaning that different sets of varieties/languages can be combined and loaded into the table. In the screenshot from Figure 3, we can see that only the European (EP) and Brazilian Portuguese (BP) varieties are selected. Note that in the table’s first column, the

prefix of the ID represents the language+variety identification. For instance, the selected example (PTEU197) is from European Portuguese, and the example immediately following is from Brazilian Portuguese. The set of controls (combo boxes) below the table allows one to define DRel constraints to be applied to the table’s fields. For example, the configuration presented states that the *discourse relation* (DR) must be *cause*, the *semantic role* (SR) is equal to *reason* and the *temporal relation* (TR) must be of *anteriority* (*Ant*). Different combinations can be set here, and different data examples will be shown accordingly in the table.

The frame of numbers appearing on the lower side of this view, entitled “Stats”, shows relevant counts and percentages according to the selections performed in the previous panel of controls. For each new selection, calculations are made, and values are shifted to the right, from (t) column toward ($t-3$). The meaning of these values depends on the path of selections the user decides to follow. For example, here, the path of selections was $DR \rightarrow SR \rightarrow TR$. Therefore, 393 in column ($t-3$) represents the total number of records loaded (for both varieties), and 108 is the number of cases from these where $DR = cause$. The 27.48% in the second line of ($t-3$) is obtained from $\frac{108}{393}$.

Finally, Figure 4 presents the feature of generating statistical distributions for a given data con-

The screenshot shows the DRIPPS application interface. At the top, there is a menu bar with 'File', 'Conditions', 'Statistics', and 'Help'. Below the menu bar, there are language selection buttons for 'EP', 'BP', 'AP', 'MP', and 'BE'. The main area is a table with columns: ID, DR, SR, CNT, POS, TR, TMC, ATMC, ATSC, and Sentence. The table contains 20 rows of annotated sentences. Below the table, there are several filter panels: 'Discourse Relation' (set to 'cause'), 'Semantic Role' (set to 'reason'), 'Connector' (set to 'All'), 'Position' (set to 'All'), 'Temporal Relation' (set to 'Ant'), 'Tense of MC' (set to 'All'), 'Aspectual type of MC' (set to 'All'), and 'Aspectual type of SC' (set to 'All'). A 'Clear' button is located to the right of these filters. Below the filters, there is a 'Selected sentence:' panel displaying a sentence in Portuguese. At the bottom, there is a 'Stats' panel showing the number of records and percentages for different temporal relations: (t) -> 46 (11,70%), (t-1) -> 72 (63,89%), (t-2) -> 108 (66,67%), and (t-3) -> 393 (27,48%).

Figure 3: The DRIPPS application to load and explore DRel corpora.

figuration loaded to the applications' table. The data configuration depends on the selected/loaded corpora and the selected constraints applied on the panel. In this particular case, we can see a graph distribution for the *Aspectual type of the SC*, for the *British English* corpus, given that the *Discourse Relation* is set to *cause*. The application allows generating several graphs like this simultaneously and for different data configurations, which enables one, for example, to compare similar phenomena on different corpora.

4 Final Remarks

In this paper, we have introduced a new language resource, DRIPPS, a corpus with an interface browser. This collection of sentences with adverbial perfect participial clauses was extracted from Portuguese varieties (European, Brazilian, Mozambican and Angolan) and British English using a web crawler specially designed and tuned for this task. This first version of DRIPPS gathers 993 APC annotated with DRel according to ISO 24617-2:8 (ISO), thus ensuring interoperability. Moreover, our annotation scheme also includes Discourse Relational Devices intervening in DRel inference, specifically connector, clauses ordering, temporal relation, tense

and aspectual types of both clauses. This new language resource comprises an interface browser enabling researchers to better study and explore the DRel phenomena in APC, comparing different Portuguese varieties and even different languages. The corpus will continue to be annotated and shared with the community so anyone can effectively analyze and explore DRel. In fact, the annotated part of DRIPPS has already allowed a wide-range study that highlighted the cross and intralinguistic variation regarding adverbial perfect participial clauses (Silvano et al., 2021). The application that we designed to explore the corpus, due to its versatility, range and the fact that it is user-friendly and intuitive, enables simple but also relevant queries intersecting several parameters.

Although the current state of knowledge about DRel and DRD and their annotation in corpora may be somewhat advanced in several languages, the same cannot be stated for Portuguese, a low-resource language. The research about DRel and the DRD that intervene in the process of inference and are relevant to the creation of automatic annotation methods must be advanced, which is the primary purpose of the current proposal. Manual annotation of these values is the first step to de-

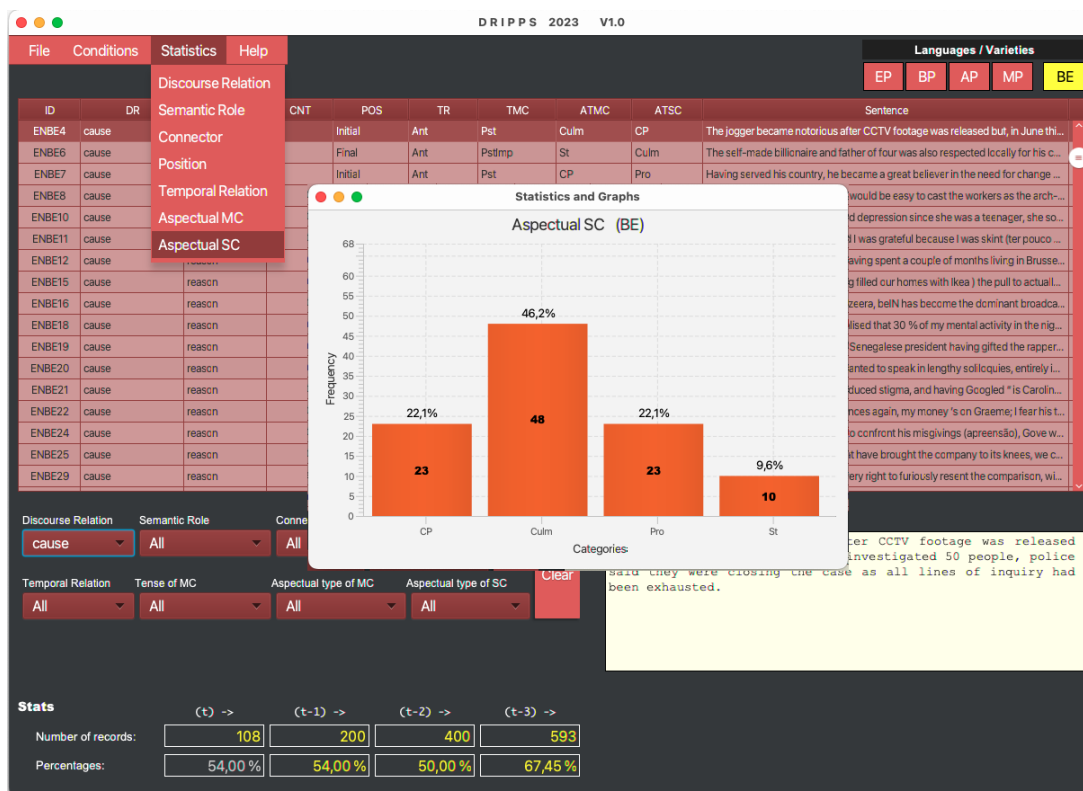


Figure 4: Selecting a statistical graph of the “Aspectual SC” distribution for the BE corpus with Discourse Relation selected on “cause”.

velop methods of semi-automatic and automatic extraction of DRel, which we intend to pursue in the future by adapting existing discourse parsers to Portuguese (e.g. Gessler et al. (2021)). Our plans for the future also include extending the annotation to more data of the current varieties/languages. To do so, we will increase the number of annotators, and, “to assess the reliability of an annotation process as a prerequisite for ensuring the correctness of the resulting annotations” (Artstein, 2017), we will not only measure inter-annotator agreement, but also conduct studies about the DRel that cause more disagreement, and the reasons for that disagreement. Lastly, we envisage making the corpus and the interface browser available in the Portulan Clarin infrastructure⁵.

Acknowledgements

National funds have funded this research through FCT – Fundação para a Ciência e a Tecnologia, I.P., within the UIDB/00022/2020 project.

⁵<https://portulanclarin.net>

References

- ISO 24617-2: 2016. Language resource management, Part 8: Semantic relations in discourse (DR-core). Standard, Geneva, CH.
- Stergos Afantenos, Nicholas Asher, Farah Benamara, Myriam Bras, Cécile Fabre, Mai Ho-dac, Anne Le Draoulec, Philippe Muller, Marie-Paule Péry-Woodley, Laurent Prévot, Josette Rebeyrolles, Ludovic Tanguy, Marianne Vergez-Couret, and Laure Vieu. 2012. An empirical resource for discovering cognitive principles of discourse organisation: the ANNODIS corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, pages 2727–2734, Istanbul, Turkey. European Language Resources Association (ELRA).
- Priscila Aleixo and Thiago Pardo. 2008. Cstnews: Um corpus de textos jornalísticos anotados segundo a teoria discursiva multidocumento cst (cross-document structure theory). Technical Report 326, Universidade de São Paulo.
- Ron Artstein. 2017. *Inter-annotator Agreement*, pages 297–313. Springer Netherlands, Dordrecht.
- Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press, United States.
- Farah Benamara and Maite Taboada. 2015. *Mapping different rhetorical relation annotations: A proposal*.

- In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 147–152, Denver, Colorado. Association for Computational Linguistics.
- Harry Bunt. 2015. [On the principles of semantic annotation](#). In *Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-11)*, London, UK. Association for Computational Linguistics.
- Harry Bunt and Rashmi Prasad. 2016. ISO DR-core (ISO 24617-8): Core concepts for the annotation of discourse relations. In *Proceedings 12th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-12)*, pages 45–54.
- Paula Cardoso, Erick Maziero, Mara Luca Castro Jorge, Eloize Seno, Ariani Di Felippo, Lucia Helena Machado Rino, Maria das Gracas Volpe Nunes, and Thiago Pardo. 2011. CSTnews-a discourse-annotated corpus for single and multi-document summarization of news texts in brazilian portuguese. In *Summarization of News Texts in Brazilian Portuguese. In the Proceedings of the 3rd RST Brazilian Meeting*.
- Lynn Carlson, Daniel Marcu, and Mary Ellen Okurovsky. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *SIGDIAL Workshop*.
- Lynn Carlson, Daniel Marcu, and Mary Ellen Okurovski. 2003. *Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory*, pages 85–112. Springer Netherlands, Dordrecht.
- Ana Maria Carvalho and Dante Lucchesi. 2016. *Portuguese in contact*, pages 41–55. Wiley Blackwell, Oxford.
- Christian Chiarcos, Bettina Klimek, Christian Fäth, Thierry Declerck, and John Philip McCrae. 2020. [On the linguistic linked open data infrastructure](#). In *Proceedings of the 1st International Workshop on Language Technology Platforms*, pages 8–15, Marseille, France. European Language Resources Association.
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.
- Sandra Collovini, Thiago Carbonel, Juliana Fuchs, Jorge Coelho, Lúcia Rino, and Renata Vieira. 2007. Summ-it: Um corpus anotado com informações discursivas visando à sumarização automática. In *V Workshop em Tecnologia da Informação e da Linguagem Humana – TIL. Proceedings of XXVII Congresso da SBC*, Rio de Janeiro.
- Iria da Cunha, Juan-Manuel Torres-Moreno, Gerardo Sierra, Luis-Adrián Cabrera-Diego, Brenda-Gabriela Castro-Rolón, and Juan-Miguel Rolland Bartilotti. 2011. [The RST Spanish treebank on-line interface](#). In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 698–703, Hissar, Bulgaria. Association for Computational Linguistics.
- Debopam Das. 2014. *Signalling of coherence relations in discourse*. Ph.D. thesis, Arts & Social Sciences: Department of Linguistics.
- Debopam Das and Maite Taboada. 2018. RST signalling corpus: A corpus of signals of coherence relations. *Language Resources and Evaluation*, 52(1):149.
- Luke Gessler, Shabnam Behzad, Yang Janet Liu, Siyao Peng, Yilun Zhu, and Amir Zeldes. 2021. [DisCoDisCo at the DISRPT2021 shared task: A system for discourse segmentation, classification, and connective detection](#). In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 51–62, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Jerry R. Hobbs. 1985. On the coherence and structure of discourse. Technical report, CSLI-85-37, Center for the Study of Language and Information.
- Nancy Ide and James Pustejovsky. 2010. What does interoperability mean, anyway? toward an operational definition of interoperability for language technology. In *Proceedings of the Second International Conference on Global Interoperability for Language Resources. Hong Kong, China*.
- Bernard Kortmann. 1995. [Adverbial participial clauses in english](#). In Martin Haspelmath & Ekkehard König, editor, *Converbs in cross-linguistic perspective. Structure and meaning of adverbial verb forms – adverbial participles, gerunds*, page 189–238. Mouton de Gruyter, Berlin.
- Ekkehard König. 1995. [The meaning of converb constructions](#). In Martin Haspelmath and Ekkehard König, editors, *Converbs in cross-linguistic perspective. Structure and meaning of adverbial verb forms – adverbial participles, gerunds*, page 1–56. Mouton de Gruyter, Berlin.
- J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.
- António Leal. 2002. O valor temporal das orações gerundivas em português. In *Actas do XVIII Encontro da Associação Portuguesa de Linguística*, page 455–464, Porto. APL.
- António. Leal. 2011. Some semantic aspects of gerundive clauses in european portuguese. In *Cahiers Chronos. From now to eternity. Amsterdam – New York: Editions Rodopi*, 22:85–113.
- Rafaella Leão. 2018. *A semântica das construções gerundivas no português europeu e no português do Brasil*. Ph.D. thesis, Universidade do Porto.
- Maria Lobo. 2003. *Aspectos da sintaxe das orações subordinadas adverbiais*. Ph.D. thesis, Universidade Nova de Lisboa.

- William Mann and Sandra Thompson. 1988. [Rhetorical structure theory: Toward a functional theory of text organization](#). *Text*, 8:243–281.
- William C. Mann and Sandra A. Thompson. 1987. [Rhetorical structure theory: A theory of text organization](#). Technical Report ISI/RS-87-190, Marina del Rey, CA: Information Sciences Institute.
- Amália Mendes and Pierre Lejeune. 2022. [CRPC-DB a discourse bank for portuguese](#). In *Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23, 2022, Proceedings*, page 79–89, Berlin, Heidelberg. Springer-Verlag.
- Amália Mendes, Deniz Zeyrek, and Giedrė Oleskevicienė. 2023. [Explicitness and implicitness of discourse relations in a multilingual discourse bank](#). *Functions of Language*, 30(1):67–91.
- Telmo Mória and Evani Viotti. 2004. [Sobre a semântica das orações gerundivas adverbiais](#). In *Actas do XX Encontro da Associação Portuguesa de Linguística*, page 715–729, Lisboa. Associação Portuguesa de Linguística.
- Margarida Petter. 2009. [O continuum afro-brasileiro do português](#). *África-Brasil: caminho da língua portuguesa*.
- Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Milt-sakaki, Livio Robaldo, Aravind K. Joshi, and Bonnie L. Webber. 2008. [The Penn Discourse Treebank 2.0](#). In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*. European Language Resources Association.
- Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. [A comprehensive grammar of the English language](#). Longman, London.
- Magdaléna Rysová, Pavlína Synková, Jiří Mírovský, Eva Hajičová, Anna Nedoluzhko, Radek Ocelák, Jiří Pergler, Lucie Poláková, Veronika Scheller, Jana Zdeňková, and Šárka Zikánová. 2016. [Prague discourse treebank 2.0](#). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.
- Ted Sanders, Vera Demberg, Jet Hoek, Merel Scholman, Fatemeh Torabi Asr, Sandrine Zufferey, and Jacqueline Evers-Vermeul. 2021. [Unifying dimensions in coherence relations: How various annotation frameworks are related](#). *Corpus Linguistics and Linguistic Theory*, 17(1):1–71.
- Purificação Silvano, António Leal, and João Cordeiro. 2019. [Algumas reflexões sobre a classificação de orações gerundivas em português europeu](#). *Revista da Associação Portuguesa de Linguística*, 5:325–247.
- Purificação Silvano, António Leal, and João Cordeiro. 2021. [On adverbial perfect participial clauses in portuguese varieties and british english](#). In Luisa Meroni Sergio Baauw and Frank Drijckoning, editors, *Current Issues in Linguistic Theory (CILT): Selected Papers from Going Romance 32*, page Chapter 14. John Benjamins Publishing, Amsterdam.
- Manfred Stede. 2004. [The Potsdam commentary corpus](#). In *Proceedings of the Workshop on Discourse Annotation*, pages 96–102, Barcelona, Spain. Association for Computational Linguistics.
- Gregory T. Stump. 1985. [The semantic variability of absolute constructions](#). Reidel, Dordrecht.
- Bonnie Webber, Markus Egg, and Valia Kordoni. 2012. [Discourse structure and language technology](#). *Natural Language Engineering*, 18(4):437–490.
- Deniz Zeyrek, Amália Mendes, Yulia Grishina, Murathan Kurfali, Samuel Gibbon, and Maciej Ogrodniczuk. 2020. [Ted multilingual discourse bank \(TED-MDB\): a parallel corpus annotated in the pdtb style](#). *Language Resources and Evaluation*, 54:587–613.
- Deniz Zeyrek, Amália Mendes, and Murathan Kurfali. 2018. [Multilingual extension of PDTB-style annotation: The case of TED multilingual discourse bank](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

A Appendix: Crawling Algorithm

The method we followed to gather sentences from the Web and build our corpus automatically is detailed here in Algorithm 1. One

Algorithm 1 – Web Crawler for Sentence Selection

```

1: Input: websites,  $W = \{w_1, w_2, \dots, w_n\}$ .
2: sentences  $\leftarrow \emptyset$ 
3: for  $w_i \in W$  do
4:    $S_i \leftarrow \text{crawlPage}(w_i, \emptyset)$ 
5:   sentences  $\leftarrow$  sentences  $\cup S_i$ 
6: end for
7: Store(sentences)
8:
9: function CRAWLPAGE( $url, lnkMem$ )
10:   $text \leftarrow \text{selectText}(url)$ 
11:   $sent \leftarrow \text{selectSentences}(text)$ 
12:  for  $u_j \in \text{subLinks}(url)$  do
13:    if  $u_j \notin lnkMem$  then
14:       $lnkMem \leftarrow lnkMem \cup \{u_j\}$ 
15:       $S_j \leftarrow \text{crawlPage}(u_j, lnkMem)$ 
16:       $sent \leftarrow sent \cup S_j$ 
17:    end if
18:  end for
19:  return sent
20: end function

```

important particularity of this algorithm is the verification of a well-formed sentence (line 10: “ $\text{selectText}(urls)$ ”) during web-page extraction, as well as the satisfaction of the linguistic patterns (line 11: “ $\text{selectSentences}(text)$ ”) pre-defined by the user. As usual, the crawler implements a recursive search method, starting with a given base URL, e.g., `www.skynews.com` or `www.expresso.pt`, and then descends into the inner⁶ hyperlink hierarchy, avoiding endless loops and repetitive content.

⁶Considering only links pointing to resources within the base URL.

Adopting ISO 24617-8 for Discourse Relations Annotation in Polish: Challenges and Future Directions

Sebastian Żurowski

Nicolaus Copernicus University in Toruń
zurowski@umk.pl

Daniel Ziembicki

University of Warsaw
daniel.ziembicki@uw.edu.pl

Aleksandra Tomaszewska

Institute of Computer Science
Polish Academy of Sciences
a.tomaszewska@ipipan.waw.pl

Maciej Ogrodniczuk

Institute of Computer Science
Polish Academy of Sciences
m.ogrodniczuk@ipipan.waw.pl

Agata Drozd

Wrocław University of Science
and Technology
agata.drozd@pwr.edu.pl

Abstract

This paper explores a discourse relations annotation project carried out under the CLARIN-PL initiative, leveraging the ISO 24617-8 standard. The goal is to boost research interoperability and foster multilingual research. Our team of three linguist-annotators tackled the annotation of a corpus spanning several genres, including e.g., literature and press articles in the Polish language. This effort was guided by a project expert and external linguists from the CLARIN-PL language technology research infrastructure. Several significant challenges emerged during the process. Ambiguities within the ISO standard's relation categories, poorly-defined definitions for certain relation categories, and the difficulty of identifying and annotating implicit discourse relations, which lack explicit discourse connectives or signaling devices, were among the key issues. To overcome these problems, we implemented strategies such as regular team meetings, collaborative annotation forms, and preliminary revisions to the annotation scheme. This paper presents the project, the annotation process, and offers initial annotation data on the discourse relations and connectives identified within the corpus. Looking forward, we discuss potential enhancements to the process, including additional revisions to the guidelines and conclude with an overview of the project's contributions and a discussion of our future development plans.

1 Introduction

As defined in the ISO-24617-8 standard, discourse relations are the relations between situations expressed explicitly or implicitly in a dis-

course. They are vital for achieving a comprehensive understanding of discourse that goes beyond the meaning of individual sentences or clauses. Discourse relations occur between units known as arguments. These arguments possess distinct names corresponding to the specific relation connecting them (for instance, one argument is called BROAD and another SPECIFIC in a relation known as ELABORATION). Arguments may or may not be linked by a connective. Connectives can be single-word (for instance *and*) or multi-word (*not only... but*). In the ISO standard, discourse relations can be classified as explicit or implicit. Explicit relations are overtly signaled in discourse, for example, with connectives (such as e.g., *however* and *and*). These connectives serve as indicators of the underlying discourse relation, assisting the annotation process. Implicit discourse relations, which play a vital role in the project and underscore the significance of the human factor in our research, lack such explicit signaling devices yet maintain a connection between the arguments. Annotating implicit relations necessitates a meticulous examination and comprehension of the samples, relying on context and the annotator's knowledge of the world as well as the organization of discourse in a given language.

Discourse relations are pivotal to the evolution of natural language processing (NLP), and have been used to develop NLP tools such as summarization, sentiment analysis, and complex question answering (ISO 24617-8:2016, 2016)¹. To sup-

¹For a comprehensive list of other applications and the correlation between discourse relations and semantic and pragmatic relations, we recommend referring

port the development of such tools, annotated resources for discourse relations have been generated through various collaborative efforts, including international initiatives. This paper presents an ongoing annotation project conducted within the CLARIN-PL consortium². In addition to a description of the project, it presents preliminary annotation statistics as well as technical challenges associated with annotating discourse relations in Polish based on practical experience of the annotators to identify possible enhancements to the process.

The project focuses on annotating discourse relations in Polish. The main objective of the annotation is to deliver the first-ever Polish discourse parser.

The project relies on a triad of components:

- the ISO 24617 guidelines (ISO 24617-5:2014, 2014; ISO 24617-8:2016, 2016) for representation of semantic relations in discourse
- knowledge gathered through the creation of the Polish subcorpus of the TED Multilingual Discourse Bank (TED-MDB) (Zeyrek et al., 2020), and
- the data and preliminary annotation of the Polish Discourse Corpus (PDC) (Heliasz, 2017)³; see more information in Section 3.1 below.

To systematically and accurately annotate discourse relations in Polish, the project employs Inforex, a web-based annotation platform (Marcinićzuk et al., 2012, 2017; Marcinićzuk and Oleksy, 2019). The system has not been prepared specifically for this work, but has been configured to meet its objectives. Annotators undertake a sequence of tasks:

1. Initial identification of discourse connectives within the samples
2. Location and labeling of relevant arguments
3. Systematic correlation of discourse connectives with their corresponding arguments

to the complete ISO-24617-8 norm, available upon payment at <https://www.iso.org/obp/ui/#iso:std:iso:24617:-8:ed-1:v1:en>.

²<https://clarin-pl.eu/index.php/en/>

³<http://zil.ipipan.waw.pl/PolishDiscourseCorpus>

4. Naming the relations

5. Approving and marking the annotations as final

2 Annotation Schemes and Standardization Efforts

Numerous annotation frameworks (presented in Table 1) have emerged over time, each possessing unique underpinnings and methodological approaches to annotate discourse relations. Hobbs' Theory of Discourse Coherence (Hobbs, 1985) introduces a catalog of 'coherence relations' and a methodology for constructing high-level tree structures. Rhetorical Structure Theory (RST) (Mann and Thompson, 1988; Carlson et al., 2002; Taboada and Mann, 2006) views texts as hierarchical, recursive tree structures, identifying 25 distinct types of relations. The Cognitive Approach to Coherence Relations (CCR) (Sanders et al., 1992) introduces an analytical framework that segments discourse relations into four key categories. Segmented Discourse Representation Theory (SDRT) (Lascarides and Asher, 2008) connects elementary discourse units in an acyclic directed graph, accommodating nonadjacent unit linkages. Lastly, the Penn Discourse Treebank (PDTB) (Miltsakaki et al., 2004; Prasad et al., 2008) stands out for its differentiation between explicit and implicit discourse markers.

Each of the frameworks offers unique insights and methodological approaches to discourse relation annotation. The primary divergences lie in their structural foundations, e.g., tree-based versus graph-based; focal points, e.g., rhetorical intent versus explicit and implicit markers; and flexibility⁴. Given this heterogeneity of existing frameworks, the ISO 24617-8:2016 standard was introduced to address discrepancies and facilitate interoperability and, through its flexible and extensible core relations, homogenize the annotation of relations in discourse to ensure compatibility across diverse annotation frameworks (ISO 24617-8:2016, 2016). Although ISO standards are a unified endeavor for global standardization, their accessibility paradoxically falls short of being fully universal as they are not freely available. To gain access to the complete norm, it is necessary to directly purchase the standard from

⁴For a deeper exploration of the differences and nuances among these theories and inventories, see e.g., (Benamara and Taboada, 2015; Hoek et al., 2021)

Table 1: Overview of Discourse Relation Annotation Schemes

No.	Short Name	Full Name
1	Hobbs' Theory	Hobbs' Theory of Discourse Coherence
2	RST	Rhetorical Structure Theory
3	CCR	Cognitive Approach to Coherence Relations
4	SDRT	Segmented Discourse Representation Theory
5	PDTB	Penn Discourse Treebank
6	ISO 24617-8:2016	Semantic annotation framework Part 8: Semantic relations in discourse, core annotation schema

the website. However, for a comprehensive understanding of this norm, one can also refer to open-access publications (Bunt and Palmer, 2013; Bunt and Prasad, 2016).

The ISO 24617-8:2016 standard, titled "Language resource management – Semantic annotation framework (SemAF) – Part 8: Semantic relations in discourse", presents an extensive framework for annotating discourse relations within linguistic corpora (ISO 24617-8:2016, 2016). It delineates a set of universally applicable discourse relations that span multiple languages. The annotation scheme put forth by the ISO standard encompasses various types of relations that can emerge in discourse, including cause-effect relations, (e.g., CAUSE), temporal (e.g., SYNCHRONY, ASYNCHRONY), CONTRAST, ELABORATION, EXEMPLIFICATION, and more (ISO 24617-8:2016, 2016).

3 Annotation

3.1 The Dataset: Polish Discourse Corpus

The dataset used in our experiments is Polish Discourse Corpus (PDC), created in a previous, preliminary phase of the project in which discourse connectives were annotated (Heliasz and Ogrodniczuk, 2019) to investigate how they are used in different types of relations.

The PDC consists of 1745 texts retrieved from the Polish Coreference Corpus (Ogrodniczuk et al., 2015), each comprising 250–350 words, extracted from documents randomly selected from the National Corpus of Polish (Przepiórkowski et al., 2012) and following the original distribution of text genres in this corpus. The size of the resource contains approximately 496,000 tokens.

3.2 Annotation Procedure

Discourse analysis has recently played a crucial role in the field of NLP, particularly in the context of experimental approaches to text parsing, which has experienced a rapid growth (Atwell et al., 2021). However, the annotation procedure is not always carried out in an appropriate manner. Indeed, the process of annotating discourse relations is a very complex task, requiring specialized linguistic knowledge and careful work from annotators.

For the purposes of our project, a team of specialists in linguistics with annotation experience was formed, comprising three individuals: a PhD in linguistics, a doctoral candidate in linguistics, and a person with a bachelor's degree in applied linguistics. The first annotator had also worked on previous test annotations, which allowed for a preliminary assessment of the quality of discourse relation marking (Heliasz and Ogrodniczuk, 2019). Additionally, the team included an experienced PhD in linguistics who provided assistance in resolving substantive problems that arose during the annotation process. The level of education of the team corresponded sufficiently to the specificity of the task. Team meetings were held once a week, allowing for regular discussion of annotation problems and the establishment of annotation rules that went beyond the instructions provided to the annotators. Before starting the annotation process, the team received detailed instructions on how to mark discourse relations. After completing the process, the obtained results were verified by checking the accuracy of a random 20% sample of annotations. This verification was carried out by people from outside the team (professional linguists associated with the CLARIN-PL infrastructure) and did not

Table 2: The summary of ISO 24617-8 relations.

ISO 24617-8 relation and corresponding connectives	Example with relation role names
CAUSE 3566 occurrences (bo, więc, jak... to...)	REASON: Las jest także olbrzymią fabryką tlenu. / <i>The forest is also a huge oxygen factory</i> CONNECTIVE: więc / <i>so</i> RESULT: zapewnia komfort oddychania / <i>it provides respiratory comfort.</i>
CONDITION 1617 occurrences (jeśli, jeżeli, gdyby)	CONNECTIVE: Jeśli / <i>If</i> ANTECEDENT: pieniądze te dostaną, / <i>if they get this money</i> CONSEQUENT: atmosfera w placówkach szpitalnych ulegnie poprawie. / <i>the atmosphere in the hospital facilities will improve.</i>
NEGATIVE CONDITION 9 occurrences (albo... albo..., chyba że, gdyby nie)	CONSEQUENT: Mamy prawo odmówić dalszych napraw i zażądać zwrotu pieniędzy, / <i>We have the right to refuse further repairs and demand a refund</i> CONNECTIVE: chyba że / <i>unless</i> NEGATED ANTECEDENT: wada nie była istotna. / <i>the defect was not significant</i>
PURPOSE 1028 occurrences (żeby, aby, by)	CONNECTIVE: Aby / <i>In order to</i> GOAL: skorygować błędy w sposobie myślenia, / <i>correct errors in the way of thinking</i> ENABLEMENT: zacznij prowadzić wykaz codziennych zajęć. / <i>start keeping a record of daily activities.</i>
MANNER 206 occurrences (poprzez, tym samym, w taki sposób, że...)	ACHIEVEMENT: Szuka się więc sposobów, jak je poprawić, / <i>So, ways are sought to improve them</i> CONNECTIVE: między innymi poprzez / <i>among other things by</i> MEANS: kojarzenie leczenia chirurgicznego z pooperacyjną chemioterapią. / <i>associating surgical treatment with postoperative chemotherapy.</i>
CONCESSION 1376 occurrences (jednak, choć/chociaż, natomiast)	EXPECTATION-RAISER: Widzimy nieraz filmy nakręcane według wybitnego utworu, / <i>We often see movies based on an outstanding work</i> CONNECTIVE: a mimo to / <i>and yet</i> EXPECTATION-DENIER: zupełnie niepodobne, przeważnie złe. / <i>completely dissimilar, usually bad.</i>
CONTRAST 3114 occurrences (a, ale, tylko, lecz)	ARGUMENT 1: Nie stoją w pierwszym szeregu, / <i>They are not at front</i> CONNECTIVE: ale / <i>but</i> ARGUMENT 2: wykonują nieraz ciężkie i niewdzięczne zadania. / <i>they often perform hard and thankless tasks.</i>
EXCEPTION 68 occurrences (inaczej, w takim razie, przeciwnie)	REGULAR: Akcje spółki są dopuszczone do obrotu na rynku regulowanym / <i>The company's shares are admitted to trading on a regulated market.</i> CONNECTIVE: za wyjątkiem / <i>except for</i> EXCLUSION: art. 8 ust. 3. / <i>Article 8(3).</i>

Table 2: The summary of ISO 24617-8 relations (continued).

ISO 24617-8 relation and corresponding connectives	Example with relation role names
SIMILARITY 278 occurrences (jeszcze, również, podobnie jak)	ARGUMENT 1: <i>Koty nie lubią pływać. / Cats don't like to swim</i> ARGUMENT 2: <i>Mają / They</i> CONNECTIVE: <i>też / also</i> ARGUMENT 2: <i>problemy ze zmianą miejsca zamieszkania. / have problems with changing their place of residence.</i> ⁵
SUBSTITUTION 451 occurrences (raczej/raczej niż, wobec tego, zamiast)	FAVOURED-ALTERNATIVE: <i>Powinna przecież promieniować światłem trwałym, / After all, it should radiate with permanent light</i> CONNECTIVE: <i>zamiast / instead of</i> DISFAVOURED-ALTERNATIVE: <i>urządzać jednorazowe fajerwerki. / organizing one-time fireworks.</i>
CONJUNCTION 17437 occurrences (i, też/także, oraz)	ARGUMENT 1: <i>Czytali gazety / They were reading newspapers</i> CONNECTIVE: <i>i / and</i> ARGUMENT 2: <i>książki. / books.</i>
DISJUNCTION 1665 occurrences (czy, lub, albo)	ARGUMENT 1: <i>Opuszczają pokój, w którym jest telewizor / They leave the room with the TV</i> CONNECTIVE: <i>lub / or</i> ARGUMENT 2: <i>przełączają kanał. / switch TV channels.</i>
EXEMPLIFICATION 609 occurrences (na przykład, jak choćby, między innymi)	SET: <i>Ksiądz ma prawo również do odpoczynku / The priest also has the right to rest</i> CONNECTIVE: <i>i np. / and, for instance,</i> INSTANCE: <i>wyjechać sobie w którąś sobotę na narty. / go skiing on some Saturday.</i>
ELABORATION 509 occurrences (właśnie, w szczególności, przede wszystkim)	BROAD: <i>Bergson był obiektem licznych ataków, / Bergson was the subject of numerous attacks,</i> CONNECTIVE: <i>w szczególności / especially</i> SPECIFIC: <i>po ogłoszeniu Ewolucji twórczej / after announcing Creative Evolution.</i>
RESTATEMENT 210 occurrences (czyli, to jest, inaczej mówiąc)	ARGUMENT 1: <i>Gdy klient nie miał już pieniędzy i przypomniał sobie o polisie, dowiadywał się w siedzibie towarzystwa o tak zwanym współczynniku wartości wykupu polisy. / When the customer had no more money and remembered the policy, he would learn at the company's headquarters about the so-called policy surrender value coefficient.</i> CONNECTIVE: <i>Innymi słowy, / In other words</i> ARGUMENT 2: <i>nie dostawał tego co wpłacił. / he did not receive what he had paid.</i>
SYNCHRONY 1092 occurrences (gdy, kiedy, tymczasem)	ARGUMENT 1: <i>W tym czasie siedzieli w oddzielnej sali / At this time, they were sitting in a separate room</i> CONNECTIVE: <i>i / and</i> ARGUMENT 2: <i>czytali gazetę. / reading a newspaper.</i>
ASYNCHRONY 2157 occurrences (aż, wreszcie, skoro)	BEFORE: <i>Córki upieką ciasta. / The daughters will bake cakes.</i> CONNECTIVE: <i>Potem / Then</i> AFTER: <i>przyjdzie czas na prezenty. / it will be time for presents.</i>

⁴ Split argument occurs when connective is interjected in the argument content.

Table 2: The summary of ISO 24617-8 relations (continued).

ISO 24617-8 relation and corresponding connectives	Example with relation role names
EXPANSION 56 occurrences	NARRATIVE: Uparła się, żebym poszedł na studia... / <i>She insisted that I go to college</i> EXPANDER: W czasie okupacji bardzo się narażała, żeby mnie uratować... / <i>During the occupation, she put herself in great danger to save me...</i>
EVALUATION 46 occurrences	SITUATION: Niewolników kazał wysłać do wiejskich ergastulów, / <i>He ordered the slaves to be sent to rural prisons</i> JUDGEMENT: co było karą straszniejszą niemal od śmierci. / <i>which was almost worse than death.</i>
FUNCTIONAL DEPENDENCE 86 occurrences	ANTECEDENT-ACT: — No jak, odpowiada wam? / <i>So, are you satisfied?</i> DEPENDENT-ACT: — Owszem, odpowiada. / <i>Yes, we are.</i>
FEEDBACK DEPENDENCE 6 occurrences	FEEDBACK-SCOPE: — A nasze dzieci są inne. / <i>But our children are different.</i> FEEDBACK-ACT: — Tak, one są inne. / <i>Yes, they are different.</i>

involve making changes to the annotations in the application, but consisted of providing feedback to the annotators, who were able to review the indicated samples again and possibly revise their original selection.

3.3 Inforex

The annotation process, outlined in 3.2, was executed using Inforex. Inforex⁶ is an online platform for constructing text corpora, developed as an integral part of the CLARIN-PL infrastructure (Marciniuk et al., 2012, 2017; Marciniuk and Oleksy, 2019). It allows parallel online access and resource sharing among multiple users. The system assists semantic annotation of texts on several levels, such as marking text references and marking word senses. It also allows for the flexible definition of custom sets of tags and relations to accommodate specific requirements. In our task, we defined a new set of discourse relations in Inforex according to the ISO standard. Importantly, Inforex is language-independent, making it relatively straightforward to replicate the substantive and technical principles of our annotation and create comparable resources in different languages.

Figure 1 presents a view of the annotator's work window in Inforex. The different colors indicate the arguments of the different relations (blue

is PURPOSE, green is ASYNCHRONY, orange is CONJUNCTION, CONTRAST or FUNCTIONAL DEPENDENCE, etc.). Numbers denote arguments of all types of all relations identified in the text numbered sequentially from the beginning of the sample. Segments highlighted in grey are connectives, which are the central elements of each relation (while it is also possible for implicit relations to exist and be labeled where the connective is not present in the text). As can be seen, Inforex allows relations to be annotated in such a way that a relation from a connective (e.g., *żeby*) is marked to the first argument of the relation (e.g., argument 11) and from the same connective to the second argument of the relation (e.g., argument 12). This is what constitutes the annotation of a single discourse relation.

3.4 Annotation Results

The annotation process offers an initial glance into the frequency of distinct discourse relations within the corpus. Initial phase statistics, as gleaned from this annotation, are detailed in Table 2. Upon initial review, certain concerns may arise due to the noticeably limited representation of certain relations. For instance, NEGATIVE CONDITION shows up in just 9 instances, while FEEDBACK DEPENDENCE is observed in a mere 6 cases. This scarcity stems from the hurdles our annotators

⁶<http://inforex-work.clarin-pl.eu>

faced when trying to apply the ISO standard definitions to the corpora samples. Identifying some of the relations within them proved to be particularly challenging. Given these circumstances, we consciously decided to sideline these problematic relations during the first phase of our work. As we kick off the second stage, our initial task will be to reevaluate and clarify definitions of discourse relations before making another attempt to recognize them within the texts. This focus includes EXPANSION and EVALUATION, in addition to the ones previously mentioned. As a result, not all relation types highlighted in Table 2 are paired with typical connectives. The assignment of specific connectives to their corresponding relationships is a task that will be addressed in the process of our ongoing analysis.

4 Using ISO Annotation Framework to Annotate Discourse Relations: Challenges

An important challenge that arises in implementing the ISO standard for annotating discourse relations is ambiguity of relation categories and unclear definitions for some of the relations. Firstly, the standard includes several relation categories that are ambiguous, making it difficult for annotators to determine which category to apply in a given context. This issue can lead to inconsistent (potentially erroneous) annotation, hindering the reliability and validity (and replicability) of research results. Secondly, some of the relation categories are not well-defined, resulting in confusion and inconsistency in the annotation process.

Thirdly, identifying and annotating implicit discourse relations also poses a challenge, although some of these relations have already been discussed in the literature ((Zikánová et al., 2019), (Demberg et al., 2019), (Hoek et al., 2018)), their labelling in the context of the ISO standard is still hampered by the lack of clear connectives/signaling devices. Accurately labeling implicit relations requires expertise and intuition on the part of the annotators, as they must rely on their knowledge of the language (especially discourse organization) and world events to identify and label these relations accurately. The following sections 4.1 and 4.2 present challenges related to distinguishing discourse and syntagmatic relations as well as discourse and semantic relations we have also encountered during the process.

4.1 Discourse Relations vs. Syntagmatic Relations

Although the syntagmatic structure of text segments has been studied quite extensively (Lüngen et al., 2010), the differences between discourse and syntagmatic relations may turn out to be much more blurred than anticipated. Syntagmatic relations exist between the elements of syntagmas and connect elements of different grammatical functions, such as predicates, subjects, complements, adjuncts, and attributes. However, they are limited to a single (simple or complex) sentence. In contrast, discourse relations can extend beyond a single sentence, linking different situations (expressed by different clauses / syntagmas) throughout the whole text, and thus making it coherent. These relations primarily indicate logical or temporal connections between situations. The challenge lies in distinguishing between a situation connected by a discourse relation and an adjunct linked to a predicate by a syntagmatic relation. Let's look at the following example:

- (1) *PL Jan kupił rower podczas dorocznego jarmarku.*
EN Jan bought the bike during the annual fair.

In cases similar to (1) annotators were not sure whether they were dealing with syntagmatic or discourse relation. This indicates that a more precise, or rather, more practical definitions of both syntagmatic and discourse relations are needed. It is possible that a lot of these relations exist alongside corresponding syntagmatic ones, but clear guidelines on how to handle them are necessary. Annotators encountered uncertainty regarding whether they should annotate discourse relations between elements such as a predicate and an adjunct within the same clause, especially when the adjunct could be interpreted as a nominalized descriptor of an independent situation.

4.2 Discourse Relations vs. Semantic Relations

Distinguishing between discourse and semantic relations can pose a challenge as the boundary between the two often appears vague and context-dependent. An example of a relation that was problematic in the annotation process is the causal relation. As we read in the ISO 24617-8 standard,

Figure 1: View of the annotator's work window in Inforex.



this relation is asymmetric, with the second argument (REASON) providing an explanation for the first argument (RESULT). Let's examine the following example from ISO 24617-8:

(2) *PL* *Być może dlatego, że wygrali, napastnicy pana Borka są bardziej wyraziści niż jego obrońcy.*

EN *Perhaps because they won, Mr. Bork's attackers come through more vividly than his defenders.*

Example 2 shows a CAUSE relation, but it could be argued that the expression *because* is a pragmatic comment that conveys the causal relation solely by its meaning. In other words, during annotation, the phenomenon that posed challenges to annotators is sometimes referred to in the literature as the 'semantic-pragmatic' distinction (Van Dijk, 1979; Miltsakaki et al., 2008).

The current annotation process allows for a preliminary overview of the frequency of individual relations in the Corpus. Table 2 presents basic statistics resulting from the first phase of annotation.

4.3 Addressing Challenges

Several solutions can be implemented to navigate the challenges encountered in adhering to the ISO standard for discourse relation annotation. First and foremost, robust teamwork and open communication between annotators and supervisors are vital to reconcile discrepancies and refine the annotation process. This would entail regular meetings and discussions, where annotators can exchange insights and pinpoint potential issues within the annotation scheme. This cooperative approach is likely to enhance the overall quality of annotations while reducing potential errors.

Secondly, to curb the subjectivity that is innate in discourse annotation tasks, double annotation and adjudication could be applied in future. This would require multiple annotators working on the same sample, with a third person, possibly a supervisor (also referred to as an 'adjudicator' or 'superannotator'), tasked with resolving any disagreements between annotators. This could serve to boost the reliability and overall quality of the annotations.

Lastly, an iterative refinement strategy can be employed to progressively enhance the annotation process. This would involve the incorporation of feedback from annotators, supervisors, and users of the annotated resources. This input, which would also encompass uncertainties and observations related to overlapping categories and challenging definitions, can then be utilized to improve the annotation guidelines, resulting in a more robust and reliable annotation scheme.

5 Towards Further Work

The annotation process has been divided into several phases, with the current phase forming a singular step within the comprehensive process. In this phase, each sample has been annotated once. Planned future phases will incorporate cross-annotation, designed to bolster data credibility and replicability. Presently, the results are under scrutiny for identification and correction of any errors or flaws.

Our annotation work has highlighted differing interpretations of relations among annotators, despite their shared expertise in the field. This variability can be partly ascribed to the broad scope of the ISO standard, which provides limited examples of sentences with distinct relations. Moreover, many phenomena observed in discourse remain relatively under-researched. Such factors can

cause annotator uncertainty, potentially impacting the quality of annotation (Hovy and Lavid, 2010; Beck et al., 2020). Yet, we anticipate persistent discrepancies among annotators in such a complex task, even with more precise annotation guidelines. This may be attributed to the inherent ambiguity and multifunctionality of many discourse relations and connectives within the text - a recognized complexity in the field (Spooren and De-gand, 2010). One interesting line of work would be to systematically gather the annotators' differing decisions and then classify these differences and possibly try to explain the reasons for the discrepancies.

The ongoing annotation phase has enabled us to identify and address potential challenges, preparing us for the subsequent round of annotation. This next phase will involve cross-annotation. Currently, we are analyzing the results to detect any errors and establish a suitable procedure for future annotation tasks.

6 Conclusions

This study represents a considerable advancement in Polish language processing, marking the successful completion of a comprehensive annotation of discourse relations. Through the course of our project, we highlighted prevalent linguistic relations which emerged as promising focal points for future investigations. The potential for optimizing annotation efficiency and quality through these findings underscores their significance.

Our exploration of the annotation process uncovered various complexities, largely attributed to the inherent subjectivity in text interpretation and the expansive remit of the ISO standard. This finding highlights the necessity of a skilled, diverse team of annotators, which is a critical factor in safeguarding data quality in linguistic research. During the project, we also navigated unique challenges related to ambiguity specific to the Polish language. One of the characteristics of the Polish language is the possible discontinuity of relational arguments. In Table 2 in the example illustrating the relation (SIMILARITY), it can be seen that argument 2 is discontinuous. Its two parts are separated by a conjunction *zaś*. There is a certain group of Polish expressions that syntactically behave in such a way that they do not need to be in front of an argument (e.g. *zaś, jeszcze, zatem*). These instances underscore the need for context-

aware annotation strategies, hinting at the future development of innovative approaches tailored to address such language-specific issues.

The paper also highlighted the theoretical distinctions between discourse, syntagmatic, and semantic relations. This observation indicates that these aspects require further exploration, which will inform future work and advance practical applications of language annotation.

Thanks to the universal recognition and global accessibility of ISO standards, the utilization of one of them in the study as an alternative to less widespread and standardized criteria significantly enhances the reliability and replicability of our findings. The only drawback is that access to the standard is not provided free of charge. However, the availability of the ISO standard in multiple languages further contributes to its broader applicability. The use of the ISO standard establishes a solid foundation for fostering cross-linguistic cooperation and strengthens the potential for future multilingual research endeavors.

In sum, our project will unveil significant insights into Polish language processing, open up promising avenues for future exploration, and lay a solid groundwork for the continuation of work in this domain. We trust that our contributions will serve as a catalyst for further research advancements and fruitful collaborations in the years to come.

Acknowledgements

The work was financed by the European Regional Development Fund as a part of the 2014–2020 Smart Growth Operational Programme, CLARIN — Common Language Resources and Technology Infrastructure, project no. POIR.04.02.00–00C002/19 and the Polish Ministry of Education and Science grant 2022/WK/09.

References

- Katherine Atwell, Junyi Jessy Li, and Malihe Alikhani. 2021. *Where Are We in Discourse Relation Recognition?* In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 314–325, Singapore and online. Association for Computational Linguistics.
- Christin Beck, Hannah Booth, Mennatallah El-Assady, and Miriam Butt. 2020. *Representation Problems in Linguistic Annotations: Ambiguity, Variation, Uncertainty, Error and Bias*. In *Proceedings of the 14th Linguistic Annotation Workshop*, pages 60–73,

- Barcelona, Spain. Association for Computational Linguistics.
- Farah Benamara and Maite Taboada. 2015. [Mapping Different Rhetorical Relation Annotations: A Proposal](#). In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 147–152, Denver, Colorado. Association for Computational Linguistics.
- Harry Bunt and Martha Palmer. 2013. [Conceptual and Representational Choices in Defining an ISO standard for Semantic Role Annotation](#). In *Proceedings Ninth Joint ISO–ACL SIGSEM Workshop on Interoperable Semantic Annotation (ISA-9)*, pages 41–50, Potsdam.
- Harry Bunt and Rashmi Prasad. 2016. [ISO DR-Core \(ISO 24617-8\): Core Concepts for the Annotation of Discourse Relations](#). In *Proceedings 12th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-11)*, pages 45–54, Portoroz, Slovenia.
- Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2002. [RST Discourse Treebank](#). LDC catalog number: LDC2002T07.
- Vera Demberg, Merel CJ Scholman, and Fatemeh Torabi Asr. 2019. [How compatible are our discourse annotation frameworks? Insights from mapping RST-DT and PDTB annotations](#). *Dialogue & Discourse*, 10(1):87–135.
- Celina Heliasz. 2017. Projekt schematu badań nad relacjami (meta)tekstowymi i narzędzia do ich analizy i opisu. Technical report, CLARIN-PL.
- Celina Heliasz and Maciej Ogrodniczuk. 2019. [Eksplicitność a implicytność w świetle analizy korpusowej \(meta\)tekstu](#). *Linguistica Copernicana*, 16:75–100.
- Jerry R. Hobbs. 1985. [On the coherence and structure of discourse](#). Technical Report No. CSLI-85-37, Center for the Study of Language and Information, Stanford University.
- Jet Hoek, Jacqueline Evers-Vermeul, and Ted J. M. Sanders. 2018. [Segmenting discourse: Incorporating interpretation into segmentation?](#) *Corpus Linguistics and Linguistic Theory*, 14(2):357–386.
- Jet Hoek, Merel Scholman, and Ted J. M. Sanders. 2021. [Is there less agreement when the discourse is underspecified?](#) In *Proceedings of the Integrating Perspectives on Discourse Annotation (DiscAnn) Workshop*, University of Tübingen, Germany.
- Eduard Hovy and Julia Lavid. 2010. [Towards a ‘science’ of corpus annotation: a new methodological challenge for corpus linguistics](#). *International Journal of Translation*, 22(1):13–36.
- ISO 24617-5:2014. 2014. [Language resource management – Semantic annotation framework \(SemAF\) – Part 5: Discourse structure \(SemAF-DS\)](#). International Organization for Standardization.
- ISO 24617-8:2016. 2016. [Language resource management – Semantic annotation framework \(SemAF\) – Part 8: Semantic relations in discourse, core annotation schema \(DR-core\)](#). International Organization for Standardization.
- Alex Lascarides and Nicholas Asher. 2008. [Segmented Discourse Representation Theory: Dynamic Semantics With Discourse Structure](#), volume 83 of *Computing Meaning. Studies in Linguistics and Philosophy*, pages 87–124. Springer Netherlands, Dordrecht.
- Harald Lüngen, Maja Bärenfänger, Mirco Hilbert, Henning Lobin, and Csilla Puskás. 2010. [Discourse Relations and Document Structure](#), volume 41 of *Linguistic Modeling of Information and Markup Languages: Contributions to Language Technology*, pages 97–123. Springer Netherlands, Dordrecht.
- William C. Mann and Sandra A. Thompson. 1988. [Rhetorical Structure Theory: Toward a functional theory of text organization](#). *Text — Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.
- Michał Marcińczuk and Marcin Oleksy. 2019. [Inforex — a collaborative system for text corpora annotation and analysis goes open](#). In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 711–719, Varna, Bulgaria. INCOMA Ltd.
- Michał Marcińczuk, Jan Kocoń, and Bartosz Broda. 2012. [Inforex – a web-based tool for text corpus management and semantic annotation](#). In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, pages 224–230, Istanbul, Turkey. European Language Resources Association (ELRA).
- Michał Marcińczuk, Marcin Oleksy, and Jan Kocoń. 2017. [Inforex — a collaborative system for text corpora annotation and analysis](#). In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2017)*, pages 473–482, Varna, Bulgaria. INCOMA Ltd.
- Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2004. [The Penn Discourse Treebank](#). In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 2237–2240, Lisbon, Portugal. European Language Resources Association (ELRA).
- Eleni Miltsakaki, Livio Robaldo, Alan Lee, and Aravind Joshi. 2008. [Sense Annotation in the Penn Discourse Treebank](#). In *Proceedings of the 9th International Conference: Computational Linguistics and Intelligent Text Processing (CICLing 2008)*, pages 275–286. Springer.
- Maciej Ogrodniczuk, Katarzyna Głowińska, Mateusz Kopeć, Agata Savary, and Magdalena Zawisławska. 2015. [Coreference in Polish: Annotation, Resolution and Evaluation](#). Walter De Gruyter.

- Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Milt-sakaki, Livio Robaldo, Aravind K. Joshi, and Bonnie L. Webber. 2008. [The Penn Discourse Tree-Bank 2.0](#). In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, pages 2961–2968, Marrakech, Morocco. European Language Resources Association (ELRA).
- Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk, editors. 2012. *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, Warszawa.
- Ted Sanders, Wilbert Spooren, and Leo G. M. Noordman. 1992. [Toward a taxonomy of coherence relations](#). *Discourse Processes*, 15(1):1–35.
- Wilbert Spooren and Liesbeth Degand. 2010. [Coding coherence relations: Reliability and validity](#). *Corpus Linguistics and Linguistic Theory*, 6(2):241–266.
- Maite Taboada and William C. Mann. 2006. [Applications of Rhetorical Structure Theory](#). *Discourse Studies*, 8(4):567–588.
- Teun A Van Dijk. 1979. [Pragmatic connectives](#). *Journal of Pragmatics*, 3(5):447–456.
- Deniz Zeyrek, Amália Mendes, Yulia Grishina, Murathan Kurfalı, Samuel Gibbon, and Maciej Ogrodniczuk. 2020. [TED Multilingual Discourse Bank \(TED-MDB\): A parallel corpus annotated in the PDTB style](#). *Language Resources and Evaluation*, 54(2):587–613.
- Šárka Zikánová, Jiří Mírovský, and Pavlína Synková. 2019. [Explicit and Implicit Discourse Relations in the Prague Discourse Treebank](#). In *Proceedings of 22nd International Conference Text, Speech, and Dialogue (TSD 2019)*, volume 11697 of *Lecture Notes in Computer Science*, pages 236–248, Cham. Springer.

An Algorithm for Pythonizing Rhetorical Structures

Andrew Potter

Computer Science & Information Systems Department

University of North Alabama

Florence, Alabama, USA

apotter1@una.edu

Abstract

Diagrams produced using Rhetorical Structure Theory can be both informative and engaging, providing insight into the properties of discourse structures and other coherence phenomena. This paper presents a deep dive into these diagrams and shows how an RST analysis can be reconceived as an emergent process. The paper describes an algorithm for transforming RST diagrams into Pythonic relational propositions and applies it to a set of RST analyses. The resulting expressions are isomorphic with RST diagrams as well as machine processable. As executable specifications of discourse structure, they support scalable applications in applied and theoretical studies. Several sample applications are presented. The transformation process itself suggests an alternative to the traditional view of rhetorical structures as recursive trees. The construction of coherence is shown to be a bottom-up synthesis, wherein discourse units combine to form relational propositions which in turn rendezvous with other relational propositions to create increasingly complex expressions until a comprehensive analysis is produced. This progressive bottom-up development of coherence is observable in the performance of the algorithm.

1 Introduction

An RST analysis is a picture of a discursive process. It shows how the elements of a text work together to support the writer's purpose. The purpose could be anything—to support the claim of an argument, to explain the result of a causal process, to bring an anecdote to a satisfying conclusion, to assure the punchline of a joke, or to solicit a donation from the reader. In a well-written text, every part plays a role, with each part

ultimately supporting the writer's intended effect. An RST analysis depicts this process, it explains how the text does what it does. A competent analysis of a well-written text is an aesthetically pleasing appreciation of the writer's mastery. This is among the strengths of RST. It is also a limitation.

Many interesting and useful things have been accomplished, thanks to RST. Among these are automated text generation, discourse parsing, summarization, machine translation, essay scoring, coherency studies, and numerous other applications. And yet it seems the diagrams that make it distinctive tend to play only a bit part in these studies. In their survey of applications of RST, for example, Taboada and Mann (2006a) found they could recount the history of achievements in RST without need for any diagrams whatsoever. It is not unusual for papers on the topic to provide only a solitary diagram used solely for the purpose of conveying the core idea of what RST is. RST diagrams may be essential in explaining the theory, but thereafter tend to be treated as dispensable. This suggests that perhaps we have yet to fully leverage the concept of RST analyses as depictions of discursive processes. Hence the motivation for this research.

If we could develop a method for transforming RST diagrams into executable code, into a notation that would be machine processable, conceptually faithful to RST, human readable, and maybe even page-count friendly, from this it might be possible to develop systems that would enable us to more deeply explore what RST is, what it has to offer, and thus enable us to look directly into the diagrams, not just as stepping stones to some other research topic, but in and of RST itself. This could lead to a deeper understanding of discursive coherence, not only as conceived by Rhetorical Structure Theory, but as conceptualized in other discourse formalisms as well.

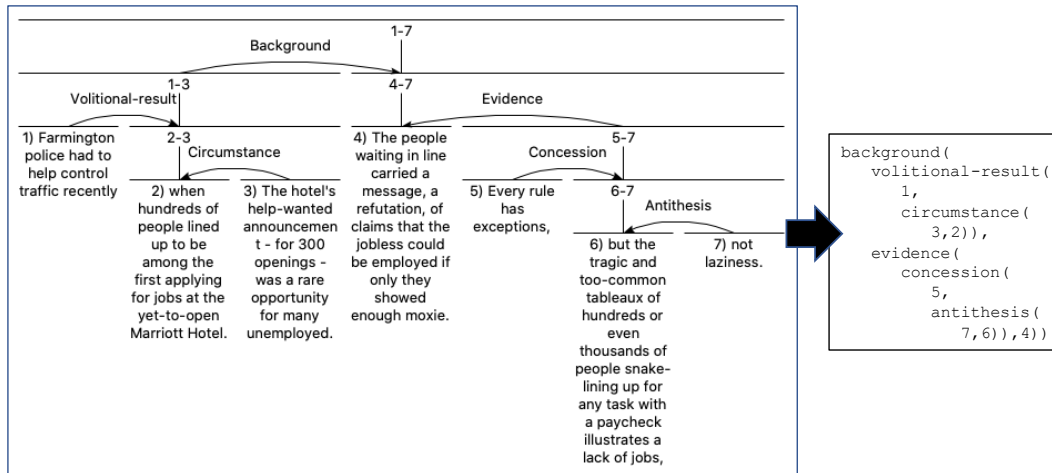


Figure 1: Pythonizing the *Not Laziness* RST analysis

The Pythonization of rhetorical structures is a process for transforming RST analyses into expressions conformant with the Python programming language, as illustrated in Figure 1. This paper describes an algorithm for making these transformations and provides direction for how these expressions can be applied to a range of research questions. I will also show how the algorithm itself sheds light on what a rhetorical structure is, how its structures come to exist, and what they mean for discursive coherence. What follows here then is a review of related literature, an overview of the motivation for developing the algorithm, and a description of the algorithm itself. This is followed by a discussion of the algorithm's potential applications and their implications. The paper concludes with a summary of the results of this study.

2 Related Research

When Rhetorical Structure Theory was originally developed by Mann and Thompson (1988) it was intended for use in automated text generation, but soon became more widely used as a descriptive theory of discourse coherence. RST is one among several theories of coherence relations; some others of note include the Penn Discourse Treebank (Webber, Prasad, Lee, & Joshi, 2019), Segmented Discourse Representation Theory (Asher & Lascarides, 2003), a taxonomic approach to coherence relations (Sanders, Spooren, & Noordman, 1992), Hobb's (1979) theory of coherence and co-reference, Polanyi's (1987) linguistic discourse model, Van Dijk's (1979) pragmatic connectives, and Grimes' (1975)

rhetorical predicates. Among the distinctive characteristics of RST are its theoretical basis and its diagrammatic technique. Its theoretical basis posits that an analysis of a text will consist of a set of schema applications, subject to the constraints of completeness, connectedness, uniqueness, and adjacency. Mann and Thompson (1988) note that the first three of these constraints are sufficient to require RST analyses to take the form of tree structures. Thus as a theory of coherence relations, RST is not limited to identifying relation pairs, but provides comprehensive specifications of the functional organization of complete texts. This in turn is reflected in the RST diagramming technique, which provides a tree-shaped rendering of the organization of the analyzed text.

During its history RST has gone through several adjustments beyond the original version (Mann & Thompson, 1987, 1988), with various extensions and adaptations (Mann & Taboada, 2005; Taboada & Mann, 2006b). Carlson and Marcu (2001) extended RST with additional relations and a somewhat different approach, putting greater emphasis on syntactic devices, with the aim of increasing analytical efficiency and scalability. The annotation guidelines defined by Stede, Taboada, and Das (2017) adhere closely to those of Mann and Thompson, with minor variations.

Relational propositions, developed by Mann and Thompson (1986) prior to and concurrently with their development of RST, are propositional analogs to RST structures, with relations being expressed as implicit assertions occurring between clauses. Mann and Thompson (2000) confined their analysis of relational propositions to discourse unit pairs, and declined to apply it to more complex

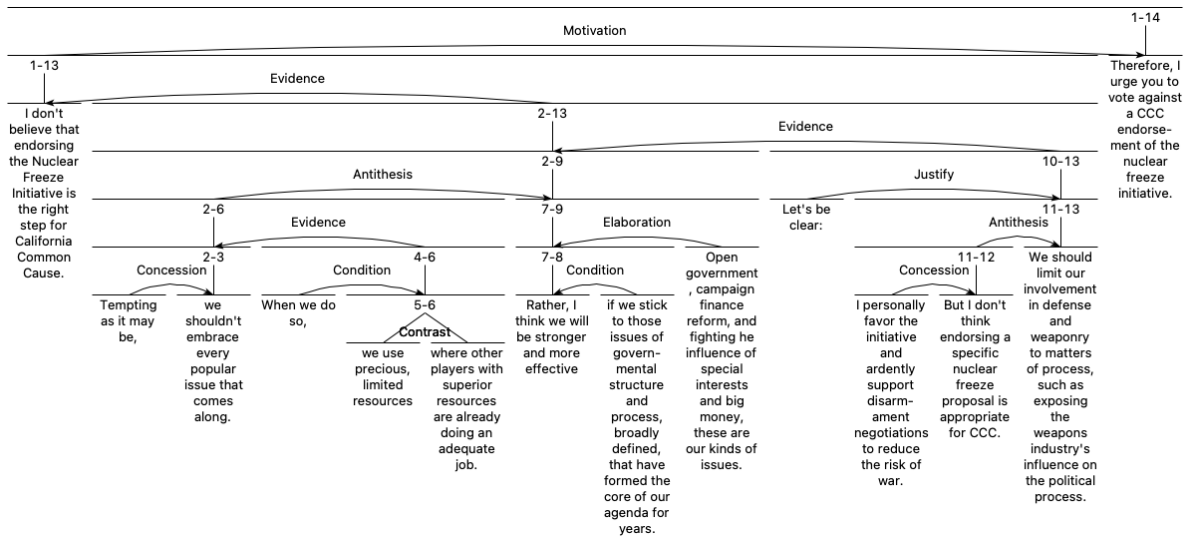


Figure 2: The *Common Cause* Analysis (Thompson & Mann, 1987)

expressions. Potter (2018) developed a notation for nested relational propositions, enabling the restatement of complete RST analyses as relational propositions. That this notation is syntactically Pythonic is fundamental to the algorithmization of RST as described in this paper.

Several tools have been developed for creating RST analyses. Among the more widely used of these are RST Tool, developed by O'Donnell (1997) and more recently rstWeb from Zeldes (2016). RST Tool is a multiplatform graphical interface for RST mark up. rstWeb is a browser-based tool developed for RST and other discourse relational formalisms. It enables annotators to work online using a browser. Both server and local versions are available. Both RST Tool and rstWeb store or export RST analyses in a common XML format.

3 Theoretical Framework

RST analyses and their respective relational propositions are structurally and semantically isomorphic, enabling transformation from one representation to the other. The interest here is in providing an automated means for transforming RST analyses into relational propositions. The motivation for doing so should be clear: while RST presents organizational properties of a text as diagrams, relational propositions present identical information in functional form. The predicates of the relational propositions may be defined as Python functions. Through transformation, the

RST diagram is redefined as an Pythonic expression. Once a diagram has been transformed, it can be supported by a set of functions implementing each of the relational predicates. That is, their implementation consists in defining a set of corresponding functions. These definitions are application specific, and dependent upon the research objective. The possibilities are open-ended. Several examples are provided in Section 5.

4 Pythonizing Rhetorical Structures

The algorithm uses an RST-Tool XML file as input and generates a Pythonic relational proposition as output. While not rocket science, its behavior has yielded some interesting observations concerning the process of discourse coherence. Therefore, a look at how the algorithm works is worthwhile. (Only the core algorithm is presented here; the complete code is being made available as an open-source project.)¹

Processing initiates at the top of the RST structure and descends recursively down each branch to the elementary discourse units. From there it constructs the leaf relational propositions and works its way back up through the structure, building the relational proposition as it goes.

Nesting structures are discovered as *span relations*. While RST-Tool uses these spans, or vertical bars, to cue visual indicators of structural subordination, for transformation they are treated as precedence operators. A span takes precedence over its satellites. So, for example, in Figure 3, the

¹ <https://github.com/anpotter/pycrst>

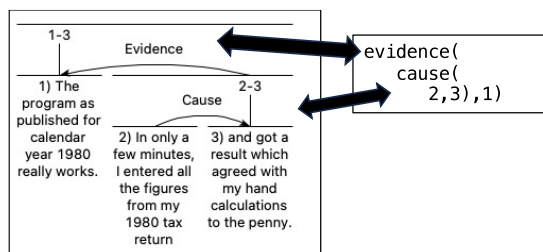


Figure 3: Span Relations as Precedence Operators

span identified as 2-3 is nested within 1-3, and therefore takes precedence over the outer span, thus defining the order of evaluation.

The core function for the transformation is simple. When called, it is passed a relational proposition object:

```
class RelProp:
def __init__(self,rel,sat,nuc,type,text):
self.rel = rel
self.sat = sat
self.nuc = nuc
self.type = type
self.text = text.strip() if text else ""
```

The algorithm's first order of business is to determine whether the relational proposition is the top span of the RST structure. If so, it simply steps down one level into the tree and makes a recursive call to the span's satellite:

```
def gen_exp(rp):
if is_top(rp) and is_span_type(rp):
return gen_exp(get_nuc(rp.sat))
```

This initiates a series of recursive calls as the function works its way down into the structure. With each call the function checks to determine whether the relational proposition under consideration is of type span. If so, it retrieves the span's satellites. If there is more than one satellite related to the span, the converge function is called to specify a convergence relation among the satellites with respect to the span:

```
elif is_span_type(rp):
if get_sat_count(rp) > 1:
exp = converge(rp)
```

When there is only one satellite, the algorithm determines whether the proposition is multinuclear. If it is, the algorithm makes a recursive call to itself for multinuclear handling. It then links the satellite to the relational proposition. Otherwise, it makes a recursive call to the satellite and links the returned value to the span's child structure. If the span has

no satellite, the satellite formats the proposition using its child structure as satellite and returns the expression:

```
else:
nuc_exp = gen_exp(get_span_nuc(rp))
sat = get_sat(rp)
if sat:
if is_multi_type(sat):
sat.nuc = nuc_exp
exp = format_rp( sat.rel,
gen_exp(sat),nuc_exp)
else:
sat_rp = get_span_nuc(sat)
if sat_rp:
sat.sat=gen_exp(sat_rp)
exp = format_rp(
sat.rel,sat.sat,nuc_exp)
else:
exp=format_rp(rp.rel,nuc_exp,rp.nuc)
```

If the relational proposition is not of type span, then it must be either a segment or a multinuclear. If it is of type segment, the algorithm first checks to determine whether it has multiple satellites, and if so, it calls the converge function to perform special handling. Otherwise, the algorithm determines whether any satellites linked to the segment are multinuclear, and makes recursive calls as needed to format the relational proposition, returning that to the caller:

```
elif is_segment(rp):
if get_sat_count(rp) > 1:
exp = converge(rp)
else:
sat = get_sat(rp)
if not sat:
exp = format_rp(rp)
elif is_multi_type(sat):
exp = format_rp(sat.rel,
gen_exp(sat), rp.sat)
else:
exp = gen_exp(sat)
```

If the relational proposition is multinuclear, the algorithm makes recursive calls for each of its nuclei and formats the results. It then determines whether the multinuclear relation has satellites, and if so, performs a convergence operation similar to that performed on the span and segment types.

For each type, the resulting expression is returned to the calling code. That is the core algorithm. It has tested successfully for 265 RST analyses including the GUM Corpus (Zeldes, 2017), the STS-Corpus (Potter, 2023), as well as a miscellany of analyses from the RST literature. Many of the analyses transformed are well over 100 units in length.

Because nesting of an expression reflects the depth of its RST structure, relational propositions can be difficult to read, so a pretty-printer was developed for post-processing. Test functions are provided to assure unit continuity and span handling. Here is the generated expression for Thompson and Mann's (1987) *Common Cause* analysis, shown in Figure 2, transformed and prettified:

```

motivation(
  evidence(
    evidence(
      justify(
        10,
        antithesis(
          concession(
            11,12),13)),
        antithesis(
          evidence(
            condition(
              4,
              contrast(
                5,6)),
            concession(
              2,3)),
          elaboration(
            9,
            condition(
              8,7))))),1),14)

```

Formatted as such, the satellite nucleus pairs align beneath their enclosing relations, and the structural depth of the discourse is indented from left to right. Multiple levels of evidence support unit 1, which then is used to provide motivation for unit 14. The relational proposition shows the rhetorical organization of the text, but unlike the diagram it does not reflect the linearity of the discourse. A relational proposition is an abstract expression of a coherence process as reenacted by the algorithm.

5 Applying Pythonized Rhetorical Structures

An application of a relational proposition consists of a set of functions that implement the relational predicates appearing in the proposition. If, for example, a relational proposition uses *evidence* and *antithesis*, the applications must provide functions by those names. The processing performed by the functions is application specific. If an application is used simply to tabulate data about a relation proposition(s), the functions may be very simple. However, the nesting of the relational propositions defines their precedence, with each nested proposition's return values being passed to its

parent. Reusable functions allow relational propositions to be treated as plug-ins within a framework. Moderately sized bulk processing can be configured by storing relational propositions as string data in Python dictionaries for runtime evaluation as Python code.

Some but not all applications are precedence sensitive. Precedence sensitive applications rely on the logic implicit in the nesting of relational propositions. For example, an application designed for a study in argument accrual may need to backtrack through discourse threads when a structural convergence is encountered. This could be used to determine the relation types of the accruing threads.

The following examples illustrate how relational propositions can be used. The first is a simple framework for measuring the frequency of argumentative relations as identified by Azar (1999). The purpose of this example is to show how readily Pythonic representations of RST analyses can be outfitted for practical applications. The second example performs an automated reduction of relational propositions to logic and then uses the logic to support examination of purported simultaneous RST analyses. The third example shows how runtime evaluations of relational propositions can be used to reenact coherence development in a discourse.

5.1 Computing an RST Metric

Using relational propositions as code requires a set of functions corresponding to the relations used in the relational proposition. Here is a set of functions for determining the Azar Score for the relations used in Thompson and Mann's (1987) *Common Cause* analysis:

```

def antithesis(*argv): return tally(argv), argv
def concession(*argv): return tally(argv), argv
def evidence(*argv): return tally(argv), argv
def motivation(*argv): return tally(argv), argv
def justify(*argv): return tally(argv), argv
def condition(*argv): return tally(argv), argv
def contrast(*argv): return tally(argv), argv
def elaboration(*argv): return tally(argv), argv
def condition(*argv): return tally(argv), argv

```

This list can be extended to include an entire RST relation set. Since every relation receives the same processing, they all call the same function:

```

def tally(argv):
    relname = sys._getframe(1).f_code.co_name
    argumentative() if relname in arg_rels \

```


$\rightarrow 6)) \rightarrow (((7 \vee 6) \wedge \neg 7) \rightarrow 6)) \wedge \neg(5 \rightarrow \neg(((7 \vee 6) \wedge \neg 7) \rightarrow 6))) \rightarrow (((7 \vee 6) \wedge \neg 7) \rightarrow 6))) \rightarrow 4)$

nucleus of the relation and unit 2, the satellite. So the relational proposition is now volitional-

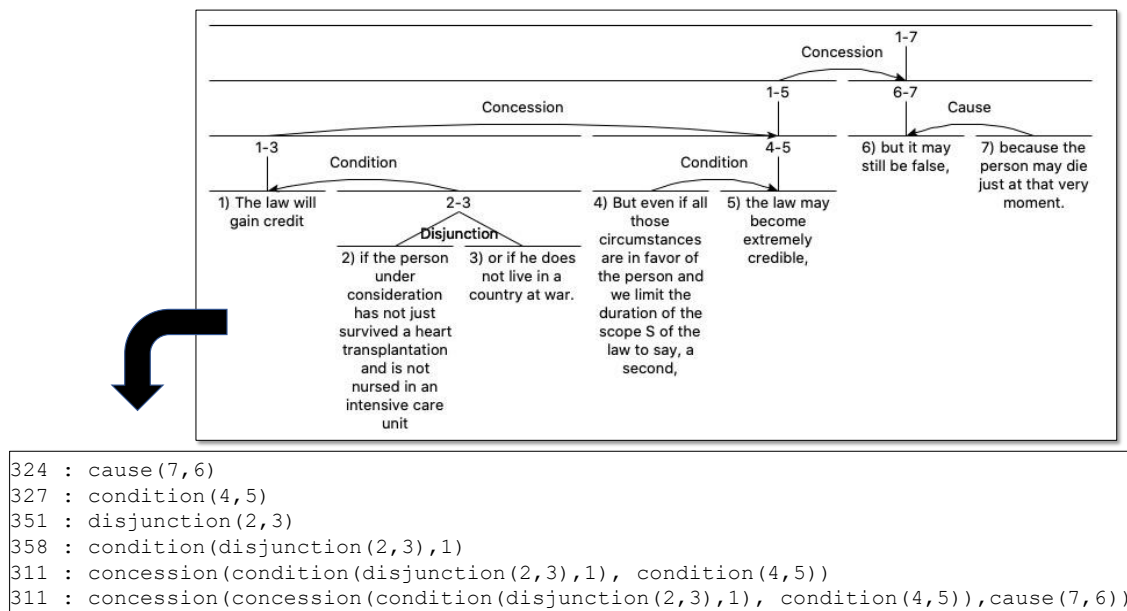


Figure 4: Reenacting the *Heart Transplant Analysis*

Potter (2018) claimed any text analyzable using RST could be reduced to propositional logic. The method described here shows the process can be fully automated. The results can be used to support fine-grained examination of RST analyses. For example, in their 1992 paper, Moore and Pollack argued that there are obvious cases where both presentational and subject matter analyses can be made of the same text. They based their claim on several examples. Here is the text of their first example:

- 1) George Bush supports big business.
- 2) He's sure to veto House Bill 1711.

Moore and Pollack say it is plausible that there is an EVIDENCE relation between unit 2, as nucleus of the relation, and unit 1, the satellite. So the relational proposition is evidence(1,2). The intended effect of EVIDENCE is that the satellite increases the reader's belief in the nucleus. For this to hold, it would therefore be necessary that the reader already believe in the satellite, since it is an assumption of the argument. The logical reduction of the relational proposition echoes this, showing unit 2 as inferred from unit 1: $((1 \rightarrow 2) \wedge 1) \rightarrow 2$.

In their second analysis of the same example, Moore and Pollack say that it is plausible that there is a VOLITIONAL-CAUSE relation between unit 1, as

cause(2,1), such that unit 2 provides a causal explanation for unit 1. As such, George Bush's support for the bill supports the inference that he supports big business: $((2 \rightarrow 1) \wedge 2) \rightarrow 1$. So in one analysis, 1 is inferred from 2, and in the other, 2 is inferred from 1. This does not affirm that multiple analyses must be supported, but rather that there are two quite different readings of the text. And once we allow arbitrary assumptions necessary for multiple decontextualized readings, all bets are off as to the correct analysis. For all we know, the bill might have been something strongly disfavored by big business, but that President Bush intended to support it anyway, making the relation between the two units CONCESSION. Similar issues arise with Moore and Pollack's second example:

- 1) Come home by 5:00.
- 2) Then we can go to the hardware store before it closes.
- 3) That way we can finish the bookshelves tonight.

The first of their analyses for this example uses the MOTIVATION relation: Finishing the bookshelves motivates going to the hardware store, and taken together these motivate coming home by 5:00: motivation(motivation(3,2),1):

$(((((3 \rightarrow 2) \wedge 3) \rightarrow 2) \rightarrow 1) \wedge (((3 \rightarrow 2) \wedge 3) \rightarrow 2)) \rightarrow 1)$

The second analysis uses the CONDITION relation: coming home by 5:00 is a condition on going to the hardware store, and together these are a condition for finishing the bookshelves: condition(condition(1,2),3), or

$$((1 \rightarrow 2) \rightarrow 3)$$

For the MOTIVATION analysis to be realizable, it is necessary that the reader accept the initial premise of the relation, the bookshelves can be finished tonight. So in one case, there is a line of reasoning leading from unit 3 to unit 1, and in the other, leading from 1 to 3. Once again, the analyses are not simultaneous. Any possibility of simultaneous analysis relies on an insufficiency of information. Decontextualized, obscure, or ambiguous texts are hard to understand, and this should be expected to impede analysis. The use of semantic relations for pragmatic purposes is identified by means of a determination of purpose, and therefore there is not really an overlap at all. If there is a problem here, it is with the limiting circumstances under which the theory is applied, not with the theory itself.

5.3 Reenacting Rhetorical Structures

The transformation algorithm can be used to reenact the process of structure formation. This process initiates with the innermost relations of each branch and works its way upward. To demonstrate this, I instrumented the algorithm with debug prints and applied it to the *Heart Transplant* analysis shown above in Figure 4. As the algorithm descends into the tree it seeks the precedence, ultimately finding it in the leaves and their

relations. These low-level relational propositions are transformed first. The algorithm continues upward, constructing more complex expressions from the bottom up, until a complete relational proposition is formulated. With each relational proposition, there is a transference of intended effect from satellite to nucleus. Without the satellite-nucleus transfer, we would have merely an empty structure. The only way to a nucleus is through its satellites. But all this is at odds with the view of RST trees as recursive.

Recursion, it has been said, is pervasive in discourse, semantically, rhetorically, structurally, grammatically, and thematically (e.g., Hwang, 1989; Muhammad, 2011; Pinker & Jackendoff, 2005; Polanyi, 1988). And of rhetorical structures, it has been widely observed that not only are they tree-shaped (Bateman, 2001; Grasso, 2002; Mann & Thompson, 1988), but that the units comprising the tree are linked to one another recursively (Das & Taboada, 2018; Demberg, Asr, & Scholman, 2019; Guerini, Stock, & Zancanaro, 2004; Peldszus & Stede, 2016; Taboada & Mann, 2006b). While these observations are structurally correct, they are functionally incomplete. As the reenactment of rhetorical structures shows, RST tree structures define themselves from the bottom up. Elementary units combine to form relational propositions and these propositions rendezvous with other propositions to create increasingly complex expressions. The tree is the result of a pragmatic process. Through this process rhetorical intentionality develops.

This becomes more obvious when analyzing a nonsensical text, where the RST linkage is discernible, but the satellite-nucleus transfers fail,

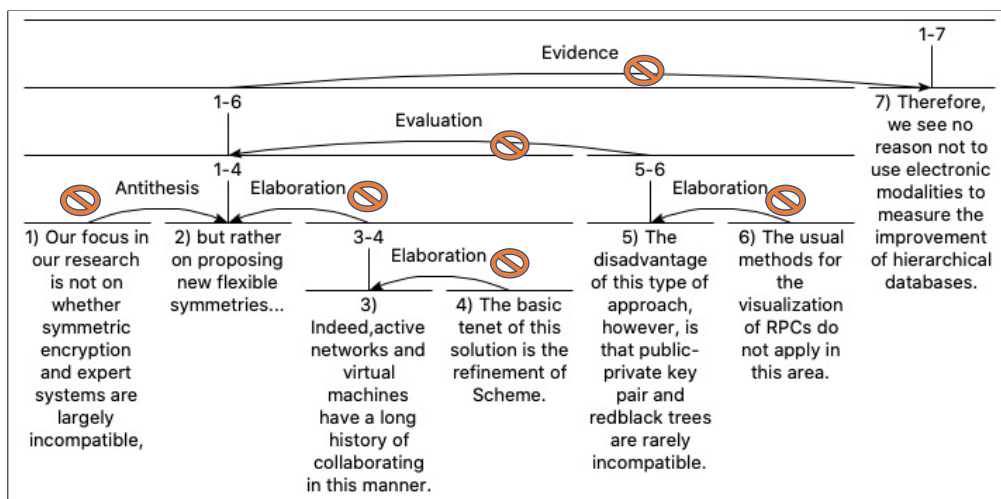


Figure 5: An Analysis of Nonsense

as shown in Figure 5. The structure is discoverable even when the intention is unachievable. Texts may be analyzable, and if so, they will be transformable and reducible, and yet at the same time nonsensical. This analysis is of a passage from a paper created using the SCIGen nonsense paper generator (Stribling, Krohn, & Aguayo, 2005). The analysis is superficially plausible, it transforms correctly, and builds up just like any other:

```
evidence(
  evaluation(
    elaboration(
      6,5),
    conjunction(
      antithesis(
        1,2),
      elaboration(
        elaboration(
          4,3,2))),7)
```

And yet the text is nonsensical. If such nonsense is analyzable, what does this say about RST? Is coherence as defined by RST merely window dressing? On the contrary, the inferences within the text, if read with attention to content, are *non sequitur* to the point of being ridiculous. The ELABORATIONS are not really elaborations, the EVALUATION is not evaluative, the EVIDENCE is not evidential. The superficiality of the analysis mirrors that of the text. For an RST analysis to be sound, the bottom-up transfer of intention from satellite to nuclei must be assured. This echoes Marcu's (2000) strong nuclearity thesis, but from a bottom-up perspective. A nucleus acquires its "strength" through its relationship with its satellite. Transference of intention upward shows that, in a coherent text, each relation subsumes its underlying structure. An RST analysis is the realization of a discursive process. The constituents of a text organize from the bottom up to realize the writer's purpose.

6 Conclusion

The algorithm presented here provides a tool for transforming RST analyses into machine processable code. As such, an RST analysis need not be regarded as an end product, but rather as a starting point for deeper investigation. Of particular interest are studies using Pythonic relational propositions to investigate threads of coherence. The algorithm is scalable to large analysis sets.

The bottom-up synthesis of relational propositions generates purely abstract renditions of coherence processes. This validates the theory of relational propositions. Relational propositions implicitly assert the intentionality between discourse units. Coherence arises out of the instantiation of these propositions, not only at the unit level but among the complex spans that bring structure to the rhetorical space. Within this space, a span is a container of an intentional effect. It is through spans that structure arises. While we may view the process from the top down, as is the tendency with RST, intentionality develops from the bottom up. The tree-structures characteristic of RST are the end-result of this process.

References

- Nicholas Asher, & Alex Lascarides. 2003. *Logics of conversation*. Cambridge, UK: Cambridge University Press.
- Moshe Azar. 1999. Argumentative text as rhetorical structure: An application of rhetorical structure theory. *Argumentation*, 13(1), 97-114.
- John A. Bateman. 2001. Between the leaves of rhetorical structure: Static and dynamic aspects of discourse organisation. *Verbum*, 23(1), 31-58.
- Lynn Carlson, & Daniel Marcu. 2001, September. Discourse tagging reference manual. Retrieved from <ftp://ftp.isi.edu/isi-pubs/tr-545.pdf>
- Debopam Das, & Maite Taboada. 2018. RST Signalling Corpus: A corpus of signals of coherence relations. *Language Resources and Evaluation*, 52(1), 149-184. doi:10.1007/s10579-017-9383-x
- Vera Demberg, Fatemeh Torabi Asr, & Merel Scholman. 2019. How compatible are our discourse annotations? Insights from mapping RST-DT and PDTB annotations.
- Floriana Grasso. 2002. Towards computational rhetoric. *Informal Logic*, 22(3), 195-229.
- Joseph E. Grimes. 1975. *The thread of discourse*. Berlin: Mouton.
- M. Guerini, O. Stock, & M. Zancanaro. 2004. Persuasive Strategies and Rhetorical Relation Selection. In *Proceedings of the ECAI Workshop on Computational Models of Natural Argument*. Valencia, Spain.
- Jerry R. Hobbs. 1979. Coherence and coreference. *Cognitive Science*, 3, 67-90.
- Shin Ja Joo Hwang. 1989. Recursion in the paragraph as a unit of discourse development.

- Discourse Processes: A Multidisciplinary Journal*, 12(4), 461-478.
- William C. Mann, & Maite Taboada. 2005, October. An introduction to rhetorical structure theory (RST). Retrieved from <http://www.sil.org/~mannb/rst/rintro99.htm>
- William C. Mann, & Sandra A. Thompson. 1986. Relational propositions in discourse. *Discourse Processes*, 9(1), 57-90.
- William C. Mann, & Sandra A. Thompson. 1987. *Rhetorical structure theory: A theory of text organization* (ISI/RS-87-190). Retrieved from Marina del Rey, CA:
- William C. Mann, & Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text - Interdisciplinary Journal for the Study of Discourse*, 8(3), 243-281.
- William C. Mann, & Sandra A. Thompson. 2000. *Toward a theory of reading between the lines: An exploration in discourse structure and implicit communication*. Paper presented at the Seventh International Pragmatics Conference, Budapest, Hungary.
- Johanna Doris Moore, & Martha E Pollack. 1992. A problem for RST: The need for multi-level discourse analysis. *Computational Linguistics*, 18(4), 527-544.
- Manaal Jassim Muhammad. 2011. *The use of Rhetorical Structure Theory in political editorials: A contrastive study of text analysis with special reference to its application as text-based generation*. University of Pune, Pune, India.
- Michael O'Donnell. 1997. RST-Tool: An RST analysis tool. In *Proceedings of the 6th European Workshop on Natural Language Generation*. Duisburg, Germany: Gerhard-Mercator University.
- Andreas Peldszus, & Manfred Stede. 2016. Rhetorical structure and argumentation structure in monologue text. In *Proceedings of the 3rd Workshop on Argument Mining* (pp. 103-112). Berlin, Germany: Association for Computational Linguistics.
- Steven Pinker, & Ray Jackendoff. 2005. The faculty of language: what's special about it? *Cognition*, 95(2), 201-236.
- Livia Polanyi. 1987. *The linguistic discourse model: Towards a formal theory of discourse structure*. Cambridge, MA: Bolt, Beranek, and Newman, Inc.
- Livia Polanyi. 1988. A formal model of the structure of discourse. *Journal of Pragmatics*, 12(5-6), 601-638.
- Andrew Potter. 2018. Reasoning between the lines: A logic of relational propositions. *Dialogue and Discourse*, 9(2), 80-110.
- Andrew Potter. 2021. Text as tautology: an exploration in inference, transitivity, and logical compression. *Text & Talk*. doi:doi:10.1515/text-2020-0230
- Andrew Potter. (2023). *STS-Corpus*. Retrieved from: <https://github.com/anpotter/STS-Corpus>
- Ted J M Sanders, W P M Spooren, & L G M Noordman. 1992. Toward a taxonomy of coherence relations. *Discourse Processes*, 15, 1-35.
- Manfred Stede, Maite Taboada, & Debopam Das. 2017. *Annotation guidelines for rhetorical structure*. Retrieved from Potsdam and Burnaby: http://www.sfu.ca/~mtaboada/docs/research/RST_Annotation_Guidelines.pdf
- Jeremy Stribling, Maxwell Krohn, & Daniel Aguayo. 2005. SCIGen - An automatic CS paper generator. Retrieved from <https://pdos.csail.mit.edu/archive/scigen/>
- Maite Taboada, & William C. Mann. 2006a. Applications of rhetorical structure theory. *Discourse Studies*, 8(4), 567-588.
- Maite Taboada, & William C. Mann. 2006b. Rhetorical structure theory: Looking back and moving ahead. *Discourse Studies*, 8(3), 423-459.
- Sandra A. Thompson. 1987. 'Concessive' as a discourse relation in expository written English. In B. Joseph & A.M. Zwickey (Eds.), *A Festschrift for Ilse Lehiste* (pp. 64-73). Columbus, Ohio: Ohio State University.
- Sandra A. Thompson, & William C. Mann. 1987. Antithesis: A study in clause combining and discourse structure. In Ross Steele & Terry Threadgold (Eds.), *Language Topics: Essays in Honour of Michael Halliday, Volume II* (pp. 359-381). Amsterdam: John Benjamins.
- Teun A. Van Dijk. 1979. Pragmatic connectives. *Journal of Pragmatics*, 447-456.
- Bonnie Webber, Rashmi Prasad, Alan Lee, & Aravind Joshi. 2019. The Penn Discourse Treebank 3.0 annotation manual.
- Amir Zeldes. 2016. rstWeb – A browser-based annotation interface for Rhetorical Structure Theory and discourse relations. In *Proceedings of NAACL-HLT 2016 (Demonstrations)* (pp. 1-5). San Diego, California: Association for Computational Linguistics.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3), 581-561.

The shaping of the narrative on migration: A corpus assisted quantitative discourse analysis of the impact of the divisive media framing of migrants in Korea

Clara Delort and EunKyoung Jo

Department of Global Korean Studies, Sogang University

35 Baekbeom-ro, Mapo-gu, Seoul 04107, South Korea

clara.delort@gmail.com - jek.cl.nlp@gmail.com

Abstract

This work explores the shaping of public opinion on migration in South Korea by utilizing BERT topic modeling (Grootendorst, 2022) which extends transformer language models to Top2Vec (Angelov, 2020) which leverages word semantic embedding to find topic vectors from documents. Data are the public discourse on Twitter and the three biggest local newspapers. The study examines the content of these topics, highlighting key themes and their implications. The findings through BERTopic modeling as a tool of discourse analysis on large data shows that, rather than a simple overall negative media narrative, the news outlets create distinctive concepts of migrants, fragmented into clustered groups, alienated from each other based on their social identities, migration status, and citizenship status. Discriminatory tropes (such as a criminalization frame and a victimization frame) predominant in the Mass Media corpus, are less salient in the New Media corpus and the Public Opinion (Tweets) corpus, where topics of compassion, human rights, union, reports of shared experiences, desire to share culture and communicate, are predominant. With the c-TF IDF formula giving the significance of words per topic, the creation of a divisive concept of refugees is visualized, with the fragmentation of one group (for example, refugees) into vastly distanced topics (either in the victimization frame, with "kid" and "refugee" in one cluster, or the criminalization frame, with "refugee" and "terrorism" in one cluster). This division in the public narrative supports the division in governmental policies. In this case, the Ministry of Justice divides asylum seekers applying for a refugee Visa into "humanitarian" or "economic" refugee categories. Asylum seekers placed in the "economic" refugee category are denied refugee status. The intertopic distance maps illustrate this shaping of divisive semantic meanings.

1 Introduction

Categorizing the recurrent topics in the public migration debate in South Korea allows us to examine the role of media in framing and depicting migrants and to understand the roots of the divisions based on social identities and citizenship status within the working class. This study's aim is to find the role of language in capitalism in shaping societal narratives and influencing perceptions by using a dynamic seeded topic modeling to categorize language data and gain insights in the discourses perpetuating capitalist structures. Scholars developed theories highlighting power dynamics, identity construction, and the importance of understanding global capitalism in the study of media representations of migrants. Stuart Hall emphasizes the role of media in constructing social hierarchies (Hall, 1997). Edward Said highlights how the media perpetuates stereotypes and exoticizes different cultures (Said, 1978). Angela McRobbie explores how media representations contribute to gendered identities and marginalize migrant women (McRobbie, 2009). Chandra Talpade Mohanty examines gendered and racialized stereotypes, including those of migrant women (Mohanty, 2003). In the context of South Korea, the three major conservative newspapers, Chosun Ilbo, JoongAng Ilbo, and DongA Ilbo, dominate the country's hard news. Smaller newspapers with varying political inclinations are also available as alternatives, but their circulation is lower. Through the discourse analysis of distinct corpora representing the mass media, the new media, and the public opinion, the role of media in the reproduction of class relations is quantitatively studied.

2 Data & Methodology

A corpus of tweets represents the public debate on migration during the 2009-2022 period. The Tweet data of migration-related Korean tweets are

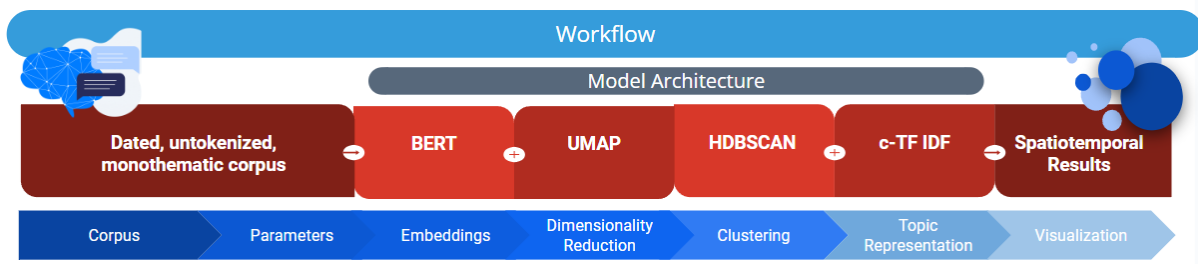


Figure 1: Workflow of the experimental design of Topic Modeling

collected using a public Twitter scraper, snsrape and tokenized using Mecab, resulting in 3 120 297 Korean tweets mentioning the following keywords: irregular immigrants, refugees, illegal immigrants, migrant workers, employment permit system, visa, migrants, immigrants, foreigners, illegal aliens, undocumented migrants, foreign workers. Only words tagged as nouns by Mecab are kept for topic modeling. Traditional methods and tools for corpus annotations such as DAMSL, DIT, RSTTool, and PDTB (Bunt, 2017) are not used. A second corpus of news articles from the local daily newspapers with the biggest daily circulation, Chosun Ilbo, Joongang Ilbo and Donga Ilbo represents the local mass media. 14 560 articles (Chosun Ilbo, $n=4,678$, Joongang Ilbo, $n=6,437$, Donga Ilbo, $n=3,445$) mentioning the same keywords are scraped. The articles were harvested over the 2009-2022 period. A third corpus of descriptions of news articles from Naver represents New Media. The Naver data were accessed using the official Naver News API and used to search for 10,338 articles mentioning the same keywords. Only the short description of each article and publishing date were obtained, as the official API limits the number of articles scraped by query to 1000 titles and the harvest to the description of the articles rather than the full text. Topic modeling algorithms are used to discover hidden semantic structures, and infer and generate coherent topics by generating contextual word and sentence vector representations. BERT (Devlin et al., 2019) is based on the encoder component of the Transformer model (Vaswani et al., 2017), which reads the text input, and uses it to then generate a language model. In addition, the Class-Based TF-IDF Procedure (Grootendorst, 2022), aggregates all the documents for each topic, to then extract the meaningful words from the entire topic. To distinguish topics from one another based on those cluster words, the class-based TF-IDF (Term Frequency - Inverse Document Frequency)

is carried out. This formula is an adaptation of the TF-IDF formula, which measures the importance of a word to a document. To obtain the importance of a word to a topic, the c-TF-IDF takes into account topic class which a document is assigned. This gives a more accurate and meaningful representation of the importance of terms within specific classes or topics, resulting in an effective topic modeling with BERTopic. Furthermore, to explore the potential hierarchical structure of the topics from the matrix created, hierarchical clustering visualization is performed. The similarity between two c-TF-IDF topics is determined by their distance, where a smaller distance indicates a higher level of similarity. In BERTopic, the merging of topics is achieved through the common linkage method “ward” (Ward Jr., 1963), or “Ward’s minimum variance method”. The formula calculates the increase in variance that would occur if two clusters were combined and compares it to the increase in variance for other potential merges. It selects the pair of clusters with the smallest increase in variance as the most similar. The tokenizer of the multilingual BERTopic model is changed to the Korean tokenizer Mecab, for a better analysis of the Korean language, and the model is fine tuned with the cleaned, dated, Korean corpus. The number of topics to extract is set to 31. In order to obtain the most coherent topics, a seeded model is performed. Seeded topic modeling is realized by giving the model a list of seed topics with keyword attributes. These guide the topic model to converge towards the topics we want to examine in the documents. However, if those topics do not exist, they will not be modeled. The detailed seed topic list is available alongside the source codes at github.com/clara1del/BERTopic-korean-tweets-newsarticles-migration-discourse. In order to integrate socio-political concepts of class struggle into the language model and combine critical discourse analysis with structural topic modeling, we design

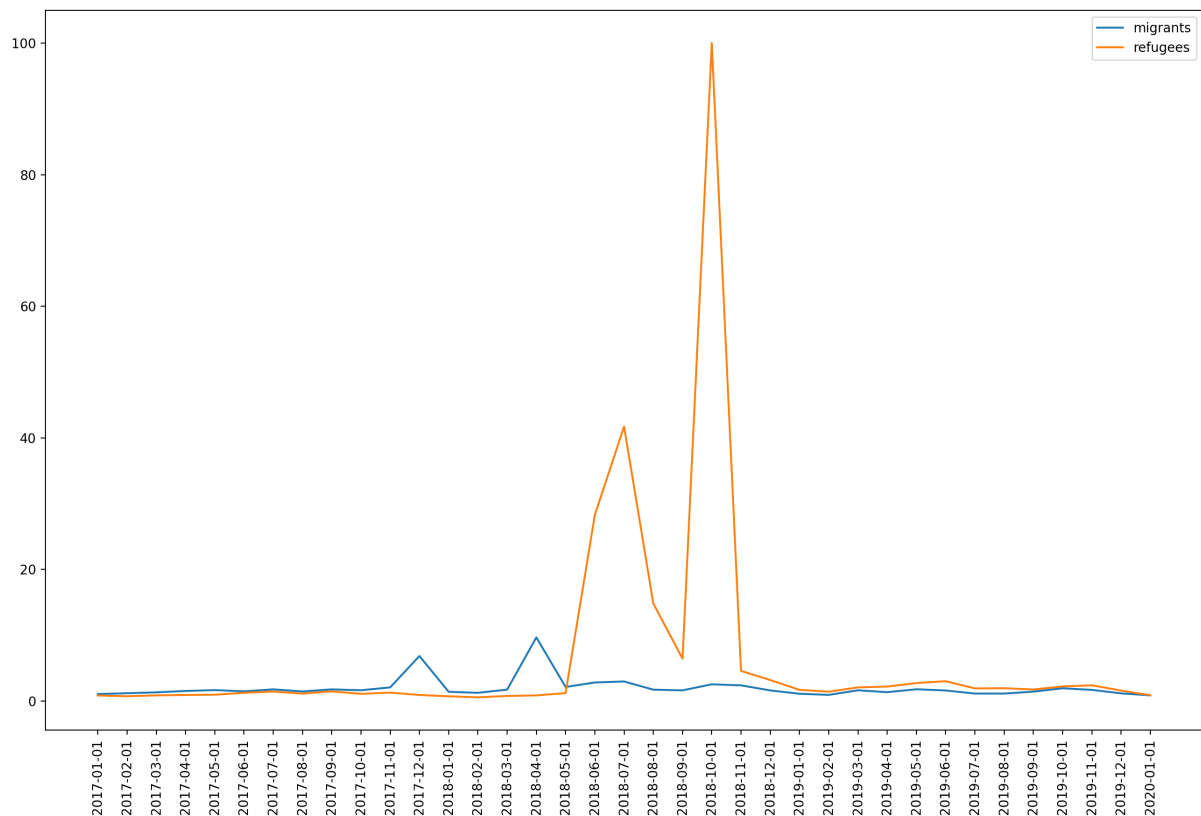


Figure 2: Trends of relative interest in “Refugees” and “Migrants” as frequency of search terms in Naver search engine

a frame of study of the migration topics, which is fed to the model as a seeded topic list. This manual guiding of topics departs from a typical non-guided topic modeling, and gives the model a deliberate perspective for a theoretically contextualized text analysis. Using the BERTopic multilingual model for topic modeling, with the MeCab tokenizer (Kudo, 2005) for the Korean language, and an added step of dynamic topic modeling, the development over time of the semantic meanings of migration related concepts in South Korea is investigated. As a quantitative method of discourse analysis, topic models offer voluminous statistical textual information, which can be used to study the structures of text in their historical and sociopolitical context. Through a study over time and a comparison between the voice of the elites and the voice of the public, we can uncover the relation between media coverage and the assumptions and values towards migrants reflected in the online discourse. Through topic modeling, the shaping of the migration narrative by the mass media, and the root of the hate on migrants is analyzed.

3 Related Work

A frame analysis based on topic modeling using LDA clustering (Pavlova and Berkers, 2022; "Galagher et al., 2017) was proposed to explore the public perception of a divisive concept. They manually defined frames and associated them with top words, which served as the basis for Latent Dirichlet Allocation clustering. This approach facilitated the identification of unique frames for discourse analysis. Building on this methodology, we adopt a similar approach by constructing a theoretical frame, a seed topic list, to extract balanced and insightful topics. A study (Nozza et al., 2022) focused on investigating language use towards specific social identities, particularly within the LGBTQIA+ community trained a model to complete sentences using LGBTQIA+ related templates and measured harmfulness scores, revealing identity-based attacks. In our work, we use another potential of the BERT model to analyze the language employed in relation to specific social identities, by studying the semantic distance between topics whose subjects are also groups of migrants defined by their social identities.

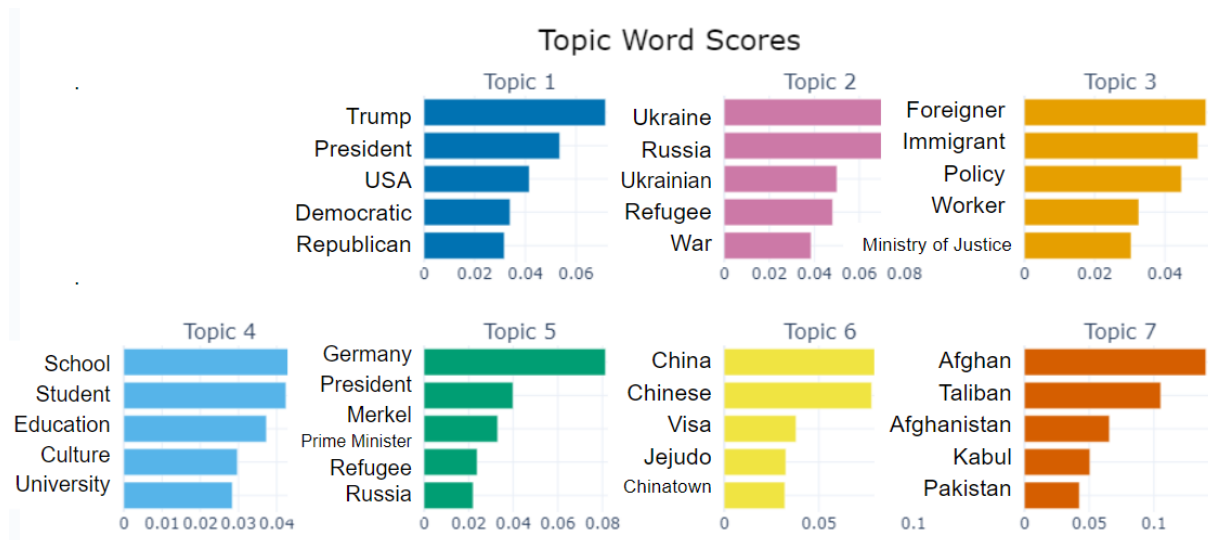


Figure 3: Barcharts of the topics in the articles harvested from Chosun Ilbo, Joongang Ilbo and Donga Ilbo

4 Analysis

Naver's search trends show frequent search terms in Naver's search engine. With 42 millions users, Naver is the most popular search engine in South Korea. Using the keyword research tool Naver Data Lab, the frequency over time of the keywords "migrants" and "refugees" searched in Naver. Figure 2 shows that a peak in relative interest in refugees in 2018 followed the arrival of asylum seekers escaping the Yemeni civil war in Jeju-do, which was heavily covered in the media, portraying the male refugees as dangerous. The public opinion of refugees worsened to the point of the organization of protests to oppose the acceptance of the asylum seekers. Figure 3 is the topic modeling of the corpus of news articles harvested from three major newspapers, Chosun Ilbo, Joongang Ilbo and Donga Ilbo, and shows distinct noteworthy frames. The top left most theme (topic 1, $n = 304$ articles) in figure 3 shows a focus on Western conservative views on migration, which are reproduced with representative keywords of the topic being: Trump, President, America, Democratic, Party, Republican, Candidate, Election, Biden, House, Congressman, Senate, illegal, immigration, Government, Exile, Minister. The third most predominant topic (topic 3, $n = 210$ articles) presents a criminalization framework of foreign workers, with the following keywords: Foreigners, Immigration, Policy, Workers, Ministry of Justice, Sojourn, Immigration policies, Employment, Expansion, Manpower, Government, Country, Budget, Population, Visa, Immigration, Employment, Illegal, Libya. The topic

describes foreign workers, but not their work conditions. Rather, "Ministry of Justice", "illegal", "Sojourn", "Visa", show a focus on their legal status. This criminalization framework is also found in topic 6 ($n = 161$ articles), with the following representative keywords: China, Visa, Jeju, Taiwan, Hong Kong, Lithuania, Italy, Smuggling, Government, Foreigner, illegal stay. The strong association between migrants and crime forms a negative sentiment. This main criminalization framework is present in all topics describing migrant workers. Topic 10 ($n = 78$) describes migrant workers, and associates them with the "illegal" term. The keywords for topic 10 are: Food, Seasons, Farmers, Vietnam, Labour, Workers, Farming, Corona, Illegal Stay, Grains, Rising, Entry, Potato. The strong association of "illegal" with even the migrants providing the country with provisions of food illustrates how criminalizing migrant workers allows for them to be exploited by the government without public outrage and resistance. Several topics describe refugees with a strong Islamophobic association with terrorism. Topic 7 ($n = 94$ articles), which describes refugees and topic 8 ($n = 87$ articles), which describes terrorism, are overlapping. The keywords for topic 7 are: Afghanistan, Taliban, Pakistan, Kabul, Refugees, Islam, Humanitarianism, US Army, Reign, Escape, Government, Stay, problem. And for topic 8 are: terror, Islam, France, refugees, Middle East, forces, Muslim, Italy, Paris, Syria, Western Country, Al Qaeda, Bomb, Religion, War. This high coverage of terrorism in the local mass media promotes a fear of terrorism in

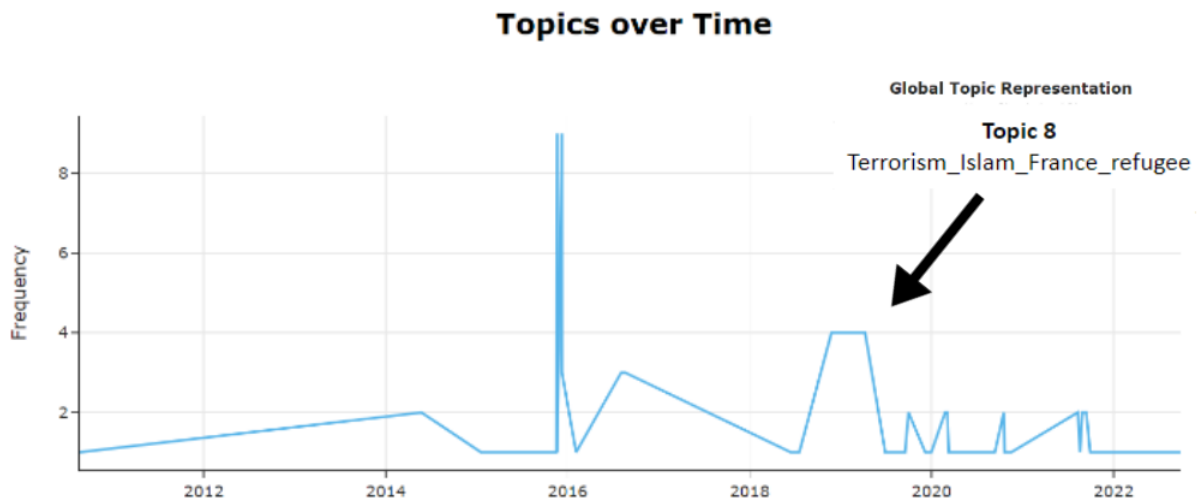


Figure 4: Frequency over time of the Topic 8 from the articles harvested from Chosun Ilbo, Joongang Ilbo and Donga Ilbo

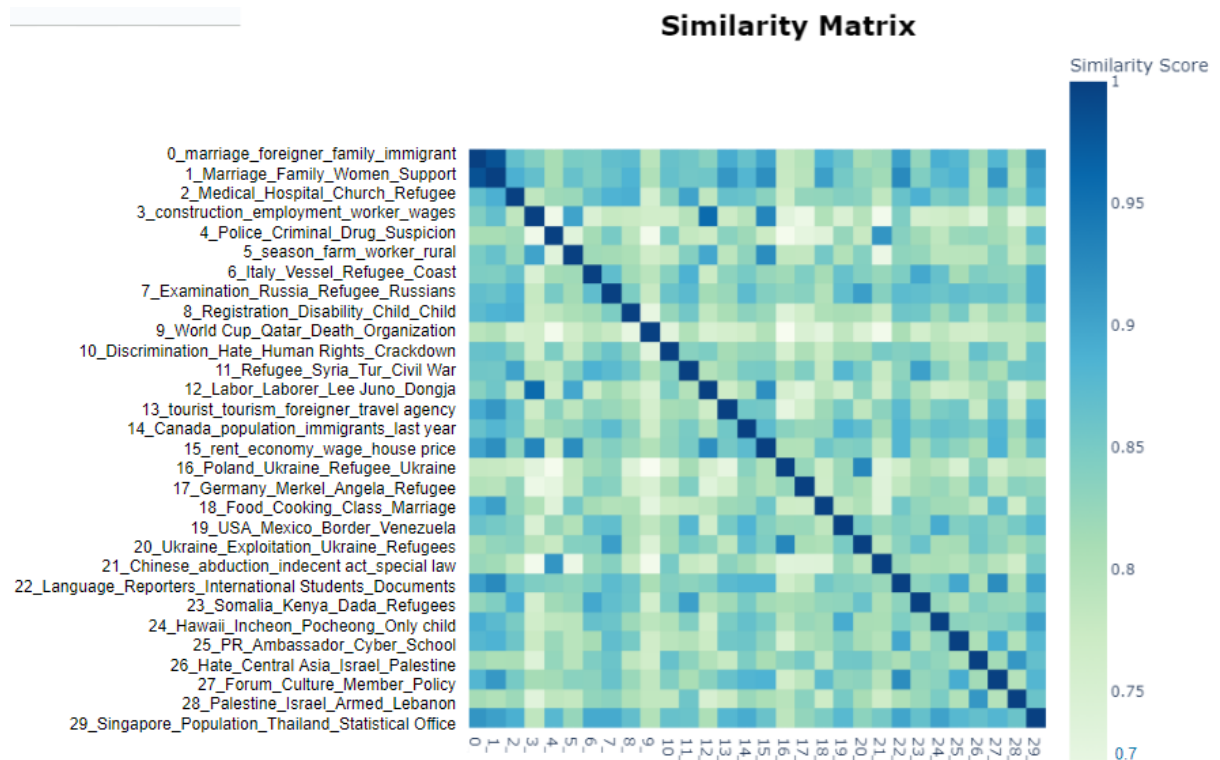


Figure 5: Similarity Matrix of the topics in the articles harvested from Naver News

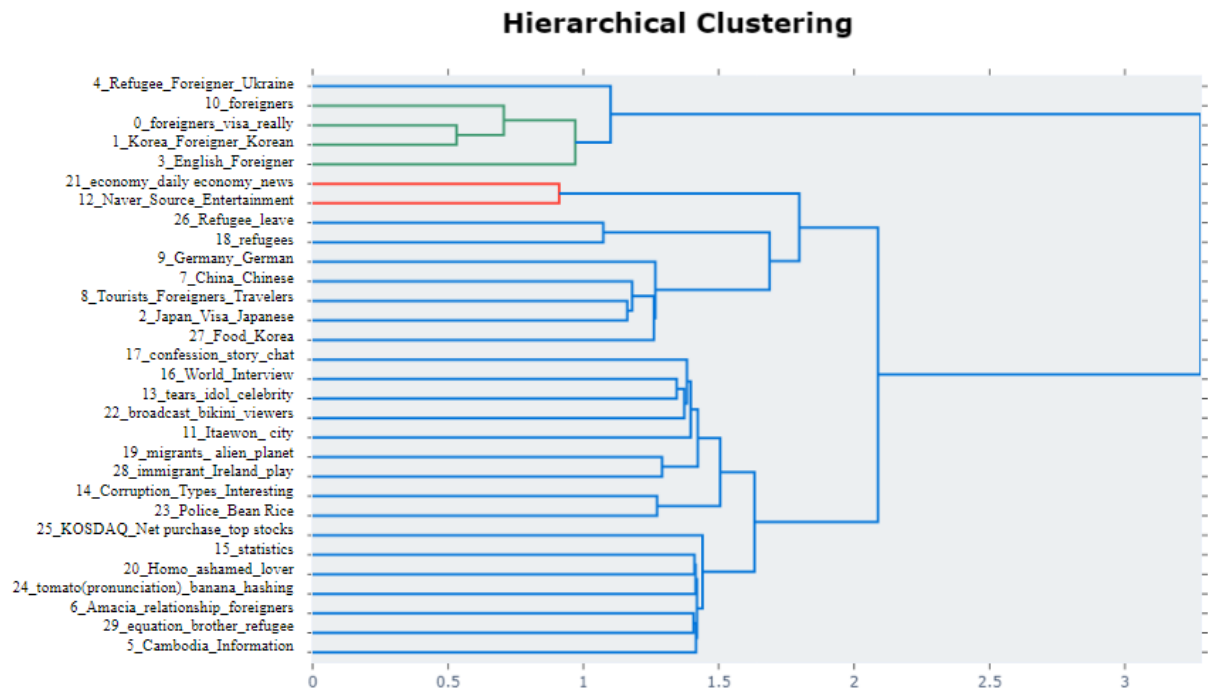


Figure 6: Hierarchical clustering of the topics in the tweets harvested in 2022

South Korea. The presence of the terrorism topic (Topic 8) in a corpus of exclusively migrant related articles, and the significance of the word “refugee” in this cluster highlights the Islamophobic association with migrants, specifically refugees, and terrorism. In figure 4, the frequency over time of the Topic 8 (Terrorism) in the mass media corpus shows how predominant it is in the media narrative on migration. Another important framework is the victimization framework, painting women migrants as victims. Topic 21 (n = 32 articles) describes migrant women with the following keywords: Women, Prostitution, Business owner, Police, Violence, suicide, victim, assault, male, report, Husband, Crime, Incident, Business, Sexual assault, Damage, punishment. Women migrants are both painted as victims of “violence”, and as criminals, with the criminalization of sex work, with “prostitution” and “police”. This victimization narrative puts women as victims of individuals, (“husband”, “male”), rather than systemic exploitation. Combined with the criminalization narrative, women migrants are distanced from claims to citizenship. A prejudiced association with drugs is also found in top 29 (m = 12 articles), grouping migrants with the following keywords: Drugs, Thailand, Possession, cultivation, firearms, production, Southeast Asia, crime, Myanmar, Suspicion, Criminal, Regulation. The Mass Media narrative shows three primordial

characteristics. First, migrants are separated into specific, and distanced groups, based on their social identities, such as gender. Then, a criminalization framework is applied, in particular to foreign workers and, or, a victimization framework, in particular to marriage migrants. Finally, an accrued coverage of Western conservative migration policies, namely USA and Germany’s policies, passes on Western conservative views on immigration. In figure 5, from the topic modeling of the New Media corpus, the criminalization of migrants through the keyword “illegal” shows a strong association of specific subgroups of migrants with illegal status. In topic 4 (n = 302 descriptions of articles), violent police intervention is justified with the following keywords: Police, Crime, Drugs, Suspicion, illegal, assault, nationality, police station, stay, Thailand, violation, police agency, arrest, police officer, foreigner, Male, Jeju. Specifically, male migrants are covered as illegal. In contrast, women migrants are associated with “support”, in topic 1 (n = 656), with the following keywords: Marriage, Family, Women, Support, Center, Education. This shows how both the criminalization frame and victimization frame restricts the rights to citizenship for both groups of migrants. Less salient topics however, do offer a coverage focusing on social justice and human rights. Topic 7 (n = 129) shows a high coverage of the situation of refugees waiting at the

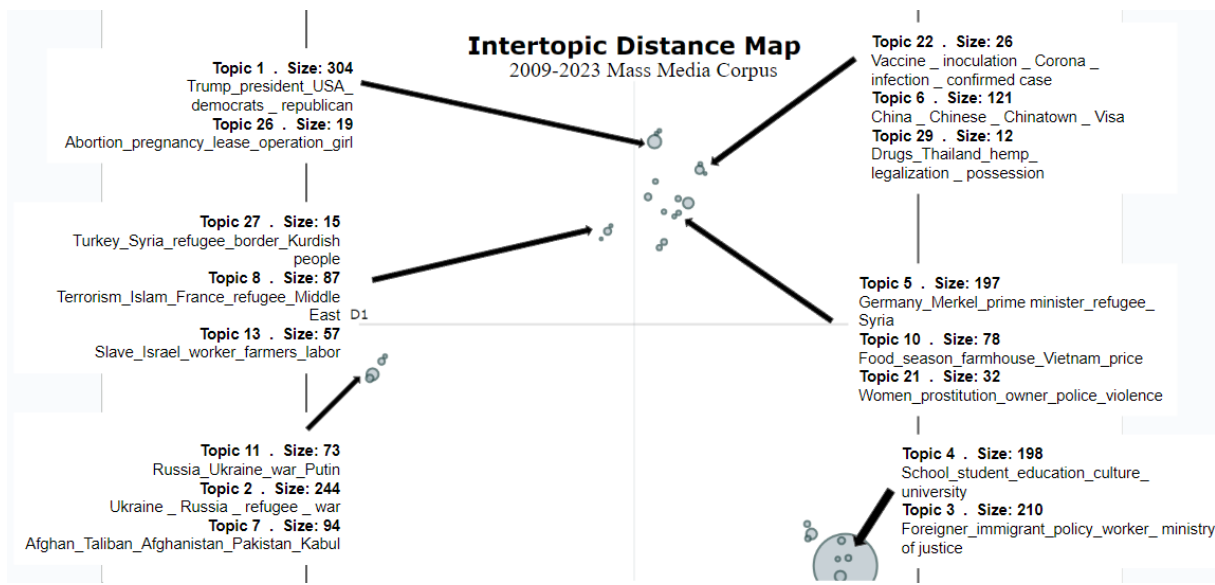


Figure 7: Intertopic Distance Map Topic in the articles harvested from Chosun Ilbo, Joongang Ilbo and Donga Ilbo

Incheon airport before being allowed to apply for the refugee status (keywords = Examination Russia Refugee Russians Conscript Ministry of Justice Incheon Recognition Litigation Airport Referral Court Forced Decision War Victory Korea Opponent cancel reject). Topic 3 (n = 392) shows a coverage of migrant workers in exploitative work conditions (keywords = Construction Employment Workers Wages Foreigners Employment site accident survivors work foreign children Juno Lee Worker Constitution Hanam Factory Manufacturing Late Payment Provision). Topic 5 (n = 261) also mentions the fatal consequences of the exploitation of migrant workers (keywords = Season Farm Worker Rural Pig Foreigner Professor Farmhouse Agriculture Organic Cadaver Batch worker remark employment farmer entry manpower shortage when work). Topic 10 (n = 113) focusing on the repressive refugee application process (keywords = discrimination hate human rights regulation registration minorities society residents government halfhuman race immigrants Refugees Illegal Equality Deportation Groups Women Respect Suggestion), and topic 12 (n = 75) even shows compassion and union, not pity, with the immigrants undergoing this administrative process (keywords = labor worker Juno Lee dongja employment illegal problem field union discrimination human rights violence environment workplace regulation condition wage relocation registration construction). In figure 6, the topic modeling of the corpus of Tweets harvested in 2022 generated several remark-

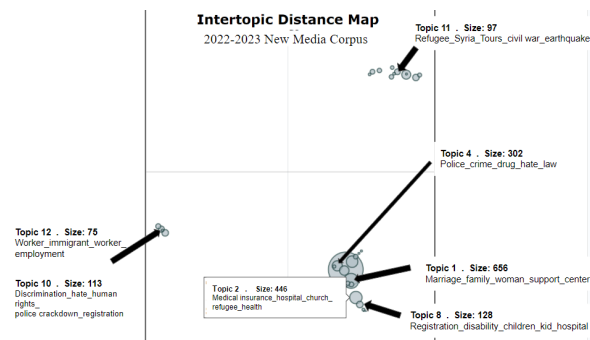


Figure 8: Intertopic distance map of the topics in the articles harvested from Naver News

able topics. The first topic (n = 7641 tweets) in the public debate on migration focuses on South Korean locals migrating to Japan (keywords = Japan Visa Japanese Tourist Visa Travel Immigration). Locals are describing their own experiences as immigrants, troubles with visa processing, administration, integration in the country. This reveals a common experience as migrants between locals and immigrants. This is a primordial source of understanding. The second topic (n= 8069) shows a desire for communication with foreigners, as class friends. (keywords = English School Speak I Today Foreigner Class Friend). The third topic (n = 7622), shows compassion with migrants in vulnerable situations (keywords = Refugees Foreigners Ukraine Women Marriage). However, the topic 26 (n = 615), with victimization keywords (keywords = Refugees Syria Ukraine United Nations Children UNICEF), presenting a focus on children, shows

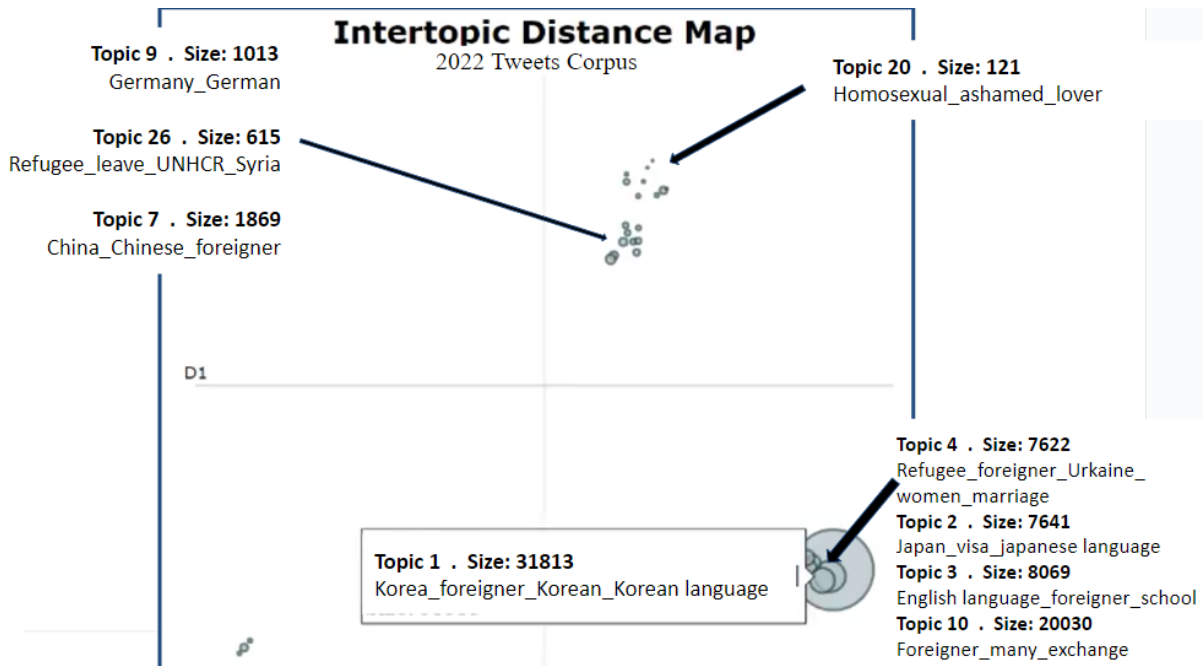


Figure 9: Intertopic distance map of the topics in the tweets harvested in 2022

how this compassion is not turned into political activism, but distracted towards pity, charity, and an individual responsibility to donate to NGOs. In figure 7, the intertopic distance map from the topic model of Mass Media articles show clusters of topics distinctively distanced from each other. On the right, migrants in charge of education are vastly separated from groups of migrants on the top right, associated with drugs. On the contrary, topics of refugees and terrorism are overlapping. The intertopic map shows the associations between refugees and Islamophobic tropes, and the fragmentation of the groups of migrants in the discourse. In figure 8, the intertopic distance map from the topic model of New Media articles shows a strong separation between refugees (on the top right of the map) and migrant workers with the description of their exploitation (bottom left of the map). The victimization frame (with “women” and “refugees”) and the criminalization frame (with “police” and “drugs”) are close. In figure 9, the intertopic distance map from the topic model of 2022 Tweets show clusters with overlapping topics on the right. Twitter users talk about their shared experiences (with visa, and as learners of English, Korean, Japanese). It is a source of union through shared experiences in the same country. On the left, separated topics are distanced based on social identities, such as nationality and sexual orientation.

5 Discussion

By framing migrants as criminals or threats to social order through the criminalization framework, the media perpetuates a narrative that justifies oppressive immigration policies and reinforces divisions within the working class. With the charitable framework, the media frames refugees and women migrants as passive victims, reducing them to non-political recipients of aid. Migrant women’s victimization in the media undermines systemic oppression: their experiences are reduced to instances of personnel, individual suffering, diverting attention from the systemic factors that contribute to their exploitation. Similarly, mass media’s appeal for charity and individual donations to aid refugees abroad, while neglecting to address the issue of visa recognition, individualizes and depoliticizes the refugee crisis, shifting the responsibility to individual acts of compassion. The mass media’s categorization of migrants into separate groups, dividing them into simplistic and stereotypical roles such as women as victims, or men as violent criminals, perpetuates a distorted narrative. By focusing on certain subgroups of migrants, the media obscures the systemic causes of migration, such as economic exploitation, political instability, and imperialist policies. This selective portrayal creates a false dichotomy of “good” versus “bad” migrants, perpetuating divisions among the working class. The study finds that the public shares experiences

with immigrants, specifically struggles with visa regulations and language learning. It does not passively accept the divisive portrayal of foreigners by the mass media, and seek alternative narrative in new media, which covers the experience of immigrants with a human rights framework. To encourage this potential for union, it is necessary to challenge the categorizations of migrants shaping the narrative in the mass media.

6 Limitations

The mono-thematic corpora were centered around the migration theme, overlapping topics remained. While the seed topic list improved the definition of topics, the majority of the data was still categorized in the topic -1, 0 and 1. Modifying the parameters of the model, particularly of the UMAP dimensionality reduction model, slightly improved this issue. The predominance of topic -1 is an important limitation in this experience, as the top three words clustered in topic -1 included “women”, “marriage” in the Mass Media corpus, and “married”, “female” in the New Media corpus. Efficiently decreasing the size of topic -1 may provide information on the shaping of the narrative on gender and migration.

Acknowledgements

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2021S1A5A8065237)

References

- Dimo Angelov. 2020. [Top2vec: Distributed representations of topics](#).
- Harry Bunt. 2017. Computational pragmatics. In Yan Huang, editor, *The Oxford Handbook of Pragmatics*, pages 326–345. Oxford University Press.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [Bert: Pre-training of deep bidirectional transformers for language understanding](#).
- Ryan J. "Gallagher, Kyle Reing, David Kale, and Greg" Ver Steeg. 2017. ["anchored correlation explanation: Topic modeling with minimal domain knowledge"](#). *"Transactions of the Association for Computational Linguistics"*, 5:529–542.
- Maarten Grootendorst. 2022. [Bertopic: Neural topic modeling with a class-based tf-idf procedure](#).
- Stuart Hall. 1997. *Representation: Cultural Representations and Signifying Practices*. Sage Publications Ltd.
- Takumitsu Kudo. 2005. [Mecab : Yet another part-of-speech and morphological analyzer](#).
- Angela McRobbie. 2009. *The Aftermath of Feminism: Gender, Culture, and Social Change*. Sage Publications Ltd.
- Chandra Talpade Mohanty. 2003. *Feminism Without Borders: Decolonizing Theory, Practicing Solidarity*. Duke University Press.
- Debora Nozza, Federico Bianchi, Anne Lauscher, and Dirk Hovy. 2022. [Measuring harmful sentence completion in language models for LGBTQIA+ individuals](#)". In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 26–34, Dublin, Ireland. Association for Computational Linguistics.
- Alina Pavlova and Pauwke Berkers. 2022. ["mental health" as defined by twitter: Frames, emotions, stigma](#). *Health Communication*, 37(5):637–647. PMID: 33356604.
- Edward Said. 1978. *Orientalism*. Pantheon Books.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#).
- Joe H. Ward Jr. 1963. [Hierarchical grouping to optimize an objective function](#). *Journal of the American Statistical Association*, 58(301):236–244.

**International Workshop
on Disinformation and
Toxic Content Analysis**

WIDISBOT: Widget to analyse disinformation and content spread by bots

Jose Manuel Camacho *
Institute of
Mathematical Sciences
(ICMAT-CSIC)

Luis Perez-Miguel
Institute for
Physical and Information
Technologies
(ITEFI-CSIC)

David Arroyo
Institute for
Physical and Information
Technologies
(ITEFI-CSIC)

Abstract

The increasing prevalence of bots poses a significant challenge for maintaining the integrity of online information. Bot campaigns have been deployed for both economic scams and political interference, making it necessary to develop a system to detect these agents and analyze their behavior. We present a scalable application designed to identify bots and to buttress the investigation of disinformation campaigns. Our intention is to provide professionals without technical expertise with an effective tool to identify and analyze content generated by bots. This will enable researchers from diverse backgrounds to study bot activity, fostering an interdisciplinary understanding of the strategies these agents use to spread disinformation, and the characteristics of their discourse. We illustrate how to use the application through a case study on COVID-19.

1 Introduction

In a world characterized by an increasing globalization and the rapid dissemination of information, many decisions are influenced by publicly accessible information obtained through online sources. In 2021, more than 50% of Twitter's users were obtaining news directly from the platform (Pew Research Center, 2021). Individuals who rely on social media for news tend to exhibit reduced engagement with news and possess limited knowledge regarding a wide range of current events (Pew Research Center, 2020). This creates an exploitable opportunity for malicious actors to manipulate public opinion or deceive unsuspecting users through disinformation, posing a threat to the 16th Sustainable Development Goal of the United Nations, which aims for an inclusive and peaceful society (Bontcheva et al., 2020).

One of these malicious agents are bots, software programs that can mimic human behavior on social

networks like Twitter. They have played a significant role in the dissemination of low credibility content (Shao et al., 2018), and their presence continues to grow within the discourse of democratic processes (Pastor-Galindo et al., 2020). Moreover, they can be combined with Large Language Models to generate counterfeit news and fabricate speech that resembles that of a human (De Angelis et al., 2023). Given the limited effectiveness of current methods for detecting non-human content (Pegoraro et al., 2023), it is crucial to adopt a different perspective. Instead of solely focusing on the accuracy of the content, an alternative approach is to identify bots based on their behavior, which can be inferred from the analysis of their metadata. Based on bot detection techniques, it is also possible to expose disinformation campaigns that have the potential to influence critical decision-making processes.

WIDISBOT ¹ has been developed to address the challenge of scrutinizing the dissemination of disinformation by bots in Twitter. This tool employs a scalable machine learning model and enables the analysis of bot discourse in tweets, making comparisons with human users participating in the same public conversations. This discourse analysis comprises the examination of sentiment, hashtags, and the usage of the most shared URLs or hashtags. Built using Streamlit ², the primary goal of this widget is to offer professionals with non-technical expertise an effective means for examining how bots propagate disinformation. It empowers them to contribute to research on these agents and enhance the field with insights from diverse disciplines. By enhancing interdisciplinary research, we facilitate the development of information consumption security frameworks and contribute to safeguard digital societies.

¹The application is available at: <https://github.com/jmcamachor1/WIDISBOT>

²<https://streamlit.io/>

*Corresponding author: josemanuel.camacho@icmat.es

2 Related works

Research on bot detection has significantly increased over the last decade, leading to the development of various methods, with supervised learning being the most widely adopted approach (Cresci, 2020). A conspicuous example of a supervised method is demonstrated in (Yang et al., 2020), where the account’s metadata is utilized to construct a scalable detector. Another popular alternative for bot detection is unsupervised learning, which does not rely on labeled datasets. An illustrative instance of this method is given in (Mazza et al., 2019), where the identification of bot accounts is constructed upon the analysis of the temporal patterns of retweeting behavior. One popular method for modeling bot behavior involves generating a string, similar to a DNA chain, that can encode different aspects of bot behavior (Cresci et al., 2017). This modeling can be exploited from both supervised and unsupervised learning methods. An additional alternative is to employ an adversarial approach (Najari et al., 2022), which mitigates the impact of evasion techniques on bot detection.

Bot detection models have been integrated into user-friendly software, making them accessible to individuals with no technical expertise. One notable example is Botometer (Sayyadiharikandeh et al., 2020), which enables users to predict the likelihood of an account being a bot by leveraging over 1200 features. Otherwise, Bot Detective (Kouvela et al., 2020) offers a web service powered by an explainable method for detecting bots. BotSlayer introduces a system with a dashboard to visualize the users who are sharing content that matches a given Twitter query (Hui et al., 2019, 2020). The system provides various metrics and allows content filtering based on entities such as hashtags, user handles, and links. One of these metrics focuses on assessing the likelihood of an account being a bot, which can be accomplished using different rules or bot detection models. Combining BotSlayer with Hoaxy enables the analysis of the spread of disinformation associated with bots and their corresponding fact-checking responses (Shao et al., 2016).

Our approach, WIDISBOT, facilitates the comparison of discourse between bots and humans within a specific conversation on Twitter. Users can input either a Twitter query or tweets IDs, enabling further analysis of tweets datasets. WIDISBOT offers an interface to visualize disparities in

discourse between automated and genuine users by applying sentiment and words frequency analysis. Additionally, WIDISBOT supports in-depth examination of fabricated content that is propagated by these entities.

3 Application description

This section presents an overview of the application’s functionalities and the machine learning (ML) models empowering them. Initially, we outline the application capabilities, followed by a description of the models. When analyzing tweets through the various functionalities, the input format requires Tweet Objects obtained via the Twitter API, and the related User Object representing the tweet author.

3.1 Functionalities

Below, we describe the application functionalities:

- *Data extraction (DE)*. It enables the retrieval of tweets by connecting to the API. Therefore, valid credentials are necessary. These can be for any version of the Twitter API (v1.1, v2). The retrieved data is then normalized in the structure of v1.1 Tweet Objects and User Objects. In particular, the user may extract tweets by ID, or via search containing a certain keyword, hashtag or URL on a specific date. This functionality is limited by Twitter API restrictions and rate limits. The generated dataset can then be used as an entry to any other WIDISBOT functionality.
- *Monitoring (M)*. It identifies which of the input tweets were generated by bots or humans. Additionally, it plots the probability distribution that the accounts that posted those tweets were bots, as well as the proportion of those accounts that were labelled as bots or humans and the number of tweets produced by each account type.
- *Forensics (F)*. Given the accounts’ usernames, it computes the likelihood of them being bots, allowing the results to be presented in an aggregated manner.
- *Sentiment analysis (SA)*. It computes the sentiment of the input tweets, displaying the human/bot sentiment distribution in both a discrete (positive-negative-neutral) and continuous fashion.

	Test datasets					
	<i>botwiki-verified</i>	<i>cresci-rtbust-2019</i>	<i>gilani-2017</i>	<i>kaiser</i>	<i>cresci-stock-2018</i>	<i>midterm-2018</i>
<i>Light (v1.1)</i>	.990	.613	.631	.944	.631	.964
<i>Light (v.2)</i>	.975	.518	.580	.936	.653	.947
<i>Botometer v3</i>	.922	.625	.689	.829	.756	.958

Table 1: AUC scores of the bot detection models on different datasets ³. The *botwiki-verified* is formed through merging datasets *botwiki-2019* and *verified-2019*.

- *Hashtag analysis (HA)*. It allows the visualization of the most frequently used hashtags by both humans and bots within the input tweets. This functionality is not case-sensitive, as bots may utilize variations of the same hashtag to promote diverse content.
- *Wordcloud (W)*. It provides a visualization of the 25 most frequent words on the tweets shared by bots and humans.
- *Analysis of spread sources (ASS)*. It displays the most shared URLs by bots and humans. It is connected to the Wayback Machine ⁴ to retrieve the content from deleted websites, as content spread by bots is often removed after a certain time. The app also provides access to Media Bias Fact Check ⁵ to determine the bias of a media and if it is a non-reliable source.
- *Analysis of discourse around hashtags (ADH)*. It enables the utilization of the functionalities *M*, *SA*, *HA*, *ASS* on tweets that contain a specific hashtag, allowing for the analysis of how the discourse surrounding the given hashtag is influenced by both bots and humans.

The classification of input accounts as bots or humans is conducted using a threshold specified by the user. A higher threshold leads to a more cautious approach by the model in determining which accounts are classified as bots. It is advisable to utilize a threshold of at least 0.51, although higher thresholds can be employed for a more conservative analysis. Additionally, the application enables users to download files with the results of the various functionalities for further analysis, either manually or in another application.

3.2 Machine Learning models

We provide details about the bot detection and sentiment analysis models integrated into the widget, powering the previous functionalities.

⁴<https://archive.org/web/>

⁵<https://mediabiasfactcheck.com/>

Bot detection The widget utilizes the *Light* model from (Antenore et al., 2022) if the input Tweet objects are in Twitter API v1.1 format. However, if the input tweets are in API v2 format, we employ an adapted version of the model that does not consider features inaccessible in API v2 but available in v1.1. These models offer scalability, requiring only a Tweet object to forecast whether an account is a bot. Table 1 demonstrates their effectiveness in detecting various types of bots. Furthermore, they achieve comparable performance to Botometer v3 (Yang et al., 2019), a widely used method for Twitter bot detection (Rauchfleisch and Kaiser, 2020). Additionally, since the model solely relies on language-agnostic features, it can predict tweets irrespective of their language.

Sentiment analyzer The app employs VADER (Hutto and Gilbert, 2014) as the sentiment analysis model. VADER utilizes a lexicon to assign scores to each word, which are subsequently combined using five rules that consider grammatical and syntactical aspects. The output is a unidimensional continuous metric (y) ranging from -1 (most negative) to 1 (most positive). To categorize y discretely, we use the thresholds provided by the authors: positive if $y > 0.05$, negative if $y < -0.05$, and neutral if $-0.05 \leq y \leq 0.05$. VADER is computationally efficient and scalable. Additionally, it performs well across various domains, particularly in analyzing microblogging content. In fact, according to (Ribeiro et al., 2016), it is an effective method for predicting three-class sentiment in social network messages.

4 Case study

This section displays how the application could be used to study bots' role on a potential disinformation campaign. For illustrative purposes, we have selected a set of 527 tweets used in experiments in (Antenore et al., 2022) from 7th February 2020 that contain the words 'Trump' and 'death toll', and their subvariants. These tweets were produced at the start of the COVID-19 pandemic when there

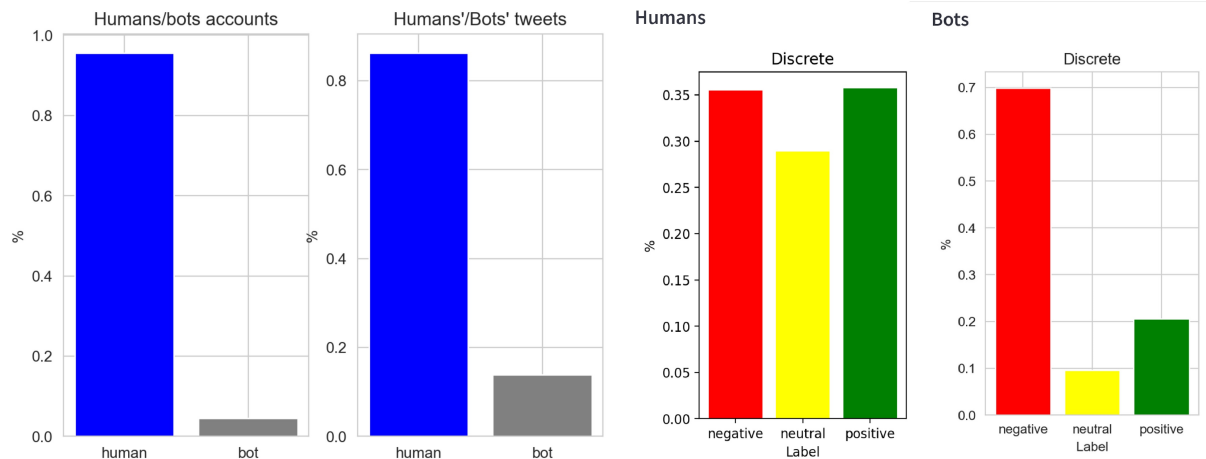


Figure 1: Screenshots of WIDISBOT output. (Left) Proportion of accounts in the subset labelled as bots/humans and the fraction of tweets produced by each type. (Right) Sentiment distribution in human/bots tweets.

was still much uncertainty about the health crisis. We aim to display how to use the widget to study whether bots intended to promote certain content by taking advantage of the crisis situation. We follow the steps below to carry out the tweets' analysis:

1. *Analysis of bot presence.* Utilising the *M* functionality, we examined the proportion of tweets produced by bots compared to humans. In Figure 1 (left) we observe that a smaller number of bots produced a larger proportion of the total tweets than humans, an indication that bots are interested to promote content in this conversation.
2. *Checking differences in sentiment.* Another indication of bot activity may be differences in the sentiment distribution between bots and humans. We used the *SA* functionality to determine if any differences were present. Specifically, as depicted in Figure 1 (right), we observed substantial discrepancies, evidence about the different content that bots and humans are sharing.
3. *Checking differences between hashtags.* Through the *HA* functionality, we examined how hashtags were used by both groups of accounts. The results for the 10 most used hashtags by bots and humans are depicted in Figure 2. We observed that bots used more hashtags and, while there was a stair-like shape in the human case, the bots

had several hashtags with the same number of occurrences. This may be an indication that bots are promoting their content using multiple hashtags in the same tweets.

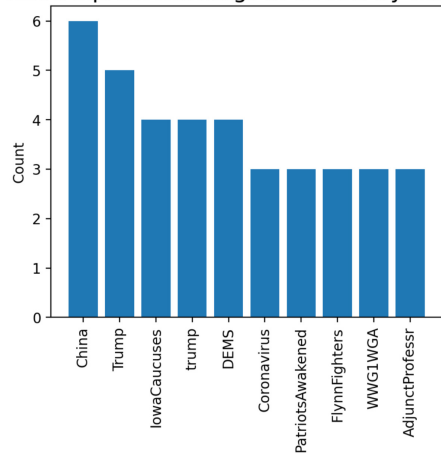
4. *Studying tweets with a certain hashtag.* We studied hashtag *#deathtoll* as it was highly shared by bots, but not at the same rate as the first six hashtags, and it was not among the most frequently used hashtags by humans. We utilized the *ASH* functionality and discovered that only one human and one bot posted tweets with the hashtag. However, the bot produced 44 tweets while the human produced only one. Furthermore, we examined the URLs shared by the bot on these tweets, observing that it shared 34 times the same URL.
5. *Analysis of the most shared URLs.* We browsed the most shared URL by the bot, finding out that it is no longer available. To check the content, we used the *ASS* functionality and retrieved the website content during the period when the tweet was produced. Figure 3 displays the website. It can be observed that some content is advertised, such as how to survive without medication or publicity about masks. Hence, we have uncovered that the identified bot was disseminating content that could potentially contribute to disinformation during the COVID-19 pandemic.

5 Discussion

This paper introduces WIDISBOT, a widget specifically developed to identify automated accounts on

⁵Datasets are accessible in <https://botometer.osome.iu.edu/bot-repository/datasets.html>

Most frequent hashtags in tweets by humans



Most frequent hashtags in tweets by bots

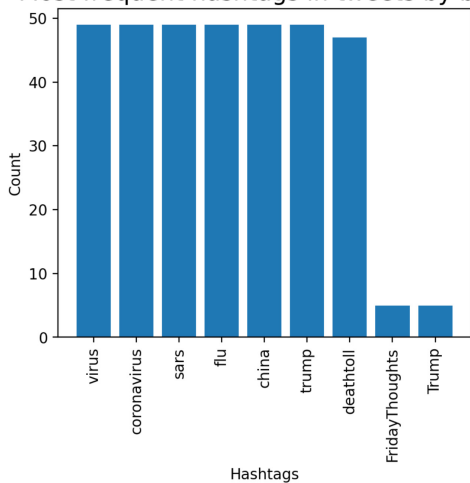


Figure 2: Ten most shared hashtags by bots and humans.

Twitter and analyze the content they promote in comparison to human users. By offering various functionalities, our aim is to provide application users with a comprehensive perspective on the information disseminated by both genuine users and bots. Additionally, we present a use case demonstrating how the widget can be utilized to uncover campaigns that potentially propagate disinformation during COVID-19 pandemic.

We have developed a user-friendly system utilizing Streamlit, which features an intuitive interface specifically designed for non-technical users, such as journalists and social scientists engaged in researching the spread of disinformation by bots. The widget demonstrates scalability and serves as an effective tool for examining disparities in content between human and automated accounts, and it is compatible with different Twitter API access. Future extensions of the widget will consist of incorporating more ML models to analyze other aspects of bot discourse, such as determin-

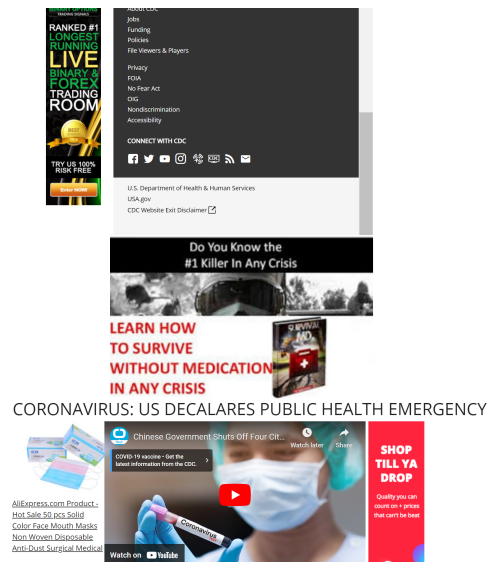


Figure 3: Screenshot of the most shared website by bots in tweets with hashtag #deathtoll, accessed through the Wayback Machine.

ing whether certain content constitutes any form of hate speech. Furthermore, it will be integrated with other applications that concentrate on identifying specific forms of misinformation, such as (Arroyo Guardado et al., 2023), in order to bolster the versatility of WIDISBOT within specific contexts.

Acknowledgements

This work was supported by the Spanish Ministry of Science program PID2021-124662OB-I00; a fellowship from "la Caixa" Foundation (ID 100010434), whose code is LCF/BQ/DI21/11860063; and Grant PLEC2021-007681 (project XAI-DisInfodemics) funded by MCIN/AEI/ 10.13039/501100011033 and by European Union NextGeneration EU/PRTR; and BBVA Foundation project AMALFI.

References

Marzia Antenore, Jose Manuel Camacho Rodriguez, and Emanuele Panizzi. 2022. A Comparative Study of Bot Detection Techniques with an Application in Twitter COVID-19 Discourse. *Social Science Computer Review*, page 08944393211073733.

David Arroyo Guardado, Gómez Espés Alberto Degli Esposti, Sara, Santiago Palmero Muñoz, and Luis Pérez-Miguel. 2023. On the design of a misinformation widget (Ms. W) against cloaked science. In *NSS 2023: 17th International Conference on Network and System Security*.

- Kalina Bontcheva, Julie Posetti, Denis Teyssou, Trisha Meyer, Sam Gregory, Clara Hanot, and Diana Maynard. 2020. Balancing act: Countering digital disinformation while respecting freedom of expression. *Geneva, Switzerland: United Nations Educational, Scientific and Cultural Organization*.
- Stefano Cresci. 2020. A decade of social bot detection. *Communications of the ACM*, 63(10):72–83.
- Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. 2017. Social fingerprinting: detection of spambot groups through DNA-inspired behavioral modeling. *IEEE Transactions on Dependable and Secure Computing*, 15(4):561–576.
- Luigi De Angelis, Francesco Baglivo, Guglielmo Arzilli, Gaetano Pierpaolo Privitera, Paolo Ferragina, Alberto Eugenio Tozzi, and Caterina Rizzo. 2023. ChatGPT and the Rise of Large Language Models: The New AI-Driven Infodemic Threat in Public Health. Available at SSRN 4352931.
- Pik-Mai Hui, Kai-Cheng Yang, Christopher Torres-Lugo, and Filippo Menczer. 2020. Botslayer: Diy real-time influence campaign detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 980–982.
- Pik-Mai Hui, Kai-Cheng Yang, Christopher Torres-Lugo, Zachary Monroe, Marc McCarty, Benjamin D Serrette, Valentin Pentchev, and Filippo Menczer. 2019. Botslayer: real-time detection of bot amplification on twitter. *Journal of Open Source Software*, 4(42):1706.
- Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.
- Maria Kouvela, Ilias Dimitriadis, and Athena Vakali. 2020. Bot-detective: An explainable twitter bot detection service with crowdsourcing functionalities. In *Proceedings of the 12th International Conference on Management of Digital EcoSystems*, pages 55–63.
- Michele Mazza, Stefano Cresci, Marco Avvenuti, Walter Quattrociocchi, and Maurizio Tesconi. 2019. Rt-bust: Exploiting temporal patterns for botnet detection on twitter. In *Proceedings of the 10th ACM conference on web science*, pages 183–192.
- Shaghayegh Najari, Mostafa Salehi, and Reza Farahbakhsh. 2022. Ganbot: a gan-based framework for social bot detection. *Social Network Analysis and Mining*, 12:1–11.
- Javier Pastor-Galindo, Mattia Zago, Pantaleone Nespoli, Sergio López Bernal, Alberto Huertas Celdrán, Manuel Gil Pérez, José A Ruipérez-Valiente, Gregorio Martínez Pérez, and Félix Gómez Mármol. 2020. Spotting political social bots in Twitter: A use case of the 2019 Spanish general election. *IEEE Transactions on Network and Service Management*, 17(4):2156–2170.
- Alessandro Pegoraro, Kavita Kumari, Hossein Fereidooni, and Ahmad-Reza Sadeghi. 2023. To ChatGPT, or not to ChatGPT: That is the question! *arXiv preprint arXiv:2304.01487*.
- Pew Research Center. 2020. Americans who mainly get their news on social media are less engaged, less knowledgeable.
- Pew Research Center. 2021. News consumption across social media in 2021.
- Adrian Rauchfleisch and Jonas Kaiser. 2020. The false positive problem of automatic bot detection in social science research. *PLoS one*, 15(10):e0241045.
- Filipe N Ribeiro, Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves, and Fabrício Benevenuto. 2016. Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5:1–29.
- Mohsen Sayyadiharikandeh, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. 2020. Detection of novel social bots by ensembles of specialized classifiers. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 2725–2732.
- Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. 2016. Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th international conference companion on world wide web*, pages 745–750.
- Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. 2018. The spread of low-credibility content by social bots. *Nature communications*, 9(1):1–9.
- Kai-Cheng Yang, Onur Varol, Clayton A Davis, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2019. Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies*, 1(1):48–61.
- Kai-Cheng Yang, Onur Varol, Pik-Mai Hui, and Filippo Menczer. 2020. Scalable and generalizable social bot detection through data selection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 1096–1103.

Debunking Disinformation with GADMO: A Topic Modeling Analysis of a Comprehensive Corpus of German-language Fact-Checks

Jonas Rieger[†] and Nico Hornig[‡] and Jonathan Flossdorf[†] and Henrik Müller[‡] and Stephan Mündges[‡] and Carsten Jentsch[†] and Jörg Rahnenführer[†] and Christina Elmer[‡]

[†]Department of Statistics, TU Dortmund University, 44221 Dortmund, Germany

{rieger, flossdorf, jentsch, rahnenfuehrer}@statistik.tu-dortmund.de

[‡]Institute of Journalism, TU Dortmund University, 44221 Dortmund, Germany

{nico.hornig, henrik.mueller, stephan.muendges, christina.elmer}@tu-dortmund.de

Abstract

In the age of (semi-) automated creation, reproduction and dissemination of misinformation, manual fact-checking can be considered as a relevant pillar of democracies. To examine the selection mechanisms of fact-checking units, the fact-checks provide a valid basis. Thus, many analyses in the field of natural language processing (NLP) regarding the spread of misinformation are based on the evaluation of fact-checks. We analyze a large German-language fact-check corpus from four specialized newsrooms over the last five years and provide scripts to reproduce the corpus and essential preprocessing steps needed to ensure comparability over time. Our topic model analysis utilizing LDA reveals a strong correlation between current events like Covid and the topics covered by fact-checks in that time. It also shows striking patterns between claims on specific topics and the ratings given by the fact-checkers. In addition, we can show that all considered fact-checking organizations focus primarily on Facebook as a source for the claims they investigate. Cross-cutting topics such as image/video analysis and data-focused fact-checking remain consistent throughout the period.

1 Introduction


In times of dynamic digital publics with significant impacts on reality, quality media cannot ignore the phenomenon of disinformation. Deliberately spreading misinformation poisons public discourse spaces (Lewandowsky et al., 2020) and undermines trust in journalistic actors and institutions by discrediting them or questioning their methods through fabricated arguments (Ognyanova et al., 2020; Giglietto et al., 2019). To counter these

negative effects, specific routines and formats have developed in journalism. Probably the best known is the fact-check, in which claims are examined for their degree of truth based on often extensive investigations (Li et al., 2022).

Due to their widespread distribution and the mostly difficult access to often incoherent platform data, it is difficult to examine disinformation campaigns in a comprehensive manner (Bastos, 2022). While, to a certain extent, the topics of the published fact-checks can be used as a proxy variable (cf., Vosoughi et al., 2018) to assess relevant disinformation campaigns, it should be taken into account that the contents of fact-checks may also reflect the media’s topic selection criteria, their working routines as well as prevailing trend topics. Consequently, a derivation to the field of disinformation campaigns can only be made to a limited extent.

In this paper, we aim to gain a deeper understanding of the topics covered by fact-checkers in Germany and Austria and their selection mechanisms with regard to the topics and origins of the claims investigated. Therefore, we built, preprocessed, analyzed and provide an extensive German-language fact-check corpus including publications from the past five years from four newsrooms specialized in this beat. The underlying research was made possible by a collaboration within the German-Austrian Digital Media Observatory (GADMO), a cooperation of fact-checkers and scientists co-funded by the European Union, see Section 1.2 for more details and related efforts.

The results show a strong relation of the fact-checks to current events — especially those with a potential for politically motivated campaigns. Clearly assignable switches in the priority topics also point to the limited resources of the news-

 Equal contribution.

rooms, as well as attention-economy effects. In addition, all fact-checking organizations focus, with varying degree, on facebook as a source for claims investigated. Cross-cutting themes, on the other hand, appear consistently throughout the period studied — for example, research on images and videos or the focus on data and figures in the fact-checks.

1.1 Related work

In the last three years, the fear of disinformation in Germany has increasingly risen (Hirndorf and Roose, 2023). Whereas in a 2021 survey around 56% indicated that they had great or very great fear, in 2023 this proportion rose to 64%. At the same time, media confidence has declined continuously over the past 8 years (Austria: 48% in 2015 → 41% in 2022, Germany: 60% in 2015 → 50% in 2022), meanwhile at least stagnating again for a few years (Newman et al., 2022).

Along with greater public awareness of the problem of disinformation, the number of fact-checking organizations worldwide has increased in recent years (Amazeen, 2020). While the Duke Reporters' Lab, which maintains a database of fact-checking organizations worldwide, counted 113 such organizations in 2016 (Graves and Cherubini, 2016), it lists 391 active groups as of May 2023¹, ten of which are located in Germany and Austria. However, the effectiveness of fact-checking in countering the belief in disinformation has been widely debated. In some cases, this has led to the conclusion that debunking has no significant effect on reducing belief in disinformation (Schwaiger, 2022). Meta-studies show that fact-checking generally has a positive effect in correcting political disinformation (Walter et al., 2020). It should be noted, however, that the effect is moderated by pre-existing beliefs, ideology and knowledge, and that the evidence on the effect on behavior and knowledge is equivocal (Ecker et al., 2022).

In addition to research on the effectiveness of fact-checking, another body of literature has focused on fact-checkers, their motivations, principles, and purposes, but “virtually no research has conducted a systematic content analysis of fact-checking” (Kim et al., 2022, p. 781). Blum (2020) therefore asks: “Who checks the fact-checkers?” (translated from German). One excep-

tion is Humprecht (2020), who analyzes a sample of eight fact-checkers from the United States, the United Kingdom, Austria and Germany with regard to the degree of source transparency provided. She finds that source transparency varies according to the level of journalistic professionalism and organizational differences. However, she uses manual quantitative content analysis, which allows for a more precise understanding of individual texts, but limits the number of observations that can be analyzed.

Automated content analysis, which enables the viewing of a larger number of texts, is used more frequently for viewing disinformation. With regard to the methodological evaluation of alternative media, topic models, such as the latent Dirichlet allocation (LDA, Blei et al., 2003), are often used. For example, Müller and Freudenthaler (2022) analyze a selection of semi-professional German language alternative media using LDA. They show that between 45% and 50% of the content is related to right-wing or populist politics. von Nordheim et al. (2021) were able to show that right-wing populist parties in countries with high media trust tend to share links with a lower source insularity if they are integrated into the party landscape (e.g., Austria), while non-integrated parties (e.g., AfD in Germany) rely more heavily on (their own) alternative media. For both type of parties, the authors were able to detect a high level of thematic insularity by using LDA.

1.2 GADMO

The basis of this study is a project funded by the European Union on combating disinformation. The German-Austrian Digital Media Observatory (GADMO) began its work at the end of 2022 and is the largest alliance of fact-checkers and academic researchers in Germany and Austria. For the first time, the leading fact-checking organizations in Germany and Austria are collaborating closely: the German Press Agency (dpa), the international news agency Agence France-Presse (AFP), the Austrian Press Agency (APA) and the non-profit independent newsroom CORRECTIV. Their work forms the core of the project and is constantly being published on the GADMO website as a new central platform for fact-checks in German².

The objectives of the GADMO project also in-

¹<https://reporterslab.org/fact-checking/>

²<https://gadmo.eu/en/gadmo-online-platform-launched/>

clude fostering media literacy, monitoring the platforms regarding overarching policies³ and researching the field of disinformation. The latter is addressed by two project partners: The Austrian Institute of Technology explores ways in which AI-driven systems can assist journalists to identify manipulated multimedia contents. The team at TU Dortmund University is dedicated to research at the interface between media and data science: On the one hand, the team is interested in fact-checkers, their selection processes, what they cover compared to traditional media and how this differs between different organizations. Therefore, we provide and analyze the German-language fact-check corpus presented in this paper. On the other hand, further work will use network analysis to investigate whether disinformation campaigns can be identified through targeted dissemination patterns⁴.

Being part of the European Digital Media Observatory (EDMO), GADMO is integrated into a Europe-wide network of media and research affiliates⁵. In addition, there are close links to projects funded in the Federal Government's research framework program on IT security, which are also intended to counteract the massive spread of disinformation⁶. In this context, the noFAKE⁷ project, also aiming at developing an assistance system for the early detection of false information, is particularly worth mentioning.

1.3 Contribution

Our contribution to research is threefold: First, we provide a corpus of about 5000 German-language fact-checks that is reproducible and extensible, thus enabling researchers to carry out further (content) analyses. This is important, as outlined in Section 1, because there is a lack of research on the texts of fact-checks and their characteristics, such as sources and topic decisions. Second, during our data collection process we identified issues such as missing (meta) data or poor comparability between different fact-checking organizations, for which we provide solutions how to address these. Third, we

³<https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation>

⁴<https://gadmo.eu/en/research-development/>

⁵<https://edmo.eu/edmo-at-a-glance/>

⁶<https://www.bmbf.de/bmbf/shareddocs/kurzmeldungen/de/2022/02/fake-news-bekaempfen.html>

⁷<https://www.forschung-it-sicherheit-kommunikationssysteme.de/projekte/nofake>

give insights into the topics being considered, the ratings being given, the sources of the claims being investigated and how these differ between different fact-checking organizations.

2 Data

Our corpus consists of data from four German-language fact-checking organizations: The German language service of Agence France-Presse (AFP), the Austrian Press Agency (APA), the non-profit newsroom CORRECTIV and the German Press Agency (dpa). In the following, we provide a brief overview of the data collecting process. All scraping and analysis scripts are available under <https://github.com/GADMO-EU/DiTox2023>.

2.1 Composition

We allocated the data in a three-step approach: As a starting point for data acquisition, we used the R (R Core Team, 2023) package `httr` (Wickham, 2022) to access a Google API referencing *ClaimReview*⁸, a tagging system that provides fact-check results and their metadata such as publication date, source, and claim rating in a structured way. In a next step, we scraped the texts corresponding to the metadata directly from the respective websites using the R package `rvest` (Wickham, 2021). As the dpa stopped using ClaimReview in July 2020 when it changed its publication platform, we also scraped the available metadata (publication date and claim). In a third step, we compared the resulting corpus with data provided by the fact-checking organizations as part of our GADMO collaboration. Finally, we restricted the corpus to fact-checks until the end of January 2023.

2.2 Cleaning

Due to the heterogeneous nature of the corpus, some cleaning was necessary. First, we removed duplicate texts identified by the same URL or the same text. In some cases, especially for fact-checks authored by CORRECTIV, we kept very similar texts if they refer to different URLs. As the dpa did not use ClaimReview throughout the whole analysis period, we identified the URL of the analyzed claim manually for most of the data. The same applies to some of the other organizations' fact-checks. In some cases, e.g., when fact-checkers have debunked a phenomenon that was widespread on social media, they did not provide a specific

⁸<https://schema.org/ClaimReview>

Period	AFP			APA			CORRECTIV			dpa		
	$ D $	$ W $	\bar{N}	$ D $	$ W $	\bar{N}	$ D $	$ W $	\bar{N}	$ D $	$ W $	\bar{N}
2018/1	132	23 983	182	.	.	.
2018/2	151	34 382	228	.	.	.
2019/1	147	33 318	227	20	3281	164
2019/2	190	58 297	307	211	31 996	152
2020/1	.	.	.	40	10 155	254	223	75 839	340	179	33 294	186
2020/2	86	35 875	417	59	17 677	300	215	84 335	392	376	70 069	186
2021/1	191	79 118	414	46	18 914	411	232	81 597	352	300	59 464	198
2021/2	185	93 891	508	36	15 273	424	238	72 155	303	340	64 800	191
2022/1	145	70 726	488	25	8942	358	234	58 087	248	323	65 948	204
2022/2	127	67 169	529	29	9669	333	238	75 604	318	384	76 414	199
2023/1	26	15 311	589	5	1207	241	34	8720	256	62	13 406	216
Total	760	362 090	476	240	81 837	341	2034	606 317	298	2195	418 672	191

Table 1: Number of fact-checks $|D|$, number of words in fact-checks $|W|$ (after all preprocessing steps), and mean number of words per fact-check \bar{N} , for the four fact-check organizations per half-year.

URL and therefore left this entry blank. Sometimes more than one URL was mentioned in the text, in which case we decided to consider only the first one mentioned. In contrast, there are fact-checks, in which no specific URL has been mentioned, but the source was given. For these cases, we decided to include the domain, e.g. *facebook.com*, in the dataset.

2.3 Preprocessing

For the later modeling of the texts we applied common preprocessing steps including lowercasing, stopword removal, punctuation removal, number removal, resolving umlauts and tokenization. Then, we kept only those words that contain at least two letters and occur at least five times in the whole dataset, which results in 27 606 vocabularies.

For referencing the set of fact-checks (cf., Section 3.1), we use the notation $D = \{D_m \mid m = 1, \dots, M\}$, where M denotes the number of all documents. Moreover, $W = \bigcup D_m$ denotes the set of all words.

Figure 1 shows how the total of 5229 fact-checks (with an average of 281 words per document, after preprocessing) are distributed among the four different organizations. Table 1 provides further insight into the distribution of fact-checks and their length over time. It can be seen that all 283 fact-checks from 2018 in our corpus were authored by CORRECTIV. We observed dpa’s first fact-checks for June 2019, from APA for February 2020, and from AFP for September 2020. The fact-checks from dpa are on average the shortest with (rela-

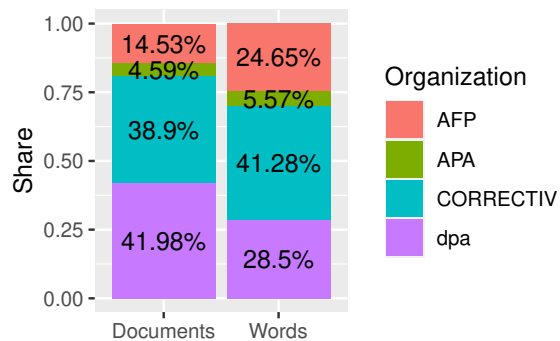


Figure 1: Share of the organizations on the total corpus of fact-checks.

tively consistently) 191 words, while AFP uses on average more than twice as many words (476) per fact-check.

3 Analysis

In the following, we use LDA as a topic model method to automatically present the thematic content from the fact-checks in an unsupervised manner. We also relate the topics identified in this way to the ratings assigned and the sources of the claims examined. Based on the findings from our data analysis we suggest further research questions for future investigations using specialized advanced NLP methods.

3.1 Topic Modeling

To analyze the given dataset, we make use of probabilistic topic modeling, which is used in many

application domains (Blei, 2012). In comparison to transformer-based methods (Vaswani et al., 2017), the modeling idea is rather intuitive: a set of documents is described by distributions of topics over time, where each word in each of these documents is assigned to one of the topics. These assignments yield word distributions for each topic, which make the topics interpretable.

Probably the best known topic model is LDA (Blei et al., 2003). The underlying probabilistic model (Griffiths and Steyvers, 2004) can be written as

$$\begin{aligned} W_n^{(m)} \mid T_n^{(m)}, \phi_k &\sim \text{Discr}(\phi_k), & \phi_k &\sim \text{Dir}(\eta), \\ T_n^{(m)} \mid \theta_m &\sim \text{Discr}(\theta_m), & \theta_m &\sim \text{Dir}(\alpha), \end{aligned}$$

where α and η are Dirichlet priors for the topic and word distributions, respectively. The number of modeled topics, K , is chosen by the user and each document is considered a bag of words set $D_m = \{W_n^{(m)} \mid n = 1, \dots, N^{(m)}\}$ with observed words $W_n^{(m)} \in \{W_1, \dots, W_V\}$. Then, $T_n^{(m)}$ describes the corresponding topic assignment for each word. Only the words are observable, while all other variables and parameters are latent. The main result, the latent word and topic distributions are represented by ϕ and θ , respectively.

For modeling topics in our German fact-check corpus, we use a reliable variant of classical LDA, estimated with the Gibbs sampler (Griffiths and Steyvers, 2004), named LDAPrototype (Rieger et al., 2022). It selects the medoid LDA — the LDA with the highest mean of pairwise similarities to all other LDAs — from a set of candidate models with independently and randomly initialized topic assignments.

We model all $M = |D| = 5229$ documents together, the vocabulary set is of size $V = 27\,606$. Since Chang et al. (2009) show that the use of common likelihood-based measures, such as perplexity, correlates poorly or even negatively with human perceptions of well partitioned topics, and Hoyle et al. (2021) show that alternative automated measures based on coherence also lead to incoherent decisions, we do not choose automated evaluation measures for parameter tuning. We tried different numbers of topics $5, \dots, 25$ showing $K = 12$ with $\alpha = \eta = 1/K$ to be appropriate in terms of granularity and coherence of topics via human eye-balling.

In the following analysis, we make use of the more reliable medoid LDA (cf., Rieger et al., 2022),

which was selected out of 100 independent replications using the R package `ldaPrototype` (Rieger, 2020).

3.2 Topics

For a better understanding of the automatically generated topics, we let human coders label them. Figure 3 shows the relative frequencies of all $K = 12$ topics in the fact-checks, per organization and overall. Accordingly, *Pictures & Videos* is the most frequently associated topic in AFP fact-checks with 21% of the words assigned to it, while 28% of the words in APA fact-checks are assigned to the topic *Laws & Legal Status*. For CORRECTIV (15% *Corona*) and dpa (12% *Quotes*), the distributions tend to be more balanced, which can to some extent be explained methodologically by the higher number of fact-checks in the analysis, raising the possibility that the smaller subcorpora realize more skewed distributions. From a contents perspective, the connection of AFP fact-checks to image content is plausible since according to their own statements they put a focus on uncovering image manipulation and deep fakes.

One advantage of topic modeling compared to traditional (hard) clustering methods is that the assignment of topics to words, which makes it a soft clustering method, allows, for example, the analysis of co-occurring topics. At the same time, this soft-clustering poses a challenge in determining a precise co-occurrence operationalization. For our analysis, we consider co-occurring topics always in reference to a dominant topic in a particular document. We understand a dominant topic per fact-check as the one that received more than half of all topic assignments in that document. The co-occurrence with other topics can then be computed using the occurrence of all other topic assignments in these associated fact-checks. Using this approach, we obtain the distributions in Figure 2, where *NA* refers to those fact-checks where no dominant topic could be determined.

It can be seen that the topics *Medicine & Health*, *Vaccination* and *Corona* strongly co-occur with each other. For all three (dominant) topics the corresponding two other topics account for about half of the co-occurring assignments. Another observation concerns the topics *Russo-Ukrainian War* and *Pictures & Videos*. While in fact-checks that thematically mainly deal with the war 37% of the remaining words are associated with the topic of im-

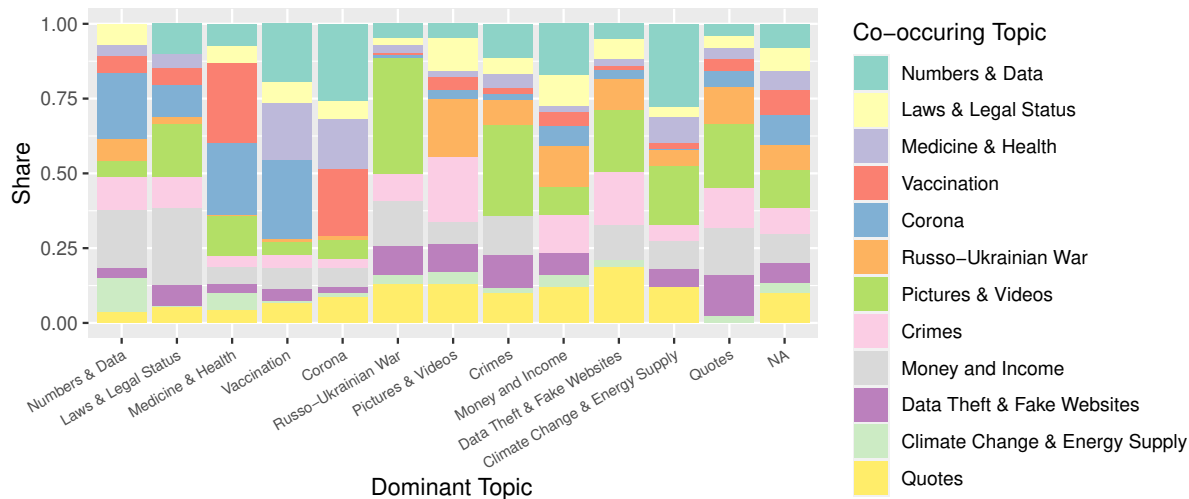


Figure 2: Co-occurring topics in the fact-checks. Dominant topics are considered as those having more than 50% of the topic assignments within the corresponding fact-check. NA refers to the absence of a dominant topic for these fact-checks.

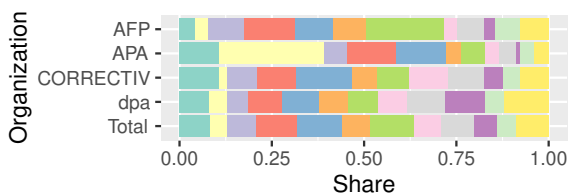


Figure 3: Distribution of the topics in fact-checks of the different organizations; cf., Fig. 2 for legend.

age manipulation, the other way around it is “only” 19%. Furthermore, as a typical side topic, *Pictures & Videos* accounts for 30% of the co-occurrences in *Crimes* fact-checks, for 21% in each of *Quotes* and *Data Theft & Fake Websites*, and for 20% in fact-checks on the topic of *Climate Change & Energy Supply*. The distribution of topics in fact-checks without a dominant topic does not show any particular peculiarities (cf., *Total* bar in Fig. 3).

In addition to the global topic distributions, the changes over time are of special interest. For this purpose, we calculate smoothed values of the number of topic assignments per day and organization using rolling sums over 90 days. To standardize the values, we divide each time series by the maximum of all smoothed values per organization. The intensity of each of the 12 topics over time is shown in Figure 4.

There is a clear focus of CORRECTIV and dpa in particular on Corona-related fact-checks in 2020. Due to the continuously high prevalence of the *Pictures & Videos* topic in AFP fact-checks, this impact is not so clearly visible for their fact-checks.

However, the topic *Vaccination* shows a clearly increased prevalence in the second half of 2021, while for APA the topic already becomes more prevalent at the beginning of 2021. The general focus of APA fact-checks on regulations by the state rather than Corona itself is also evident, which in turn explains the high share of this topic *Laws & Legal Status* in Fig. 3. With the start of the war in February 2022, all organizations show a shift in the prioritization of their fact-checks toward the topic *Russo-Ukrainian War*. Overall, the dpa shows the most balanced distribution of topics over the entire period, while the APA shows the clearest focus on one of the modeled topics (cf., Fig. 3).

3.3 Ratings

The analysis of the checked claims’ ratings in the fact-checks is only possible for AFP and CORRECTIV, since APA and dpa do not use a rating scale, but only free-text ratings. Manual review and comparison of the ratings with the textual ratings revealed that there may be occasional incorrect entries. For instance, there was one observation with a rating of 5 and a textual rating of “falsch” (incorrect), while, in general, the AFP fact-checks ratings range from 1-5, with 1 for incorrect and 5 for correct. By correcting this one observation from 5 to 1, AFP fact-checks only realize ratings 1–3 and NA (1: 557, 2: 115, 3: 67, NA: 21). In Figure 5, the distributions of the ratings in the AFP fact-checks are presented depending on the topic.

According to this, AFP fact-checks assigned to

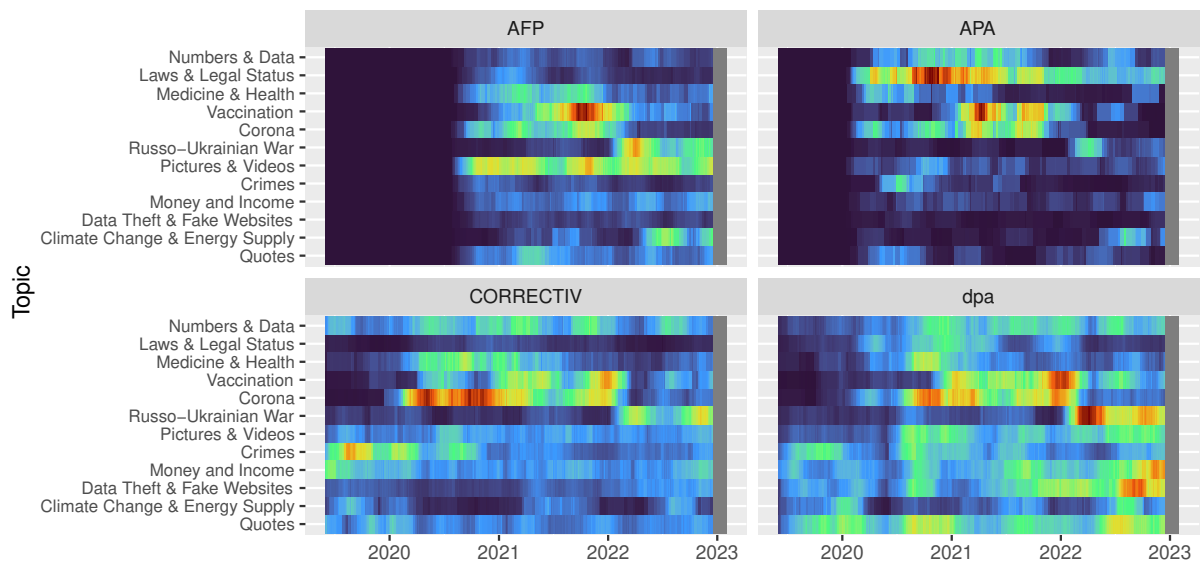


Figure 4: Topic intensity in published fact-checks over time per organization. Values were calculated based on a 90-day rolling window and normalized with the maximum value per organization.

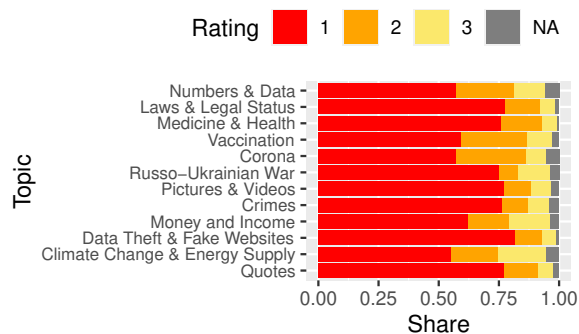


Figure 5: Distribution of the AFP ratings per topic.

the topic *Data Theft & Fake Websites* obtain in over 75% of the cases the lowest possible rating. This topic is thus most often associated with *incorrect* rated claims. Overall, it can be seen that for all topics more than 50% of the corresponding fact-checks obtain rating 1, which can be explained by the global concentration of this rating (73% of the fact-checks). The greatest tendency of a topic to less pronounced degrees of disinformation, i.e., ratings of 2 and 3, can be observed for *Climate Change & Energy Supply*.

In contrast, fact-checks by CORRECTIV are rated on a broader scale of a total of 7 levels identified by us. It is known that CORRECTIV has used a new scale for their rating from October 16, 2020. In this context, the textual ratings *missing context* and *unproved* were added to the scale, which correspond to 4 in the new rating scheme. Table 2

gives the list of textual ratings that occur, their frequencies, and their associated numerical ratings in ClaimReview. The left column in bold reflects the ratings we merged from the old and new schemes.

A manual investigation of individual fact-checks has shown that the numerical rating 2 is also associated with the textual ratings *falscher Kontext* (wrong context) and *manipuliert* (manipulated). Moreover, the ratings *missing context* are also found in fact-checks with the (merged) rating 3, 4, and rarely 6; for all especially for fact-checks before the change of the scheme.

Accordingly, Figure 6 shows that the category *missing context* in light blue has been assigned frequently since its implementation, almost completely replacing *partially incorrect* ratings for some topics. The figure shows the distribution of the ratings over time in relation to the topic. For some topics, the rating 5 temporarily reaches over 50% of the assignments.

A striking pattern is the high number of NA values during the Covid pandemic period. We explain this as a result of the inability to check the associated claims conclusively and reliably and because the existing scale did not contain the required rating. With the implementation of rating 5, no more NA values occur.

It is notable that assignments to the topic *Data Theft & Fake Websites* occur in up to 50% of cases from fact-checks about claims that are purely fictional. Over time, it also becomes apparent that

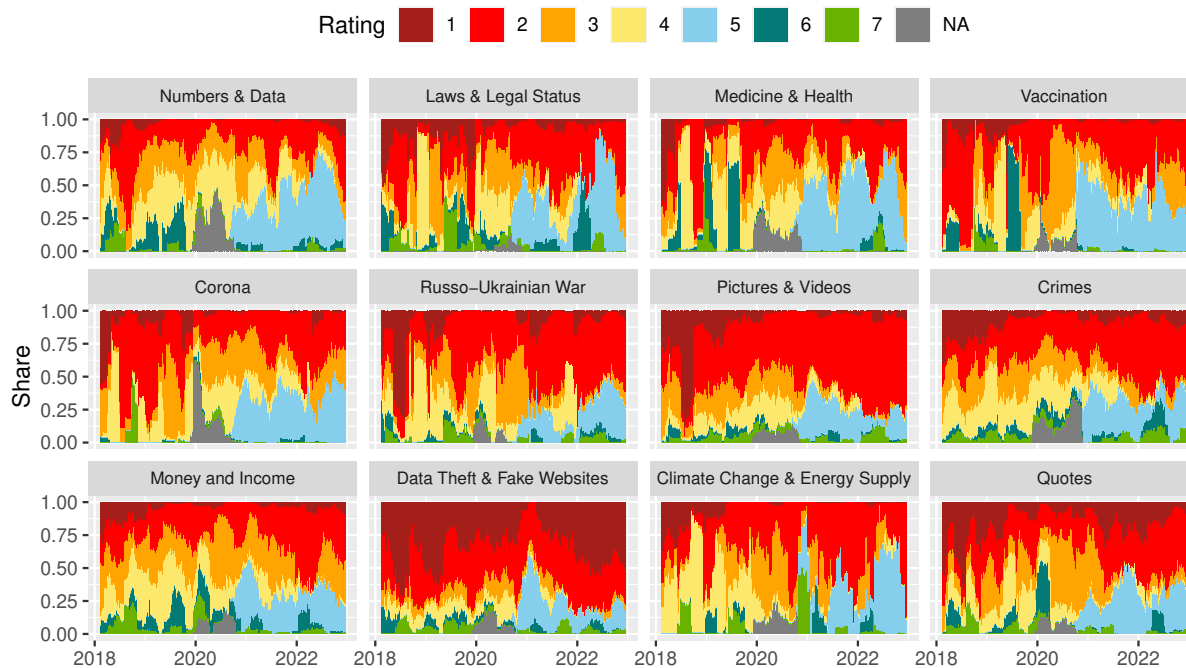


Figure 6: Distribution of the processed merged CORRECTIV ratings (cf., Table 2) per topic over time.

Our	Textual rating	Old	New	$ D $
1	frei erfunden (purely fictional)	1	0	261
2	falsch (incorrect)	2	1	733
3	größtenteils falsch (largely incorrect)	3	2	306
4	teilweise falsch (partially incorrect)	4	3	249
5	fehlender Kontext* (missing context)	.	4	294
6	größtenteils richtig (largely correct)	5	7	75
7	richtig (correct)	6	8	77
NA	.	.	.	54

Table 2: Number of CORRECTIV fact-checks in relation to our processed merged ratings **1** to **7** and **NA**. Until Oct. 15, 2020, an old rating scheme was used, after that a new one. *also includes “unbelegt” (unproved).

Pictures & Videos, beginning in 2021 and probably also due to the co-occurrences in fact-checks on the topic of *Russo-Ukrainian War*, is associated considerably more frequently with false claims from 2022 onward. For the latter topic, we observe an

abrupt increase in severe disinformation (ratings 1 & 2) at the beginning of the war.

The topic that is overall less strongly associated with false claims (ratings 1 & 2), but more with misleading claims (3–5) and partly also with correctly rated (6 & 7) claims is *Numbers & Data*. An interpretation is that it seems easy to make a statement with only a few erroneous information or an incorrect integration of percentage, relative or absolute numbers, which either already contains a misinterpretation or consciously accepts this misinterpretation by the reader.

3.4 Domain

We investigated which websites were the source of the claims that were fact-checked. As Table 3 shows, Facebook is the dominant source of claims, accounting for almost 3579 of the 5229 fact-checks in our corpus. This is not surprising, given that three of the four fact-checking organizations examined in this paper cooperate with Meta/Facebook: CORRECTIV since 2017⁹, dpa since early 2019, and AFP since 2020. The other 1650 entries are spread across a number of other sites, with only Twitter having more than 200 entries. An NA en-

⁹<https://correctiv.org/faktencheck/ueber-uns/2018/12/17/ueber-die-kooperation-zwischen-correctiv-faktencheck-und-facebook/>

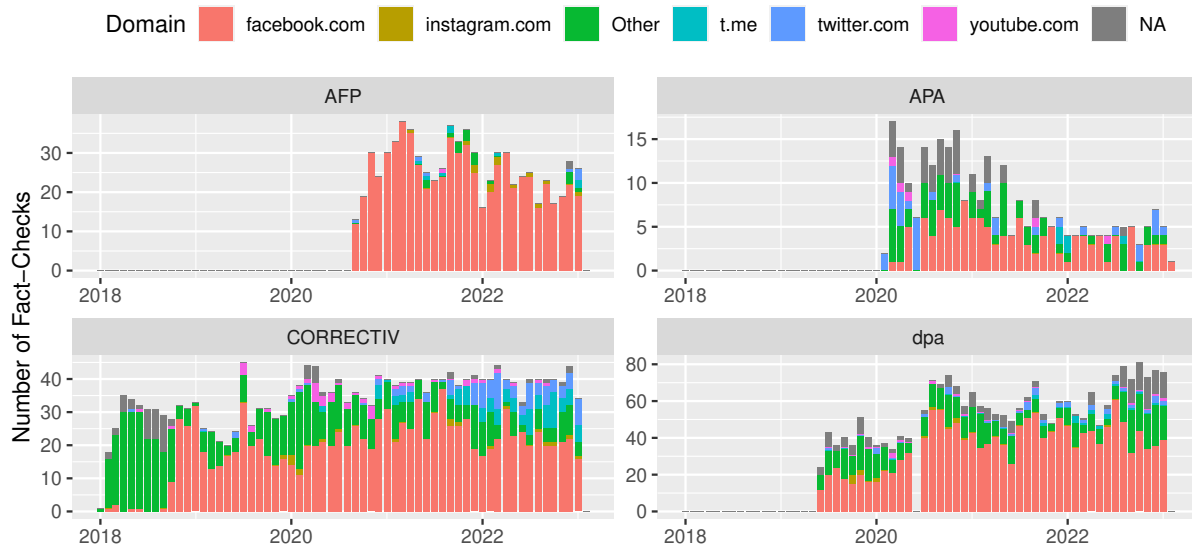


Figure 7: Number of fact-checks per month, organization and the source of the claim.

Domain	AFP	APA	CORR.	dpa
facebook.com	714	130	1156	1579
twitter.com	7	33	108	63
t.me (Telegram)	8	5	95	25
youtube.com	1	5	46	20
instagram.com	13	0	19	24
anonymousnews.org	0	0	27	18
journalistenwatch.com	0	0	21	13
wochenblick.at	1	2	22	4
report24.news	3	4	13	4
reitschuster.de	0	0	9	12
truth24.net	0	0	17	4
Other	11	23	448	226
NA	2	38	53	203
Total	760	240	2034	2195

Table 3: Number of fact-checks per organization depending on the source of the claim.

try often indicates that a fact-check is dealing with a general phenomenon or a claim that is widely spread in different variations. In some cases, it also indicates that the claim was not made by a website or social media platform, for example when politicians make a claim in a public speech.

Figure 7 shows the distribution of claim sources over time for each fact-checking organization. A striking aspect is the almost absolute dominance of Facebook as a source of claims checked by AFP. This contrasts in particular with the APA, which has a greater variance in sources but also does not work with Facebook. They also have relatively more fact-checks with an NA entry as the source. The share of Facebook as a source for claims checked by COR-

RECTIV starts to rise significantly a few months before they start cooperating with Facebook. Nevertheless, both CORRECTIV and dpa also look for other sources of disinformation besides Facebook. Still, the effect of Meta’s funding is visible and raises media economics questions about the funding of fact-checking and the incentives that come along.

We also examined which claim sources are associated with particular topics. Figure 8 shows that Telegram has the largest share of the topic *Russo-Ukrainian War*. This supports the findings of a report by the Ukrainian analytical platform Vox Ukraine and its fact-checking section Vox checks, in which the authors show how widespread Russian propaganda is on Telegram (Vox Check, 2022). The other platforms have different focuses: While Facebook, Instagram and Twitter have similar topic shares, the topic *Corona* has by far the largest share on Youtube. The focus on Corona can also be seen on the non-platform domains *report24.news* and *reitschuster.de*, which also have high shares of assignments to the topic *Vaccination*. *Truth24.net* focuses on the topic *Crimes*, which contains many statements with a xenophobic or racist tone, as it deals with real or faked crimes that are (sometimes erroneously) blamed on migrants.

Reitschuster.de and *truth24.net* also stand out when looking at the ratings given to them by CORRECTIV (see Figure 9). The “lack of context” rating was given relatively more often to the non-platforms than to the platforms whose claims were

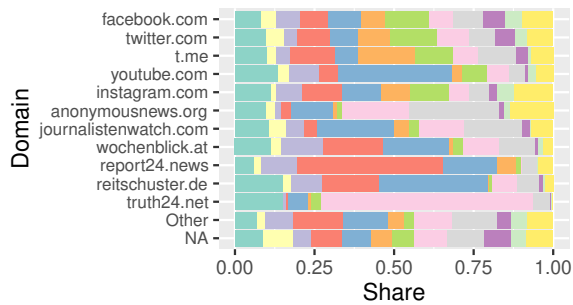


Figure 8: Distribution of the topics in fact-checks depending on the source of the claim; cf., Fig. 2 for legend.

more likely to be rated as incorrect or largely incorrect by the fact-checkers. However, the analysis of claims that do not originate from Facebook should be treated with caution. There are two reasons for this: First, as mentioned above, the number of claims from platforms other than Facebook is much lower, and even lower for the non-platforms. Their observations are therefore much more likely to be highly sensitive to outliers. Second, claims associated with the platforms may have originally been made by other sites that either posted their articles themselves, e.g., on Facebook, or had their articles shared by other users.

4 Conclusion

The topic model analysis using LDA on a dataset of 5229 German-language fact-checks from AFP, APA, dpa and CORRECTIV in the period from 2018 to January 2023 shows that in 2020 all four organizations — unsurprisingly — have a strong focus on (various) Covid related topics. In addition, there is a smooth transition to more mentions of words related to vaccination, resulting in *Vaccination* being the top topic in 2021. Then, at the beginning of 2022, a sudden shift of attention to the Russo-Ukrainian war can be identified. In particular, AFP increasingly combines fact-checks on this topic with visual content checks. At the same time, AFP fact-checks consistently result in negative ratings, and CORRECTIV rarely publishes fact-checks with (partially) positive ratings as well. For the analysis of CORRECTIV’s ratings, it is important to merge the ratings of the old and new scales in a meaningful way to avoid false conclusions.

4.1 Discussion

Facebook claims are clearly checked most frequently (> 68%). The distribution over time

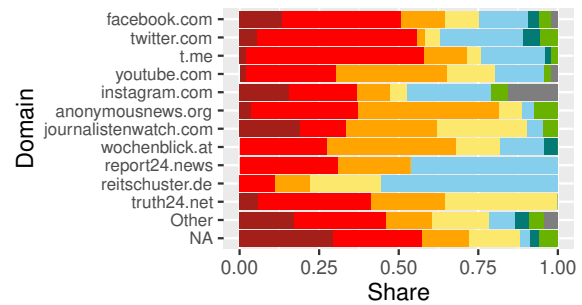


Figure 9: Distribution of the processed merged CORRECTIV ratings (cf., Table 2) depending on the source of the claim; cf., Fig. 6 for legend.

suggests that this might also be due to funding from Meta’s (now also including Instagram) fact-checking program. Survey data collected from 93 organizations worldwide show that Meta’s third party fact-checking program is still the leading funding source in 2022 with 45.2%, while grants cover 29.0% (IFCN, 2023).

This raises several questions: What is the direction of the cause-effect relationships? Is there an unfavourable bias towards current news topics or particular sources? And what consequences can result from this? On the one hand, one can propose that more (independent) money is necessary to ensure a broader attention of the fact-checkers and to slightly loosen the focus from Facebook. It could be a strategic decision that claims that *also* circulate on Facebook are preferably associated with itself. On the other hand, it can be assumed that most claims are in fact circulating on Facebook, so maybe this is not even a restriction of the thematic range for the general debunking.

4.2 Limitations

The distribution of ratings of AFP shows that often claims are checked for which it is likely in advance that they are false due to the focus on manipulated pictures and videos. This indicates a prioritization of resources and raises the question whether additional financial resources would lead to a better coverage of all *checkworthy* claims, and not only certain misinformation.

In principle, checked sources are still often used as a proxy for topical disinformation spread. Humprecht (2019), for example, uses fact-checks to distinguish between the spread of disinformation in the United States, the United Kingdom, Germany, and Austria. This raises the question to what extent fact-check corpora are representative for dis-

information spread. At the same time, there are other approaches to form disinformation corpora, e.g., based on less *trustworthy* sources, identified using NewsGuard¹⁰ scores (Carrella et al., 2023).

Since we focused on topic modeling in the present analysis, the findings are mainly limited to their inductive character (Chen et al., 2023). Nevertheless, we can extract research questions for further analysis.

4.3 Further Research

Further analyses should take into account the challenges and pitfalls of misinformation research (Altay et al., 2023), according to which, for example, misinformation is by no means just a social media phenomenon. Rather, other digital as well as offline media are also prone to misinformation. This is especially important when creating a reference disinformation dataset, which can be used to analyze under-fact-checked topics. By including a reference *quality* media dataset, the relation and the dissemination of (dis)information between low and high quality media can be analyzed. With the help of modern large language models (cf., Grottendorst, 2022; Conneau et al., 2020), it is possible to measure and compare differences in terms of the stance, sentiment and intensity of statements in typical quality media, alternative media, and fact-checks.

Acknowledgements

This research is co-funded by the European Union. We thank the German Press Agency (dpa), the international news agency Agence France-Presse (AFP), the Austrian Press Agency (APA) and the non-profit independent newsroom CORRECTIV for their cooperation regarding the plausibility of the data.



Co-funded by
the European Union

References

Sacha Altay, Manon Berriche, and Alberto Acerbi. 2023. Misinformation on misinformation: Conceptual and methodological challenges. *Social Media + Society*, 9(1).

¹⁰<https://www.newsguardtech.com/>

Michelle A. Amazeen. 2020. Journalistic interventions: The structural factors affecting the global emergence of fact-checking. *Journalism*, 21(1):95–111.

Marco Bastos. 2022. Editorial: Five challenges in detection and mitigation of disinformation on social media. *Online Information Review*, 46(3):413–421.

David M. Blei. 2012. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Roger Blum. 2020. Fakten, Fake News und Wahrheitssuche: Wer checkt die Faktenchecker? pages 251–260. Nomos Verlagsgesellschaft mbH & Co. KG.

Fabio Carrella, Alessandro Miani, and Stephan Lewandowsky. 2023. IRMA: the 335-million-word italian coRpus for studying misinformAtion. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Dubrovnik, Croatia. Association for Computational Linguistics.

Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *NIPS: Advances in Neural Information Processing Systems*, volume 22, pages 288–296. Curran Associates Inc.

Yingying Chen, Zhao Peng, Sei-Hill Kim, and Chang Won Choi. 2023. What we can do and cannot do with topic modeling: A systematic review. *Communication Methods and Measures*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Ullrich K. H. Ecker, Stephan Lewandowsky, John Cook, Philipp Schmid, Lisa K. Fazio, Nadia Brashier, Panayiota Kendeou, Emily K. Vraga, and Michelle A. Amazeen. 2022. The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1(1):13–29.

Fabio Giglietto, Laura Iannelli, Augusto Valeriani, and Luca Rossi. 2019. ‘fake news’ is the invention of a liar: How false information circulates within the hybrid news system. *Current Sociology*, 67(4):625–642.

Lucas Graves and Federica Cherubini. 2016. *The Rise of Fact-Checking Sites in Europe*. Reuters Institute for the Study of Journalism.

- Thomas L. Griffiths and Mark Steyvers. 2004. [Finding scientific topics](#). *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235.
- Maarten Grootendorst. 2022. [BERTopic: Neural topic modeling with a class-based TF-IDF procedure](#). *arXiv*.
- Dominik Hirndorf and Jochen Roose. 2023. [Welchen Nachrichten kann man noch trauen? Angst vor Desinformation und Vertrauen in öffentlich-rechtliche Medien — repräsentative Umfragen](#). Monitor — Wahl- und Sozialforschung. Konrad-Adenauer-Stiftung.
- Alexander Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Lee Boyd-Graber, and Philip Resnik. 2021. [Is automated topic model evaluation broken? The incoherence of coherence](#). In *NeurIPS: Advances in Neural Information Processing Systems*.
- Edda Humprecht. 2019. [Where ‘fake news’ flourishes: a comparison across four western democracies](#). *Information, Communication & Society*, 22(13):1973–1988.
- Edda Humprecht. 2020. [How do they debunk “fake news”? A cross-national comparison of transparency in fact checks](#). *Digital Journalism*, 8(3):310–327.
- IFCN. 2023. [State of the Fact-Checkers 2022](#). International Fact-Checking Network (IFCN) at Poynter.
- Hyun Suk Kim, Yoo Ji Suh, Eun-mee Kim, Eunryung Chong, Hwajung Hong, Boyoung Song, Yena Ko, and Ji Soo Choi. 2022. [Fact-checking and audience engagement: A study of content analysis and audience behavioral data of fact-checking coverage from news media](#). *Digital Journalism*, 10(5):781–800.
- S. Lewandowsky, L. Smillie, D. Garcia, R. Hertwig, J. Weatherall, S. Egidy, R.E. Robertson, C. O’connor, A. Kozyreva, P. Lorenz-Spreen, Y. Blaschke, and M. Leiser. 2020. [Technology and Democracy: Understanding the influence of online technologies on political behaviour and decision-making](#). Publications Office of the European Union.
- Jianing Li, Jordan M. Foley, Omar Dumdum, and Michael W. Wagner. 2022. [The power of a genre: Political news presented as fact-checking increases accurate belief updating and hostile media perceptions](#). *Mass Communication and Society*, 25(2):282–307.
- Philipp Müller and Rainer Freudenthaler. 2022. [Right-wing, populist, controlled by foreign powers? Topic diversification and partisanship in the content structures of German-language alternative media](#). *Digital Journalism*, 10(8):1363–1386.
- Nic Newman, Richard Fletcher, Craig T. Robertson, Kirsten Eddy, and Rasmus Kleis Nielsen. 2022. [Reuters Institute Digital News Report 2022](#). Reuters Institute for the Study of Journalism.
- Katharine Ognyanova, David Lazer, Ronald E. Robertson, and Christo Wilson. 2020. [Misinformation in action: Fake news exposure is linked to lower trust in media, higher trust in government when your side is in power](#). *Harvard Kennedy School (HKS) Misinformation Review*.
- R Core Team. 2023. [R: A Language and Environment for Statistical Computing](#). R Foundation for Statistical Computing, Vienna, Austria.
- Jonas Rieger. 2020. [ldaPrototype: A method in R to get a prototype of multiple latent Dirichlet allocations](#). *Journal of Open Source Software*, 5(51):2181.
- Jonas Rieger, Carsten Jentsch, and Jörg Rahnenführer. 2022. [LDAPrototype: A model selection algorithm to improve reliability of latent Dirichlet allocation](#). Preprint available at Research Square.
- Lisa Schwaiger. 2022. [Gegen die Öffentlichkeit — Alternative Nachrichtenmedien im deutschsprachigen Raum](#). transcript.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *NIPS: Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Gerret von Nordheim, Jonas Rieger, and Katharina Kleinen von Königslöw. 2021. [From the fringes to the core — an analysis of right-wing populists’ linking practices in seven EU parliaments and Switzerland](#). *Digital Journalism*.
- Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. [The spread of true and false news online](#). *Science*, 359(6380):1146–1151.
- Vox Check. 2022. [Disinformation about Ukraine in Russian and pro-Russian Telegram channels](#).
- Nathan Walter, Jonathan Cohen, R. Lance Holbert, and Yasmin Morag. 2020. [Fact-checking: A meta-analysis of what works and for whom](#). *Political Communication*, 37(3):350–375.
- Hadley Wickham. 2021. [rvest: Easily Harvest \(Scrape\) Web Pages](#). R package version 1.0.2.
- Hadley Wickham. 2022. [httr: Tools for Working with URLs and HTTP](#). R package version 1.4.4.

Exploring Intensities of Hate Speech on Social Media: A Case Study on Explaining Multilingual Models with XAI

Raisa Romanov Geleta,¹ Klaus Eckelt,² Emilia Parada-Cabaleiro^{1,3,4} and Markus Schedl^{1,3}

¹Institute of Computational Perception, Johannes Kepler University Linz, Austria

²Institute of Computer Graphics, Johannes Kepler University Linz, Austria

³Human-centered AI Group, Linz Institute of Technology (LIT), Austria

⁴Department of Music Pedagogy, Nuremberg University of Music, Germany
raisa.geleta@gmail.com, klaus.eckelt@jku.at

Abstract

Hate speech on social media platforms has grown to become a major problem. In this study, we explore strategies to efficiently lessen its harmful effects by supporting content moderation through machine learning (ML). In order to present a more accurate spectrum of severity and surmount the constraints of seeing hate speech as a binary task (as typical in sentiment analysis), we classify hate speech into four intensities: no hate, intimidation, offense or discrimination, and promotion of violence. For this, we first involve 31 users in annotating a dataset in English and German. To promote interpretability and transparency, we integrate our ML system in a dashboard provided with explainable AI (XAI). By performing a case study with 40 non-experts moderators, we evaluated the efficacy of the proposed XAI dashboard in supporting content moderation. Our results suggest that assessing hate intensities is important for content moderators, as these can be related to specific penalties. Similarly, XAI seems to be a promising method to improve ML trustworthiness, by this, facilitating moderators' well-informed decision-making.

1 Introduction

The rapid growth of hate speech is a worrying problem that has been brought on by the immediate nature of social media (Mollas et al., 2022). Effectively limiting hate speech has become more difficult due to its wide impact and quick propagation (United Nation, 2023). Therefore, given the pressing need to address this issue, investigating efficient techniques and methodologies able to reduce its negative consequences has become crucial. By analyzing hate speech detection methods and the potential for XAI to improve transparency and interpretability, our study intends to support these initiatives.

Hate speech is typically characterized in research studies as either being hateful or not, i. e., in binary terms (Aluru et al., 2021; Deshpande et al., 2022; Duwairi et al., 2021; Roy et al., 2020; Plaza-del Arco et al., 2021; Del Vigna et al., 2017). Nonetheless, there have been instances where more nuanced classifications have been examined (Ibrohim and Budi, 2019; Mollas et al., 2022; Del Vigna et al., 2017). To get over this limitation, we adopted the levels by Olteanu et al. (2018), which include three unique intensities: intimidation, offense or discrimination, and promotion of violence. In addition, we included “no hate” to account for situations in which hate speech traits are not present.

Through the design science research (DSR) methodology (Peffer et al., 2007), we create an artifact that engages humans in the evaluation of hate speech, i. e., a dashboard to support social media content moderation. Inspired by Bunde (2021), our dashboard (depicted in Figure 1) includes novel features, such as a hate speech detection algorithm based on Universal Language Model Fine-Tuning, SHapley Additive exPlanation (SHAP) (Lundberg and Lee, 2017) text heat mapping, text similarity, and a four-level hate speech intensity scale. Our dashboard enables moderators to comprehend and explore the underlying assumptions of the machine learning (ML) model's predictions, by this assisting them in making well-informed decisions.

We aim to answer two Research Questions:

RQ1: Are intensities of hate speech an important factor to be considered in content moderation?

RQ2: Is XAI a successful way to support moderators' judgment of social media content?

2 Related Work

A well-defined, linguistically nuanced, and intergroup-relationship-aware concept is required for an automated approach to be precise (Fortuna

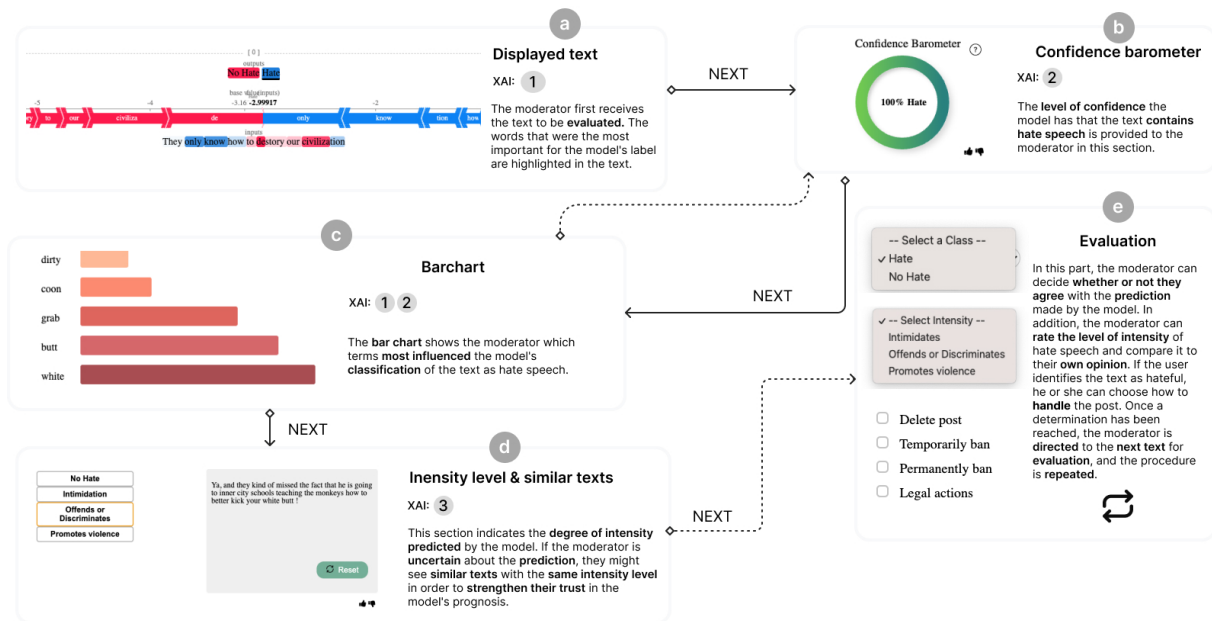


Figure 1: Chart flow diagram of the moderator's journey through the XAI dashboard.

and Nunes, 2018). Amongst the number of definitions proposed in the literature, Nobata et al. (2016) identifies hate speech as speech that disparages and attacks a group based on characteristics like ethnicity, religion, gender, or sexual orientation. Fortuna and Nunes (2018) defines it as language that criticizes or disparages groups based on particular traits: depending on the linguistic style, it might provoke violence or hate. Despite the attempts, hate speech detection is still limited by the lack of a distinct and widely accepted definition.

Besides the conceptual problems of defining hate speech, technical difficulties in detecting it include differences in training datasets as well as biases in ML algorithms (MacAvaney et al., 2019). In addition, developing a uniform method to identify hate speech is further impaired by the different laws regarding the right to free speech from different nations (United Nation, 2023). Still, the urgency of effectively combating hate speech on social media has led to the development of a variety of ML techniques aiming to automatically identify it. One approach for transparent hate speech detection is Masked Rationale Prediction (MRP), introduced by Kim et al. (2022). MRP uses context-relevant tokens and unmasked rationales to anticipate masked human rationales in order to reduce bias and increase explainability. To detect hate speech on Twitter, Zhang et al. (2018) devised a C-GRU, which combines a CNN and a gated recurrent network (GRU), while Khan et al. (2022) introduced a deep learning model called BiCHAT that combines contextual word representation, deep CNN,

BiLSTM, and hierarchical attention to successfully detect hate speech in Twitter.

Despite the promising outcomes, the application of ML in detecting hate speech presents still limitations. Nobata et al. (2016) emphasized that some forms of hate speech are not sufficiently investigated. Furthermore, it is well-known that ML models are affected by biases that negatively impact the decision-making process (Molnar, 2022). The lack of transparency of many ML models makes it more difficult to spot and correct such biases. Due to this, works like the one Mehta and Passi (2022) and Bunde (2021) have started looking at the possibility of using XAI to enhance the interpretability of hate speech recognition systems.

3 Methods

3.1 Dataset of Hate Speech

Since it has been shown that hate speech recognition through ML can be affected by the target language (Aluru et al., 2021), we investigate two languages in our study. In order to create a meta-corpus of hate speech in English and German, we collected pre-existing hate speech datasets in both languages, which included GermEval¹ (Wiegand, 2019), hasoc-fire-2020² (Dowlagar and Mamidi, 2021), UCSM-DUE GHSR³ (Ross et al., 2016) and those by Davidson et al. (2017) and de Gibert et al. (2018). From each language, a total of 1,500

¹<https://github.com/uds-lsv/GermEval-2018-Data>

²<https://github.com/suman101112/hasoc-fire-2020>

³https://github.com/UCSM-DUE/IWG_hatespeech_public

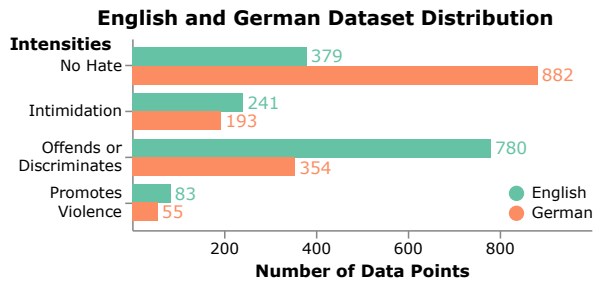


Figure 2: Distribution of annotator classifications for each intensity in English and German languages.

texts were randomly selected and annotated according to the labels proposed by Olteanu et al. (2018). Texts that contained only links or a username were removed, resulting in 1,437 and 1,476 samples for English and German, respectively. We reached out to potential annotators using social media sites including Instagram, Facebook, and Github. 31 contributors (18 males, 13 females, in a 26-35 age range) took part in the annotation process. A user interface was developed using Streamlit to enable users to annotate the data according to the hate intensity values. The application’s source code is freely accessible.⁴

Before taking part in the experiment, the annotators were required to agree to the participation terms, which stipulated that their anonymous responses would be used for scientific research.⁵ Each participant was instructed on the task before annotating a minimum of 10 samples in the chosen language. The annotators were requested to identify the level of hate expressed in the text through a forced-choice test. They could choose one of the following intensities: (i) *no hate*, (ii) *intimidation*, (iii) *offends or discriminates*, (iv) and *promotes violence*. The distribution of annotations across intensities and languages, shown in Figure 2, is highly imbalanced, which we expect to affect the ML performance. Compared to the other labels, the most extreme intensity *promotes violence* was chosen by far fewer times in both languages. The majority of German data was rated as *no hate*, whereas the majority of the English data was rated as *offends or discriminates*.

3.2 Dashboard

We developed an XAI dashboard⁶ that supports multi-lingual evaluation to enhance content moder-

⁴https://github.com/Raisarom/Streamlit_AnnotationApp

⁵The procedures used in this research were carried out in accordance with the tenets of the Declaration of Helsinki.

⁶<https://github.com/Raisarom/Hate-Speech-Detection-Dashboard-with-XAI>

ation strategies for safer online communities. Figure 1 depicts the interaction flow in the moderation dashboard. The first section (Fig. 1a) displays the input text, predicted label, and highlights the words that contributed to—or against—the prediction with a heatmap based on the words’ SHAP values (Lundberg and Lee, 2017). We additionally calculate the predicted probabilities’ entropy, with higher values indicating greater certainty, to assess the ML model’s trustworthiness with the *Confidence barometer* (Fig. 1b) (Bogert, 2021). The bar chart in Figure 1c ranks the words most influential on the classification of *hate* or *no hate*. By visualizing the trustworthiness of the model and highlighting important words, users can make informed decisions and develop a deeper understanding of the underlying model.

The next section of the dashboard (Fig. 1d) displays the text’s hate speech intensity and similar texts classified with the same intensity. A nearest neighbor search identifies text samples of similar content and hate intensity. These samples for the predicted intensity provide contextual information to enhance moderator precision.

The moderator can then evaluate the model’s prediction and determine whether or not they concur with it (Fig. 1e). If the text is identified as non-hateful, the dashboard automatically directs the moderator to the next text. If the text is identified as hate speech, the moderator is prompted to select the level of hate speech intensity and decide on the appropriate action to take against the person who posted the text. The moderator can also rate the usefulness of the XAI methods and provide feedback by selecting the thumbs-up or thumbs-down icon next to each method (Fig. 1 1-4).

3.3 User Study

To test the XAI dashboard along with other evaluation methodologies we performed a user study with 40 volunteers (26 male, 14 female). Most of them were university students ($n = 34$) and around half Austrian ($n = 22$); the rest of participants were spread amongst 11 nationalities. Due to the imbalanced distribution, the potential effect of these attributes will not be evaluated. The individuals who exhibited the greatest level of skill in their particular languages were intentionally allocated to either the German or English cohort.

The goal of the user study was to assess whether different evaluation methodologies influence mod-

erators' decisions (see Figure 3). With evaluation methodologies, we refer to the underlying methods used to assign a hate label (suggested to the moderator) to a given text (presented to the moderator for evaluation). Four evaluation methodologies were assessed: A) labels suggested by a human; B) labels suggested by AI; C) labels suggested by a human who revised AI ratings; D) labels from AI assessed through the XAI dashboard. For each language, 10 participants were randomly assigned to each group. Their task was to act as "moderators" i. e., for a given text they would get a suggested label, and subsequently they were requested to rate the text. In case of disagreement w. r. t. the suggested label, they were requested to indicate the appropriate intensity of hate. To ensure an objective evaluation, moderators did not know to which group they were assigned.

3.4 ML Models Implementation

We implemented a system able to distinguish first between hate and no hate speech; subsequently between three fine-grained intensities (intimidates, offends, and promotes violence). Due to the limited size of our dataset, pre-trained Hate Bert Models from Huggingface were used to classify the data into hate and no hate, individually for each language (see Section 4). We also evaluated a multilingual Hate Bert model to test the machine's capacity to classify both languages together. The pre-trained models were fine-tuned with our re-annotated data of the respective language, or both languages for the multilingual model. The annotated data was also used to train several ML algorithms to additionally identify the hate intensity in the texts. These algorithms included Random Forest (RF), Support Vector Machine (SVM), Naive Bayes (NB), Fasttext classifier, and a Dummy classifier used as baseline to evaluate the performance of the other classifiers. We opted for this two-step approach to leverage the information of the pre-trained models to improve the overall detection of hate speech and focussed on traditional algorithms instead of deep learning models due to the small size of the dataset and its imbalanced character.

Before training the models, the data was preprocessed following standard techniques in text processing, such as lowercase conversion, punctuation removal, stop-word removal, and lemmatization. The model performance will be evaluated in terms of precision, recall, F1 score, and accuracy metrics.

4 Results

4.1 ML Accuracy

In this study, separate BERT models were trained for each language to predict two output labels: hate and no hate. An approximate data split of 75-10-15 was aimed for, with slight deviations due to efforts to create a balanced test dataset. The distribution of sequence lengths in the dataset was examined to determine the optimal `max_length` for tokenization. The corresponding `AutoTokenizer` from the pre-trained BERT models was used, and the models were trained using `CrossEntropyLoss` and the Adam optimizer. Class weights were calculated based on the class distribution in the training set and added to the `CrossEntropyLoss` function to balance the contribution of each class during training. A scheduler was employed to adjust the learning rate during training. The training parameters provided by Liu et al. (2019) were followed.

In order to recognize the intensity of hate, we also trained a different model for each language. Due to space constraints only the optimal hyperparameters for the Random Forest classifier (which achieved best results) are given. According to the conducted grid search, the parameters were: `max_depth` \in 20, `min_samples_leaf` \in 2, `min_samples_split` \in 2 and `n_estimators` \in 100.

Table 1 shows the best performance by the pre-trained Hate BERT models for each language. While we also considered a model trained solely on English data,⁷ the Multilingual-hatespeech-robacofi⁸ Model (M-BERT) obtained the highest accuracy of 72% for the English dataset. The Bert-base-german-cased-hatespeechGermEval18Coarse⁹ Model (BERT-GER) achieved an accuracy of 68% in the German dataset. Overall, the Multilingual Bert Model outperformed the German one, especially in terms of precision and recall for the English data. Still, both models demonstrated comparable F1-scores. Among all the classifiers for hate speech intensity, the RF classifier achieved the highest accuracy with 38% on the English dataset and 48% on the German one. Note that in a three-class problem, these results, although low, are still above chance.

⁷<https://huggingface.co/Hate-speech-CNERG/dehatebert-mono-english>

⁸<https://huggingface.co/Andrazp/multilingual-hate-speech-robacofi>

⁹<https://huggingface.co/deepset/bert-base-german-cased-hatespeech-GermEval18Coarse>

Model Type	Dataset	Label	Precision	Recall	F1-Score	Accuracy
M-BERT	English	Hate	0.76	0.64	0.69	0.72
		No Hate	0.69	0.8	0.74	
BERT-GER	German	Hate	0.68	0.69	0.69	0.68
		No Hate	0.69	0.67	0.68	

Table 1: Performance of BERT models on English and German datasets for hate speech detection.

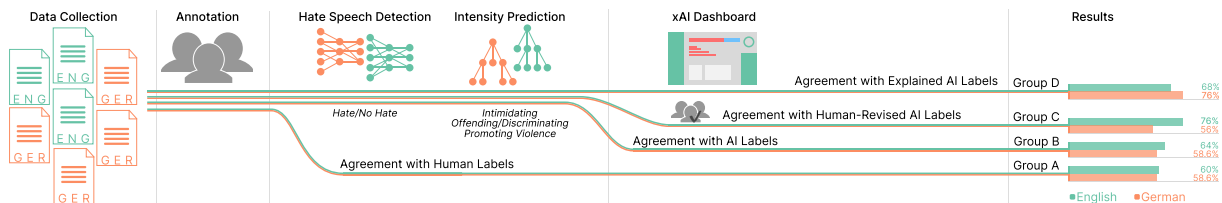


Figure 3: Design and results of the study comparing evaluation methodologies on the German and English datasets.

4.2 Dashboard Evaluation

We assessed the percentages of agreement within and across groups in order to evaluate each evaluation methodology’s efficacy. The findings of our user case study are shown in Figure 3, along with the percentages of matches for each category and language. Groups A and B exhibited similar rates of agreement for the German group, however, Group C had a somewhat lower rate. With 76%, Group D had the highest level of agreement. The outcomes were a little different for the English group: Group A had the lowest match rate followed by Group B and Group D. The greatest match rate was in Group C with 76%.

Groups D and C had quite high agreement percentages. The results from Group D suggest that the dashboard’s extra explanations enhance participants’ confidence in their choices. Still, the results from Group C, highlight the importance of involving a person in the decision-making process.

Additionally, we looked into how the severity of hate speech related to moderator action. Spearman correlation indicated a smaller link between the intensity of hate speech and moderator actions in German ($r \approx 0.19$) than in English ($r \approx 0.54$).

5 Discussion and Limitations

The BERT model’s inferior accuracy is probably due to the small amount of annotated data (about 1,450 data points), which constitutes one of the main limitations of our work. Indeed, larger datasets are often needed to attain the best performance for deep learning models like BERT, as shown in previous works (Saleh et al., 2023). Concerning the classification of hate intensity, the imbalance of our dataset contributed further to the

low ML accuracy. There were remarkably few annotated data points, especially for the “promotes violence” category. Indeed, obtaining high-quality annotations for hate speech is a well-known problem, already highlighted by previous works (Del Vigna et al., 2017).

The outcomes from the user study revealed that there was a prominent bias toward political hate speech in the German data. This may, indeed restrict the usability of the German model in non-political hate speech, which highlights the need of collecting high-quality and representative dataset across multiple languages and contexts. Similarly, although the majority of study participants agreed with the utilized intensities, they also proposed adding others such as irony or sarcasm, which should be considered in the future research.

6 Conclusions

Concerning RQ1, our study shows that, especially for English, low hate intensities were generally related to moderator actions of low severity, such as *delete post* or *temporary ban*, while a higher hate intensity was mostly linked to permanent bans. This suggests that hate speech intensity might be a criteria to undertake specific moderator actions. Concerning RQ2, our results from the German data indicate that XAI improves the decision-making capabilities of moderators, as shown by a higher agreement with respect to the other methods.

We showed that defining hate speech in terms of intensities, as well as developing XAI tools, are both promising ways to improve the quality and effectiveness of online-content moderation, by this making the internet a safer place for everyone.

Acknowledgements

A special thanks to all of our participants. Without them, this study would have not been possible.

References

- S. S. Aluru, B. Mathew, P. Saha, and A. Mukherjee. 2021. **A Deep Dive into Multilingual Hate Speech Classification**. In *Proc. ECML PKDD*, pages 423–439.
- K. Bogert. 2021. **Notes on Generalizing the Maximum Entropy Principle to Uncertain Data**. *arXiv*.
- E. Bunde. 2021. **AI-Assisted and Explainable Hate Speech Detection for Social Media Moderators – A Design Science Approach**. In *Proc. HICSS*, pages 1264–1273.
- T. Davidson, D. Warmley, M. Macy, and I. Weber. 2017. **Automated Hate Speech Detection and the Problem of Offensive Language**. *Proc. ICWSM*, pages 512–515.
- O. de Gibert, N. Perez, A. García-Pablos, and M. Cuadros. 2018. **Hate Speech Dataset from a White Supremacy Forum**. In *Proc. ALW2*, pages 11–20.
- F. Del Vigna, A. Cimino, F. Dell’Orletta, M. Petrocchi, and M. Tesconi. 2017. **Hate Me, Hate Me Not: Hate Speech Detection on Facebook**. In *Proc. ITASEC*, pages 86–91.
- N. Deshpande, N. Farris, and V. Kumar. 2022. **Highly Generalizable Models for Multilingual Hate Speech Detection**. *arXiv*.
- S. Dowlagar and R. Mamidi. 2021. **HASOCone@FIRE-HASOC2020: Using BERT and Multilingual BERT models for Hate Speech Detection**. *arXiv*.
- R. Duwairi, A. Hayajneh, and M. Quwaider. 2021. **A Deep Learning Framework for Automatic Detection of Hate Speech Embedded in Arabic Tweets**. *Arab. J. Sci. Eng.*, 46:4001–4014.
- P. Fortuna and S. Nunes. 2018. **A Survey on Automatic Detection of Hate Speech in Text**. *ACM Comput. Surv.*, 51(4):1–30.
- M. O. Ibrohim and I. Budi. 2019. **Multi-Label Hate Speech and Abusive Language Detection in Indonesian Twitter**. In *Proc. ALW*, pages 46–57.
- S. Khan, M. Fazil, V. Sejwal, M. Alshara, R. Alotaibi, A. Kamal, and A. Baig. 2022. **BiCHAT: BiLSTM with deep CNN and hierarchical attention for hate speech detection**. *J. King Saud Univ. Comput. Inf. Sci.*, 34:4335–4344.
- J. Kim, B. Lee, and K. Sohn. 2022. **Why Is It Hate Speech? Masked Rationale Prediction for Explainable Hate Speech Detection**. In *Proc. COLING*, pages 6644–6655.
- Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. **RoBERTa: A Robustly Optimized BERT Pre-training Approach**. *arXiv*.
- Scott M Lundberg and Su-In Lee. 2017. **A Unified Approach to Interpreting Model Predictions**. In *Proc. NIPS*, pages 4765–4774.
- S. MacAvaney, H. Yao, E. Yang, K. Russell, N. Goharian, and O. Frieder. 2019. **Hate Speech Detection: Challenges and Solutions**. *PLoS one*, 14(8):e0221152.
- H. Mehta and K. Passi. 2022. **Social Media Hate Speech Detection Using Explainable Artificial Intelligence (XAI)**. *Algorithms*, 15:291.
- I. Mollas, Z. Chrysopoulou, S. Karlos, and G. Tsoumakas. 2022. **ETHOS: A Multi-Label Hate Speech Detection Dataset**. *Complex & Intell. Syst.*, pages 1–16.
- C. Molnar. 2022. *Interpretable Machine Learning*, 2nd edition.
- C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang. 2016. **Abusive Language Detection in Online User Content**. In *Proc. WWW*, pages 145–153.
- A. Olteanu, C. Castillo, J. Boy, and K. Varshney. 2018. **The Effect of Extremist Violence on Hateful Speech Online**. In *Proc. ICWSM*.
- K. Peffers, T. Tuunanen, M. Rothenberger, and S. Chatterjee. 2007. **A Design Science Research Methodology for Information Systems Research**. *J. Manage. Inf. Syst.*, 24(3):45–77.
- F. M. Plaza-del Arco, M. D. Molina-González, L. A. Ureña-López, and M. T. Martín-Valdivia. 2021. **Comparing Pre-Trained Language Models for Spanish Hate Speech Detection**. *Expert Syst. Appl.*, 166:114120.
- B. Ross, M. Rist, G. Carbonell, B. Cabrera, N. Kurowsky, and M. Wojatzki. 2016. **Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis**. In *Proc. NLP4CMC*, volume 17, pages 6–9.
- P. K. Roy, A. K. Tripathy, T. K. Das, and X.-Z. Gao. 2020. **A Framework for Hate Speech Detection Using Deep Convolutional Neural Network**. *IEEE Access*, 8:204951–204962.
- H. Saleh, A. Alhothali, and K. Moria. 2023. **Detection of Hate Speech using BERT and Hate Speech Word Embedding with Deep Model**. *Appl. Artif. Intell.*, 37(1).
- United Nation. 2023. **UN Strategy and Plan of Action on Hate Speech**. <https://www.un.org/en/hate-speech>.
- M. Wiegand. 2019. **GermEval-2018 Corpus (DE)**. heidata. V1.
- Z. Zhang, D. Robinson, and J. Tepper. 2018. **Detecting Hate Speech on Twitter Using a Convolution-Gru Based Deep Neural Network**. In *Proc. ESWC*, pages 745–760.

Assessing Italian News Reliability in the Health Domain through Text Analysis of Headlines

Luca Giordano

UNIOR NLP Research Group
University of Naples "L'Orientale"
giordanoluca.uni@gmail.com

Maria Pia di Buono

UNIOR NLP Research Group
University of Naples "L'Orientale"
mpdibuono@unior.it

Abstract

Fake news detection and fact checking represent challenging research areas in Natural Language Processing (NLP), especially in the health domain, which presents specific characteristics to be dealt with. On the one hand, online sources have become one of the main channels to retrieve health-related information. On the other hand, most of the time such online information suffers from lack of quality and requires domain-specific knowledge to be assessed. Therefore, the spread of untrustworthy health-related content urges to be mitigated since it may represent a threat for lives.

To this aim, we develop a domain-specific annotated dataset suitable for training automatic systems to assess Italian news reliability. Our proposal tries to overcome some of the limitations of the available datasets by applying an in-depth text analysis to obtain a more fine-grained reliability assessment in the health domain.

1 Introduction

Lately, the use of online sources for retrieving health information has become widespread, and thus an important source of medical advice (Dai et al., 2020). Particularly, social media platforms (SMPs) seem to be one of the most preferred channels to search and share information, especially in the health domain (Chen et al., 2018). As proved by several scholars (e.g., Finney Rutten et al. (2019); Basch et al. (2017)), the Internet and SMPs represent the main source of information for adults and also adolescents that are active users and searchers for online health information (Greškovičová et al., 2022).

Nevertheless, online health information is affected by several limitations with reference to its quality (Melchior and Oliveira, 2022). The lack of quality in information may generate two main types of untrustworthy content, namely disinformation and misinformation (Lazer et al., 2018).

Nowadays, fighting the spread of untrustworthy and low-quality content through fake news detection and/or fact checking represents one of the main challenges to be faced. This is particularly true in the medical domain because such untrustworthy health-related content threaten lives (Anoop et al., 2020).

The Covid-19 pandemic has exacerbated the problem and brought out the need for gold standard datasets and predefined benchmarks for automated approaches, which have been neglected before that, as revealed by Viviani and Pasi (2017).

In fact, the scarcity of comprehensive resources, mainly datasets, for fake health news detection slows down the development of novel approaches devoted to detect misinformation and disinformation within this domain (Dai et al., 2020).

Still, the development of resources suitable for assessing information and news in the health domain is far to be fully satisfied, mainly with reference to some domain-specific aspects and languages.

For this reason, in this paper we present a domain-specific annotated dataset suitable for training automatic systems to assess Italian news reliability. Our proposal tries to overcome some of the limitations of the available datasets and to propose a more fine-grained assessment of health-related news, achieved through an in-depth text analysis. Our main contributions are three: (i) proposing a set of stylometric, lexical, and sentiment features to assess news reliability; (ii) developing a domain-specific dataset for the Italian language¹; (iii) providing a first baseline for the developed dataset.

The rest of the paper is organized as follows. In the next section, we present studies which are relevant to our analysis, referring mainly to the development of datasets for fake news detection. In Section 3, we introduce our methodology, our dataset and the feature set. In Section 4 we explain the experimen-

¹The dataset is publicly available at <https://github.com/unior-nlp-research-group/TRADISAN>.

tal setup and present the results. Finally in Section 5 conclusion and future work are discussed.

2 Related Work

The majority of studies published and resources made available focus on a binary classification of the veracity of English news at document-level (that is, an overall veracity rating either True or False for the whole news), although tested by means of different kinds of analysis (such as a range of linguistic features, e.g., Choudhary and Arora (2021); Kasseropoulos and Tjortjis (2021), sentiment analysis, e.g., Alonso et al. (2021) and others). As shown in D’Ulizia et al. (2021), out of the 27 datasets surveyed in the paper, 14 present a binary veracity classification (such as Shu et al. (2020); Tacchini et al. (2017)), while only 4 of them a three-way rating scale (such as Thorne et al. (2018)) and 6 a four-way one (such as Santia and Williams (2018)). Furthermore, 22 out of 27 are monolingual English datasets, only 2 are focused on the Health domain (Posadas-Durán et al., 2019; Jwa et al., 2019) and all of them are annotated at document-level.

Although in Bonet-Jover (2022) the classification proposed is still binary (Reliable/Unreliable), it is noteworthy that the author works on Spanish and that the annotation proposal is focused on the individual annotation of different structural and content elements of the news, therefore going beyond the document-level of analysis.

Regarding the Italian language, to the best of our knowledge, there seems to exist only one publicly accessible dataset of Italian news annotated according to their veracity value, namely HoaxItaly (Pierri et al., 2020): it is a dataset composed of 1.2M tweets referring to 37k Italian news in total, divided into 3566 fact-checked true news and 32,686 fake news. However, the news domain is generic, the assessment is binary and at document-level.

With reference to the set of features typical of trustworthy and untrustworthy news respectively, several studies highlight different kinds of linguistic patterns.

In Biyani et al. (2016) the authors show that the degree of informality of a webpage, as measured by different metrics, is a strong indicator of it being a clickbait, that is an article with a misleading headline, exaggerating the content on the landing page. The amount of superlatives, quotes, exclamations, upper case letters, question marks and other indi-

cators are used as features for a machine-learning model which achieves a 74.9% F1 score in predicting clickbaits.

Horne and Adali (2017) apply a set of linguistic features to three datasets in order to analyze the language of news articles in the political domain. They show that stylistic features such as the length of the article, the use of punctuation, the amount of personal pronouns, nouns and adverbs, the lexical redundancy of the text and others, applied both to the headline and to the body of the news, can help distinguish between real and fake news. Their findings are mostly confirmed by Shrestha and Spezzano (2021), who conduct a reproducibility study, and in addition show that also other factors, such as emotion and readability features are helpful in the fake news detection task.

In Rashkin et al. (2017) the authors show that features such as the amount of swear words, hedge words, sexual-related words, negations, superlatives and others appear to be typical of fake political news, while a frequent use of numbers, money-related words, assertive expressions and comparatives appear to be typical of true political news.

Greškovičová et al. (2022) show that seemingly minor editorial elements, such as poor grammar or boldface, in addition to the presence of superlatives, clickbaits and appeal to authority in health-related messages, which are all typical elements of untrustworthy news, influence and distort the perception of the credibility of news among secondary school students.

3 Methodology

Dai et al. (2020) identified several challenges that have to be addressed in fake health news detection, as they are specific of this domain. In fact, fake health news may require specialized knowledge to be recognized more than fake news in other domains.

Furthermore, health news are also easier to be manipulated, in that they can be easily transformed into misinformation or disinformation just by stating the association as causation or mixing up the absolute risk and relative risk, which, as Dai et al. (2020) point out, require just minor modifications of the true information.

Thus, the proposed methodology tries to combine the identification of trustworthy sources together with the integration of linguistic and sentiment features selected by means of an in-depth analysis. To

our aims, we adopt the criterion of reliability instead of veracity, to distinguish untrustworthy news from trustworthy ones and assume that stylometric, lexical and sentiment-based characteristics can be representative of the degree of news reliability.

As first step, we collect a list of news sources (i.e., online newspapers) which have been classified as trustable or untrustable by Newsguard², Media Bias/Fact Check³, Bufale.net⁴ and Butac⁵, two international and two italian fact checking organizations which, among other activities, publish analyses and reports on news sources' trustworthiness. Furthermore, we take into account the data and analysis provided in the Digital News Report 2022 for Italy published by the Reuters Institute for the Study of Journalism⁶. Therefore, we create two lists of sources, respectively a *trustworthy* list and an *untrustworthy* list (Table 1).

We use these sources to extract a set of health-related news, using the classification by categories provided by the newspapers themselves together with a topic-label based extraction. This allows us to come up with a list of both trustworthy and untrustworthy news. Then, we perform a linguistic analysis to select a set of features that are representative of news reliability.

3.1 Data Collection

The list of trustworthy sources is made up of 12 Italian news outlets (e.g., Il Sole 24 Ore⁷, la Repubblica⁸, ANSA⁹), while the list of untrustworthy sources is made up of 26 Italian news outlets (e.g., Voxnews¹⁰, Dionidream¹¹, Byoblu¹²) for a total amount of 38 news sources (Table 1).

In order to collect the data from our sources, we write Python scripts tailored to each news outlet in order to scrape the news content. We exploit the Python libraries *pandas*¹³, *requests*¹⁴,

²<https://www.newsguardtech.com/it/>

³<https://mediabiasfactcheck.com/>

⁴<https://www.bufale.net/>

⁵<https://www.butac.it/>

⁶<https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2022/italy>

⁷<https://www.ilsole24ore.com/>

⁸<https://www.repubblica.it/>

⁹<https://www.ansa.it/>

¹⁰<https://voxnews.info/>

¹¹<https://dionidream.com/>

¹²<https://www.byoblu.com/>

¹³<https://pandas.pydata.org/>

¹⁴<https://requests.readthedocs.io/en/latest/>

*beautifulsoup*¹⁵ and *newspaper3k*¹⁶, which stem from machine-learning and data science. We aim at extracting the URLs of each article in the health-related categories of the news outlets and through those we extract the news content, that is the article's source, date of publication, headline, body of text and links to its images, if any (Table 2).

Then, we remove broken links and articles with missing information, as well as duplicate articles from the same source. We also remark a potentially interesting phenomenon: 28 articles among the ones extracted from the trustworthy sources and 17 among the ones from the untrustworthy sources present an identical headline, despite being published by different sources. This might suggest plagiarism among news outlets. We keep these articles in our dataset since they might be significant, although we are aware that the presence of duplicates might affect the training data. Nevertheless, they represent a small part within the total amount of data. From the trustworthy list we keep a total of 9.973 news, which amount to 156.372 sentences and 4.925.379 tokens (we adopt the default AntConc token definition "Character Classes"¹⁷); from the untrustworthy list we keep a total of 22.128 news, which amount to 611.433 sentences and 17.648.641 tokens. Therefore, the corpus is made up of a total of 32.101 news published between November 1999 and February 2023, and it amounts to 767.805 sentences and 22.574.020 tokens (Table 3). To the aim of the present analysis we consider just news headlines, which amount to a total of 351.104 tokens.

3.2 Linguistic Analysis

In order to select the features suitable for our news assessment, we perform an initial analysis of our corpus to identify a first set of linguistic aspects denoting (un)reliability. We adopt a method which includes a top-down approach, namely applying features already used by other scholars for other languages and domains (see Section 2), and a bottom-up approach, that is we analyse the dataset and collect features that arise from our set of news.

¹⁵<https://beautiful-soup-4.readthedocs.io/en/latest/>

¹⁶<https://newspaper.readthedocs.io/en/latest/>

¹⁷<https://laurenceanthony.net/software/antconc/releases/AntConc4011/help.pdf>, p.13

Trustworthy	Untrustworthy	
Salute.gov	Eticamente.net	Vacciniinforma
ISS	Raffaele Palermo News	eVenti Avversi
la Repubblica	Nexus Edizioni	ByoBlu
il Post	Scienza e Conoscenza	Come Don Chisciotte
Vaccinarsi.org	COMILVA	VoxNews
il Fatto Quotidiano	Vivo in Salute	Mag24
TPI	The Living Spirits	Dagospia
AGI	Il Paragone	Filosofia e Scienza
ANSA	Database Italia	controinformazione.info
Focus	Ingannati	Disquisendo
Il Sole 24 Ore	CheSuccede	Essere Informati
Corriere della Sera	SocialBuzz!	Eurosalus
	Silenzi e Falsità	Dionidream

Table 1: Data Sources

ID	Source	Date	Headline	Text	Image	URL
3762	la Repubblica	2019/04/12	Fagioli e spinaci tengono lontano il tumore della vescica	...	Image1.jpg	...
9626	Il Sole 24 Ore	2023/01/12	Più contagi, non casi più gravi e lo scudo dei vaccini: ecco perché la variante Kraken non deve fare paura	...	Image1.jpg	...
15526	ByoBlu	2022/09/21	“BILL GATES HA GESTITO IL COVID PER ARRICCHIRSI”: ORA SE NE ACCORGE ANCHE IL MAINSTREAM	...	Image1.jpg	...
18104	VoxNews	2021/04/09	RECORD DI MORTI SPALMATI: 718 IN 24 ORE, 9 APRILE SCORSO ANNO ERANO STATI 612	...	Image1.jpg	...

Table 2: Examples of Trustworthy (IDs 3762 and 9626) and Untrustworthy (IDs 15526 and 18104) Entries from our Corpus

List	# News	# Sentences	# Tokens
Trust.	9.973	156.372	4.925.379
Untrust.	22.128	611.433	17.648.641
TOTAL	32.101	767.805	22.574.020

Table 3: Corpus Description

We obtain a total number of 31 features (Table 4) accounting for three different levels of analysis, namely stylometry, lexicon, and sentiment.

Stylometric Features The stylometric features we take into account refer to sentence and word length (by characters), the use of uppercase style, the frequency of consecutive question and exclamation marks, frequency of quotes, double quotes and single quotes, ellipses and direct discourse. We also compute the amount of typos through a customized Contextual Spell Checker¹⁸, a deep-learning based Noisy Channel Model Spell Algorithm trained on the PAISÀ Corpus¹⁹, one of the largest publicly available corpora of Italian Web texts, licensed under Creative Commons.

¹⁸Towards Data Science - Training a Contextual Spell Checker for Italian Language

¹⁹<https://www.corpusitaliano.it/>

The number of words written in uppercase, the number of long words (understood as being longer than 6 characters) and the number of typos are all weighted values accounting for the length of the sentence.

Lexical Features The lexical features we compute are the number of adverbs, comparatives, superlatives, currency-related words (such as *dollar*), negative adverbs, nouns, proper nouns, adjectives, possessive adjectives other than the 1st and 2nd singular and digits. Additionally, we exploit the Revised HurtLex (Tontodimamma et al., 2022), a lexicon of offensive, aggressive, and hateful words divided into 17 categories in over 50 languages in order to compute the number of occurrences of such words in the corpus. In the revised version, every Italian headword is annotated with an offensiveness level score, derived by applying an Item Response Theory model to the ratings provided by a large number of annotators (Tontodimamma et al., 2022). Therefore, we also compute the total offensiveness score of the sentence based on the scores of the words contained in it.

Furthermore, we also count the occurrences of

domain-specific *buzzwords*, understood by the definition provided by the Cambridge Dictionary: "a word or expression from a particular subject area that has become fashionable by being used a lot, especially on television and in the newspapers"²⁰. For this purpose, we compile a gazetteer of 73 words and phrases extracted from the top 300 keywords in the corpus sorted by likelihood and from the top ranking bigrams and trigrams sorted by frequency. Some examples of buzzwords in our gazetteer are *vaccino* (vaccine), *covid*, *coronavirus*, *sintomi* (symptoms), *immunità di gregge* (herd immunity), *lockdown*, *AIDS*, *green pass*, *vaiolo delle scimmie* (monkeypox) and *no vax*. We assume that Covid-19 global impact, urgency, and relevance as a major health crisis have led to a significant concentration of Covid-19-related keywords in the corpus, despite the pandemic started only in 2020, while the corpus contains news up to 1999. This might be evidence of the impact of the pandemic on news production in Italy. Therefore, we choose to keep this statistical bias in our buzzwords gazetteer as well.

All lexical features, except for the offensiveness score, are weighted values accounting for the length of the sentence.

Sentiment Features Additionally, we exploit the adoption of sentiment-related features. This comes from the fact that several scholars (Alonso et al., 2021; Bhutani et al., 2019; Ajao et al., 2019) have recognized that the polarity and strength of sentiments expressed in text can improve the results in fake news and rumor detection tasks.

Thus, we apply NRC Emotion Intensity Lexicon (Mohammad and Turney, 2013) to detect and evaluate the presence of emotions-related words within the texts, such as *anger*, *joy*, and *trust*. In fact, we notice that news from the untrustworthy sources are characterized by a more frequent use of words associated with negative emotions, such as *anger*, e.g., Example (1), while trustworthy news tend to express more positive emotions, such as *joy* or *trust*, e.g., Example (2).

Source: Disquisendo - ID: 26847

Il governo italiano ha dichiarato GUERRA agli italiani. OBBLIGO VACCINALE che passa da 4 a 12 e fino a 16 anni!! SVEGLIAAAAA!! (The Italian government has declared WAR on the Italians.

²⁰<https://dictionary.cambridge.org/it/dizionario/inglese/buzzword>

COMPULSORY VACCINATION goes from 4 to 12 and up to 16 years!! WAKE UUUUUP!! Source: La Repubblica - ID: 4148

Lo smartwatch? Può salvare la vita (letteralmente) (The smartwatch? It can (literally) save lives) Furthermore, through SentITA (Nicola, 2018) we also consider the sentiment polarity of the headlines, as untrustworthy news tend to present a mostly negative polarity while trustworthy news a mostly positive one (Shrestha and Spezzano, 2021).

3.3 Reliability Assessment

We perform an analysis of news headlines from both trustworthy and untrustworthy sets, according to the aforementioned features and use these results to define a textual model. The textual model characterizes the set of untrustworthy news headlines and presents the following linguistic aspects:

- Longer headlines (by characters);
- Frequent use of uppercase style;
- Presence of consecutive question and exclamation marks;
- Higher frequency of ellipses, typos, double and single quotes (but less direct discourse);
- Higher frequency of adverbs, superlatives, first person singular pronouns and negative adverbs;
- Limited use of comparatives, currency-related words, nouns, adjectives, second person singular pronouns and digits;
- Higher frequency of words and phrases from the HurtLex lexicon and a higher offensiveness score;
- Higher frequency of proper nouns;
- Slightly higher frequency of buzzwords;
- Lower frequency of lexical items related to trust and joy.

Then, the textual model is employed to assess the headline reliability.

3.4 Dataset Creation

On the basis of such methodology, we create a dataset which contains the information related to the textual model for assessing reliability (Figure 1). The selected features are annotated according to their pertaining level, that is stylometric (*styl*), lexical (*lex*), and sentiment (*sent*).

Stylo-metric	Lexical	Sentiment
char_count	adverb_count_w	nrc_anger_w
uppercase_word_w	comp_w	nrc_trust_w
long_w_w	superl_count_w	nrc_joy_w
consecutive_question_count	currency_w	opos
consecutive_excla_count	rev_hurtlex_count_w	oneg
quotes_count	hurtlex_score	
dou_quotes_count	neg_adverbs_count_w	
single_quote_count	noun_count_w	
ellipses_count	prop_noun_count_w	
direct_discourse	adj_count_w	
typo_count_w	adj_poss_others_w	
	1st_pers_sing_w	
	2nd_pers_sing_w	
	digits_w	
	buzzwords_count_w	

Table 4: List of the 31 Reliability Features

```
{
  "id": 16733,
  "source": "Come Don Chisciotte",
  "date": "2020-10-13 17:03:25+00:00",
  "url": "https://comedonchisciotte.org/la-farsa-dei-tamponi-e-degli-asintomatici/",
  "headline": "La FARSA dei Tamponi e degli Asintomatici",
  "styl": [
    {"char_count": 35.0, "uppercase_word_w": 0.142857, "long_w_w": 0.285714,
     "consecutive_question_count": 0, "consecutive_excla_count": 0, ...}
  ],
  "lex": [
    {"adverb_count_w": 0.0, "comp_w": 0.0, "superl_count_w": 0.0,
     "currency_w": 0.0, "rev_hurtlex_count_w": 0.0, ...}
  ],
  "sent": [
    {"nrc_anger_w": 0.1428571428571428, "nrc_trust_w": 0.0, "nrc_joy_w": 0.0,
     "opos": 0.021332342, "oneg": 0.9845031}
  ]
}
```

Figure 1: Annotation example for the headline ID: 16733 *La FARSA dei Tamponi e degli Asintomatici* (The FRAUD of Swabs and Asymptomatic patients), source: Come Don Chisciotte

In addition to this annotation at title-level, we also provide the dataset with additional annotations (Table 5), such as lemmatization (L), Part-of-speech tagging (PoS), Inside–Outside–Beginning chunk-tagging and (IOB) Named Entity Recognition tagging (NER). We test the annotated dataset performing an experiment to evaluate the results from some of the most common classifiers.

4 Experiment

We conduct a series of experiments to test our hypothesis, i.e. the assumption that stylometric, lexical and sentiment-based features can be suitable for assessing news reliability. Therefore, the main aim of these experiments is to test how fit our feature set is for an automatic assessment of news

reliability. Although the final goal is a fine-grained (multi-class) automatic reliability annotation of the whole dataset, for the sake of these experiments and its contextual aim (i.e., testing the feature set and the generalizability of the results for the dataset annotation, rather than the classification granularity and performance *per se*), we assume that every article from the untrustworthy and trustworthy lists make up only two separate classes, therefore configuring it as a binary classification problem.

Since the dataset is imbalanced, we perform an undersampling process, i.e. we extract a sample of random untrustworthy news equal to the (smaller) subset of trustworthy news (9973 samples). We end up with two equally sized subsets which amount to a total of 19946 samples. We justify the under-

H	ID	Token	L	PoS	IOB	NER
18192	1	AstraZeneca	AstraZeneca	PROPN	B	ORG
18192	2	vietato	vietare	VERB	O	–
18192	3	in	in	ADP	O	–
18192	4

Table 5: Annotation example extracted from the headline ID: 18192 *AstraZeneca vietato in Germania sotto ai 60 anni, Merkel: “Impossibile nascondere l’insicurezza”* (AstraZeneca banned in Germany under 60 years old, Merkel: “Impossible to hide uncertainty”), source: VoxNews

sampling since the final number of samples is still a considerable amount. Finally, we do not stratify the sampling process neither on date of publication, nor source of provenance nor any other factor since we aim at a subset as randomized as possible. After the random undersampling, the subset of untrustworthy news contains 2 of the 17 duplicates, while the subset of trustworthy news keeps all its original 28 duplicates.

Environmental Setup All code was written and compiled in Python 3.10 on Linux Ubuntu 23.04 and several packages and libraries were exploited, such as *pandas*, *NumPy*²¹, *SpaCy*²², *NLTK*²³, *Transformers*²⁴, *scikit-learn*²⁵, *fastText*²⁶ and *PyTorch*²⁷.

The Neural Network runs on an NVIDIA GeForce RTX™ 3060 Laptop GPU with CUDA v12.0.

Feature Selection In order to reduce computational cost, avoid overfitting, increase generalizability, and contribute to the explainability of the models, we apply statistical-based feature selection techniques, aiming at reducing the number of input variables to only those that have the strongest relationship with the target variable (Butcher and Smith, 2020). We adopt a filter-based univariate feature selection method. In detail, since we are dealing with numerical input variables and categorical output variables, we perform an analysis of variance (ANOVA) to compute the ANOVA correlation coefficient (F-value). ANOVA test is used to compare the means of different groups on a dependent variable and to determine whether the difference in group means is due to random variation or if they

represent true population differences. Its assumptions are independence, homogeneity of variances of the residuals and a normal distribution (Butcher and Smith, 2020). We assume that each feature is independent from the other, and, since we conduct the analysis on two equally big subsets built *ad-hoc*, we can also assume feature homogeneity (Sawyer, 2009).

Regarding normality of distribution, several scholars, e.g., Lumley et al. (2002); Ghasemi and Zahediasl (2012), show that with large sample sizes the distribution of data can be ignored, as the potential violation of the normality assumption does not cause problems. Moreover, the adoption of the ANOVA test is justified due to its robustness under conditions of non-normally distributed data, as proved by Schmider et al. (2010) and Blanca Mena et al. (2017). Since ANOVA test can be suitable for both normal and non-normal distributions, especially with large sample sizes and our sample size amounts to 19946 samples, we choose not to test normality and to perform directly the ANOVA test. Features are then sorted in descending order by the F-value computed with the ANOVA test to determine the importance. We choose to consider the topK features that have an F-value of more than 100. We therefore keep the top 13 features (Table 6).

Classification We conduct a series of experiments, testing five different machine-learning classifiers (namely, Logistic Regression, Decision Tree, Multinomial Naive-Bayes, Random Forest and LinearSVC) and, for BERT, a Multi-Layer Perceptron (MLP) with different input combinations and different word embedding techniques (namely, GloVe, fastText, and pre-trained BERT Base). We split the data in 90:10 training and testing ratio and make sure that all the duplicates are always only in the training set, since, as stated in Section 3.1, they might have been generated through a process we want to take into account. We then perform a cross-

²¹<https://numpy.org/>

²²<https://spacy.io/>

²³<https://www.nltk.org/index.html>

²⁴<https://huggingface.co/docs/transformers/index>

²⁵<https://scikit-learn.org/stable/>

²⁶<https://fasttext.cc/>

²⁷<https://pytorch.org/>

#	Top Features	F-value
1	prop_noun_count_w	1068.08
2	uppercase_word_w	832.12
3	char_count	630.26
4	dou_quotes_count	393.81
5	ellipses_count	387.03
6	single_quote_count	367.05
7	quotes_count	338.06
8	direct_discourse	223.11
9	noun_count_w	193.67
10	typo_count_w	172.22
11	long_w_w	170.29
12	oneg	159.22
13	hurtlex_score	128.98

Table 6: Top features calculated using ANOVA

validation on the training set, i.e. we split it into 10 train/validation subsets, while the test set remains unaltered. In each iteration, the training set is used for training while the validation set for validation. The performance measure reported is then the average of the values computed in the loop. For the MLP, the cross-validation is performed directly in the training loop, while, for the ML classifiers, through a GridSearchCV²⁸ technique implemented via scikit-learn, which also allows us to perform a hyperparameter optimization for every ML classifier. The cross-validation parameter is set to 10 folds. We then use the best estimator obtained for the classification task on the test set.

First, we try a classification taking only the whole 31 numerical features as input, without any word vector representation.

Then we ignore the features and classify the data only with three different word embedding techniques; we first try GloVe, then fastText and finally an italian XXL Bert Base transformer based model pre-trained on the whole italian Wikipedia, OPUS corpus and the italian subset of OSCAR corpus, for a total amount of 13,138,379,147 tokens²⁹. Being a Base model, it is made up of 12 layers of transformers block with a hidden size of 768 and 12 self-attention heads and has around 110M trainable parameters.

Then, we combine the different word embeddings with all 31 features.

²⁸https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

²⁹<https://huggingface.co/dbmdz/bert-base-italian-cased>

Finally, we classify the data with a combination of the different word embeddings and only the top 13 features we obtained from the feature selection process. We implement the MLP with PyTorch: the pooled output of the BERT encoder is used as input, the dropout rate is set at 0.5, the activation function is ReLu, the optimizer Adam, the loss function CrossEntropy, and we found that the optimal number of epochs is 6. When combining BERT with the features, the linear layer takes as input a tensor of length equal to the pooled output of BERT + the number of features.

Results The results (Table 7) show that, as expected, state-of-the-art BERT is the best model, achieving an F1 score of 0.855 alone, 0.884 when combined with the top 13 features. A classification based exclusively on our entire feature set achieves an F1 score of 0.70, while with only the top 13 it decreases to 0.68. Although the score is slightly lower, it is noteworthy that less than half of the original feature set were used. This emphasizes the importance of the feature selection process, and this must be taken into account for the dataset annotation, for example by assigning different weights to different features. The use of our features in all settings (alone, in combination with word embeddings and with BERT) improves the results, although slightly. The improvement is more considerable for fastText word embeddings than GloVe. These results show that this feature set can be a starting point for assessing Italian news reliability in the health domain.

5 Conclusion and Future Work

In this paper, we present our preliminary work on the automatic reliability assessment of Italian news in the health domain. Our methodology is based on the use of trustworthy and untrustworthy sources and the definition and selection of a set of stylistometric, lexical and sentiment features suitable for detecting misinformation and disinformation within health-related content. We believe that our approach can help improving the explainability of classification models thanks to our in-depth linguistic analysis. In addition, we also believe that the research community will be able to further exploit our annotated dataset to build upon this resource.

As future work, we intend to investigate further the linguistic features as well as the integration of information from external knowledge bases in order to check content manipulation. We also plan

Model	Classifier	P_{MacroAVG}	R_{MacroAVG}	F1
All Features	RandomForest	0.70	0.70	0.70
Top13 Features	RandomForest	0.68	0.68	0.68
fastText	LinearSVC	0.76	0.76	0.76
fastText + All Features	LinearSVC	0.78	0.78	0.78
fastText + Top13 Features	LinearSVC	0.79	0.79	0.79
GloVe	LogisticRegression	0.78	0.78	0.78
GloVe + All Features	LogisticRegression	0.79	0.79	0.79
GloVe + Top13 Features	LogisticRegression	0.79	0.78	0.79
BERT _{BASE}	Multi-Layer Perceptron	0.855	0.855	0.855
BERT _{BASE} + All Features	Multi-Layer Perceptron	0.871	0.871	0.871
BERT_{BASE} + Top13 Features	Multi-Layer Perceptron	0.887	0.884	0.884

Table 7: Experiment Results

to extend our analysis to the whole news content and assign different weights to the features on the basis of their relevance and other linguistic and stylistic considerations related to this specific domain. Finally, we will investigate the integration of social media-related aspects, such as news network propagation, reach and engagement.

Acknowledgements

Luca Giordano has been supported by Borsa di Studio GARR "Orio Carlini" 2022/23 - Consortium GARR, the National Research and Education Network.

Maria Pia di Buono has been supported by Fondo FSE/REACT-EU - Progetti DM 1062 del 10/08/2021 "Ricercatori a Tempo Determinato di tipo A) (RTDA)". Azione IV.4 - Dottorati e contratti di ricerca su tematiche dell'innovazione/Azione IV.6 - Contratti di ricerca su tematiche Green.

The authors would like to thank Raffaele Manna for his support.

References

- Oluwaseun Ajao, Deepayan Bhowmik, and Shahrzad Zargari. 2019. Sentiment aware fake news detection on online social networks. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2507–2511. IEEE.
- Miguel A Alonso, David Vilares, Carlos Gómez-Rodríguez, and Jesús Vilares. 2021. Sentiment analysis for fake news detection. *Electronics*, 10(11):1348.
- K Anoop, P Deepak, and VL Lajish. 2020. Emotion cognizance improves health fake news identification. In *IDEAS*, volume 2020, page 24th.
- CH Basch, Patricia Zybert, Rachel Reeves, and CE Basch. 2017. What do popular youtubetm videos say about vaccines? *Child: care, health and development*, 43(4):499–503.
- Bhavika Bhutani, Neha Rastogi, Priyanshu Sehgal, and Archana Purwar. 2019. Fake news detection using sentiment analysis. In *2019 twelfth international conference on contemporary computing (IC3)*, pages 1–5. IEEE.
- Prakhar Biyani, Kostas Tsioutsoulis, and John Blackmer. 2016. "8 amazing secrets for getting more clicks": Detecting clickbaits in news streams using article informality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.
- M José Blanca Mena, Rafael Alarcón Postigo, Jaume Arnau Gras, Roser Bono Cabré, and Rebecca Bendayan. 2017. Non-normal data: Is anova still a valid option? *Psicothema*, 2017, vol. 29, num. 4, p. 552-557.
- Alba Bonet-Jover. 2022. Veracity vs. reliability: Changing the approach of our annotation guideline.
- Brandon Butcher and Brian J Smith. 2020. Feature engineering and selection: A practical approach for predictive models: by max kuhn and kjell johnson. boca raton, fl: Chapman & hall/crc press, 2019, xv+297 pp., \$79.95 (h), isbn: 978-1-13-807922-9.
- Liang Chen, Xiaohui Wang, and Tai-Quan Peng. 2018. Nature and diffusion of gynecologic cancer-related misinformation on social media: analysis of tweets. *Journal of Medical Internet Research*, 20(10):e11515.
- Anshika Choudhary and Anuja Arora. 2021. Linguistic feature based learning model for fake news detection and classification. *Expert Systems with Applications*, 169:114171.
- Enyan Dai, Yiwei Sun, and Suhang Wang. 2020. Ginger cannot cure cancer: Battling fake health news with a comprehensive data repository. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 853–862.

- Arianna D'Ulizia, Maria Chiara Caschera, Fernando Ferri, and Patrizia Grifoni. 2021. Fake news detection: a survey of evaluation datasets. *PeerJ Computer Science*, 7:e518.
- Lila J Finney Rutten, Kelly D Blake, Alexandra J Greenberg-Worisek, Summer V Allen, Richard P Moser, and Bradford W Hesse. 2019. Online health information seeking among us adults: measuring progress toward a healthy people 2020 objective. *Public Health Reports*, 134(6):617–625.
- Asghar Ghasemi and Saleh Zahediasl. 2012. Normality tests for statistical analysis: a guide for non-statisticians. *International journal of endocrinology and metabolism*, 10(2):486.
- Katarína Greškovičová, Radomír Masaryk, Nikola Synak, and Vladimíra Čavojská. 2022. Superlatives, clickbaits, appeals to authority, poor grammar, or boldface: Is editorial style related to the credibility of online health messages? *Frontiers in Psychology*, page 5056.
- Benjamin Horne and Sibel Adali. 2017. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 759–766.
- Heejung Jwa, Dongsuk Oh, Kinam Park, Jang Mook Kang, and Heuseok Lim. 2019. exbake: Automatic fake news detection model based on bidirectional encoder representations from transformers (bert). *Applied Sciences*, 9(19):4062.
- Dimitrios Panagiotis Kasseropoulos and Christos Tjortjis. 2021. An approach utilizing linguistic features for fake news detection. In *Artificial Intelligence Applications and Innovations: 17th IFIP WG 12.5 International Conference, AIAI 2021, Hersonissos, Greece, June 25–27, 2021, Proceedings 17*, pages 646–658. Springer.
- David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science*, 359(6380):1094–1096.
- Thomas Lumley, Paula Diehr, Scott Emerson, and Lu Chen. 2002. The importance of the normality assumption in large public health data sets. *Annual review of public health*, 23(1):151–169.
- Cristiane Melchior and Mirian Oliveira. 2022. Health-related fake news on social media platforms: A systematic literature review. *new media & society*, 24(6):1500–1522.
- Saif M Mohammad and Peter D Turney. 2013. Nrc emotion lexicon. *National Research Council, Canada*, 2:234.
- Giancarlo Nicola. 2018. Bidirectional attentional lstm for aspect based sentiment analysis on italian. *EVALITA Evaluation of NLP and Speech Tools for Italian*, 12(108).
- Francesco Pierri, Alessandro Artoni, and Stefano Ceri. 2020. Hoaxitaly: a collection of italian disinformation and fact-checking stories shared on twitter in 2019. *arXiv preprint arXiv:2001.10926*.
- Juan-Pablo Posadas-Durán, Helena Gómez-Adorno, Grigori Sidorov, and Jesús Jaime Moreno Escobar. 2019. Detection of fake news in a new corpus for the spanish language. *Journal of Intelligent & Fuzzy Systems*, 36(5):4869–4876.
- Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2931–2937.
- Giovanni Santia and Jake Williams. 2018. Buzzface: A news veracity dataset with facebook user commentary and egos. In *Proceedings of the international AAAI conference on web and social media*, volume 12, pages 531–540.
- Steven F Sawyer. 2009. Analysis of variance: the fundamental concepts. *Journal of Manual & Manipulative Therapy*, 17(2):27E–38E.
- Emanuel Schmider, Matthias Ziegler, Erik Danay, Luzi Beyer, and Markus Bühner. 2010. Is it really robust? *Methodology*.
- Anu Shrestha and Francesca Spezzano. 2021. Textual characteristics of news title and body to detect fake news: a reproducibility study. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43*, pages 120–133. Springer.
- Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188.
- Eugenio Tacchini, Gabriele Ballarin, Marco L Della Vedova, Stefano Moret, and Luca De Alfaro. 2017. Some like it hoax: Automated fake news detection in social networks. *arXiv preprint arXiv:1704.07506*.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.
- Alice Tontodimamma, Lara Fontanella, Stefano Anzani, and Valerio Basile. 2022. An italian lexical resource for incivility detection in online discourses. *Quality & Quantity*, pages 1–19.

Marco Viviani and Gabriella Pasi. 2017. Credibility in social media: opinions, news, and health information—a survey. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, 7(5):e1209.

Cross-Lingual Transfer Learning for Misinformation Detection: Investigating Performance Across Multiple Languages

Oguzhan Ozcelik^{1,2,†}, Arda Sarp Yenicesu^{2,†}, Onur Yildirim^{2,†},
Dilruba Sultan Haliloglu^{2,†}, Erdem Ege Eroglu^{2,†}, and Fazli Can²

¹Aselsan Research Center, Ankara, Turkey

²Computer Engineering Department, Bilkent University, Ankara, Turkey

{oguzhan.ozcelik, sarp.yenicesu, o.yildirim, sultan.haliloglu}@bilkent.edu.tr,
ege.eroglu@ug.bilkent.edu.tr,
canf@cs.bilkent.edu.tr

Abstract

Detection of misinformation on social media requires human-annotated datasets to achieve truthful results. However, the annotation process is time-consuming due to the difficulty of labeling the veracity of the claims. Furthermore, most of the annotated misinformation detection datasets in the social media domain predominantly reside in English. To overcome this problem, we investigate the performance of cross-lingual transfer learning for misinformation detection across various languages, including, Arabic, Chinese, Turkish, and Polish. For this purpose, we analyze three different experimental setups on multilingual pre-trained language models in five natural languages (English, Arabic, Chinese, Turkish, and Polish). The results show that the multi-lingual mDeBERTa model can be applicable with fine-tuning in a widely-used language, i.e., English, and tested on a low-resource Turkish language with a successful recovery ratio, i.e., the metric shows the percentage of the recovered baseline score. For each model, we observe higher and more robust transfer ability between Polish and Arabic. Furthermore, it is possible to claim that contextual similarities outweigh language similarities, due to unsuccessful transfer learning ability between the English-Polish language pair.

1 Introduction

With the extensive use of social media, assessing the credibility of news has become a demanding task as the community is exposed to a substantial amount of information. Moreover, with the success of transformer-based auto-regressive models, it becomes challenging for a human reader to determine the reliability of the source of news (Hsu and Thompson, 2023). To overcome this issue, large language models (LLMs) become more popular to determine the veracity of a given news article

(Kaliyar et al., 2021). However, it is challenging to develop a robust task-dependent LLM in low-resource languages due to the limitations of the training corpus. In this work, we will conduct detailed experiments to observe the cross-lingual transfer learning in the misinformation detection domain across various languages. Our study provides insight into which natural languages can be adapted to others, where the target domain limits the availability of an organized dataset.

Constructing a misinformation detection dataset is a challenging task as it requires human experts in the corresponding domain to annotate the disputed news (Shu et al., 2017a). Therefore, our experimental procedure employs multilingual pre-trained models to explore the transfer abilities of natural languages. The motivation of this study is to show how state-of-the-art approaches perform in low-resource languages when the source data is a widely-spoken language, i.e., English. Thus, we discuss the ways to choose a source language for a target language when the target language is limited in resources¹.

Misinformation detection can be performed on both noisy social media posts (Shu et al., 2017b) and well-written news articles (Wang, 2017). A common approach is training a classifier for a human-annotated dataset and predicting the veracity classes on a test collection. However, if a natural language has limited sources, the implementation and up-to-dateness of the proposed methods turn out to be an issue for that language.

1.1 Research Questions

To combat misinformation when there is a data limitation problem, we answer the following research questions:

¹During this study, we use the “low-resource language” term for the misinformation detection task. Although a language has limited resources in the misinformation detection task, it can be high-resourced for other natural language processing problems.

[†]These authors contributed equally to this work

Table 1: **The available annotated misinformation detection datasets in English, Arabic, Chinese, Turkish, and Polish languages.** The referenced datasets are composed of social media (Twitter or Weibo) texts. (*) Note that the table is not totally comprehensive. In other words, there may be some datasets that have been overlooked, especially in English.

Language	No.	Available Datasets*
English 🇺🇸	17	(Kochkina et al., 2018), (Ma et al., 2016), (Derczynski et al., 2017), (Ma et al., 2017), (Shu et al., 2020), (Gorrell et al., 2019), (Nguyen et al., 2019), (Nguyen and Yu, 2021), (Dai et al., 2020), (Cui and Lee, 2020), (Dharawat et al., 2022), (Li et al., 2020), (Patwa et al., 2021), (Alam et al., 2021), (Cheng et al., 2021), (Dadkhah et al., 2023), (Toraman et al., 2022a)
Arabic 🇸🇦	3	(Haouari et al., 2020), (Alam et al., 2021), (Hadj Ameer and Aliane, 2021)
Chinese 🇨🇳	1	(Yang et al., 2021)
Turkish 🇹🇷	1	(Toraman et al., 2022a)
Polish 🇵🇱	1	(Jarynowski, 2020)

RQ-1: Can we use widely-spoken high-resource language, such as English, as a source language in misinformation for low-resource target languages?

RQ-2: Which low-resource source language can be a better candidate for a high-resource target language in terms of transfer ability of misinformation detection task among the pairs of English, Chinese, Arabic, Turkish, and Polish?

1.2 Contributions

There are several studies conducted, including but not limited to cross-lingual data on fake news detection task (Arif et al., 2022; Du et al., 2021; Chu et al., 2021). However, there are a very few misinformation detection studies involving low-resource languages, such as Turkish (Toraman et al., 2022a) and Polish (Jarynowski, 2020). To the best of our knowledge, our study is the first to investigate the transfer ability across aforementioned languages in misinformation detection. Our contributions are the following:

- This is the first misinformation detection study that explores the transfer ability including Turkish and Polish languages.
- Our investigation aims to determine the most effective multilingual model for effectively transferring the task of misinformation detection across different languages.

The rest of the paper is organized as follows, in Section 2, we briefly introduce previous studies conducted in the area of misinformation detection and cross-lingual transfer learning. In Section 3, we

formulate our problem in detail. Our approach to investigating the transfer ability of misinformation detection in various languages is given in Section 4. Section 5 describes the datasets we use in our experiments. In Section 6, we describe the experimental setup and then provide the results we obtain in Section 7. We discuss the experimental results in Section 8. Next, we provide limitations and ethical considerations in Section 9. Finally, Section 10 concludes the paper.

2 Related Work

We review previous works in terms of datasets, misinformation detection, and cross-lingual transfer learning studies.

2.1 Datasets

Table 1 summarises the incomplete list of datasets that can be used for misinformation detection in various domains, e.g., politics (Kochkina et al., 2018), public health (Cui and Lee, 2020), and so on. All of these datasets consist of social media posts, which resemble an informal way of presenting information. From Table 1, we observe that English covers the majority of the studies in the misinformation/disinformation area; hence, we decided to acknowledge English as a high-resource language as opposed to others (Arabic (Haouari et al., 2020), Chinese (Yang et al., 2021), Turkish (Toraman et al., 2022a), and Polish (Jarynowski, 2020)). Note that we also accept Arabic as a high-resource language for this study since there is more than one misinformation detection dataset in the Arabic language.

2.2 Misinformation Detection

Misinformation detection has become an important task, due to the ease of reaching and sharing content with the popularity of social media. There are different approaches to solving this detection problem. For instance, Helmstetter and Paulheim (2018) propose an ensemble method to predict fake news in a weakly supervised manner. Their ensemble model includes both traditional machine learning approaches like SVM (Vapnik, 1999), and Naive Bayes. De et al. (2021) utilize a transformer-based model, using BERT (Devlin et al., 2018) as the backbone, for multilingual fake news detection. Their dataset consists of news articles collected from various news websites with translated versions to low-resource languages such as Vietnamese. Monti et al. (2019) use a geometric deep-learning method to identify fake news in a dataset collected from Twitter, a widely-used social media platform. Graph neural networks are employed to distinguish fake news (Meyers et al., 2020). Social contexts are also used as a supportive feature for news content in a transformer-based architecture (Raza and Ding, 2022).

2.3 Cross-lingual Transfer Learning

Limited resources in some languages for a specific task, such as misinformation detection, require the emergence of cross-lingual studies. Probabilistic methods for cross-lingual information retrieval are investigated (Nie et al., 1999; Xu et al., 2001). A recurrent neural network-based approach is utilized to investigate multilingual analysis for limited data (Can et al., 2018). Moreover, Sun et al. (2021) employ a multilingual response generation layer and a cross-lingual knowledge retrieval layer to handle the language barrier in the context of the conversation. Besides, studies based on transfer learning in terms of few-shot learning are carried out to overcome the limited data problem (Hardalov et al., 2022).

Some studies utilize additional extracted features from external multi-lingual sources. Wen et al. (2018) utilizes an approach for rumor verification, employing multimedia content and external information in other news platforms. They achieve good performance with the use of extracted features. Dementieva and Panchenko (2021) propose a feature called "cross-lingual evidence" to be utilized in fake news identification. This feature is based on the idea "if a news is true, the facts mentioned in

different languages should be identical". They report that the state-of-art models that use this feature perform better than their default versions. (Hamouchi and Ghogho, 2022) propose a framework for fake news detection employing external pieces of evidence searched by the web to verify the veracity of the news in multilingual datasets.

3 Problem Formulation

Suppose we have a misinformation dataset in target language $F_T = \{(N_i^T, L_i^T)\}_i^{|F_T|}$ with $|F_T|$ microblog-veracity pairs, where for all i , N_i^T refers to a tweet with veracity label L_i^T . The veracity, L_i^T , represents whether a microblog includes true information or false information as a binary variable (Eq. 1).

$$L_i^T = \begin{cases} 1 & \text{if } N_i^T \text{ includes true claim} \\ 0 & \text{if } N_i^T \text{ includes false claim} \end{cases} \quad (1)$$

We also have a collection of social media datasets, C_S , in other source languages:

$$C_S : [F_1 = \{(N_i^1, L_i^1)\}_i^{|F_1|}, \dots, F_z = \{(N_i^z, L_i^z)\}_i^{|F_z|}]_\gamma^{|C_S|} \quad (2)$$

In Eq. 2, $|C_S|$ refers to the number of available misinformation detection datasets in other source languages we accessed, and each F refers to a dataset in other source languages. Each dataset, similar to the F_T consists of microblog-veracity pairs. γ is used for indexing the datasets in the C collection.

We will have a multilingual model set, $H = \{\{h(N)\}_m^{|C_S|+1}\}_k^K$ which has $K \times (|C_S| + 1)$ pre-trained models. Each $h(N)$ represents a multilingual language model focusing on one of the source languages or the target language while using a pre-trained multilingual model, e.g., mBERT (Devlin et al., 2018). For the target language and other languages, there are $|C_S| + 1$ models (There are $|C_S|$ source languages and 1 target language.), and for each of them there are K different multilingual model architecture, i.e. $K = 3$ for mBERT (Devlin et al., 2018), XLM-R (Conneau et al., 2019), and mDeBERTa (He et al., 2020). For the F_T and $C_{S\gamma}$ for all γ , $H = \{h(N)\}_m^{|C_S|+1}$ will be fine-tuned using aforementioned pre-trained multilingual models in source languages which is the language used in F_M during the fine-tuning of the h_m .

Given F_T , C_S , and H , we want to find cross-lingual transfer ability on misinformation detection in the target language. To find this transfer ability,

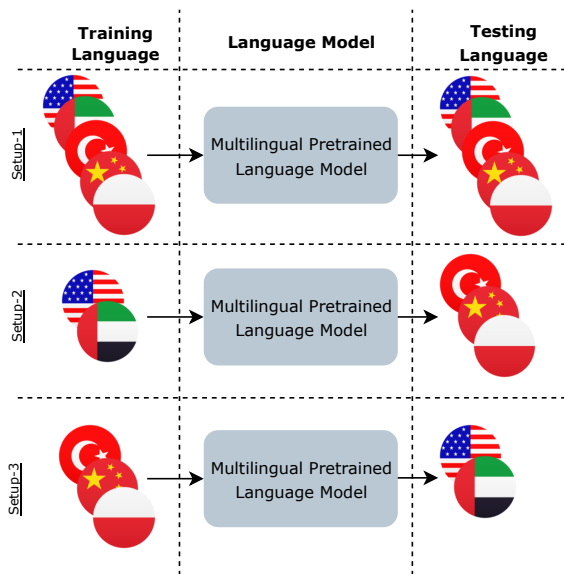


Figure 1: The illustration of our experimental methodology. (Setup-1) shows when a model is trained and tested in the same language for a specific task. (Setup-2) indicates when the language is crossed, i.e., training on a widely-used high-resource language (i.e., English or Arabic) and tested on low-resource languages. (Setup-3) simply represents when the model is trained and tested on low-resource, and high-resource languages, respectively.

first, we will evaluate h_t on F_T , the target dataset, e.g., CHECKED (Yang et al., 2021) if the target language is Chinese and achieve an F1 score, $F1_{Target}$. Then, we will repeat the same evaluation for all h_γ , where $h_\gamma \neq h_t$ on F_T and achieve a separate F1 score $F1_\gamma$, where $h_{t\gamma}$ is fine-tuned using C_{S_γ} . To evaluate the transfer ability of a language model, we employ relative zero-shot transfer ability (Turc et al., 2021) and call it “recovery ratio” following the study (Toraman et al., 2022b). We use the recovery ratio between the target language and the remaining languages from the C_S collection. The recovery ratio is formulated as in Eq. 3.

$$\text{Recovery Ratio}_\gamma = \frac{F1_\gamma}{F1_{Target}} \quad (3)$$

Finally, we will use these Recovery Ratio $_\gamma$ scores to compare and analyze the transfer learning ability of each source language in C_S to a target language.

4 Method

We investigate the transfer learning ability across five different languages, namely English, Chinese, Arabic, Turkish, and Polish. Particularly, we con-

duct analysis on a single NLP task, namely, misinformation detection. In order to find which language is a better choice when language transfer is required, we fine-tune pre-trained multilingual mBERT (Devlin et al., 2018), XLM-R (Conneau et al., 2019) and mDeBERTa (He et al., 2020) models in source languages, and predict the truthfulness of tweets (True or False) in a target language. The performance of the language transfer ability is evaluated on models via recovery ratio over baselines, where the baselines are the models fine-tuned and tested on the same language. In other words, we assume that the best performance occurs when the source and target language are the same. Thus, we use the baseline score as the denominator in Eq. 3.

We provide an illustration (see Figure 1), to explain our methodology for the experimental procedure. When a multilingual model is fine-tuned and tested in the same language, it yields promising results. However, for low-resource languages, such as Turkish, there are a few available data collections for specific problems, e.g., misinformation detection. This motivates cross-language studies to explore which widely spoken language can fit into a language if there is a lack of data collection in that language.

This methodology provides us an opportunity to empirically find the ability to transfer information from a high-resource source language to a low-resource target language while giving some valuable insights about hidden transfer mechanisms such as geopolitical influence on a language, shared vocabulary between languages, the impact of an alphabet on a language, and contextual similarities regardless of language differences.

5 Dataset

In this study, we use the English and Turkish microblogs from the splits of the MiDe-22 dataset (Toraman et al., 2022a), Chinese from the CHECKED dataset (Yang et al., 2021), Arabic from the AraCOVID19-MFH dataset (Hadj Ameur and Aliane, 2021) and Polish from Andrzej’s dataset (Jarynowski, 2020). MiDe-22 is a tweet collection of misinformation domains, including various topics such as the Russo-Ukraine War, COVID-19, refugees, and so on, while CHECKED, AraCOVID19-MFH and Andrzej’s only contain microblogs about COVID-19. For all datasets, we only use the true and false labeled social media posts in our experiments. The main statistics of the

Table 2: The main statistics of the datasets used in this study. The values are microblog counts for True labeled and False labeled microblogs.

Languages	English (🇺🇸)		Chinese (🇨🇳)		Arabic (🇸🇦)		Turkish (🇹🇷)		Polish (🇵🇱)	
Datasets	MiDe-22		CHECKED		AraCOVID19-MFH		MiDe-22		Andrzej’s	
Splits	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
True	576	151	1,408	352	320	80	533	136	377	95
False	1,381	348	276	68	1,609	402	1,379	353	84	21
Total	1,957	499	1,684	420	1,929	482	1,912	489	461	116

datasets used in this study are given in Table 2.

6 Experimental Approach

In this study, we first define three experimental procedures. Then we utilize different multilingual pre-trained language models. We provide the details in the following sections.

6.1 Experimental Procedure

The experimental procedure consists of three types of setup (see Figure 1):

Setup-1 : When a model is trained and tested in the same language. (e.g., English → English)

Setup-2 : When a model is trained on a widely-used source language and tested on a low-resource language. (e.g., English → Turkish)

Setup-3 : When a model is trained on a low-resource source language and tested on high-resource languages. (e.g., Polish → Arabic)

In order to obtain a reference point for the recovery ratio metric, we construct “Setup-1”. We assume that if a language model is trained and tested on the same language, its score is the maximum reference point to be achieved. Then, we implement “Setup-2” to answer **RQ-1**. Next, we use “Setup-3” for **RQ-2**. In order to investigate the better source language candidate, and transfer ability across languages, we evaluate recovery ratio metrics by employing the results of “Setup-1”.

6.2 Language Models

We utilize three different multilingual pre-trained language models. The motivation behind choosing multilingual models is to have language knowledge of our studied languages in the pretraining corpus of these models. Thus, we can observe whether a specific task (in this study, the task is misinformation detection) can be learned via these models. The models are the following:

mBERT: BERT (Devlin et al., 2018) (Bidirectional Encoder Representations from Transformers) architecture serves as the foundation for the multilingual model known as mBERT. BERT was trained using Wikipedia and the Book Corpus dataset, which includes more than 10,000 books of various genres. To learn embedded representations of texts in many languages, this model is trained in a broad range of languages. mBERT can be used to process texts in several languages and for tasks like classification and translation because it supports multiple languages.

XLm-R: Cross-lingual Language Model - RoBERTa is what the acronym XLM-R (Conneau et al., 2019) stands for. A sizable pre-training dataset that included numerous huge, multilingual texts were used to train this model. Indeed, 100 languages from 2.5TB of filtered CommonCrawl data were used as its pre-training material. In order to learn embedded representations of multilingual texts, XLM-R employs an unsupervised learning technique. This makes it possible to identify semantic connections and commonalities across several languages.

mDeBERTa: Multilingual Decoding-enhanced BERT with Disentangled Attention is referred to as mDeBERTa (He et al., 2020). This model improves the BERT and RoBERTa (Zhuang et al., 2021) models using disentangled attention and enhanced mask decoder.

6.3 Experimental Setup

During the experiments, we use Hugging Face (Wolf et al., 2020) library to fine-tune Transformer-based language models. We choose learning rate $5e-5$, batch size 16, the number of epochs 10, and maximum sequence length 128, following the study (Toraman et al., 2022a). During the training of the models, we employ an NVIDIA RTX A400. We use stratified five-fold cross-validation where the

Table 3: **Experimental results of Setup-1.** Column notations for metrics: precision (P), recall (R), and weighted F1 score (F1). Five-fold average precision, recall, and weighted F1 scores are reported.

Models Datasets/Metrics	mBERT			XLM-R			mDeBERTa		
	P	R	F1	P	R	F1	P	R	F1
MiDe-22-EN 🇺🇸	0.879	0.880	0.879	0.724	0.806	0.758	0.884	0.881	0.882
AraCOVID19-MFH 🇮🇸	0.998	0.998	0.998	0.997	0.997	0.997	0.997	0.997	0.997
CHECKED 🇨🇳	0.991	0.991	0.991	0.996	0.996	0.996	0.996	0.996	0.996
MiDe-22-TR 🇹🇷	0.894	0.895	0.894	0.885	0.886	0.885	0.902	0.901	0.901
Andrzej’s 🇵🇱	0.771	0.790	0.771	0.809	0.834	0.794	0.770	0.820	0.787

statistics of train and test splits are given in Table 2.

7 Experimental Results

We report the results obtained for Setup-1 in Table 3. Out of three multilingual language models, mDeBERTa produces higher F1 scores in English, Chinese, and Turkish datasets. On the other hand, mBERT performs better in Arabic, and XLM-R does it in the Polish language. The results are very high for Chinese and Arabic, with around 99% F1 scores. This is possible because these datasets are specifically on one topic, i.e., COVID-19. However, the Polish dataset is also in the COVID-19 domain but the models perform lower in Polish when compared to Chinese and Arabic. We may claim that Chinese and Arabic datasets are easier to detect misinformation possibly having biased patterns in texts.

From Table 4, we observe gray-highlighted cells, which are the average of weighted F1 scores on five-fold splits when source and target language are the same, i.e., Setup-1. For **RQ-2**, it can be seen that the mBERT model produces the highest score when it is trained in Arabic (a well-resourced language) and tested in Polish (a low-resource language) with a 95% recovery ratio. Similarly, the mDeBERTa achieves the highest score for the Arabic-Polish pair. For RQ-3, the XLM-R model produces the highest recovery ratio, 95%, with the Turkish-English pair. The rest of the experimental results are given in Section 4.

8 Discussion

In our studies, we use five languages from four different language families: Altaic (Turkish), European (English and Polish), Zhou (Chinese), and Sámi (Arabic). This separation gives us a fair ground for our experiments. In Table 4, we observe that the transfer ability from English to Turk-

ish is higher than in any other source language. On average, we achieve an 84% recovery ratio for this transformation which suggest that English can be used as a source language for the Turkish language in a task-oriented setting, (**RQ1**). However, the transformation from English (as a high-resource language) to other low-resource languages except Turkish is not successful, and we arguably claim that this difference is due to contextual differences between the datasets used for the study where the datasets used for English and Turkish languages combined similar topics from the Russo-Ukraine War, COVID-19, refugees, etc., while others only focus on COVID-19, (**RQ1**). Moreover, even though Polish and English are in the same language family, the transfer performance between these two languages is low compared to some other pairs that contain Polish and English as either the target or the source. The reason behind these relatively lower scores between Polish and English can be due to the context of the data which supports our previous claim.

On the other hand, relatively lower results can be observed in the transfer ability of Arabic and Turkish, even though there are a lot of borrowed words. Another observation is the good transfer ability of Arabic to Chinese and vice versa. Since the Arabic and Chinese datasets both contain social media posts only about COVID-19, the performances of all models are better when these two languages are used as the source and the target languages. This also clearly shows that the domain of the data is essential and has an impact on the performance. This claim can be supported by the transfer ability performance from the Turkish language to the English language, where this transformation achieved an 88.3% recovery ratio on average of three models by utilizing similar misinformation domains. To conclude, if the domain of the data is similar, any low-resource language can be used as a source

Table 4: **The results of cross-lingual fake news experiments (Setup-2 and Setup-3).** Gray-highlighted cells are the weighted average of F1 scores in the same source and target languages retrieved from Table 3. The other cells represent the column-based recovery scores corresponding to the given source language. The best recovery ratios are given in bold for each target language. The recovery scores are computed specifically for the models, i.e., the denominator is the gray-highlighted cell in the column of a model. For instance, the F1 score is 0.879 when the source and target are English (see Table 3); also, when the source is Chinese and the target is English the F1 score is 0.519. Thus, the recovery ratio (Eq. 3) of Chinese→English is $\frac{0.519}{0.879} = 59\%$. The results are used to answer **RQ-1** and **RQ-2**.

Model	Source/Target	English	Chinese	Arabic	Turkish	Polish
mBERT	English	0.879	13%	17%	82%	38%
	Chinese	59%	0.991	20%	59%	48%
	Arabic	25%	75%	0.998	42%	95%
	Turkish	80%	68%	24%	0.894	40%
	Polish	40%	80%	77%	43%	0.771
XLM-R	English	0.758	16%	9%	80%	23%
	Chinese	77%	0.996	7%	68%	15%
	Arabic	25%	79%	0.997	30%	92%
	Turkish	95%	46%	10%	0.885	36%
	Polish	49%	78%	77%	43%	0.794
mDeBERTa	English	0.882	55%	39%	90%	49%
	Chinese	67%	0.996	12%	68%	20%
	Arabic	31%	82%	0.997	41%	93%
	Turkish	90%	70%	38%	0.901	59%
	Polish	38%	81%	86%	39%	0.787

language for a high-resource target language, e.g., English and Arabic in our study. For example, Polish can be used as a source language for Arabic, and Turkish can be used as a source language for English, (**RQ2**).

We conclude that multilingual Transformer-based models, e.g., mDeBERTa, performs well even if the source language is different from the target language. These promising results show that a multilingual model can be used for a low-resource language, although the target language is not available in terms of training resources.

9 Limitations and Ethical Consideration

In this section, we discuss the limitations and challenges encountered in our study, including the scarcity of non-English misinformation detection datasets, the binary labeling approach, and the difficulties associated with using microblog text from social media platforms.

9.1 Datasets

Due to the limited availability of non-English misinformation detection social media datasets, we had to combine multiple datasets focusing on different topics and collected at different time peri-

ods. This diversity in the datasets could potentially introduce bias into our research. Ideally, a multilingual dataset collected during the same time period and on the same topic would be preferable for observing the transfer ability between languages. However, due to the limitations of misinformation detection datasets in low-resource languages, we were unable to create such a setup.

9.2 Labels

The datasets we utilized have binary labels in terms of veracity. While this approach provides a simple and straightforward way to label data, it may oversimplify the complexity of misinformation and disinformation. Binary labels do not account for different levels of reliability and accuracy. Furthermore, they may fail to capture cultural and sociopolitical variations, thereby limiting the model’s ability to generalize well to different contexts.

9.3 Usage of Microblog Text

Texts obtained from social media platforms can be noisy and contain a mixture of multiple languages within a single text. Additionally, the quality of these texts can be low. These factors can pose challenges to language transfer ability and can decrease

the accuracy of misinformation/disinformation detection. Moreover, inherent biases present in social media platforms can also influence the model and introduce bias into its predictions.

9.4 Ethical Consideration and Possible Use Cases

This paper acknowledges and addresses several ethical considerations inherent in the research and development of fake news detection. Privacy and data protection are of utmost importance, and user data and personal information are treated with strict confidentiality throughout the research process. Moreover, we acknowledge broader societal impacts of misinformation detection such as the potential for censorship, and the effects of trust on social media.

We also anticipate that the experimental setup investigated throughout the paper can be used for other NLP problems. The transfer learning ability across multiple languages in other problems, e.g., rumor or stance detection and emotion recognition (Küçük and Can, 2020), need to be studied for further possibilities.

10 Conclusion

In order to observe cross-lingual few-shot transfer skills between languages, we carried out a number of experiments. For this purpose, multiple languages were used in a cross-lingual transfer learning structure employing multilingual pre-trained models. In this way, we provide a comparative examination of the performance of state-of-the-art methods for the misinformation detection task. We believe that this study will help future NLP researchers who plan to use the low-source language datasets in their cross-lingual study by giving them insight.

We observe that English can be used as a source language for the Turkish language depending on the dataset domain. Our most important observation is the context of the data is essential and we observe relatively better results for the transfer abilities between languages whose datasets are in the same domain. In future work, we will include other languages, such as Czech and Finnish, to observe the effects of agglutinative patterns of those languages between Turkish. We also plan to improve our study into several social media platforms, such as Facebook posts and Instagram content to investigate the effect of the social media domain on the datasets.

References

- Firoj Alam, Fahim Dalvi, Shaden Shaar, Nadir Durani, Hamdy Mubarak, Alex Nikolov, Giovanni Da San Martino, Ahmed Abdelali, Hassan Sajjad, Kareem Darwish, et al. 2021. Fighting the covid-19 infodemic in social media: a holistic perspective and a call to arms. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 913–922.
- Muhammad Arif, Atnafu Lambebo Tonja, Iqra Ameer, Olga Kolesnikova, Alexander Gelbukh, Grigori Sidorov, and Abdul Gafar Manuel Meque. 2022. Cic at checkthat! 2022: multi-class and cross-lingual fake news detection. *Working Notes of CLEF*.
- Ethem F. Can, Aysu Ezen-Can, and Fazli Can. 2018. [Multilingual sentiment analysis: An rnn-based framework for limited data](#).
- Mingxi Cheng, Songli Wang, Xiaofeng Yan, Tianqi Yang, Wenshuo Wang, Zehao Huang, Xiongye Xiao, Shahin Nazarian, and Paul Bogdan. 2021. [A covid-19 rumor dataset](#). *Frontiers in Psychology*, 12.
- Samuel Kai Wah Chu, Runbin Xie, and Yanshu Wang. 2021. [Cross-language fake news detection](#). *Data and Information Management*, 5(1):100–109.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.
- Limeng Cui and Dongwon Lee. 2020. Coaid: Covid-19 healthcare misinformation dataset. *arXiv preprint arXiv:2006.00885*.
- Sajjad Dadkhah, Xichen Zhang, Alexander Gerald Weismann, Amir Firouzi, and Ali A. Ghorbani. 2023. [TruthSeeker: The Largest Social Media Ground-Truth Dataset for Real/Fake Content](#).
- Enyan Dai, Yiwei Sun, and Suhang Wang. 2020. [Ginger cannot cure cancer: Battling fake health news with a comprehensive data repository](#). *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):853–862.
- Arkadipta De, Dibyanayan Bandyopadhyay, Baban Gain, and Asif Ekbal. 2021. [A transformer-based approach to multilingual fake news detection in low-resource languages](#). *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 21(1).
- Daryna Dementieva and Alexander Panchenko. 2021. Cross-lingual evidence improves monolingual fake news detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 310–320.

- Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. [SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours](#). In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76, Vancouver, Canada. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Arkin Dharawat, Ismini Lourentzou, Alex Morales, and ChengXiang Zhai. 2022. [Drink bleach or do what now? covid-hera: A study of risk-informed health decision making in the presence of covid-19 misinformation](#). *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1):1218–1227.
- Jiangshu Du, Yingtong Dou, Congying Xia, Limeng Cui, Jing Ma, and S Yu Philip. 2021. Cross-lingual covid-19 fake news detection. In *2021 International Conference on Data Mining Workshops (ICDMW)*, pages 859–862. IEEE.
- Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. Semeval-2019 task 7: Rumoureval 2019: Determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854, United States. Association for Computational Linguistics.
- Mohamed Seghir Hadj Ameer and Hassina Aliane. 2021. [Aracovid19-mfh: Arabic covid-19 multi-label fake news & hate speech detection dataset](#). *Procedia Computer Science*, 189:232–241.
- Hicham Hammouchi and Mounir Ghogho. 2022. [Evidence-aware multilingual fake news detection](#). *IEEE Access*, 10:116808–116818.
- Fatima Haouari, Maram Hasanain, Reem Suwaileh, and Tamer Elsayed. 2020. Arcov19-rumors: Arabic covid-19 twitter dataset for misinformation detection. *arXiv preprint arXiv:2010.08768*.
- Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022. [Few-shot cross-lingual stance detection with sentiment-based pre-training](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10729–10737.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.
- Stefan Helmstetter and Heiko Paulheim. 2018. [Weakly supervised learning for fake news detection on twitter](#). In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 274–277.
- T. Hsu and S. A. Thompson. 2023. [Disinformation researchers raise alarms about a.i. chatbots](#). (Accessed: 02-Apr-2023).
- Andrzej Jarynowski. 2020. [A dataset of media releases \(Twitter, News and Comments, Youtube, Facebook\) from Poland related to COVID-19 for open research](#).
- Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. 2021. Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia tools and applications*, 80(8):11765–11788.
- Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2018. All-in-one: Multi-task learning for rumour verification. *arXiv preprint arXiv:1806.03713*.
- Dilek Küçük and Fazli Can. 2020. [Stance detection: A survey](#). *ACM Comput. Surv.*, 53(1).
- Yichuan Li, Bohan Jiang, Kai Shu, and Huan Liu. 2020. Mm-covid: A multilingual and multimodal data repository for combating covid-19 disinformation. *arXiv preprint arXiv:2011.04088*.
- Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI'16*, page 3818–3824. AAAI Press.
- Jing Ma, Wei Gao, and Kam-Fai Wong. 2017. [Detect rumors in microblog posts using propagation structure via kernel learning](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 708–717, Vancouver, Canada. Association for Computational Linguistics.
- Marion Meyers, Gerhard Weiss, and Gerasimos Spanakis. 2020. Fake news detection on twitter using propagation structures. In *Disinformation in Open Online Media: Second Multidisciplinary International Symposium, MISDOOM 2020, Leiden, The Netherlands, October 26–27, 2020, Proceedings 2*, pages 138–158. Springer.
- Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M Bronstein. 2019. Fake news detection on social media using geometric deep learning. *arXiv preprint arXiv:1902.06673*.
- Minh Nguyen and Zhou Yu. 2021. [Improving named entity recognition in spoken dialog systems by context and speech pattern modeling](#). In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 45–55, Singapore and Online. Association for Computational Linguistics.

- Tam Nguyen, Matthias Weidlich, Bolong Zheng, Hongzhi Yin, Nguyen Hung, and Bela Stantic. 2019. [From anomaly detection to rumour detection using data streams of social platforms](#). *Proceedings of the VLDB Endowment*, 12:1016–1029.
- Jian-Yun Nie, Michel Simard, Pierre Isabelle, and Richard Durand. 1999. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 74–81.
- Parth Patwa, Shivam Sharma, Srinivas Pykl, Vineeth Guptha, Gitanjali Kumari, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2021. Fighting an infodemic: COVID-19 fake news dataset. In *Combating Online Hostile Posts in Regional Languages during Emergency Situation*, pages 21–29, Cham. Springer International Publishing.
- Shaina Raza and Chen Ding. 2022. Fake news detection based on news content and social contexts: a transformer-based approach. *International Journal of Data Science and Analytics*, 13(4):335–362.
- Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. [Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media](#). *Big Data*, 8(3):171–188.
- Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017a. [Fake news detection on social media: A data mining perspective](#). *SIGKDD Explor. Newsl.*, 19(1):22–36.
- Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017b. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.
- Weiwei Sun, Chuan Meng, Qi Meng, Zhaochun Ren, Pengjie Ren, Zhumin Chen, and Maarten de Rijke. 2021. [Conversations powered by cross-lingual knowledge](#). In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 1442–1451, New York, NY, USA. Association for Computing Machinery.
- Cagri Toraman, Oguzhan Ozelik, Furkan Şahinuç, and Fazli Can. 2022a. Not good times for lies: Misinformation detection on the russia-ukraine war, covid-19, and refugees. *arXiv preprint arXiv:2210.05401*.
- Cagri Toraman, Furkan Şahinuç, and Eyup Yilmaz. 2022b. Large-scale hate speech detection with cross-domain transfer. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2215–2225.
- Iulia Turc, Kenton Lee, Jacob Eisenstein, Ming-Wei Chang, and Kristina Toutanova. 2021. [Revisiting the primacy of english in zero-shot cross-lingual transfer](#). *CoRR*, abs/2106.16171.
- Vladimir Vapnik. 1999. *The nature of statistical learning theory*. Springer science & business media.
- William Yang Wang. 2017. [“liar, liar pants on fire”: A new benchmark dataset for fake news detection](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.
- Weiming Wen, Songwen Su, and Zhou Yu. 2018. Cross-lingual cross-platform rumor verification pivoting on multimedia content. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3487–3496.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Jinxi Xu, Ralph Weischedel, and Chanh Nguyen. 2001. [Evaluating a probabilistic model for cross-lingual information retrieval](#). In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, page 105–110, New York, NY, USA. Association for Computing Machinery.
- Chen Yang, Xinyi Zhou, and Reza Zafarani. 2021. [Checked: Chinese covid-19 fake news dataset](#). *Social Network Analysis and Mining (SNAM)*.
- Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. [A robustly optimized BERT pre-training approach with post-training](#). In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

A First Attempt to Detect Misinformation in Russia-Ukraine War News through Text Similarity

Nina Khairova^{1,2✉} and Bogdan Ivasiuk³ and Fabrizio Lo Scudo⁴
Carmela Comito⁵ and Andrea Galassi³

¹ Umeå University, Sweden

² National Technical University “Kharkiv Polytechnic Institute”, Ukraine

³ DISI University of Bologna, Bologna, Italy

⁴ DEMACS, University of Calabria, Rende, Italy

⁵ ICAR-CNR, Rende, Italy

ninakh@cs.umu.se

Abstract

The paper focuses on misinformation detection in established global news outlets’ texts covering significant and well-known events of the Russian-Ukraine war. We created the RUWA dataset and applied unsupervised ML approaches as the first dimension of misinformation detection. We consider several different aspects of semantic similarity identification of the articles from various regions in order to confirm the hypothesis that if the news covering the same event from the outlets of various regions over the world are similar enough it means they reflect each other or, instead, if they are completely divergent it means some of them are likely not trustworthy.

1 Introduction

Since the 2016 U.S. presidential election, passing through the U.K. Brexit referendum and the COVID-19 pandemic, misinformation is becoming one of the more significant problems of Modern Society (Zhou and Zafarani, 2020). Two major reasons for this relate to the huge amount of people relying mainly on online sources to get their information and news and the high speed of information spreading via the Internet. Large-scale misinformation campaigns carried out by a big corporation, a political party, or even a government of a certain country can affect various social, economic, and political events. Usually, these kinds of campaigns involve socially sensitive domains such as elections, coronavirus, or military operations and can not only threaten public security and social stability but even affect the results of elections and wars. This has been especially evident in the coverage of the current Russia-Ukraine war when misinformation has become a part of an information war and propaganda activities. The information warfare strategy has a twofold goal: the first is to manipulate the attitudes of people directly involved in the

war, and the second is to modify societies’ opinions of other countries (Thomas, 2014; Theohary, 2018).

To this effect, since the beginning of the Russian invasion of Ukraine, misleading information has been spreading online on social media and by many media outlets. Wide dissemination of misinformation was made possible by two main factors: assessing the truthfulness of facts is highly complex to war events, and news outlets are often inclined to lower the bar of the fact-checking process to provide information as quickly as possible. (Claudia et al., 2021). In this context, careful human-made fact-checking is thus not always possible. However automatic misinformation detection can not always help as well due to lacking labeled benchmark datasets of the particular domain, which relates to the war or military conflicts. While previous works regarding the automatic detection of misinformation do exist, they typically address specific domains, and to the best of our knowledge, little progress has been made regarding the domain of armed military conflicts.

The aim of our work is to analyze and compare news from several established outlets in an unsupervised fashion. Drawing similarities and differences between sources could facilitate future work on fact-checking aimed at establishing finding patterns of reliability of sources and information truthfulness. Specifically, we compare full texts, titles, meaningful sentences, and perform a sentiment analysis.

For this purpose, we create a novel dataset of news in English related to the Russian-Ukrainian war and release it publicly.¹ While we observed some relevant patterns and similarities, our results are not conclusive. Moreover, we find that the number of articles available for each source and

¹https://github.com/ninakhairova/dataset_RUWA

the length of such articles strongly influence the outcome. Nonetheless, we deem our results useful for future works on this topic.

2 Related work

Machine Learning and Deep Learning methods require a large amount of labelled data to effectively train. Applying automatic misinformation detection approaches based on supervised machine learning methods is reasonably common (Capuano et al., 2023; Agrawal et al., 2021). However, in order to use the methods that provide good results it is necessary to train them on specific domain data, which are not available in this context.

We can distinguish several major approaches to misinformation labeling. Most of existed labeled datasets containing political news and some other kinds of news are manually labeled (Silverman et al., 2016; Wang, 2017) or utilize fact-checking websites such as PolitiFact or GossipCop (Shu et al., 2020). For instance, a corpus that is described in Choudhary and Arora (2021) comprises 1,627 articles that were manually fact-checked by professional journalists from BuzzFeed. In some cases, the real news was extracted from a special group of trustworthy sources, while the fake news was extracted from sources of the fake news list like "Business Insider's Zimdars Fake news list" (Janicka et al., 2019). One more approach to annotating the fake news dataset was applied to the AMT dataset (Potthast et al., 2018), which contains 480 articles annotated as fake and true. While fake news articles were imitated by journalists intentionally, the real news was obtained from outlets of several domains.

In general, there are only a few labeled misinformation detection datasets that cover war and military topics (Salem et al., 2019). Furthermore, designing such kind of dataset becomes a much more challenging task due to the fact that the dataset must be created during the ongoing war when actual fact-checking is impossible, there is a good chance of the existence of a bias in various information sources, and so-called "fog of war" effect always can be inherent.

3 Data

Following the requirements of fake news corpus information balance (Rubin et al., 2016; Golbeck et al., 2018), We create a novel dataset called "RUWA" (Russian-Ukraine WAR), composed of

several media outlets from Ukraine, Russia, European, Asia, and the USA. We selected nine of the most information-significant events of Russia's Invasion of Ukraine and aligned articles in English language from all the outlets according to these events. The list of events includes widely-known events such as "The Bucha massacre" and "Sinking of the warship Moskva". To collect articles from the selected news outlets we applied a keyword-based research strategy, conditioned by specific time intervals and topic classification of the sites rubrics. We identified about 100 keywords, which range from geographical names (e.g., Bucha or Olenivka), specific buildings names (e.g., Kramatorsk train station or Mariupol theatre), organizations names (e.g., Red Cross), prominent individual names (e.g., Zelenskyi, Putin), to proper nouns and phrases (e.g., Nuclear Power Plant).

Currently, the RUWA dataset includes more than 16,500 news articles covering the Russian-Ukraine war events that occurred from February 2022 to September 2022. Table 1 shows the article distributions by selected news outlets and events.

4 Methodology of Analysis

Being aware of the complexity of assessing the truthfulness of facts for war events in the absence of the necessary resources to carry out a *journalistic*-oriented process of fact-checking, we decide to relax the problem to assess the veracity of reported facts. We assume that the news reported by news outlets located in the two countries that are directly involved in the conflict can be expected to be highly different. Discrepancies can be substantial up to the point of denying events such as a bombing of residential areas or civilian killings. Additionally, we assume that even though events reported by selected trustworthy independent news agencies and media should be accurate, however, their narrative perspective can remain not neutral.

Thus, as the first dimension of analysis, we focus on textual similarity, comparing the news and assessing if they have a similar meaning. We want to establish whether the news covering the same event from the outlets of various regions over the world are similar enough to indicate they reflect each other or, instead, they are completely divergent and consequently some of them are likely, not trustworthy. We will consider and aggregate several similarity measures that represent many different aspects (Hövelmeyer et al., 2022).

Source	Azovstal	Beginning	Bucha	Nuclear Plant	Prisoners	Railway	Moskva Sinking	Supermarket	Mariupol Theatre	Total
Al Jazeera	23	143	79	186	31	16	34	32	56	600
BBC	22	137	34	236	22	16	17	41	25	550
Censor.Net	826	1730	397	747	117	749	31	173	324	5094
News Front	10	28	18	16	7	7	5	2	1	94
NBC News	8	155	86	129	13	29	36	13	37	506
Reuters	68	924	143	649	23	16	38	15	133	1993
Russia Today	32	14	102	485	236	12	15	22	22	940
Ukrinform	827	3359	570	925	129	601	22	153	163	6749
Total	1816	6490	1429	3373	578	1468	175	436	761	16526

Table 1: RUWA Dataset Selected News outlets and War events

4.1 The similarity between articles based on pre-trained vectors

As the first dimension of analysis, we focus on pairwise evaluating the semantic similarity of all outlets' articles, aggregating all the articles from the same source as a single textual document. As textual encoder, we use FastText (Mikolov et al., 2018).

4.2 Similarity between the title of articles

Authors and correspondents of news agencies and media try to aggregate a major idea of an article, its narrative, or its specific message in the title. Therefore, we analyze similarities between articles over the same topic and use a hierarchical method to aggregate them into similarities between sources. We match each title of every article covering the particular event of the one source with comparable articles titles of the other source. Then we average the similarity scores of titles of two sources that cover the same event and thus we obtain a score similarity for the higher level of the hierarchy, namely for two sources. More formally, our purpose is to obtain a measure of similarity between two sources based on sets of articles titles covering the same event

4.3 Similarity between semantically meaningful sentences

Even if news articles carry different narratives, and contain different informational messages, their semantic similarity score based on the semantics of words or even semantics sentences, can be close enough. Obviously, this is due to the fact that all news articles include a lot of close-meaning sentences or phrases like "correspondent claimed" or 'it seems not obvious' and so on. In order to compare more semantically concentrated texts that only focus on the information of a particular event we extract sets of sentences from all articles of a source that describe only military and close-to-military

actions regarding this particular event.

We utilize two approaches to compare the semantic similarity of such kinds of sentences. In the first one, we process only the sentences that contain keywords related to the considered event. For the second, we add additional knowledge via the lists of verbs that represent the actions involved in certain events. In order to generate such lists, we primarily based on the open list of words associated with the Russian-Ukrainian war from Solopova et al. (2023) and supplemented it with the verbs obtained from the articles. We selected only verbs that relate to a military domain and a given event from all the verbs extracted from the texts. For instance, for the "Moskva sinking" event the list of verbs related to the event includes more than 120 verbs. We also experiment with pre-processing, namely stemming and stop word removal.

4.4 Sentiment Analysis

Given an event for each media outlet, we compute the sentiment analysis for each article concerning that event. We performed sentence-level sentiment analysis and computed the article's overall sentiment by averaging the sentiment of every single sentence. Sentiment analysis has been performed using a statistical approach based on a Convolutional Neural Network for Sentence Classification (Kim, 2014) provided within the NLP toolkit STANZA (Qi et al., 2020).

Due to the linguistic journalist style and jargon, most sentences used within the articles do not provide valuable insights. Hence, we perform a preliminary step and restrict our analysis to a subset of all sentences we consider more informative. To assess the informativeness of a sentence, we employ a keyword-based approach. For each event, we collect all the articles related to that event and rely on TF-IDF to identify the most "significant" words. Then, we maintain only the sentences containing the extracted keywords for each article.

5 Results and discussion

5.1 Leveraging the pre-trained vectors

The experiment confirms our hypothesis. It shows that the semantic similarities between the outlets' texts of countries involved in the conflict (e.g., Censor.net and RT) and websites articles texts of other countries (e.g., Reuters and The Guardian) are less than the similarity of all other considered sites among themselves for almost all events. Also, the semantic similarity coefficients do not have a significant difference, ranging from 91% to 99%.

This can be explained primarily by the special military topic of the news, which is not stipulated by the lexis of the linguistic models. In addition, articles covering the same events may produce different narratives or real and fake facts, but their semantics remain the same.

Table 2 shows the pairwise cosine semantic similarity coefficients for articles of all outlets for the "Sinking of the Moskva" topic based on fastText's subword pre-trained vector from Facebook AI.

5.2 The articles headlines comparison

Leveraging the pre-trained FastText model for headlines' semantic similarity score calculation produces more distributive semantic similarity scores than for full-text articles. However, we observe that the headlines of articles on the same topic and belonging to the same outlet also produce relatively low similarity values, so we can not regard this approach as accurate.

Table 3 shows the example of the distribution of the pairwise cosine semantic similarity coefficients for articles headlines of all outlets for the "Sinking of the Moskva" topic.

We assume that there are a few reasons for this. First of all, the result of handling the titles of the articles depends on the size of the dataset even more than the processing of the articles' full texts. However, in the case of some websites for some events, we do not have a large number of articles (Table 1). Secondly, the effectiveness of the approach based on the semantic similarity of titles may depend on the quality and informativeness of the headlines themselves and their compliance with a particular event. But based on the considered domain we can assume that titles often not only call or describe an event but also reflect the ongoing tensions that can include the authors' biased opinions and feelings.

5.3 Use of extra knowledge for semantic similarity detection

As we mentioned in Section 4.3, we utilize keywords and military action verbs to supplement semantic similarity calculation with additional knowledge about an event. Leveraging sentences that contain keywords related to the considered event enables producing more specific and directly related to the subject of the event texts. However, this inevitably entails losing a large amount of information. Using extra knowledge via the lists of verbs that represent the actions involved in certain events allows us to determine the semantic similarity of news articles, focusing more on the semantic content of articles regarding a particular event. Table 4 shows the example of the semantic similarity for selected sentences that include action verbs for the "Sinking of the Moskva" topic.

The last experiment most explicitly confirms our hypothesis that the semantic similarity coefficient between established outlets of countries involved in the war from two different sides is the smallest. Consequently, we can assume that the value of the semantic similarity coefficient can correlate with producing some other information about the same event that can be identified as misinformation

5.4 Sentiment Analysis

As described in Section 4.4, we perform the sentiment analysis of each document at the sentence level. This is due to the issues Sentiment analysis tools have when working at the document level (Behdenna et al., 2018). In an attempt to mitigate such issues, we decided to perform our analysis at the sentence level and collect the result by simply counting the occurrences for the three classes: *Negative*, *Neutral*, and *Positive*. For each source, we thus aggregate the sentiment counting over all the sentences of the collected articles that focus on a specific event. In Table 5, we report the sentiment analysis made with the NLP toolkit STANZA for the event "Sinking of the Moskva".

Table 5 shows that most Neutral sentences are a common trait among all the sources. That is an expected result due to the journalistic nature of the analyzed documents, which might also be considered a potential noise source for any downstream task. We thus attempted to mitigate that by restricting our analysis to only the sentences that report event-specific keywords, assuming that such sentences would be more suitable to contain potential

	The Guardian	Reuters	Al Jazeera	Censor	CNN	Ukrinform	Russia Today
The Guardian	100%	99.7%	99.9%	94.7%	99.9%	99.8%	99.5%
Reuters	99.7%	100%	99.7%	94.8%	99.6%	99.6%	99.6%
Al Jazeera	99.9%	99.7%	100%	94.5%	99.9%	99.7%	99.4%
Censor	94.7%	94.8%	94.5%	100%	94.8%	95.1%	93.5%
CNN	99.9%	99.6%	99.9%	94.8%	100%	99.8%	99.3%
Ukrinform	99.8%	99.6%	99.7%	95.1%	99.8%	100%	99.3%
Russia Today	99.5%	99.6%	99.4%	93.5%	99.3%	99.3%	100%

Table 2: The semantic similarity for articles of all outlets for the “Sinking of the Moskva” topic based on fastText’s pre-trained vectors

	The Guardian	Reuters	Al Jazeera	Censor	CNN	Ukrinform	Russia Today
The Guardian	75.6%	72.5%	72.5%	74.3%	74.5%	70.7%	69.7%
Reuters	72.5%	77.5%	69.5%	77.2%	72.2%	71.5%	75.5%
Al Jazeera	72.5%	69.5%	70.7%	75.0%	68.9%	64.6%	70.3%
Censor	74.3%	77.2%	75.0%	92.5%	75.2%	66.7%	80.9%
CNN	74.5%	72.2%	68.9%	75.2%	76.6%	70.5%	70.5%
Ukrinform	70.7%	71.5%	64.6%	66.7%	70.5%	81.7%	68.5%
Russia Today	69.7%	75.5%	70.3%	80.9%	70.5%	68.5%	97.5%

Table 3: The semantic similarity for articles headlines of all outlets for the “Sinking of the Moskva” topic

misinformation. We report the results in Table 6. We hypothesize that such a sentence subset could provide more representative information to assess potential source polarization.

6 Conclusion

Creating high-quality resources about a controversial topic such as the Russian-Ukrainian war is a challenging task. In this work, we presented a novel dataset about the conflict, by identifying specific events and imposing a set of constraints on the selection of articles. In our view, such constraints should guarantee a better semantic alignment among articles from news sources, which in turn should facilitate subsequent tasks, such as media bias and misinformation detection. Such a dataset can provide a rich perspective of the different journalistic narrations of the Russian-Ukrainian war and support future research.

Additionally, as a first attempt to detect misinformation in Russia-Ukraine war news, we applied the text similarity approach and Sentiment Analysis. We analyzed the advantages and disadvantages of several approaches to comparing the semantic similarity of news covering the same event in various established outlet news sources.

We also we demonstrated that even though sentiment analysis alone may not be sufficient for misinformation detection, it can provide useful insights that can be combined with other techniques to improve detection accuracy.

We hope that our study contributes to the further development of unsupervised ML approaches to misinformation detection in established outlets news articles.

Acknowledgements

This work was supported by the EU H2020 ICT48 project “Humane AI Net” under contract #952026.

References

- Chetan Agrawal, Anjana Pandey, and Sachin Goyal. 2021. A survey on role of machine learning and nlp in fake news detection on social media. In *GUCON*, pages 1–7. IEEE.
- S. Behdenna, F. Barigou, and G. Belalem. 2018. *Document level sentiment analysis: A survey*. *EAI Endorsed Transactions on Context-aware Systems and Applications*, 4(13):e2.
- Nicola Capuano, Giuseppe Fenza, Vincenzo Loia, and Francesco David Nota. 2023. Content based fake news detection with machine and deep learning: a systematic review. *Neurocomputing*.
- Anshika Choudhary and Anuja Arora. 2021. *Linguistic feature based learning model for fake news detection and classification*. *Expert Systems with Applications*, 169.
- Padovani Claudia, Giuliano Bobba, Baroni Alice, Marinella Belluati, Cecilia Biancalana, Bomba Mauro, Fubini Alice, Marrazzo Francesco, Rega Rossella, Ruggiero Christian, et al. 2021. Italy: A highly regulated system in search of equality. *The Media for Democracy Monitor 2021*, pages 315–386.

	The Guardian	Reuters	Al Jazeera	Censor	CNN	Ukrinform	Russia Today
The Guardian	-	16.8%	40.9%	18.6%	24.7%	25.2%	17.6%
Reuters	16.8%	-	14.7%	17.6%	10.1%	7.0%	19.4%
Al Jazeera	40.9%	14.7%	-	15.5%	24.6%	21.5%	16.4%
Censor	18.6%	17.6%	15.5%	-	7.2%	9.3%	8.1%
CNN	24.7%	10.1%	24.6%	7.2%	-	42.8%	9.7%
ukrinform	25.2%	7.0%	21.5%	9.3%	42.8%	-	11.5%
Russia Today	17.6%	19.4%	16.4%	8.1%	9.7%	11.5%	-

Table 4: The semantic similarity for selected sentences including action verbs for the “Sinking of the Moskva” topic

Source	Articles	Sentences	Negative (%)	Neutral (%)	Positive (%)
Al Jazeera	34	2501	25.83	72.53	1.64
BBC	17	545	30.64	65.32	4.04
Censor	31	314	26.43	69.43	4.14
News Front	5	264	28.03	68.56	3.41
Reuters	15	253	33.99	64.03	1.98
Russia Today	15	463	27.43	68.25	4.32
Ukrinform	22	300	25.0	71.33	3.67

Table 5: Sentiment analysis results for the event “Sinking of the Moskva”.

Source	Articles	Sentences	Negative (%)	Neutral (%)	Positive (%)
Al Jazeera	34	328	31.1	67.99	0.91
BBC	17	96	36.46	57.29	6.25
Censor	31	26	34.62	65.38	0.0
News Front	5	12	33.33	66.67	0.0
Reuters	15	14	71.43	28.57	0.0
Russia Today	15	19	63.16	36.84	0.0
Ukrinform	22	27	40.74	59.26	0.0

Table 6: Sentiment analysis on the TF-IDF filtered sentences for the event “Sinking of the Moskva”.

Jennifer A. Golbeck, Matthew Louis Mauriello, Matthew Louis Mauriello, Keval H Bhanushali, Christopher Bonk, Mohamed Amine Bouzaghane, and Cody Buntain. 2018. [Fake news vs satire: A dataset and analysis](#). In *WebSci*, volume 5, pages 17–21.

Alica Hövelmeyer, Katarina Boland, and Stefan Dietze. 2022. [Simba at checkthat!-2022: Lexical and semantic similarity based detection of verified claims in an unsupervised and supervised way](#). In *Working Notes of CLEF*, volume 3180, pages 511–531. CEUR-WS.org.

Maria Janicka, Maria Pszona, and Aleksander Wawer. 2019. [Cross-domain failures of fake news detection](#). *Computación y Sistemas*, 23:1089–1097.

Yoon Kim. 2014. [Convolutional neural networks for sentence classification](#).

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhersch, and Armand Joulin. 2018. [Advances in pre-training distributed word representations](#). In *LREC*.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. [A stylometric inquiry into hyperpartisan and fake news](#). In *ACL*, pages 231–240. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. [Stanza: A python natural language processing toolkit for many human languages](#).

Victoria Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. 2016. [Fake news or truth? using satirical cues to detect potentially misleading news](#). In *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, pages 7–17. Association for Computational Linguistics.

Fatima K. Abu Salem, Al Feel Roaa, Shady Elbassuoni, Jaber Mohamad, and Farah May. 2019. [Fa-kes: A fake news dataset around the syrian war](#). In *ICWSM*, volume 13, pages 573–582.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. [Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media](#). *Big Data*, 8(3):171–188.

Craig Silverman, Lauren Strapagiel, Hamza Shaban, Ellie Singer-Vine, and Rong Jeremy. 2016. [Hyperpartisan facebook pages are publishing false and misleading information at an alarming rate](#). In *Buzzfeed News*.

Veronika Solopova, Oana-Iuliana Popescu, Christoph Benz Müller, and Tim Landgraf. 2023. [Automated multilingual detection of pro-kremlin propaganda in newspapers and telegram posts](#). *Datenbank-Spektrum*, pages 1–10.

Catherine A Theohary. 2018. [Information warfare: Issues for congress](#). *Congressional Research Service*, pages 7–5700.

Timothy Thomas. 2014. [Russia’s information warfare strategy: Can the nation cope in future conflicts?](#) *The Journal of Slavic Military Studies*, 27(1):101–130.

William Yang Wang. 2017. [“liar, liar pants on fire”: A new benchmark dataset for fake news detection](#). In *ACL*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Xinyi Zhou and Reza Zafarani. 2020. [A survey of fake news: Fundamental theories, detection methods, and opportunities](#). *ACM Computing Surveys (CSUR)*, 53(5):1–40.

**Linking Lexicographic and
Language Learning
Resources (4LR)**

Unlocking the Complexity of English Phrasal Verbs and Polysemes: An Analysis of Semantic Relations Using A-Level Vocabulary Items

Kohei Takebayashi

Tokyo University of Foreign Studies
takebayashi.kohei.v0@tufs.ac.jp

Yukio Tono

Tokyo University of Foreign Studies
y.tono@tufs.ac.jp

Abstract

This two-part study aims to explore semantic transformations exhibited by English phrasal verbs (PVs) and polysemous verbs. Despite the prevalence of PVs in English communication, L2 learners of English have a noticeable tendency to avoid PVs in favour of their one-word equivalents. In order to overcome this avoidance, this research argues that PVs may serve as significant building blocks for developing learners' vocabulary knowledge. To this end, this study explores the possibility of utilising PVs as a bridge between semantic representations of A-level verbs and those of B/C-level verbs as defined by the CEFR. To ascertain the vocabulary levels of verbs found in the most common PVs, a corpus of PV textbooks (size = 3.5 million tokens) was compiled, and frequency data of word pairs composed of verbs and particles were extracted. Also, pairs of PVs and their single-word verb equivalents (SVs) were retrieved from a thesaurus. After producing a list of [PV – SV] pairs, the vocabulary levels of the SVs found on the list were identified in accordance with the English Vocabulary Profile in order to investigate the extent to which PVs can replace their SV counterparts. In addition, both PVs' semantic transparency and the degree of semantic transformation between PVs and their SV equivalents were examined. This study will demonstrate how PVs have the potential to serve as a bridge between A-level and B/C-level verbs, and a selected group of PVs will make a significant impact on the expansion of the range of meaning related to verb semantics. Furthermore, a similar methodology was applied to the investigation of sense relations and semantic transparency exhibited by polysemous A-level verbs in relation to their synonymous SVs. The findings of this study show that verb semantics display a cline of transparency, and learners' deconstruction effort of various senses displayed by polysemes may be facilitated by the semantic precision provided by higher-level SVs.

1 Introduction

The phrasal verb structure is a unique aspect of the Germanic languages (Dagut & Laufer, 1985; Darwin & Gray, 1999), playing an essential role in everyday English communication. However, phrasal verbs' ambiguity in semantic transparency, as well as their irregular syntactic features continue to confuse learners of English worldwide, leading them to choose single-word verbs instead in their production (Siyanova & Schmitt, 2007). Despite the challenges posed by phrasal verbs (PVs) and learners' tendency to avoid PVs in their output in favour of single-word verbs, the current study aims to validate the claim that learners' familiarity with the constituent verbs found in most PVs and their inclination for choosing single-word verbs are two attributes that, together, could offer a more effective means of developing vocabulary. Approaches taken for this study was threefold. First, a corpus of approximately 3.5 million tokens consisting of text data derived from 25 English phrasal verbs textbooks was compiled in order to identify what forms of PVs have been perceived to be most common and essential for learners by teachers and materials developers. Second, by assigning the vocabulary levels indexed in the English Vocabulary Profile presented by the English Profile Programme (Saville & Hawkey, 2010) (hereafter the EVP) to the constituent verbs of the PVs extracted from the corpus, the proportion of basic-level verbs that were present in the form of PVs was identified. And lastly, the EVP's vocabulary levels were assigned to single-word synonyms which were found to correspond to PVs

in order to quantitatively comprehend the relationship between PVs and their single-word verb equivalents in relation to vocabulary level development and its utility in formulating vocabulary learning strategies that would be of help to learners. The study aims to discern the leverage of A-level lexical verbs and the limitations thereof in an effort to obtain quantitative findings which may lend support to a realisation of more efficient or effective vocabulary building strategies for beginner-level learners, and ultimately to the encouragement of vocabulary acquisition among foreign language learners as a whole.

2 Review of Related Literature

PVs are generally recognised as informal or colloquial in tone, occurring 2,000 times per million words in fiction and conversation (Biber et al., 1999). Consequently, PVs have been deemed somewhat inappropriate in academic prose and formal registers, and thus the use of single-word verbs of Graeco-Latin origin in place of PVs has been perceived to be more acceptable and encouraged among learners of academic discipline (Coxhead & Byrd, 2007). However, the ubiquity of PVs has been recognised throughout the use of the English language in which learners are expected to see one PV construction for every 150 English words they encounter (Gardner & Davies, 2007), making a strong case that learners would benefit greatly from their familiarity with the characteristics and utility of PVs (Siyanova & Schmitt, 2007). PVs are considered difficult for learners to acquire due to their structural features generally reserved for the Germanic languages (Dagut & Laufer, 1985; Gilquin, 2015) and semantic complexity arising from idiomaticity and polysemy (Hulstijn & Marchena, 1989; Laufer & Eliasson, 1993; Liao & Fukuya, 2002). Efforts in the creation of PV wordlists have been made to enhance the accessibility of PVs for learners (Biber et al., 1999; Gardner & Davies, 2007; Liu, 2011). In a more recent attempt to reduce the total number

of meanings of PVs to be introduced down to a manageable size based on frequency criteria, Garnier and Schmitt (2015) succeeded in producing the Phrasal Verb Pedagogical list, or more commonly known as the PHaVE list. The list contains 150 most essential PVs, as well as carefully selected definitions for the PVs based on the percentage of usages covered by these definitions. For example, the PV *take off* is given three definitions on the list with the meaning of *removing something* showing 41% of usage while the meanings of *leaving or departing suddenly* and *leaving the ground* showing 28.5% and 14% respectively. By giving priority to high frequency meanings and disregarding the rest, the study succeeded in lowering the number of meanings for the 150 PVs to be listed down to a manageable size of 288.

2.1 Research Questions

From the review of the literature, certain points could be made in regard to English PVs. First, PVs are an indispensable part of English communication without which fluid verbal interaction as well as adequate comprehension would be near impossible. Second, since PVs are one of the distinct features of the Germanic languages that are highly polysemous and often rather figurative, L2 users of English are likely to avoid using PVs in their production in preference to the safer alternative of single-word verbs, consequently making “nonnatives sound stilted and unnatural in speech” (Siyanova & Schmitt, 2007). Finally, even though efforts have been made for the production of pedagogical wordlists of essential PVs, no wordlists have incorporated the utility of vocabulary level classification provided by the English Vocabulary Profile to the discernment of the relationship between PVs and their single-word equivalents (hereafter SVs). The current study explored PVs from three perspectives, touching upon the vocabulary levels of lexical verbs found in common PVs, their convertibility into SVs, and polysemy exhibited by common PVs. Special attention was paid on

vocabulary level progression that was expected to occur as the vocabulary level of the constituent verbs in PVs became more advanced. To this end the following research questions were addressed:

1. What are the CEFR vocabulary levels of constituent verbs found in common PVs?
2. To what extent can PVs be converted into single-word verbs, and what are the CEFR vocabulary levels of those single-word verbs?

3 Method

3.1 A list of common PVs

For the purpose of inquiring into the vocabulary levels of constituent verbs found in common PVs, compiling a list of word combinations potentially capable of forming PVs, i.e., Verb + Particle + Preposition (if any), was necessary. Furthermore, the list would be required to contain details about what combinations were considered most essential for learners by teachers and materials developers, as well as information about vocabulary levels of the constituent verbs in the word combinations that would be found in the compilation process. To this end, three steps were taken. First, a corpus of approximately 3.5 million tokens containing texts from phrasal verbs textbooks was compiled. Second, by using a special pattern matching query for extracting word combinations of Verb + Particle + Preposition (if any), frequency data of all possible combinations from the corpus was extracted. And third, vocabulary levels which correspond to the constituent verbs in the word combinations as defined by the CEFR were assigned to the verbs.

3.1.1 Corpus compiled for the study

In order to determine what combinations of lexical verbs and particles have been considered to be essential for learners by educators, a total of 25 textbooks which had been published specifically for the purpose of introducing and describing the utility of English phrasal verbs were assembled and converted into PDF files (see Appendix A for the list of the textbooks collected for this study). The texts

contained in the PDF files were stored in the corpus manager, Sketch Engine (Kilgariff et al., 2014), and consequently a corpus of approximately 3.5 million tokens was compiled. For convenience, the corpus will be referred to as the Common English Phrasal Verbs Corpus throughout this paper (hereafter CEPVC). Since the CEPVC was designed to contain texts specifically produced for the description of the most useful and common English phrasal verbs, it was presumed that frequency data extracted from the corpus would accurately demonstrate what PVs had been judged to be most common and essential for learners by educators.

3.1.2 Data extraction using CQL

With the aim of extracting specific word combinations from the CEPVC as part of the second step of the process, a special code or query language was used. The query language applicable to Sketch Engine is termed Corpus Query Language (hereafter CQL) (Jakubíček et al., 2010), and is used to set criteria for words, part-of-speech, positions, etc. that would be necessary for accurate data extraction. Since word combinations which would form PVs were of interest in the current study, the following CQL as in (1) was applied to the data extraction process, which proceeded to look for the word combinations of [any lexical verbs except for *be verbs*] + [adverbs or particles or prepositions] + [prepositions (if any)] contained in the CEPVC. Although some textbooks were found to introduce transitive phrasal verbs with an object inserted in between the verb and the particle with abbreviations such as *sb* and *sth* for *somebody* and *something* respectively, (for example, *take sb out*), the majority of phrasal verbs were not introduced in this fashion. And therefore, the inclusion of *sb* (or *somebody/someone*) and *sth* (or *something*) into the CQL was deemed unnecessary in this investigation.

(1) CQL:

```
[tag="V.*"&!tag="VB.?"] [tag="RB|RP|IN"] [tag="IN"]?
```

3.1.3 Filtering out non-PVs

The result from the data extraction via the aforementioned CQL was exported to a spreadsheet in Microsoft Excel. Since the data contained numerous combinations that did not qualify as PVs, a certain screening measure against non-PV combinations was necessary. To this end, only particles were selected immediately following verbs by means of the filter function in MS Excel to ensure that verbs exclusively followed by appropriate particles would remain in the data. Furthermore, relative frequency (per million) was restricted to “5 or above” to filter out those combinations that were theoretically only present in a few textbooks. Moreover, since the data obtained after the filtering process with the particles still contained word combinations such as *know about* and *study at* which would not function as PVs, another measure of identifying questionable verbs (i.e., *know*, *study*, *learn*, etc.) was performed, and their validity as PVs was examined and rejected by comparing example phrasal verbs entries in dictionaries. Finally, careful attention was paid to the deletion of several combinations that contained the particle *to* which included such combinations as *go to*, *need to*, *want to*, etc., for they did not qualify as PVs.

3.1.4 Preparation of vocabulary levels

For the assignment of vocabulary levels to the verbs extracted from the corpus as part of the third step of the process, the CEFR level classification as defined by the English Vocabulary Profile (EVP) of the English Profile Programme (Saville, 2010), was referenced. The data pertaining to verbs in the EVP database was searched online (English Profile, n.d.), and was tabulated in a spreadsheet in Microsoft Excel. Since multiple proficiency levels are given to a verb in the EVP due to the polysemous nature of high-frequency English verbs, certain measure of removing duplicates was necessary. For example, the verb *make* is presented to cover five different levels, ranging from A1 to C1, in the EVP depending on its semantic complexity in given contexts. Since forms, rather than meanings, were

of interest at this stage of the study, duplicates were removed while keeping the least difficult level assigned to each verb for further analysis. Therefore, the level A1 remained tagged to the verb *make* in this study. By applying the same logic to all verbs found in the EVP, the current study proceeded to reduce the 2,317 verbs originally catalogued in the EVP to a total of 1,324 unique verbs.

3.1.5 Assigning vocabulary levels to the verbs

As part of the third step of compiling a list of common PVs, the verbs found in the data were put through the process of vocabulary level assignment by the computer programming language R (R Core Team, 2022) and the application of the open-source package Tidyverse (Wickham et al., 2019). With the help of Tidyverse, the constituent verbs in the word combinations extracted from the CEPVC and their corresponding EVP vocabulary levels were tied together for the completion of the three-step process of compiling a list of common PVs. The relative frequency on the list was to indicate the number of times per million tokens the particular combinations of verbs and particles would appear in the introductions, definitions, and example-sentences in the textbooks. The current study assumes that the higher the frequency the more likely that educators would regard the word combinations as essential PVs.

3.1.6 Counting items

By making use of the table function in R, the occurrences of each vocabulary level across the range of A1 to C2 associated with the verbs present on the list were tallied, revealing the extent of representation held by each vocabulary level in the extracted data. This process allowed the investigation to quantitatively discern the overall vocabulary levels of constituent verbs found in common PVs. The value of the minimum relative frequency (hereafter RF) was incrementally raised from 5 to 7, and eventually to 10 to determine the degree of change in representation held by each vocabulary level as a

function of RF. An increase in RF would mean that a smaller number of PVs would remain on the list, but the remaining PVs would be more common. It was expected that the degree of representation held by the occurrences of A-level verbs in the common PVs would increase as the PVs became more common.

3.2 Collecting synonyms

For the purpose of investigating PV's convertibility into SVs, synonyms from *the Oxford Thesaurus of English* (Oxford University Press, 2006) (hereafter OTE) were collected digitally by means of using all verbs catalogued in the EVP (i.e., 1324 verbs) as search words. All text data containing synonyms that corresponded to each search word in the thesaurus was saved as an individual text file. Furthermore, all text files collected in this manner were processed with the help of computational efficiency provided by the programming language Python (Van Rossum & Drake, 2009) such that multiword synonyms including PVs and single-word synonyms were separated into two different files. The file containing multiword synonyms was further processed in a similar fashion to the filtering procedure of non-PVs described in 3.1.3, appropriate particles were used to extract possible PVs from the multiword synonyms. The search words and the extracted synonymous PVs were then tabulated in a spreadsheet side by side as a list, and vocabulary levels were assigned to all verbs present in the list following the same procedure with R described in 3.1.5. Consequently, a comprehensive list of single-word verbs catalogued in the EVP and their synonymous PVs with vocabulary levels corresponding to all verbs present in the list was generated.

3.2.1 Counting the types of synonyms

Maximising the filter function in MS Excel enabled the specification of particular PVs based on the vocabulary levels of their constituent verbs. This, in turn, facilitated the search capability for the corresponding SVs of those specified PVs.

Consequently, specification of PVs whose constituent verbs belonged to A1 level in accordance with the EVP allowed a search for SVs which corresponded to PVs composed of A1-level verbs. The SVs identified in the process were extracted and had their duplicates removed such that types of SVs synonymous with PVs composed of A1-level verbs were revealed. Since each type of SV had been assigned a vocabulary level, it was made possible to group together the SVs based on their vocabulary levels. The number of SVs contained in each group was measured in comparison to the number of verbs contained in each level group of the EVP to calculate the percentage of representation exhibited by the SVs in each level group. Furthermore, the total number of SVs attained in the process was compared with the total number of verbs catalogued in the EVP (i.e., 1324 verbs) to reveal the extent to which PVs composed of A1-level verbs can be converted into SVs in relation to the total number of verbs listed in the EVP. The same procedure was performed on PVs composed of A1 & A2-level verbs to reveal the convertibility of PVs composed of A-level verbs into SVs. PVs composed of B1-level verbs as well as B2-level verbs were cumulatively added to the process, ultimately revealing the convertibility of PVs composed of all four levels ranging from A1 to B2 into SVs.

4 Results

4.1 A list of common PVs

The application of the CQL to the extraction of word combinations that would form PVs from CEPVC resulted in over 41,000 items which included such word combinations as *do not* and *see also*. After following the procedure of filtering out non-PVs by specifying particles following the verbs, restricting the relative frequency to "5 per million or above", and assigning vocabulary levels to the remaining

verbs with the help of R, a frequency list of 1,402 common PVs was created as shown in Table 1.

Ranking	Verb Level	Verb	Particle	Preposition	RF/million
1	A1	go	out	-	311
2	A1	go	on	-	276
3	A1	look	at	-	236
4	A1	come	in	-	233
5	A1	use	in	-	232
⋮	⋮	⋮	⋮	⋮	⋮
1398	C2	spark	off	-	5
1399	B2	cooperate	with	-	5
1400	C2	refrain	from	-	5
1401	C2	patch	up	-	5
1402	B2	sneeze	at	-	5

Table 1: PVs extracted from CEPVC.

4.1.1 Levels of verbs in common PVs

In pursuit of determining the degree of representation held by each vocabulary level associated with the verbs found in common PVs, the number of occurrences of each vocabulary level across the range from A1 to C2 present in the frequency list were tallied with the help of the table function in R. The investigation proceeded to increase the minimum relative frequency (RF) from 5 to 7, and ultimately to 10 to assess the degree of change in representation held by each vocabulary level as a function of RF. The result shows that 67% of 1,402 common PVs at a minimum RF of 5 were of A-level verb constructions. The degree of representation held by PVs composed of A-level verbs increased to 70% at a minimum RF of 7 with 999 common PVs. Finally, it was found that at a minimum RF of 10, the total number of common PVs stood at 724, and 76% of the PVs were composed of A-level verbs. The result shows that, on average, more than 70% of common PVs are composed of A-level verbs, which quantitatively confirms the intuitive notion that most PVs are constructions of basic-level vocabulary as shown in Table 2.

Level	Relative Frequency (per million)					
	~ 5		~ 7		~ 10	
A1	585	42%	442	44%	344	48%
A2	346	25%	262	26%	206	28%
B1	281	20%	187	19%	118	16%
B2	126	9%	82	8%	43	6%
C1	24	2%	10	1%	5	1%
C2	40	3%	16	2%	8	1%
Total	1402		999		724	

Table 2: Vocabulary levels of verbs in common PVs.

4.2 Convertibility of PVs to SVs

All 1324 single-word verbs registered in the EVP were used as search words for the collection of their synonymous PVs from the OTE. The identified PVs were then tabulated in a spreadsheet alongside their corresponding search words. Vocabulary levels were assigned to all verbs present in the list for the creation a comprehensive list of SVs and their corresponding PVs, which resulted in 10,899 entries as shown in Table 3. Consequently, 72 A1-level verbs and 88 A2-level verbs were determined to be capable of forming PVs. For the PVs composed of A1-level verbs, 909 unique SVs were identified as synonymous with such PVs, representing 69% of all verbs listed in the EVP. Furthermore, a total of 1073 unique SVs or 81% of the verbs catalogued in the EVP were found to be synonymous with PVs composed of A1&A2-level verbs. The addition of PVs composed of B1 and B2-level verbs to the PVs of A-level verb constructions only increased the percentage of SVs to 85% and 87% respectively. Interestingly, an addition of PVs composed of C-level verbs did not change the overall percentage of SVs synonymous with PVs. This finding shows that A-level verbs (i.e., A1 & A2-level verbs) are already capable of producing verb semantics delivered by more than 80% of verbs listed in the EVP when combined with particles, and that PVs composed of higher-level verbs account for less than 10% of verb semantics unrepresented by PVs composed of A-level verbs. The finding also shows that the A-level verbs found in the PVs, which represent 12% of the verbs in the EVP, have a significant impact or leverage in representing verb semantics which are supposedly confined in B/C-level single-word verbs. The breakdown of SVs and their corresponding vocabulary levels is shown in Table 4.

No.	SV Level	SV	Meanings	PV Verb	Particle	Verb Level
1	A1	answer	1	come	back	A1
2	A1	answer	1	get	back to	A1
3	A1	answer	1	write	back	A1
4	A1	answer	2	defend	oneself against	B1
1	1	1	1	1	1	1
10896	C2	yield	4	comply	with	C1
10897	C2	yield	4	consent	to	C2
10898	C2	yield	4	go	along with	A1
10899	C2	yield	4	submit	to	B2

Table 3: A comprehensive list of SVs and PVs

EVP	A1	A2	B1	B2	C1	C2	TTL
	85	116	290	366	212	255	1324
PV	72						72
SV	85%	90	214	245	141	145	909
	87%	78%	74%	67%	67%	57%	69%
PV	72	88					160
SV	85%	76%	58%				12%
	79	100	251	283	172	188	1073
	93%	86%	87%	77%	81%	74%	81%
PV	72	88	168				328
SV	85%	76%	58%	308	180	199	25%
	81	103	258	308	180	199	1129
	95%	89%	89%	84%	85%	78%	85%
PV	72	88	168	178			506
SV	85%	76%	58%	49%	183	205	38%
	82	104	259	313	183	205	1146
	96%	90%	89%	86%	86%	80%	87%
PV	72	88	168	178	61		567
SV	85%	76%	58%	49%	29%		43%
	82	104	259	313	184	205	1147
	96%	90%	89%	86%	87%	80%	87%
PV	72	88	168	178	61	103	670
SV	85%	76%	58%	49%	29%	40%	51%
	82	104	261	317	184	207	1155
	96%	90%	90%	87%	87%	81%	87%

Table 4: SVs' vocabulary levels corresponding to PVs of varied vocabulary levels.

4.3 Exploring polysemy

By rearranging the comprehensive list of SVs and PVs shown in Table 3 such that PVs were listed alongside their single-word synonyms, the levels of semantic complexity exhibited by polysemous PVs became accessible through the means of SVs and the vocabulary levels assigned to them. An example case with the PV *go through* is shown in Table 5 where four meanings contained in *go through* are expressed in the form of SVs. The first meaning of the PV is expressed in 11 unique SVs whose vocabulary levels range from A2 to C2. Even though assigning vocabulary levels to verb semantics can be difficult and is up to a certain level of subjectivity, the observation that B2-level verbs constitute most representation of the meaning suggest that the first sense of the PV belongs to the vocabulary level of B2, or at least belong to an intermediate level. Furthermore, the semantic level of the second meaning of *go through* can be determined by the SV that best captures the notion of *using up something* even though this type of judgement requires statistical analysis of intuition to overcome the inevitable subjectivity. As a by-

product of investigating PVs, the study was able to produce a list of single-word verbs and their corresponding single-word synonyms, and it was determined that the same logic of determining levels of semantic complexity by means of single-word synonyms can be applied to the investigation of polysemous single-word verbs. An example case is shown in Table 6 where the verb *colour* is presented to have four distinct meanings. By observing the concentration of B2 and C-level verbs being synonymous with the semantics imparted by the third and fourth meanings of *colour*, it can be intuitively determined that the latter two senses held by *colour* belong to an advanced vocabulary level.

Level	PV Verb	Particle	Meaning	SV	Level
A1	go	through	1	receive	A2
A1	go	through	1	stand	A2
A1	go	through	1	experience	B1
A1	go	through	1	face	B1
A1	go	through	1	bear	B2
A1	go	through	1	endure	B2
A1	go	through	1	suffer	B2
A1	go	through	1	tolerate	B2
A1	go	through	1	undergo	C1
A1	go	through	1	sustain	C2
A1	go	through	1	withstand	C2
A1	go	through	2	spend	A2
A1	go	through	2	waste	B1
A1	go	through	2	consume	B2
A1	go	through	2	exhaust	C1
A1	go	through	2	squander	C2
A1	go	through	3	check	A2
A1	go	through	3	search	B1
A1	go	through	3	inspect	C1
A1	go	through	4	study	A1
A1	go	through	4	check	A2
A1	go	through	4	consider	B1
A1	go	through	4	analyse	B2
A1	go	through	4	examine	B2
A1	go	through	4	inspect	C1
A1	go	through	4	scan	C1

Table 5: Polysemous PVs expressed in SVs

Level	Word	Meaning	Synonym	Level
A1	colour	1	paint	A1
A1	colour	1	stain	C2
A1	colour	2	blush	B2
A1	colour	3	affect	B2
A1	colour	3	influence	B2
A1	colour	3	poison	B2
A1	colour	3	distort	C1
A1	colour	3	twist	C1
A1	colour	4	bend	B2
A1	colour	4	disguise	B2
A1	colour	4	strain	B2
A1	colour	4	distort	C1
A1	colour	4	enhance	C1
A1	colour	4	exaggerate	C1
A1	colour	4	overdo	C1

Table 6: Polysemy expressed in synonyms.

4.4 Interchangeability of single-word verbs

Table 6 demonstrates that A1-level single-word verbs behave in a similar way to that of PVs composed of A1-level verbs. By following the same methodology described in 3.2.1, it was determined that A1-level single-word verbs, which represent 6% of the entire verbs listed in the EVP, are

interchangeable with 61% of unique verbs catalogued in the EVP. Furthermore, 79% of all verbs presented in the profile were determined to be synonymous with A-level single-word verbs (i.e., A1 and A2 combined) as shown in Table 7. The results indicate greater expressiveness of PVs in comparison with single-word verbs since the semantic representation exhibited by PVs composed of A1-level verbs as measured by the number of corresponding SVs was 69% or 8% greater than that of their single-word counterparts. Interestingly, however, the interchangeability with synonyms exhibited by single-word verbs belonging to a range of vocabulary levels from A1 to B2 collectively demonstrated a 94% coverage of all verbs catalogued in the EVP, indicating that the majority of semantics required in communication can be accomplished by employing single-word verbs of up to level B2. The finding also shows that the number of SVs corresponding to PVs was capped at 87% of all verbs listed the EVP even with the inclusion of B2-level verbs as constituent verbs, while single-word verbs were seen to outperform PVs in their expressiveness after passing the B2-level threshold as determined by the number of corresponding synonyms.

EVP	A1	A2	B1	B2	C1	C2	TTL
Verbs	84	116	290	366	212	255	1324
Synonyms	99%	77%	91%	91%	91%	91%	91%
Verbs	84	113	273	356	212	255	1324
Synonyms	99%	97%	94%	94%	94%	94%	94%
Verbs	84	106	265	324	188	225	1192
Synonyms	99%	91%	91%	89%	89%	88%	90%
Verbs	84	113	273	356	212	255	1324
Synonyms	99%	97%	94%	94%	94%	94%	94%

Table7: Interchangeability of single-word verbs with their single-word synonyms of varied vocabulary levels

5 Discussion

5.1 Summary of major findings

The current study has explored characteristics of PVs from three specific perspectives: vocabulary levels, convertibility, and

polysemy. In investigating the vocabulary levels of common PVs, the majority of constituent verbs found in the 1,402 common PVs were judged to be of A-level classification as 70% of common PV-forms were found to be combinations of an A-level lexical verb and a particle occasionally followed by a preposition regardless of modifications made to minimum relative frequencies. Furthermore, certain lexical verbs and particles were found to be particularly productive in the formation of PVs. The top 20 most productive lexical verbs (i.e., *go, come, get, run, look, move, fall, take, keep, put, walk, stay, work, live, pull, grow, make, stand, bring, and hold*) of which *live* is the only B-level verb were collectively capable of forming 21.8 PVs on average at a minimum relative frequency of 5 per million, while the top 3 (i.e., *come, go, and get*) demonstrated their capability of producing 53 PV-forms on average. Likewise, the top 10 most frequent particles (i.e., *up, out, in, on, off, down, for, back, away, and with*) were each found to be part of more than 100 PV-forms on average while the top 3 (i.e., *up, out, and in*) being components of 169 PVs on average. The convertibility of PVs to SVs was investigated by means of synonyms contained in the OTE. Subsequently, PVs composed of A-level verbs (i.e., A1 and A2 combined) were found to be synonymous with 81% of all single-word verbs from a wholistic range of vocabulary levels from A1 to C2 catalogued in the EVP, prompting the study to conclude that PVs not only function as a bridge between vocabulary levels (i.e., PVs composed of A-level verbs acting as a bridge between A and B levels in semanticity specifically, while PVs composed of B-level verbs bridging between B and C level verb semantics), but also as “a free pass” allowing access to various tiers of semantic representations. In addition, the degree of verb semantics delivered solely by PVs composed of B-level verbs was determined to be relatively modest accounting for less than 10% of semantics unrepresented by PVs composed of A-level verbs, signalling the significance of A-level verbs in expressiveness when combined with particles.

Furthermore, polysemy exhibited by polysemous PVs and single-word verbs as well as possible vocabulary level classification of their various semantics was explored by considering the utilisation of single-word synonyms and their assigned vocabulary levels. Attempts at assigning vocabulary levels to verb semantics can be vulnerable to criticism as a high degree of subjectivity would inevitably be involved. However, the current study has successfully suggested a method that utilizes synonyms and their assigned vocabulary levels to provide a more objective approach in determining the levels of difficulty among the various semantics exhibited by polysemous verbs. Finally, the investigation into the interchangeability of single-word verbs with their single-word synonyms indicated that the semantic expressiveness exhibited by PVs composed of A-level verbs was greater than that of A-level single-word counterparts, while single-word verbs' semantics became greater than those of PVs after crossing the B2-level threshold.

5.2 Pedagogical implications

The current study has succeeded in incorporating the utility of vocabulary level categorisation brought forward by the EVP into clarifying the hitherto vague notion of *high-frequency* or *common* often associated with the descriptions of PVs. By observing the results obtained from the current study which indicate that PVs are vastly synonymous with single-word verbs, it stands to reason that learners would avoid PVs when the safer alternative of using single-word equivalents is readily available without taking the risk of misinterpretations and idiomaticity associated with PVs. Admittedly, PVs are not indispensable for conveying one's intentions, and single-word verbs are often more preferred in certain registers. However, since PVs are extremely common in spoken English, complete disregard for PVs in the classroom could inhibit learners' ability to comprehend details provided in situations where the use of PVs would

be more appropriate, which are ubiquitous in the English-speaking community. With the knowledge from the current study that more than 70% of common PVs are composed of A-level lexical verbs, as well as the fact that 81% of single-word verbs catalogued in the EVP (or 1,073 single-word verbs) could be expressed by at least one PV composed of an A-level verb (see Table 4), certain measure of incorporating the utility of both PVs and single-word verbs into learners' lexical development could be proposed. For example, compilation of wordlists that display the relationship between PVs and SVs (single-word counterparts) could be considered. Table 5 could be such a wordlist that conveys PVs' semantic relations to their SVs, clearly demonstrating that higher-level single-word verbs could be expressed by A1-level verbs when combined with particles. By observing such wordlists, learners could clarify the meaning of newly encountered C-level verbs such as *exhaust* and *squander* in Table 5 by referring to their lower-level synonyms (e.g., *waste* and *consume*), or to their PV counterpart (i.e., *go through*), which could be construed as more semantically transparent. Additionally, semantically opaque versions of *go through*, such as the one listed as Meaning No. 4 in Table 5, could be familiarised with the help of transparency provided by B-level single-word verbs such as *consider*, *analyse*, and *examine*. Likewise, a collection of semantically opaque PVs such as *put up with* and *take after* could be listed and have their meanings clarified by the semantic concreteness provided by their single-word counterparts included in an example list shown in Table 8. Such a list could motivate learners to learn not only the meanings of ambiguous PVs expressed in SVs, but also the fact that A-level verbs such as *take* and *stand* could suggest *to tolerate* or *to endure*. Furthermore, the symbiotic relationship between PVs and SVs could be put to good use so as to eliminate the need for placing L1 translations alongside target

words which may only encourage memorisation of translated texts rather than the semantics of the target English words themselves. For instance, research has shown that access to external information such as dictionaries and glosses, as well as repeated exposure, foster the formation of form-meaning relationships within learners' lexicon (Hulstijn et al., 1996). Therefore, instead of relying on L1 translations, glosses that provide single-word equivalents of basic vocabulary levels corresponding to target vocabulary items may be proposed. Table 5 also indicates the potential efficiency in learning when PVs are used to good advantage as the 19 unique SVs on the list could easily be expressed by only one English phrasal verb presumably without the need for L1 translations since the verbs *go* is undoubtedly already known by learners. Furthermore, the current study has succeeded in identifying 169 single-word verbs catalogued in the EVP that are unexchangeable with PVs (see Appendix B). Such verbs include *change* and *walk*, and further study into why such verbs do not possess PV counterparts may shed light on more effective approaches to teaching and learning PVs.

Level	Single-word	→	PV Verb	Particle	Preposition	Verb Level
A1	take	→	put	up	with	A1
A2	stand	→	put	up	with	A1
B1	accept	→	put	up	with	A1
B1	support	→	put	up	with	A1
B2	tolerate	→	put	up	with	A1
B2	swallow	→	put	up	with	A1
B2	endure	→	put	up	with	A1
B2	bear	→	put	up	with	A1
B1	suggest	→	take	after	-	A2
B2	recall	→	take	after	-	A2
C1	resemble	→	take	after	-	A2

Table 8: Single-word verbs corresponding to PVs

6 Conclusion

This paper has demonstrated the utility of incorporating the vocabulary level classification provided by the EVP into investigating several characteristics of English phrasal verbs. By replacing such expressions as *high-frequency* and *common* with more precise account of CEFR-based level specifications such as, *A-level*, the current study succeeded in shedding light on the

multi-faceted nature of phrasal verbs which involved vocabulary levels, convertibility to single-word verbs, and polysemy. The study has empirically confirmed the intuitive notion that phrasal verbs are combinations of basic-level verbs and particles with corpus-informed quantitative data which could be of use in encouraging learners to adopt phrasal verbs into their repertoire. Furthermore, the study has confirmed that certain symbiotic relationships between phrasal verbs and single-word verbs in vocabulary learning could be established and put to use in creating materials for pedagogical purposes. The effectiveness of phrasal verbs in assisting the development of learner vocabulary is a topic of further research. Moreover, the account of semantic transparency exhibited by phrasal verbs cannot be detached from subjectivity, which may complicate efforts in classifying what is transparent and what is opaque and in placing them along the cline of semantic transparency. However, the quantitative information regarding PVs obtained from the current study suggests that more than four fifths of all verbs indexed in the EVP have at least one PV counterpart composed of a basic-level lexical verb, and therefore, more learning resources other than L1 translations that take full advantage of PVs in vocabulary learning could be proposed and put to good use. In other words, English phrasal verbs could be one untapped resource that have been shunned by learners for too long. Further research into the relationship between phrasal verbs and single-word verbs may hold the key to drastically reducing the workload that learners have to handle, or *deal with*, when furthering their lexical knowledge.

References

- Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., & Quirk, R. (1999). Longman grammar of spoken and written English. Longman.

- Coxhead, A., & Byrd, P. (2007). Preparing writing teachers to teach the vocabulary and grammar of academic prose. *Journal of Second Language Writing*, 16(3), 129–147. <https://doi.org/10.1016/j.jslw.2007.07.002>
- Dagut, M., & Laufer, B. (1985). Avoidance of phrasal verbs—a case for contrastive analysis. *Studies in Second Language Acquisition*, 7(1), 73–79. <https://doi.org/10.1017/s0272263100005167>
- Darwin, C. M., & Gray, L. S. (1999). Going after the phrasal verb: An alternative approach to classification. *TESOL Quarterly*, 33(1), 65. <https://doi.org/10.2307/3588191>
- Gardner, D., & Davies, M. (2007). Pointing out frequent phrasal verbs: A corpus-based analysis. *TESOL Quarterly*, 41(2), 339–359. <https://doi.org/10.1002/j.1545-7249.2007.tb00062.x>
- Garnier, M., & Schmitt, N. (2015). The phave list: A pedagogical list of phrasal verbs and their most frequent meaning senses. *Language Teaching Research*, 19(6), 645–666. <https://doi.org/10.1177/1362168814559798>
- Gilquin, G. (2015). The use of phrasal verbs by French-speaking EFL learners. A constructional and collostructional corpus-based approach. *Corpus Linguistics and Linguistic Theory*, 11(1). <https://doi.org/10.1515/cilt-2014-0005>
- Hulstijn, J. H., & Marchena, E. (1989). Avoidance. *Studies in Second Language Acquisition*, 11(3), 241–255. <https://doi.org/10.1017/s0272263100008123>
- Hulstijn, J. H., Hollander, M., & Greidanus, T. (1996). Incidental Vocabulary Learning by Advanced Foreign Language Students: The Influence of Marginal Glosses, Dictionary Use, and Reoccurrence of Unknown Words. *The Modern Language Journal*, 80(3), 327–339. <https://doi.org/10.1111/j.1540-4781.1996.tb01614.x>
- Jakubíček, M., Kilgarriff, A., McCarthy, D., & Rychlý, P. (2010). Fast Syntactic Searching in Very Large Corpora for Many Languages. *Proc PACLIC* (Vol. 24, pp. 741–747). Japan.
- Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten Years on. *Lexicography*, 1(1), 7–36. <https://doi.org/10.1007/s40607-014-0009-9>
- Laufer, B., & Eliasson, S. (1993). What causes avoidance in L2 learning. *Studies in Second Language Acquisition*, 15(1), 35–48. <https://doi.org/10.1017/s0272263100011657>
- Liao, Y., & Fukuya, Y. J. (2002). Avoidance of phrasal verbs: The case of Chinese learners of English. *Language Learning*, 54(2), 193–226. <https://doi.org/10.1111/j.1467-9922.2004.00254.x>
- Liu, D. (2011). The most frequently used English phrasal verbs in American and British English: A multicorpus examination. *TESOL Quarterly*, 45(4), 661–688. <https://doi.org/10.5054/tq.2011.247707>
- Oxford University Press. (2006). *Oxford Thesaurus of English*.
- R Core Team (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>. *Statistical Computing*, Vienna, Austria. URL <https://www.R-project.org/>.
- Saville, N., & Hawkey, R. (2010). The english profile programme – the first three years. *English Profile Journal*, 1. <https://doi.org/10.1017/s2041536210000061>
- Sivanova, A., & Schmitt, N. (2007). Native and nonnative use of multi-word vs. one-word verbs. *IRAL - International Review of Applied Linguistics in Language Teaching*, 45(2). <https://doi.org/10.1515/iral.2007.005>
- Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.
- Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Golemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen

TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." /Journal of Open Source Software/, *4*(43), 1686. [doi:10.21105/joss.01686](https://doi.org/10.21105/joss.01686) .

A Appendices

Appendix A. The List of 25 Textbooks Referenced for the Compilation of the CEPVC

- Booth, T., & Davies, B. F. (2021). English for everyone: English phrasal verbs. DK.
- Burdine, S., & Barlow, M. (2008). Business phrasal verbs and collocations. Athelstan.
- Cambridge Univ. Press. (2006). Cambridge Phrasal Verbs Dictionary.
- Colby, H. E. (2014). 150 Useful English collocations, idioms, and phrasal verbs. BookBaby.
- Dixson, R. J. (2004). Essential idioms in English: Phrasal verbs and collocations. Longman.
- Emir, M., & Allans, R. (2019). Advanced English: Idioms, Phrasal Verbs, Vocabulary and Phrases: 700 Expressions of Academic Language. Independently Published.
- Errey, M. (2007). 1000 Phrasal Verbs in Context. Teflgames.com.
- Flockhart, J., & Pelteret, C. (2012). Work on your phrasal verbs: Master the 400 most common phrasal verbs. Collins.
- Gul, I. (2020). 1000+ Phrasal Verbs With Meanings and Sentences. TheEnglishLover.com.
- Harrison, J. (2003). Phrasal verbs. Stanley.
- Hart, C. W. (2020). Phrasal Verbs. Barrons Educational Services.
- Makar, A. (2020). 100 Practical English Phrasal Verbs. Independently Published.
- McCarthy, M., & O'Dell, F. (2017). English Phrasal Verbs in Use Advanced Book with Answers: Vocabulary Reference and Practice. Cambridge University Press.

- McIntosh, C. (2006). Oxford Phrasal Verbs Dictionary for learners of English. OUP Oxford.
- Melvin, J. (2018). Phrasal Verbs and Idioms in Context.
- Melvin, J. (2019). Phrasal Verbs : Practice Tests. English Language Academy.
- Mordaunt, O. G., & McGuire, M. (2020). Phrasal Verbs for English Language Learners. Cambridge Scholars Publishing.
- Parkinson, D. (2007). Really Learn 100 Phrasal Verbs: Learn the 100 most frequent and useful phrasal verbs in English in six easy steps. OUP Oxford.
- Roche, M. (2020). Master English Collocations & Phrasal Verbs. Roche English Language Publishing.
- Sandford, G. (2012). Amazingly Easy Phrasal Verbs! Praski Publishing .
- Shepherd, D., Wagland, M., & Pinkney, R. (2015). Easy Phrasal Verbs: Learn English Through Conversations. Shaer Publishing.
- Smith, D. B. (2020). English Phrasal Verbs Ultimate Collection.
- Spears, R. A. (2006). McGraw-Hill's Dictionary of American idioms and Phrasal verbs. McGraw-Hill.
- Spears, R. (2008). McGraw-Hill's Essential Phrasal Verbs Dictionary. McGraw-Hill.
- Wyatt, R. (2006). Check Your English Vocabulary for Phrasal Verbs and Idioms. A & C Black.

Appendix B. EVP Verbs With No PV Equivalents

A1

be; change; walk

A2

boil; brush; camp; download; email; lend; matter; point; snow; surf; text; thank

B1

apologise; barbecue; blog; breathe; clap; cycle; deserve; fax; film; fry; grill; guide; hitchhike; iron; lock; owe; own; rebuild; sew; skate; ski; smell; smile; star; sunbathe; type; unpack; upload; vote

B2

alter; bark; benefit; blink; bookmark; bounce;
coach; compromise; cruise; debit; doubt; enable;
enquire; entitle; envy; fine; frighten; Google;
gossip; guarantee; harm; kneel; link;
misunderstand; participate; photograph; poison;
pollute; punch; reward; rewrite; rip; rule; sentence;
sneeze; sob; specialize; splash; spray; stare; steer;
stroke; suspect; switch; terrify; unlock; whisper;
whistle; yawn

C1

alternate; commute; distort; generalize; grade; hop;
insert; modify; narrow; oblige; outnumber;
outrage; presume; price; privatize; readjust;
recharge; recreate; redevelop; relocate; rethink;
scare; simplify; sip; smuggle; starve; summarize;
surge

C2

amend; arch; blackmail; bond; cling;
commemorate; diagnose; dice; drift; exemplify;
filter; fluctuate; frown; gasp; gesture; giggle; glare;
glue; grin; haul; hum; maximize; merit;
misinterpret; misplace; moan; murmur; nest;
overlap; pat; redistribute; reign; restructure; rhyme;
riot; scar; shape; shrug; shudder; speculate; spit;
sprinkle; spur; squeak; stain; vaccinate; weep; wink

Towards a Unified Digital Resource for Tunisian Arabic Lexicography

Elisa Gugliotta¹ and Michele Mallia¹ and Livia Panasci^{2*}

1. Istituto di Linguistica Computazionale, Consiglio Nazionale delle Ricerche;

2. Sapienza University of Rome

1. {firstname.lastname}@ilc.cnr.it, 2. livia.panasc@outlook.it

Abstract

This paper presents our work on linking language tools for Tunisian Arabic, focusing on a lexicographic database and a corpus of informal written texts. This work on Tunisian Arabic is an ongoing pilot study, while our wider goal is to create resources for various under-resourced languages. We outline a methodology that emphasises open science principles, leveraging existing language resources and NLP tools for standardisation and annotation. Our approach ensures reproducibility and benefits other researchers. We share annotated data on a digital platform and release NLP tools on a dedicated repository. Our work aligns with FAIR principles, facilitating open and effective research on under-resourced languages.

1 Introduction

This paper describes a research methodology for the study of under-resourced languages, presenting it through the exemplification of a pilot study we are conducting on Tunisian Arabic dialect (TA). Therefore, the work is part of a wider project aiming at supporting studies on under-resourced languages using both quantitative research methods, such as statistical analysis and Deep Learning techniques, and qualitative research methods, such as Linguistics and Dialectology. The lack of computational resources, such as annotated corpora, language models, and digital lexicons, to name a few, has been a major roadblock to the processing of under-resourced languages. Usually, these languages have a poor tradition of linguistic studies: to a few ancient written sources correspond few analyses on lexicography, morphology, phonetics, etc. Moreover, it lacks communication between scientific sectors: different research areas, such as Digital Humanities and Dialectology, hardly converge

* All three authors collaborated on the project. For academic purposes, E. Gugliotta is responsible for sections 2, 3.2, 4.2, 5.2, 6; M. Mallia for sections 2 (Step 3), 5.1; L. Panasci for sections 1, 3.1, 4.1, 5 (introduction).

and collaborate in the study of under-resourced languages. Consequently, the studies that have been carried out remain isolated and underexploited. On the contrary, only a comprehensive approach can reflect the dynamism and complexity of a language, by preserving the quality of linguistic data at all stages of data processing, from identification and selection, collection, pre-processing, processing, analysis, annotation and data fruition. For what concerns Arabic dialects, i.e. Colloquial Arabic (CA), to which TA belongs, the limited availability of data is one of the main reasons why these varieties are still defined as under-resourced.¹ At the same time, the specificity of the multilingual realities of the Arab countries, with special reference to the diglossic situation,² makes building corpora of CA a challenge. CA has always been a predominantly oral language, very few written texts have been recorded and texts prior to the 20th century are extremely rare.³ There is no standardised writing system, the studies that have been conducted so far have often focused on specific aspects of the language and have almost never been connected with each other. Linguistic research that has been conducted in the past often did not respect strict methodological criteria (for example, not reporting the number of informants, their age, or geographical origin). It is for all these reasons that, although in the last decades the building of linguistic corpora for Arabic has incredibly increased (Darwish et al., 2021) and although a number of CA corpora has recently been released (see Section 3.2), these corpora cannot support wide linguistic analysis.

Therefore, our project, whose ultimate goal is to connect and make linguistic data on under-resourced languages easily available by users, has as its first step the data collection. To collect

¹For details on the causes that lead some languages to be defined as under-resourced, see Pretorius and Soria (2017).

²See Ferguson (1959); Versteegh (2014); Owens (2006); Abboud-Haggar (2006); Sayahi (2014).

³The CA literature is really rare: see Davies (2006).

data, we exploit existing resources, i.e. ancient (dialectological sources from the 19th century to the present) and modern (corpora of authentic written TA), which, although originally created for very different purposes, come together to present more complete and detailed data possible.⁴

In Section 2 we present the main aims of our project, while in Section 3 we start reporting on the pilot study, by outlining different kinds of work and data available for TA. In Section 4, we describe the linguistic resources employed for our study (a lexicographic database TA-Italian and vice versa and a TA corpus). These were previously created for specific purposes, that we are currently normalising in terms of content and format standardisation. Such data will be released through a digital platform aimed at providing access to linguistic information and facilitating complex queries, which would undoubtedly be a milestone in this domain. At the same time, computational tools built to process these data will be made available through a dedicated repository.⁵ In Section 5 we outline the project methodology stages applied to the pilot study so far. Indeed, our ultimate goal is to unify a big amount of TA data (described in Section 4), to be employed for future studies, in different fields (NLP, Digital Humanities, Linguistics and Dialectology).⁶ With this aim, we devised a methodology inspired by the principles of the data economy, sustainability of research and the FAIR principles of open science.⁷ Finally, in Section 6, we discuss our conclusions and future works.

2 General Project Aims

The macro-objective of this project is to develop and put into practice a hybrid methodology that could strongly contribute to the current state of research on under-resourced languages, starting from Arabic dialects. Following open science principles, the methodology aligns with transparency, collaboration, and accessibility. Such methodology is organized in three steps. In Step 1, existing linguistic resources are compiled using freely available tools, corpora, glossaries, and dictionaries from the scientific community, promoting openness. The work of Step 2 adheres to open science principles. In fact, text standardization and annotation are realised

by using NLP tools. This enables work reproducibility and allows other researchers to exploit our tools and methodology. In Step 3, annotated data and NLP tools are provided, emphasizing open data. Overall, the methodology adheres to the FAIR principles (Wilkinson et al., 2016; De Jong et al., 2018), promoting Findability, Accessibility, Interoperability, and Reusability of linguistic resources and data, facilitating open and effective research on under-resourced languages.⁸ Since our ultimate goal is to advance research on different under-resourced languages, at the end of Step 3 there is a recursive cycle to start the process again (Step 1) with a new under-resourced language or language variety.

Step 1. Resource Compilation: Economizing Data. This first work stage is based on the concept of ‘data economy’ rather than ‘creation from scratch’. It aims to identify existing linguistic tools, corpora, glossaries, and dictionaries available among the scientific community in various formats and for different purposes. Such resources are often underutilized after their initial creation and use (Macchiarelli, 2023). This is because, once used for the purposes for which they were created, they are not maintained, extended, or adapted to standards that would allow their use by audiences other than those imagined at the time of their creation (Pretorius and Soria, 2017). We will use any available resources that we become aware of, such as resources created for other purposes, like corpora created for sentiment analysis, which perhaps do not have fine-grained grammatical annotations. We will be in charge of the annotation of these data. Our first objective is to retrieve these resources, promoting data sustainability, and standardise them into a unified format (Step 2).

Step 2. Standardisation and Annotation: Enhancing Linguistic Insights. This stage also includes text normalisation and the semi-automatic annotation of linguistic features is done using existing tools. Text normalisation ensures consistency and prepares the text for subsequent processing. In the analysis of under-resourced language data, we consider morpho-syntactic information crucial for disambiguating semantically challenging elements extracted from the production context (Jarrar et al., 2022; Nahli et al., 2023). For this reason, we train (and release at the end of Step 3) morphological embeddings for each language (Cotterell and Schütze,

⁴See Section 4 for linguistic resources description.

⁵At this link: <https://github.com/LinguaeVerse>.

⁶About cooperation, use, sustainability of language data in these fields, see Fišer and Witt (2022).

⁷See Section 2 for further details on these topics.

⁸For further information on the FAIR principles, please see <https://www.go-fair.org/fair-principles/>.

2015).⁹ To produce morpho-syntactic annotations we can exploit existing tools, such as a Multi-Task architecture created for TA data annotation (Gugliotta et al., 2020). Such an architecture can learn linguistic insights from small, noisy data (Gugliotta and Dinarelli, 2023). Thus, it can be useful for processing multiple varieties of CA, starting with the varieties most similar to TA (the target language of our pilot study), such as the North African varieties.

Step 3. Providing Data: Enabling Further Studies. Finally, the last work stage focuses on providing annotated data to support further studies in this direction. The annotated data will be available through a digital platform that supports queries from researchers interested in linguistic and lexicographic studies on the collected texts. This, together with the release of annotated data and pre-trained morphological embeddings, could greatly facilitate the preservation and digital accessibility of these languages, thereby fostering cultural and linguistic diversity in the digital world.

On morphological embeddings. In this phase, we investigate the incorporation of morphological knowledge in word embeddings, to capture semantic and morphological similarities. Training such embeddings for the under-studied language would have several utilities. They would ease the annotation of additional data; they would help in lexical and ontological modeling of the language resources underlying the digital platform (see below). Finally, we could release a tool with great potential, which under-resourced languages generally lack, and which we could easily investigate from the data annotated in Step 2. After an initial phase of evaluating the available models (see Sezerer and Tekir, 2021), we will train on the already annotated data a model capable of generating embeddings combining morphemes, POS-tags and lemmas.¹⁰

Concerning our pilot study on TA, Yagi et al. (2022), shows that the evaluation metrics for Arabic embedding models need to take into consideration the morphological characteristics of the language. Moreover, Salama et al. (2018) emphasize the incorporation of morphological analysis in the training of word embedding models, given the

⁹Morphological embeddings are numerical representations of morphemes or morphological units in a language, embedded in a continuous vector space (Cotterell and Schütze, 2015). For completeness, see also Bengio et al. (2003). For further information on morphological embeddings, please see below.

¹⁰Using morphemes for word embeddings in morphologically rich languages is useful to encode more semantic information (Romanov and Khusainova, 2019).

morphological complexity of the Arabic language. The drive to exploit word embeddings for Arabic NLP has been matched by efforts to annotate Arabic texts with Linked Data. Bouziane et al. (2020) present a comprehensive framework for annotating Arabic texts with Linked Data. This kind of annotated data becomes a precious resource for training more sophisticated NLP models, contributing to the larger goal of making CA texts more accessible, less ambiguous, and more useful in various NLP applications, such as information retrieval, word sense disambiguation and other related areas.

On the digital platform. Such a platform is intended not only as a tool for conducting queries but also as an aggregator of information, particularly focusing on under-resourced languages. One of the salient features of the platform will be its capacity to perform complex queries through data correlation. This is essential for extracting nuanced information and recognizing patterns within the data (Alhafi et al., 2019). By enabling users to create complex queries that integrate data from multiple sources, the platform facilitates simultaneous analysis of the two data sources (querying both via the central Analysis Node, see Figure 1). This advanced capability helps researchers derive more meaningful insights by leveraging the combined power of integrated data.

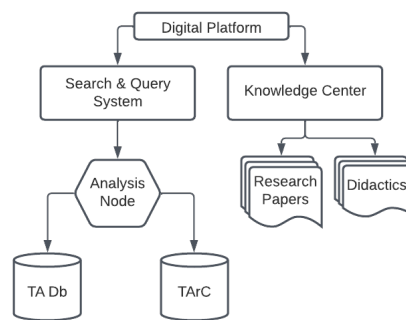


Figure 1: The digital platform general structure

To understand the type of digital platform we plan to implement, we refer to similar work on Arabic language, this is the one of Jarrar and Amayreh (2019). This lexicographic search engine is constructed atop the most extensive Arabic multilingual database, facilitating users in searching and retrieving translations, synonyms, definitions, and more.¹¹ Similar to this work, our platform will be developed with cutting-edge features and in alignment with the

¹¹The search engine can be accessed at <https://ontology.birzeit.edu>.

recommendations and best practices of the World Wide Web Consortium (W3C) for publishing data on the web. Additionally, our digital platform will serve as a comprehensive repository, aggregating diverse types of information related to the study of the under-studied language. It will encompass a wide range of resources such as recipes, travel blogs, and other existing information on the under-studied language. By incorporating this diverse information, our platform is intended to provide a holistic and rich source of data for researchers and others interested in discovering languages and cultures. Furthermore, with the texts and information collected on our platform, it will be possible to develop teaching materials based on authentic data (*Didactics* in Figure 1). Regarding the *Analysis Node*, in Figure 1, this module is understood as the one in which the matching process between the data collected in the two instruments is performed. In the case of the TA data, this process will be based on the *root* level information.¹² Moreover, the platform will adhere to the W3C's OntoLex-Lemon RDF model,¹³ emphasizing our dedication to ensuring standardisation and interoperability.

After Step 3: Milestones and Takeaways. This methodology can be applied to different languages, allowing the expansion of research and application of the results obtained. By repeating these three steps for different languages or language varieties, it is possible to extend the application of the hybrid methodology and advance research in a wide range of language contexts with scarce resources. This cycle helps to create a sustainable data ecosystem and improve linguistic knowledge for under-resourced languages.

3 Tunisian Arabic State-of-the-Art

This section presents the state-of-the-art of digital and non-digital resources available for TA, the subject of our pilot study.

¹²See the subsections 5.1 and 5.2 for more information about the root level.

¹³Resource Description Framework (RDF) is a standard model for data interchange on the web. It allows for the integration of various sources with different structures and makes it easier for machines to understand the semantics of the information. Lemon (Lexicon Model for Ontologies) is a model based on RDF and designed for representing lexical information relative to ontologies. It allows for the representation of a wide range of linguistic structures necessary for the development of NLP applications. <https://www.w3.org/2016/05/ontolex/>.

3.1 Available Non-digital Resources

As mentioned above, dealing with Arabic dialects means having access to a very limited number of written sources. In fact, mainly for identity reasons, Arab speakers normally have a strong hierarchical perception of the languages they speak: on the one hand, Standard and Koranic Arabic represent the high register of the language, used in written texts and in formal and non-spontaneous situations; on the other hand, dialect is perceived as a lower register, sometimes even vulgar, and it is the language of everyday life, spontaneity and orality (Boussofara-Omar, 2006). From this, it clearly follows that, over the centuries, the documents which had to be preserved and which deserved the written form, were essentially composed in the highest register of the diglottic *continuum*, i.e. in Koranic/ Standard/ Literary Arabic. However, Arabs have always had the local dialect as native language, and have always expressed themselves orally in this variety. As a consequence, there are very few written sources that report ancient dialect lexicon, linguistic traces of which are mostly found in the phenomena of loan and interference and in Middle Arabic (an intermediate variety product of the interference of the Modern Standard Arabic (MSA) and the CA¹⁴). In short, this means that as far as Arabic dialects are concerned, and specifically TA, it is virtually impossible to have access to primary sources prior to the 21st century. It was only in the contemporary era that Arabic dialects started to be used in digital informal communication (Caubet, 2019), providing the first appearance of sizable linguistic data of CA. However, evidences of a previous linguistic stage is found in dialectological studies, mostly performed by European researchers, starting from the 19th century. Among them, there are the works included in the lexicographic database which will be described extensively in Section 4.1. To cite some of the works that can be considered sources of TA lexicon prior to the current period, we can mention pioneering studies such as the Maghrebi (i.e. North African) Arabic dictionary by Beaussier et al. (2006), the TA grammar and glossary by Stumme (1896) and the impressive description of Takrouna's Arabic by Marçais (1961). It is also necessary to mention dictionaries and manuals

¹⁴Middle Arabic is described in more detail by Lentin (2008, 216) as 'the language of numerous Arabic texts distinguished by its linguistically (and therefore stylistically) mixed nature, as it combines standard and colloquial features with others of a third type, neither standard nor colloquial'.

for French students published in the early 20th century (such as, for example, the works of Nicolas ((s.d.); Jourdan (1913)). These pioneering studies represent almost the only evidence of linguistic stage that otherwise would have been forgotten. But precisely because they are forerunners, all these studies present various problems: e.g. it is sometimes not clear which linguistic variety they refer to and they do not always use accurate transcriptions of CA phonetics. For this reason, it is necessary to compare them with further sources: more recent and accurate dialectological studies (such as Behnstedt (1998, 1999); Ritt-Benmimoun (2014)), manuals for foreign students published in recent years (such as: Ben Ammar and Vacchiani (2016); Durand and Tarquini (2023)) but also, and above all, with primary sources, i.e. interviews on field and authentic exchanges in social networks.

3.2 Available Digital Resources

Concerning digital platforms for dictionaries or lexicons of TA, to the best of our knowledge, there are only the *Linguistic dynamics in the Greater Tunis Area: a corpus-based approach* (TUNICO) (Dallaji et al., 2020) and the *Tunisian Arabic Corpus* (TAC) (McNeil, 2018).¹⁵ The first makes available through a digital platform a Tunisian dictionary and a corpus of data associated with accurate linguistic information. TUNICO data are encoded in a Latin-based transcription and can be searched using a search bar. Instead, TAC collects raw texts, encoded in not-normalised Arabic script. TAC texts can be observed by search queries based on three different systems: *Exact*, *Stem*, and *Regex*. The first two require an Arabic-encoded input, while the third one requires the users to transliterate the input by following a modified version of the Buckwalter transliteration system.¹⁶ These tools are useful for language analysis, although they present some difficulties in their use. With regard to the processing and the study of CA in the NLP field, there is a trend in recent years to produce a multitude of CA corpora that has allowed for progress in the study of CAs. In the case of TA, among the various recently released corpora we can mention a corpus of Facebook comments, manually annotated for sentiment analysis

¹⁵See also: <https://www.livelingua.com/arabic/courses/tunisian> and <https://derja.ninja/>.

¹⁶Further information on TAC query system at page: <https://www.tunisiya.org/help/>. Buckwalter transliteration system at <http://www.qamus.org/transliteration.htm>.

(TSAC) (Mdhaaffar et al., 2017) and a parallel corpus of TA-MSA, the TD-COM corpus, extracted from social networks (Kchaou et al., 2022).¹⁷ Another downloadable corpus for TA is the Tunisian Arabizi Corpus (TArC), released by Gugliotta and Dinarelli (2022) and described in Section 4.2. Finally, we should mention some multi-dialectal resources that include TA among other CA varieties. One of these is PADIC (Meftouh et al., 2018), a parallel corpus of six CAs. Another one is MADAR (Bouamor et al., 2014), which consists of a parallel corpus of the CA of 25 Arab cities, including cities of Tunisia (Tunis and Sfax). The same corpus has recently been released in CODA orthography (Habash et al., 2018) by Eryani et al. (2020).

Although a number of corpora have been produced, TA is still considered an under-resourced language. It is possible that the solution to the complexity of CA (morphological and orthographic, due to the absence of standards and a situation of multilingualism, diglossia, etc.), does not lie solely in the amount of data, processed according to universally valid methodologies for all languages. As a very simple example, each of the mentioned resources was created for a specific purpose and consequently represents a portion of the linguistic reality of TA. These are indeed valuable resources, but not sufficient for a complete mapping of this language. Moreover, each resource, including TUNICO and TAC, presents its own language encoding system, based on Latin or Arabic script. Perhaps there is a need to develop a methodology suited to the case of under-resourced languages and thus aim more than ever to preserve data quality. In the next section, we will explain how our contribution attempts to investigate this possibility.

4 Linguistic Resources Description

4.1 The TA Lexicographic Database

TA is a rich and composite language, which fully reflects the history and culture of a country located in the center of southern Mediterranean coast, known since ancient times as a land of human as well as linguistic passage and exchange (Marçais, 1950; Baccouche, 2009). TA has a varied lexical composition, due to the coexistence of a main Arabic linguistic stratum (Hilali, pre-Hilali and Classical Arabic); adstrate languages (such

¹⁷Other resources, released by the same Arabic NLP group, are available at <https://sites.google.com/site/anlprg/corpora-corpus?authuser=0>.

as Berber, Punic, Greek, and Latin) and many superstrate languages (such as Spanish, Lingua Franca¹⁸, Turkish, Italian, French and English).¹⁹ In addition, all these elements are combined with diglossia (with Standard Arabic) and bilingualism (with French).²⁰ In order to record at least a part of the lexical richness of TA and attempt linguistic analysis, it was first of all necessary to create a tool for registering the lexicon available in the TA bibliographic sources: this tool is the TA lexicographic database (Panasci, 2021), consisting of 13,800 headwords and 5,600 Arabic roots and focused on diachronic and diatopic variation in the TA lexicon. To date, the database collects all the lexical entries of ten glossaries, two papers and three dictionaries²¹ representing about a century and a half of Tunisian linguistic history and various local dialects. The oldest source is in fact a grammar written in 1896 (Stumme) and the most recent one is a 2017 paper on Tunis jargon (Labidi). Moreover, the database contains dialects representative of various areas of the country, such as the dialect of the capital, Tunis (Ben Ammar and Vacchiani, 2016), that of a coastal city such as Susa (Talmoudi, 1981), or a Bedouin dialect of the South of the country, such as that of the Marazig tribe (Boris, 1958). To build the lexicographic database, all headwords have been translated into Italian and they have been marked with an abbreviation designating the reference source of the entry. The individual words referring to a specific meaning were compared with each other, adopting a criterion that highlighted the diachronic evolution of the language (that is, an insertion of the occurrences in the sources from the oldest to the most modern). To make the material more enjoyable for the reader, it has been organized in the structure of an Italian-TA dictionary, i.e. with the entries inserted in alphabetical order, as well as in the structure of a TA-Italian dictionary, i.e. according to the traditional Arabic language setting of radical letters. Finally, the database entries present additional information (when available): etymology of the word, diatopic collocation,

semantic shifts, obsolescences, linguistic register, etc. Below are two examples of entries, the first one taken from the Italian-Tunisian database, the second one from the Tunisian-Italian database.

Camaleonte s.m. *omm əl-buʔa* AN11/ *umm əl-būyya* JJ13; *bu keššēš* GB58; *bu hremba* [dim. *bu hrēmba*] GB58; (Mağārba, ai confini tra Tripolitania e Cirenaica e Warfella, a Ovest dei Mağārba) *herba* GB58; (Wargemma, confederazione tribale tra Gabès e Médénine, e Rbāye⁴, nomadi della zona del Oued Šūf) *herbēya* GB58; *tata* MQ2002

حناك *hank* [pl. *hnāk*] MH77 **palato**; (pan-maghrebino) *hanek* [pl. *aḥnāk-*; + art. *l-hank/ laḥnek*, pl. *laḥnāk-*; + pron. suff. *hanki*] JQ61/ (pan-maghrebino) *hnak* [pl. *aḥnāk-*; + art. *l-hank/ laḥnek*, pl. *laḥnāk-*; + pron. suff. *hanki*] JQ61/ *hank* [pl. *hnāk*] MH77 **mascella (umana); guancia**; *h^anek* [pl. *aḥnēk-*; + pron. suff. I pers. sing. *hen^eki*; + pron. suff. III pers. f. sing. *h^anekha*] GB58 **mandibola** – *daqq əl-hank* JQ62 **idiom. parlare di futilità; straparlare; hanka** JQ62 **esperienza di vita; maḥannek** JQ62 **espetto**

Figure 2: TA Lexicographic Database Sample

Figure 2 shows how the database works. In the first case, all the occurrences for the meaning of "chameleon" in the various sources are reported. The entries are followed by the reference abbreviation (e.g. AN11 represents (Nicolas, (s.d.)) and they are in chronological order. The diatopic variation is highlighted (e.g. the lexical variants for the term in the different tribes of southern Tunisia are specified). In the second case, instead, all the occurrences found in the sources for the Arabic root *hnk* are reported. The order of appearance of the terms is the traditional one of Arabic dictionaries (first the ten forms of the verb appear, then the nouns, etc.). In this case the geographical location of a term (the word for "jaw" or "cheek") is highlighted and an example of an idiomatic expression is given.

4.2 Tunisian Arabizi Corpus (TArC)

TArC gathers texts from various informal digital writing contexts, such as blogs, forums, and Facebook, including rap song lyrics shared on dedicated forums. The collection of these texts aims to investigate Arabizi, a Latin script encoding used in informal online communication. Additionally, the inclusion of rap song lyrics allows for a comparative analysis of both the Arabic and Latin script encoding systems in TA.²² Together with the texts, were publicly available, also some metadata of the authors

²²TArC data are available at <https://github.com/eligliotta/tarc>.

¹⁸With Lingua Franca we refer to the Italian-based pidgin spoken in the regencies of Tunis, Tripoli and Algiers during the Ottoman rule (Cifoletti, 2004).

¹⁹See: Baccouche (1994).

²⁰See Daoud (2007).

²¹The TA lexicographic database sources include Ben Abdelkader et al. (1977); Ben Alaya and Quitout (2010); Ben Ammar and Vacchiani (2016); Bevacqua (2008); Boris (1958); Jourdan (1913); Labidi (2017); Marçais and Hamrouni (1977); Nicolas ((s.d.); Quéméneur (1961a,b, 1962); Quitout (2002); Stumme (1896); Talmoudi (1981).

of texts were collected. These are their provenience, age-range and gender (Gugliotta, 2022).

TArC data have been semi-automatically annotated with various linguistic information at word-level, by means of a neural Multi-Task Architecture (MTA) (Gugliotta et al., 2020).²³ These annotation levels are shown in Table 2 and consist of text normalisation into CODA-*Star* orthography in Arabic script (Habash et al., 2018), sub-tokenisation, POS-tagging and lemmatisation. To avoid transliterating code-switching into Arabic script, the initial annotation level of TArC data is token classification, which, as shown in Table 1, consists of three classes: *Foreign*, *Arabizi* and *Emotag*. The *Emotag* class encompasses para-textual elements like emoticons and smileys that are not intended for transliteration. Only the tokens classified as *Arabizi* have been annotated with the linguistic information. The formalism employed for Part-of-Speech tagging is the one of the Penn Arabic Treebank (Maamouri et al., 2004), while lemmas are also encoded in CODA-*Star*. Below we report some information on TArC data.

Token Class	TArC - Total of Lemmas: 5,063				
	Blogs	Forums	Facebook	Rap	Total
<i>Arabizi</i>	5,978	6,026	11,833	7,680	31,517
<i>Foreign</i>	707	5,873	3,624	1,010	11,214
<i>Emotag</i>	7	10	600	1	618
Tokens	6,692	11,909	16,057	8,691	43,349
Sentences	366	755	3,162	515	4,798

Table 1: *The Tunisian Arabizi Corpus*

5 Resources Integration

The two linguistic tools described in the previous section, despite having the same variety of CA as their subject, namely TA, are very different. It is precisely in their diversity that their complementarity and the usefulness of their combination lies. In fact, the lexicographic database was created to observe the variation of TA at the diachronic and diatopic level, thus, it mainly collects lemmas through secondary sources. Instead, TArC collects authentic texts encoded in a non-standardised writing system, known as Arabizi. This is shown in Example 1, where the first line consists of the original text in Arabizi encoding; the second line is the transcription of the oral reconstruction of the same sentence; and the third line is its translation. This sentence, in TArC is provided with the annotation levels shown in Table

²³This is available at <https://gricad-gitlab.univ-grenoble-alpes.fr/dinarelm/tarc-multi-task-system>.

2, where the sentence is reported in Arabic script (normalisation in CODA-*Star*), in the first column. In the following columns, we can observe how the sentence has been processed at the sub-tokenisation, POS-tagging and lemmatisation levels.

- (1) *Tdaweb zebda wtzidha lil farina*
/t-ðaw:əb əz-zəbda w-t-zīd-hā l-əl fārīna/
'Melt the butter and mix it with the flour'.

CODA	Tokeniz.	POS	Lemma
تذوّب	تذوّب	CV2S-CV	ذوّب
الزبدة	الزبدة	DET+NOUN-NSUFF_FEM_SG	زبدة
وتزیدها	وتزیدها	CONJ+CV2S-CV+	زاد
		CVSUFF_DO:3FS	
لال	لهال	PREP+DET	ل
فارينة	فارينة	NOUN-NSUFF_FEM_SG	فارينة

Table 2: TArC Annotation Levels

The lexicographic database provides specific information about individual entries (always in the lemmatic form): diatopic and diachronic variation, etymology, semantic changes, etc. In order to give an excerpt of them, we report in the following example, the information collected at the voice /fārīna/ 'flour'.²⁴

- (2) **Flour** s.f. [< ita. or lingua franca *farina*] *fērīna* HS1896/ (2) [coll. *fērīnē*] BAR77/ *farina* AW2010/ *fērīna* AV2016; *dqīq* AN11/ *dqīq* JJ13/ *dqīq* MH77/ *dqēq* MH77 – fine flour *degīq* GB58 – flour (probably of soft wheat) purchased already ground *fārīnē* GB58 – idiom. “add water, add flour...” (phrase to be used during an anecdotal narrative, signifying that it was a never-ending enterprise) *zīd əl-ma zīd əd-dqīq* MH77

From Example 2, it is possible to see how, from the nineteenth century to the present, the concept is mainly expressed by a loanword from Italian (Stumme, 1896) or from the Lingua Franca (Cifoletti, 2004, 234): *fārīna*. The loanword would seem to have supplanted the Arabic *dqīq*, although we find the latter in some of the database's supplies, both with the meaning of 'flour' and 'fine flour'.

²⁴The abbreviations in order are: HS1896: (Stumme, 1896); BAR77: (Ben Abdelkader et al., 1977); AW2010: (Ben Alaya and Quitout, 2010); AV2016: (Ben Ammar and Vacchiani, 2016); AN11: (Nicolas, (s.d.); JJ13: (Jourdan, 1913); MH77: (Marçais and Hamrouni, 1977); GB58: (Boris, 1958).

The database thus allows hypotheses to be made: most likely the two terms must have coexisted for a long time (Stumme in the late 19th century recorded *fārīna* for Tunis; Nicolas and Jourdan in the early 20th century reported only *dqīq*), perhaps as diatopic variants or perhaps with specialization of meaning, as was the case in the 1950s in Marazig speech,²⁵ in which *fārīna* was merely the product of soft wheat already ground, and as reconstructed by Cifoletti (1998, 152) for Tunis, where with the entry of the loanword into common parlance, *dqīq* came to mean ‘semolina’. Finally, the database (MH77: (Marçais and Hamrouni, 1977)) provides an idiomatic expression related to the concept of ‘flour’: *zīd əl-ma zīd əd-dqīq*.

From these examples, we can clearly see how the integration of these two resources can yield a tool that is unique in its completeness. In fact, together they can provide lexicographic, etymological, diachronic and diatopic information plus examples from real native usage occurrences and morpho-syntactic information of such sentences. In the following section, we explain how we were able to link the information of these tools.

5.1 Analysis and Conversion of Lexicographic Data structure

In the context of this research project focused on the management of under-resourced Arabic dialects, we elected to devise and implement a scraping tool specifically designed to delve into a dictionary’s intricacies, extract pertinent data, and utilize this information for subsequent linguistic analyses and potential cross-referencing with other linguistic data sets. This decision stemmed from the realization of the untapped potential housed within these lexicographic structures, often layered and dense with information but largely inaccessible due to their static presentation. To accomplish this ambitious task, we deployed a carefully constructed script that meticulously parsed the dictionary, illuminating its structure on an entry-by-entry basis. The cornerstone of our process was a .docx file, the format of the lexicographic database. The document was formatted according to specific standards that allowed us to codify a system of rules for data extraction, rules contingent on the elements’ location within each entry. The algorithm’s cornerstone was the identification and extraction of the Italian definition

²⁵Recorded in the dictionary of Boris (1958), corresponding to the abbreviation GB58.

within each entry, typically represented as a distinct bold string. Once this key piece of information was located, the algorithm triggered a systematic reverse sequence search designed to uncover other elements. This exploratory process, proceeding backwards from the definition, focused on locating: 1) the source reference indicating the individual or group responsible for proposing the hypothesis; 2) any enclosed morphological information presented within square brackets (see Figure 2). This could include TA variants trailed by morpho-syntactic data such as part-of-speech and further grammatical information; 3) As shown in Figure 2, the TA lemma tethered to the root, which is encoded in Arabic characters. Instead, the lemma, a central component of each entry, is rendered in italics with specific unicode characters. Furthermore, it’s noteworthy that multiple variants can be linked to a single semantic interpretation within this structure. Upon extraction, the raw data underwent a transformation process designed to adapt it into a data structure capable of reflecting the inherent relationship and interlinking between disparate elements dispersed across the corpus. This was a vital aspect of the project as we frequently encountered references to other dictionary entries and cross-references that needed to be retained to maintain the richness of the dataset. Given the nature of

```
{
  "root": "شلشل",
  "definitions": [
    {
      "meaning": "casco spogliato della
maggior parte dei suoi datteri",
      "occurrences": [
        {
          "lemma": "šəɫšūɫ",
          "source": "GB58",
          "variations": [],
          "additional_data": [
            {
              "text": "pl. šalāšīl"
            }
          ]
        }
      ]
    }
  ],
  "examples": [
    {
      "tun": "wəgəθa kunət sēreḥ ‘ala
šəɫšūɫ, hāk el‘əbse",
      "source": "GB58",
      "ita": "idiom. all’epoca ero pastore
per il conte di Šalšūɫ, quel tirchio (modo di
dire per designare un avaro);"
    }
  ]
},
  "references": []
}
```

Figure 3: A TA dictionary entry encoded in JSON

the source document and the complexities involved in the extraction process, it was inevitable that we would encounter a certain degree of noise within the data. This noise could manifest as characters not belonging to the target alphabet, misplaced punctuation marks, or other elements that deviated from the expected data type. To address these issues, we developed a series of rules using regular expressions, specifically designed to identify and control such anomalies, effectively cleansing the dataset.

The result of this comprehensive process was a script capable of extracting a substantial volume of data from the source dictionary. Nevertheless, we acknowledge that a completely automated process remains elusive due to the possibility of errors and irregularities inherent in the data. Consequently, a degree of manual data cleansing is still necessary. For instance, it's not uncommon to encounter text segments belonging to another lemma embedded within a definition, a complication arising from inconsistencies in formatting. While our script currently lacks the functionality to extract or classify morpho-syntactic categories or the etymological and additional information often found within dictionary entries, we view these as areas for future development rather than limitations. We are actively working on enhancements designed to incorporate these elements into the script, thereby adding another layer of richness to the extracted data. As we continue to refine and develop this tool, our focus is shifting toward addressing the broader challenges associated with data extraction for the creation of accessible and interoperable lexical resources. This ongoing endeavor aligns with our commitment to the FAIR principles. By enhancing our capacity to extract and utilize the rich data contained within lexicographic resources, we believe we can significantly contribute to the field of under-resourced language studies.

5.2 Corpus Annotation extension

Considering the different encoding employed for the level of lemmatisation of the two tools (scientific transliteration for the lexicographic database and normalisation in CODA-Star for TArC), we discarded lemmas as a common key between the two tools to be put into communication. Since, on the other hand, the lexicographic database is provided with an annotation level of the root from which the recorded lemma is derived, it decided to use the root as the first key element for joining the linguistic

tools. To produce this additional annotation layer, we investigated the functionality of the CAMEL Tools (Obeid et al., 2020).²⁶ This is a suite of Arabic NLP tools, such as lemmatisers, tokenisers and POS-tagger, and provides also roots. However, among the databases provided with CAMEL Tools (MSA, Egyptian Arabic and Gulf Arabic databases), only the database for the MSA, according to our tests, provides roots. Annotating the Tunisian Arabizi data, collected in TArC, with an MSA database, clearly assumes difficulties in identifying tokens. However, as shown in Table 3, the results were not unsatisfactory, in terms of quality. This is mainly because TArC has been normalised to CODA-Star, an Arabic character encoding, MSA-like. In fact, as input to the Camel morphology analyser, we provided the lemma annotation level of each TArC token, by excluding the tokens classified as *foreign* and *emotag*, and the tokens POS-tagged as punctuation (*PUNC*), numerals (*NOUN_NUM*) or proper nouns (*NOUN_PROP*). The excluded tokens amount to 9,363 tokens, thus, the total of lemmas provided to the Camel analyser was 33,986.²⁷ In Table 3 we report the results of Camel Tools on TArC data.

<i>Total of TArC token provided: 33,986</i>		
<i>Not Found</i>	Wrong Annotation	Correct Annotation
4,017	6,056	23,913

Table 3: Results of CAMEL Tools on TArC data

The table shows that 4,017 tokens were not recognised at all by the analyser (*Not Found* in Table 3). In some other cases (**Wrong Annotation**), the morphological analyser provided a root instead, based on MSA, but this was incorrect in the case of TA, as shown in Example 2. These cases amount to 6,056. The cases of **Correct Annotation**, on the other hand, amount to 23,913.

- (3) *al boulis*
 āl- būlis
 بيليس ال fr:police [Camel root]
 بيليس ال fr:police [Correct root]
 ‘The policeman’.

Considering the linking functionality envisaged for this level of TArC annotation, while manually

²⁶These are available at https://github.com/CAMEL-Lab/camel_tools.

²⁷As shown in Table 1, the total tokens of TArC are 43,349. These correspond to an amount of 5,063 unique, non-repeated, lemmas.

validating the roots automatically generated, we took some decisions based on the lexicographic database characteristics. When a lemma results from the combination of different words (as in the case of *blāš*, ‘without’, which is the fusion of *b-*, *lā* and *šy?*), the database records the TA lemma both as it is (*blāš*) and pointing to its components. Therefore, by validating TARc roots, we left these tokens as they are, instead of reducing them to their etymological components.

Finally, after the manual correction and integration of the *Not Found* and **Wrong Annotation** occurrences, respectively, the number of unique roots in TARc amounts to 1356.²⁸ The 76.3% of these (1034 unique roots) are matching with the lexicographic database roots.

6 Conclusions and Future Work

In this paper, we described our work on linking two linguistic tools previously created for different purposes. This work concerns Tunisian Arabic, and the resources we are working on are a large lexicographic database and a corpus of informal written texts from digital contexts. We explained the characteristics of these linguistic tools and how we managed to link them by enhancing their content. The work described is an ongoing pilot study, part of a larger project involving the development of resources for under-resourced languages. We described the methodology we developed for these types of languages. We outlined how this methodology adheres to the principles of open science, emphasizing transparency, interoperability and accessibility of data. Our project involves the use of existing language resources using tools, corpora, glossaries and dictionaries freely available among the scientific community. We deal with standardisation and morpho-syntactic annotation of texts with NLP tools. These ensure the reproducibility of our methodology. By sharing both the annotated data and the tools we create, other researchers will benefit from our work. The annotated data will be made available through a freely accessible digital platform. The NLP tools will be released on a repository dedicated to the project. Overall, the work described is in line with the FAIR principles, facilitating open and effective research on under-resourced languages.

²⁸For *unique root*, we mean the roots counted only once.

References

- Soha Abboud-Haggar. 2006. Dialects: Genesis. In *Encyclopedia of Arabic Language and Linguistics*, volume I, pages 613–622. Brill, Leiden – Boston.
- Diana Alhafi, Anton Deik, Elhadj Benkhelifa, and Mustafa Jarrar. 2019. *Usability Evaluation of Lexicographic e-services*. In *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*. IEEE.
- Taïeb Baccouche. 1994. *L'emprunt en arabe moderne*. Académie tunisienne des sciences des lettres et des arts, Beït al-Hikma.
- Taïeb Baccouche. 2009. Tunisia. In *Encyclopedia of Arabic Language and Linguistics*, volume IV, pages 571–577. Brill, Leiden – Boston.
- Marcelin Beaussier, Mohamed Ben Cheneb, and Albert Lentin. 2006. *Dictionnaire Pratique Arabe-Français (Arabe Maghrébin)*. Ibis Press, Paris.
- Peter Behnstedt. 1998. Zum Arabischen von Djerba (Tunesien) I. *Zeitschrift für Arabische Linguistik*, 35, pages 52–83.
- Peter Behnstedt. 1999. Zum Arabischen von Djerba (Tunesien) II: Texte. *Zeitschrift für Arabische Linguistik*, 36, pages 32–65.
- Rached Ben Abdelkader et al. 1977. *Peace Corps English-Tunisian Arabic Dictionary*. Peace Corps, Washington D.C.
- Wahid Ben Alaya and Michel Quitout. 2010. *L'Arabe tunisien de poche – Guide de conversation*. Assimil France, Chennevières sur Marne Cedex.
- Hager Ben Ammar and Valérie Vacchiani. 2016. *Parler tunisien fissa! Une méthode originale pour apprendre l'arabe tunisien en 6 mois*. Editions Arabesques, Tunis.
- Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A Neural Probabilistic Language Model. *The Journal of Machine Learning Research*, 3:1137–1155.
- Massimo Bevacqua. 2008. Osservazioni sul linguaggio dei giovani tunisini. *Il filo di seta – Studi arabo-islamici in onore di Wasim Dahmash*, pages 11–24.
- Gilbert Boris. 1958. *Lexique du parler arabe des Marazig*. Imprimerie nationale – Librairie C. Klincksieck, Paris.
- Houda Bouamor, Nizar Habash, and Kemal Oflazer. 2014. A Multidialectal Parallel Corpus of Arabic. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC)*, pages 1240–1245.
- Naima Boussofara-Omar. 2006. Diglossia. In *Encyclopedia of Arabic Language and Linguistics*, volume I, pages 629–637. Brill, Leiden – Boston.

- Abdelghani Bouziane, Djelloul Bouchiha, and Nouredine Doumi. 2020. Annotating Arabic texts with linked data. In *2020 4th International Symposium on Informatics and its Applications (ISIA)*, pages 1–5. IEEE.
- Dominique Caubet. 2019. Vers une littérature numérique pour la darija au maroc, une démarche collective. In Catherine Miller, Alexandrine Barontini, Marie-Aimée Germanos, Jairo Guerrero Guerrero, and Christophe Pereira, editors, *Studies on Arabic Dialectology and Sociolinguistics. Proceedings of the 12th International Conference of AIDA*. Livres de l'IREMAM.
- Guido Cifoletti. 1998. Osservazioni sugli italianismi nel dialetto di Tunisi. *Incontri linguistici*, 21:137–153.
- Guido Cifoletti. 2004. *La lingua franca barbaresca*. Il Calamo.
- Ryan Cotterell and Hinrich Schütze. 2015. **Morphological Word-Embeddings**. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1287–1292, Denver, Colorado. Association for Computational Linguistics.
- Ines Dallaji, Ines Gabsi, Stephan Procházka, and Karlheinz Mörth. 2020. A Digital Dictionary of Tunis Arabic-TUNICO (ELEXIS). *Slovenian language resource repository CLARIN.SI*.
- Mohamed Daoud. 2007. The Language Situation in Tunisia. *Language planning and policy in Africa, Vol. II: Algeria, Côte d'Ivoire, Nigeria and Tunisia*, pages 256–308.
- Kareem Darwish, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Husein T Al-Natsheh, Houda Bouamor, Karim Bouzoubaa, Violetta Cavalli-Sforza, Samhaa R El-Beltagy, Wassim El-Hajj, et al. 2021. A panoramic survey of natural language processing in the arab world. *Communications of the ACM*, 64(4):72–81.
- Humphrey Davies. 2006. Dialect Literature. In *Encyclopedia of Arabic Language and Linguistics*, volume I, pages 597–604. Brill, Leiden – Boston.
- FMG De Jong, Bente Maegaard, Koenraad De Smedt, Darja Fišer, Dieter Van Uytvanck, et al. 2018. Clarin: Towards FAIR and responsible data science using language resources. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3259–3264.
- Olivier Durand and Maura Tarquini. 2023. *Corso di arabo tunisino. Manuale di comunicazione con grammatica ed esercizi. Livelli A1-B2 del Quadro Comune Europeo di Riferimento per le Lingue*. Ulrico Hoepli Editore S.p.A., Milano.
- Fadhl Eryani, Nizar Habash, Houda Bouamor, and Salam Khalifa. 2020. A spelling correction corpus for multiple Arabic dialects. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4130–4138.
- Charles A. Ferguson. 1959. Diglossia. *Word*, 15(2):325–340.
- Darja Fišer and Andreas Witt. 2022. *CLARIN: The Infrastructure for Language Resources*, volume 1. Walter de Gruyter GmbH & Co KG.
- Elisa Gugliotta. 2022. *Realization of a Tunisian Arabish Corpus with use within the scope of NLP-Natural Language Processing*. Ph.D. thesis, Sapienza University of Rome and Université Grenoble Alpes.
- Elisa Gugliotta and Marco Dinarelli. 2022. Tarc: Tunisian arabish corpus first complete release. In *13th Conference on Language Resources and Evaluation (LREC 2022)*.
- Elisa Gugliotta and Marco Dinarelli. 2023. An empirical analysis of task relations in the multi-task annotation of an arabizi corpus. *Accepted paper for the 4th Conference on Language, Data and Knowledge*.
- Elisa Gugliotta, Marco Dinarelli, and Olivier Kraif. 2020. Multi-task sequence prediction for Tunisian arabizi multi-level annotation. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 178–191.
- Nizar Habash, Fadhl Eryani, Salam Khalifa, Owen Rambow, Dana Abdulrahim, Alexander Erdmann, Reem Faraj, Wajdi Zaghoulani, Houda Bouamor, Nasser Zalmout, et al. 2018. Unified guidelines and resources for Arabic dialect orthography. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC)*.
- Mustafa Jarrar and Hamzeh Amayreh. 2019. **An Arabic-Multilingual Database with a Lexicographic Search Engine**. In *Natural Language Processing and Information Systems*, pages 234–246. Springer International Publishing.
- Mustafa Jarrar, Fadi A. Zaraket, Tymaa Hammouda, Daanish Masood Alavi, and Martin Waahlsch. 2022. Lisan: Yemeni, Iraqi, Libyan, and Sudanese Arabic dialect corpora with morphological annotations. *arXiv preprint arXiv:2212.06468*.
- J. Jourdan. 1913. *Cours normal et pratique d'Arabe Parlé – Vocabulaire – Historiettes – Proverbes – Chants – Dialecte Tunisien, 4e édition*. Éditions Bouslama, Tunis.
- Saméh Kchaou, Rahma Boujelbane, and Lamia Hadrich Belguith. 2022. Hybrid pipeline for building Arabic Tunisian dialect-standard Arabic neural machine translation model from scratch. *ACM Transactions on Asian and Low-Resource Language Information Processing*.
- Emna Labidi. 2017. L'artisanat traditionnel à Tunis – Sa terminologie et son lexique. *Tunisian and Libyan Arabic Dialects – Common Trends – Recent Developments – Diachronic Aspects*, pages 147–160.

- Jérôme Lentin. 2008. Middle Arabic. In *Encyclopedia of Arabic Language and Linguistics*, volume III, pages 215–224. Brill, Leiden – Boston.
- Mohamed Maamouri, Ann Bies, Tim Buckwalter, and Wigdan Mekki. 2004. The Penn Arabic Treebank: Building a large-scale annotated Arabic corpus. In *NEMLAR conference on Arabic language resources and tools*, volume 27, pages 466–467. Cairo.
- Agnese Macchiarelli. 2023. Sinergie fra vedph e cnr-ilc in termini di condivisione della conoscenza e sostenibilità dei progetti digitali. In *DH. 22–Digital Humanities. Per un confronto interdisciplinare tra saperi umanistici a 30 anni dalla nascita del World Wide Web*. L’Erma di Bretschneider.
- Aberrahmân Marçais, William Guîga. 1961. *Textes arabes de Takroûna par William Marçais et Abderrahmân Guîga – II – Glossaire – Contribution à l’étude du vocabulaire arabe, Tome I – VIII*. Imprimerie Nationale – Centre National de la Recherche Scientifique – Librairie Orientaliste Paul Geuthner, Paris.
- Philippe Marçais and M.-S. Hamrouni. 1977. *Textes d’arabe maghrébin*. Librairie d’Amérique et d’Orient, Adrien Maisonneuve – J. Maisonneuve, succ., Paris.
- William Marçais. 1950. Les parlers arabes. *Initiation à la Tunisie*, pages 195–219.
- Karen McNeil. 2018. Tunisian Arabic corpus: Creating a written corpus of an ‘unwritten’ language. *Arabic corpus linguistics*, 30.
- Salima Mdhaffar, Fethi Bougares, Yannick Esteve, and Lamia Hadrich-Belguith. 2017. Sentiment analysis of Tunisian dialects: Linguistic resources and experiments. In *Third Arabic Natural Language Processing Workshop (WANLP)*, pages 55–61.
- Karima Meftouh, Salima Harrat, and Kamel Smaïli. 2018. Padic: Extension and new experiments. In *7th International Conference on Advanced Technologies ICAT*.
- Ouafae Nahli, Elisa Gugliotta, Nadia Khelif, and Giulia Benotto. 2023. Advancing dialectal analysis: Annotating corpora and building lexical resources for Arabic dialects. *Forthcoming*.
- Alfred Nicolas. (s.d.) [1911]. *Dictionnaire Français-Arabe, Idiome Tunisien*. Imprimeur - Éditeur Frédéric Weber, Tunis.
- Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadh Eryani, Alexander Erdmann, and Nizar Habash. 2020. *CAMEL tools: An open source python toolkit for Arabic natural language processing*. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France. European Language Resources Association.
- Jonathan Owens. 2006. *A Linguistic History of Arabic*. Oxford: Oxford University Press.
- Livia Panasci. 2021. *Studi lessicali sull’arabo di Tunisia*, volume I, II, III, IV. PhD Thesis in Civilizations of Asia and Africa, Sapienza University of Rome, Rome.
- Laurette Pretorius and Claudia Soria. 2017. Introduction to the Special Issue. *Language resources and evaluation*, 51:891–895.
- Michel Quitout. 2002. *Parlons l’arabe tunisien: Langue & culture*. L’Harmattan, Paris - Budapest - Torino.
- Jean Quéméneur. 1961a. Notes sur quelques vocables du parler tunisien figurant au Supplement de A. Lentin - 1ère partie. *Revue de l’I.B.L.A.*, Vol. 24/93, pages 1–22.
- Jean Quéméneur. 1961b. Notes sur quelques vocables du parler tunisien figurant au Supplement de A. Lentin - 2ème partie. *Revue de l’I.B.L.A.*, Vol. 24/94, pages 167–181.
- Jean Quéméneur. 1962. Glossaire de dialectal 1942-1962. *Revue de l’I.B.L.A.*, Vol. 25/100, pages 325–367.
- Veronika Ritt-Benmimoun. 2014. *Grammatik des arabischen Beduinendialekts der Region Douz (Südtunesien)*. Harrassowitz Verlag, Wiesbaden.
- Vitaly Romanov and Albina Khusainova. 2019. Evaluation of Morphological Embeddings for the Russian Language. In *Proceedings of the 2019 3rd International Conference on Natural Language Processing and Information Retrieval*, pages 144–148.
- Rana Aref Salama, Abdou Youssef, and Aly Fahmy. 2018. Morphological word embedding for Arabic. *Procedia computer science*, 142:83–93.
- Lotfi Sayahi. 2014. *Diglossia and language contact: Language variation and change in North Africa*. Cambridge University Press.
- Erhan Sezerer and Selma Tekir. 2021. A Survey on Neural Word Embeddings. *arXiv preprint arXiv:2110.01804*.
- Hans Stumme. 1896. *Grammatik des tunisischen arabisch: nebst Glossar*. JC Hinrichs.
- Fathi Talmoudi. 1981. *Texts in the Arabic Dialect of Sūsa (Tunisia): Transcription, Translation, Notes and Glossary*. Acta Universitatis Gothoburgensis, Göteborg.
- Kees Versteegh. 2014. *Arabic language*. Edinburgh University Press.
- Mark Wilkinson, Michel Dumontier, I. Jsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, et al. 2016. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific data*, 3(1):1–9.
- Sane Yagi, Ashraf Elnagar, and Shehdeh Fareh. 2022. A Benchmark for Evaluating Arabic Word Embedding Models. *Natural Language Engineering*, pages 1–26.

Bridging Corpora: Creating learner pathway across texts

Hugh Paterson III

University of North Texas / Denton, Texas
 University of Oregon / Eugene, Oregon
 Drexel University / Philadelphia, Pennsylvania
 i@hp3.me

Bret Mulligan and Anna Lacy and Patricia Guardiola

Haverford College / Haverford, Pennsylvania
 bmulliga, alacy, pguardiola@haverford.edu

Abstract

The Bridge, a linked data application supporting curriculum development is presented. It was developed with Latin in mind, but has been extended to Greek as well. It quickly helps instructors and students find new vocabulary words in newly assigned texts, based on texts they have already encountered in their curriculum.

1 Introduction

In this paper we present *The Bridge*, a linked data application, started in 2014 (Pistone, 2020) with on-going development designed for use by participants in language pedagogy processes.¹ *The Bridge* and its supporting tool-chains facilitate web-based interactions with texts as instructors and students navigate the learning and acquisition of new lexical items.

The Bridge is written in Python 3. It uses Python-based Natural Language Processing on texts to lemmatize them and then link lemmas across texts. The user interface allows users to query and receive reports regarding lexeme similarity across several selected texts. In this way, instructors, grounding their curriculum in texts, can map out the new vocabulary from text to text as they craft lesson plans. Likewise learners can look for new-to-them words, on the basis of the texts they have already been exposed to. In this way, learner pathways can be “charted” based on texts learners have already encountered. Our success in facilitating the acquisition of Latin and Greek has led us to believe that the application can be used in more languages than just English, Latin, and Greek. The code running *The Bridge* is available via Github.²

¹<https://bridge.haverford.edu>

²<https://github.com/HCDigitalScholarsHip/FastBridge>

2 CEFR Applicability

Measuring an individual’s language proficiency and language-learning progress is important for a host of reasons. The *Common European Framework of Reference for Languages* (CEFR) is a standard developed and widely used in the European Union for language competency description (Council of Europe, 2001). It is applied in the context of language proficiency assessment and language-learning curriculum development. Given the market position of the EU and its national languages, CEFR carries a significant presence in the area of language competency certification and language pedagogy, especially in the government and business sectors. Other systems for indicating language competencies have been mapped to CEFR. For example, the Cambridge English Scale used in the UK³ and the dominant system in the USA, the *American Council on the Teaching of Foreign Languages* (ACTFL) system (American Council on the Teaching of Foreign Languages, 2016). In contrast to the ACTFL system, which is designed primarily for assessing oral language fluency, the framework consists of a set of competency descriptions covering the areas of speaking, reading, and writing.⁴ The CEFR competencies are laid out in progressively increasing capabilities from the perspective of the pedagogical trajectory found in curriculum of commonly taught languages (CTL). CTLs are languages which have generally undergone substantial language development activities (Fishman, 1968; Ferguson, 1968). For example, languages such as English, German, Chinese, Russian, and Italian all have strong ethno-linguistic populations and are

³<https://www.cambridgeenglish.org>

⁴<https://www.coe.int/en/web/common-european-framework-reference-languages/table-1-cefr-3.3-common-reference-levels-global-scale>

languages that benefit from national-government level support. They are also marked by being used in communities that engage in intergenerational transmission. It is easy to apply the CEFR competencies to CTLs because they frequently rank at 0 or 1 on the *Expanded Graded Intergenerational Disruption Scale* (EGIDS) (Lewis and Simons, 2010; Bickford et al., 2015). That is, language use occurs in all the scenarios outlined in CEFR. However, for languages which score at a level between EGIDS 8a and 10, it is harder to consistently apply the CEFR competencies, assessments, and associated pedagogical methods. There are several reasons for this which vary by circumstances. Many of the *Less Commonly Taught Languages* of the world are also technologically under-resourced and do not yet have significant literary materials. Therefore, measuring language competency on the basis of a person's reading skills in a language as required by CEFR presents a challenge. In other cases—such as sign languages, endangered languages, and languages of antiquity (LA)—oral user communities do not exist. It is a challenge to prove CEFR B1 level competency under the requirement: “Can deal with most situations likely to arise whilst travelling in an area where the language is spoken”. These language use contexts appear to be at odds with the CEFR presumed relationship between oral/aural methods of communication and the written/reading methods of communication. More recent work has helped extend CEFR concepts to sign languages (Council of Europe, 2018). However, as sign languages are not the only non-oral languages, challenges exist in aligning curriculum and assessments to CEFR for endangered languages and LAs. Unlike many endangered languages, LAs such as Ancient Greek, Latin, Classical Chinese, Hittite, or Ancient Egyptian have large exploitable corpora. Endangered languages and LAs also differ in that LAs often have a significant educational presence but lack communities with current oral communication practices; although some argue that even for LAs, oral-first approaches support learners more effectively (Buth, 2020; Halcomb, 2020). Curriculum developers working with more commonly taught languages also use texts. Some have mapped texts or corpora according to a CEFR level (Xia et al., 2016; Wilkens et al., 2018) even though mapping text to CEFR levels and student capabilities to specific texts is challenging (Escobar-Acevedo et al., 2022). Using graded texts has some draw-

backs as texts are not the same as performative communication which CEFR is supposed to be assessing. Nevertheless, it has long been the practice for the languages of antiquity to be taught through the use of texts—without the requirements for oral competency, and literacy in some language has been a presumed foundational competency.

3 Instructional Goals and Classroom Context

Our current classroom context involves the instruction of languages of antiquity through text based approaches. Considering both communicative (oral/aural/signed) and text based approaches, a rather uncontroversial assertion is that sufficient vocabulary acquisition is essential if a language learner is to gain fluency in the new language. This is true whether a student's learning environment prioritizes *Comprehension* or *Skill-Building* in fostering language acquisition (Krashen, 2017). Vocabulary knowledge is not sufficient for comprehension, as cultural context, grammar, and discourse structures also need to be acquired. Ultimately, successful language learners must possess an operational vocabulary that allows them to understand a text (or utterance). This common-sense observation is well-supported by research into second language acquisition in several languages. Vocabulary knowledge is repeatedly claimed as the single best predictor of reading comprehension (Hu Hsueh-chao and Nation, 2000; Stæhr, 2008). Within the context of English, Chall (1958, 156–158) showed that vocabulary difficulty accounts for as much as 80% of the variability in reading scores, far outpacing syntactical elements. While these findings have been supported by research in inflected languages—e.g., on German (Röthlisberger et al., 2023)—the effect in highly-inflected historical languages like Latin and Ancient Greek remains to be assayed. For instructors focused on fostering successful reading of historical languages, these robust findings strongly suggest the importance of matching reading activities with lexical knowledge.

Yet the reading and instruction of many historical languages are on the horns of a dilemma. These languages often comprise vast corpora—in the case of Latin estimated at over a trillion words—yet a typical Latin student might engage texts totaling just a few tens of thousands of words (or a mere 0.000002% of the total corpus). Within this small slice, novice readers routinely move directly

from fabricated Latin in textbooks to difficult historical texts, whose reading grade level is akin to college-level texts (Gruber-Miller and Mulligan, 2022). To attain full comprehension, readers must typically know 95 to 98% of the words in that text (Hu Hsueh-chao and Nation, 2000). Yet many novice readers routinely know only 25% of the words in commonly-taught texts. While the statistics vary across language fields, the overarching concerns are the same. Instructors and independent learners have begun to pay attention to this dilemma, but lacked accurate and easily accessible tools to help them bridge the gap between their individual lexical knowledge and the lexical competence expected by the target text,⁵ as other tools routinely provide full vocabularies. These often automatically-generated and so prone to provide inaccurate information, especially for homonyms and inflected forms.

4 The Bridge

While *The Bridge* currently exists and can be exemplified by use cases, it is also undergoing active development based on classroom support needs.

4.1 Example use case

Imagine a class in which students completed an elementary sequence in the language using a standard textbook (e.g., *Wheelock's Latin* Wheelock and LaFleur, 2011), but turned to reading a historical text after finishing only 36 of the 40 chapters in the textbook (a common scenario, either because instructors run out of school year or because the final chapters of textbooks often present less common grammatical constructions that can be glossed in reading). Imagine this same class aimed to read the open-access version of Nepos' *Life of Hannibal* at Dickinson College Commentaries (DCC).⁶ The DCC version of the text includes vocabulary, but only other words that are not among the 997 most common words in Classical Latin that it has identified as the DCC Latin Core. Students using Wheelock have been exposed to a core vocabulary of 829 words (fewer if, as in our imagined scenario they have not yet finished the book); yet only 489 of these are also in the DCC Latin Core. Thus instructors who wished to know what words were known and unknown for their student would have

⁵Here we mean a competence with a finer granularity than CEFR competencies imply.

⁶<https://dcc.dickinson.edu/nepos-hannibal/chapter-1>

a great deal of time consuming work to identify words for their students—or cast them to the lexical wolves and let them fend for themselves, which will almost certainly lead them to use suboptimal resources that provide both too much and inaccurate lexical support. Also, while it might be useful to know the global vocabulary needed for Nepos, our instructor and students might instead wish to focus only on the first assignment.

The Bridge can quickly produce exactly this list. The first chapter of Nepos' *Life of Hannibal* contains 77 unique words. By default, *The Bridge* list appears with macrons but one can easily toggle between macronized and unmacronized entries. One can display basic English definitions or more full definitions—or create a practice or self-quiz list by removing the dictionary entries or definitions entirely. One can also reveal more information about each word, its importance in the text, or its frequency in Latin more generally. One can reveal the first time every word appears in the text—and sort by that appearance, creating a running vocabulary for each sub-division of the text. One can reveal the number of appearances in the entire text (and also sort), creating a quick reference for those words that will reappear frequently or are unique within the text [toggle up/down]. One can reveal the part of speech; and add a link to powerful open-source dictionaries like *Logeion*, connecting our list with an authoritative lexical resource. Finally, one can also reveal the rank of the word within the *Bridge Corpus*, which boasts over 1.5 million words in a range of poetry and prose from antiquity to neo-Latin texts.⁷ Every column of data is sortable.

But what makes *The Bridge* such a powerful tool is that it empowers users to customize the words that appear in the list. To return to our original scenario, students were not reading Nepos 1 with no lexical knowledge but having (supposedly) mastered vocabulary from the first 36 chapters of Wheelock. Instead of 77 words, there are only 25 unfamiliar words—still too many to expect students to divine from context but a much more manageable set, if one were to seek to prepare students to encounter them. But, of course, DCC commentaries already assume that students will not know any words that are not already among the 997 most common Latin words. So one could create a list

⁷Currently there are about 300 Greek and Latin texts, textbooks segments, and core vocabulary lists.

that shows only those words in the DCC that also appear in this section of our reading. This returns a list of the 22 words (17 if we exclude proper nouns) that could be the foundation for preparatory activities—a supplemental list. One can also use the *The Bridge* to create a list of the 55 words in the text that students have already seen for review or assessment purposes.

This process can then be sequenced as students continue to read and gain familiarity with new words. To take another possible scenario: imagine students are engaging with text in the Advanced Placement Program (AP)⁸ selections of the *Aeneid*—or to better align with a typical weekly assignment, the first 100 lines of *Aeneid, Book 1*. One could construct a vocabulary list by excluding multiple sources of vocabulary: say, (1) the 50 most common Latin verbs; (2) the 400-most common words in the DCC Latin core; (3) all of the words from the *Cambridge Latin Course* textbook (Cambridge School Classics Project, 1998); and (4) any word that appeared in a text that you have already read, e.g., *Catullus 1* and the AP selections of *Caesar's Gallic Wars*. The resulting vocabulary list results in a useful learning aid.

The Bridge lists can be further customized using morphological filters: e.g., a list of just nouns, or just 3rd declension nouns, 3rd declension nouns and adjectives, or a list that excludes proper names (or just proper names). These lists can be printed or exported (as CSV files) for further manipulation or transfer to a flashcard program, question bank, or other media.

4.2 Usage

The Bridge has been well reviewed (Pistone, 2020) and has seen significant use among classicists. Usage growth beyond Haverford College resulted in over 24,000 unique user sessions in 2022.

4.3 Active development

To support this lexical tool, we are further developing *The Bridge* ecosystem to enable users to: (1) encode texts for analysis in this and other digital ecosystems; (2) analyze and compare the readability of texts; and (3) discover readable texts

⁸The *Advanced Placement Program* is a commercial educational program available through secondary schools in the United States. Passing students are generally given university level credit for course completion. The AP Latin curriculum is well known by classicists in the United States. <https://apcentral.collegeboard.org/courses/ap-latin/course/ap-latin-reading-list>

for data-informed lesson plans, syllabi, and curricula. Integration with Linking Latin (LiLa)⁹ and its scheme is part of ongoing NEH grant funded work. The current vision for *The Bridge* ecosystem includes *Bridge/Lemmatizer*, *Bridge/Stats*, and *Bridge/Oracle*.

4.3.1 Bridge/Lemmatizer

Bridge/Lemmatizer will be a web-based environment, allowing more rapid, accurate, and detailed lexical and syntactic encoding of texts, and facilitating collaboration by faculty, students, and other contributors. Lemmatizers can be optimized for different languages. Our plan is to enable different lemmatizers for different language requirements.

4.3.2 Bridge/Stats

Bridge/Stats will be a web-based dashboard that displays information about lexical and syntactic difficulty—i.e., readability—for texts, and the effect that user-defined knowledge has on textual readability for one or more texts and/or sections based on their (1) generic readability; and (2) readability that factors in personalized lexical knowledge using metrics such as: (a) word length; (b) word frequency, or the prevalence of very common words; (c) lexical sophistication, or the percentage of rarer words; (d) lexical variation, or the variety of different words; (e) hapax legomena, or words that appear only once; (f) the corpus frequency of rare and/or unknown words; (g) the number of words per sentence; and (h) the number and length of subordinate clauses.

4.3.3 Bridge/Oracle

Bridge/Oracle will be a web-based app that allows users to discover lexically readable texts in the Bridge Corpus by revealing the authors, texts, and passages that have the highest percentage of familiar vocabulary alongside basic readability data, with users selecting the author(s), text(s), or genre(s) they would like to explore and then indicate their known vocabulary by selecting textbooks used, lists mastered, and texts previously read.

5 Conclusion

Early development of *The Bridge* ecosystem has focused on Latin but its framework has been designed to be language agnostic. This allows the development of Latin to serve as a model system for the longer-term goal of supporting the teaching and

⁹<https://lila-erc.eu>

accessibility of other languages, beginning with Ancient Greek and then other historical languages. This can be further extended to other commonly taught modern languages, across a global spectrum. *The Bridge Readability Apps* will be designed for use with any language for which Natural Language Processing (NLP) resources exist, creating the potential of use cases far beyond its initial target audiences at schools, colleges, and universities around the world.

Acknowledgements

Bret Mulligan is *The Bridge* project director. Hugh Paterson III's involvement is sponsored via the *LEADING Data Science Fellowship* through the *Institute of Museum and Library Services* (IMLS) RE-246450-OLS-20. Michael Rabayda and over twenty other contributors have made *The Bridge* possible. Their names are in the project credits.¹⁰ We are grateful to Paul Unger for editorial comments and to an anonymous reviewer for relevant suggestions.

References

- American Council on the Teaching of Foreign Languages. 2016. *Assigning CEFR Ratings to ACTFL Assessments*. American Council on the Teaching of Foreign Languages, Alexandria, Virginia.
- J. Albert Bickford, M. Paul Lewis, and Gary F. Simons. 2015. *Rating the vitality of sign languages*. *Journal of Multilingual and Multicultural Development*, 36(5):513–527.
- Randall Buth. 2020. The Role of Pronunciation in New Testament Greek Studies. In David Alan Black and Benjamin L. Merkle, editors, *Linguistics and New Testament Greek: Key Issues in the Current Debate*, pages 169–194. Baker Academic, a division of Baker Publishing Group, Grand Rapids, Michigan.
- Cambridge School Classics Project. 1998. *Cambridge Latin Course*, 4th edition. Cambridge University Press, Cambridge, UK.
- Jeanne Sternlicht Chall. 1958. *Readability: An Appraisal of Research and Application*. Number 34 in Educational Research Monographs. The Ohio State University Bureau of Educational Research, Columbus, Ohio.
- Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press; [CoE] Modern Languages Division, Strasbourg, France.
- Council of Europe. 2018. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment — Companion Volume with New Descriptors*. Language Policy Programme, Education Policy Division, Education Department, Council of Europe, Strasbourg, France.
- Adelina Escobar-Acevedo, Josefina Guerrero-García, and Rafael Guzmán-Cabrera. 2022. *A Model Text Recommendation System for Engaging English Language Learners: Facilitating Selections on CEFR*. *Rupkatha Journal on Interdisciplinary Studies in Humanities*, 14(3):1–8.
- Charles A. Ferguson. 1968. Language Development. In Joshua A. Fishman, Charles A. Ferguson, and J. Das Gupta, editors, *Language Problems of Developing Nations*, pages 27–36. Wiley and Sons, New York.
- Joshua A. Fishman. 1968. Language Problems and Types of Political and Socio-Cultural Integration: A Conceptual Postscript. In *Report on the Ninth International Conference on Second Language Problems, Tunis, 24–27 April*. English-Teaching Information Centre, London, England.
- John Gruber-Miller and Bret Mulligan. 2022. *Latin Vocabulary Knowledge and the Readability of Latin Texts: A Preliminary Study*. *New England Classical Journal*, 49(1):80–101.
- T. Michael W. Halcomb. 2020. Living Language Approaches. In David Alan Black and Benjamin L. Merkle, editors, *Linguistics and New Testament Greek: Key Issues in the Current Debate*, pages 147–168. Baker Academic, a division of Baker Publishing Group, Grand Rapids, Michigan.
- Marcella Hu Hsueh-chao and Paul Nation. 2000. *Unknown vocabulary density and reading comprehension*. *Reading in a Foreign Language*, 13(1):403–430.
- Stephen Krashen. 2017. *The Case for Comprehensible Input*. *Language Magazine*, July.
- M. Paul Lewis and Gary F. Simons. 2010. *Assessing Endangerment: Expanding Fishman's GIDS*. *Revue roumaine de linguistique*, 55(2):103–120.
- Amy Pistone. 2020. *Review: A Digital Tool that Helps Teachers Generate Latin and Greek Vocabulary Lists*. *Society for Classical Studies Blog*.
- Martina Röthlisberger, Christoph Zangger, and Britta Juska-Bacher. 2023. *The role of vocabulary components in second language learners' early reading comprehension*. *Journal of Research in Reading*, 46(1):1–21.
- Lars Stenius Stæhr. 2008. *Vocabulary size and the skills of listening, reading and writing*. *Language Learning Journal*, 36(2):139–152.

¹⁰<https://bridge.haverford.edu/about/people>

Frederic M. Wheelock and Richard A. LaFleur. 2011. *Wheelock's Latin*, 7th edition. The Wheelock's Latin Series. Collins Reference, New York.

Rodrigo Wilkens, Leonardo Zilio, and Cédric Fairon. 2018. *SW4ALL: A CEFR Classified and Aligned Corpus for Language Learning*. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. *Text Readability Assessment for Second Language Learners*. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22, San Diego, CA. Association for Computational Linguistics.

**PROfiling LINGuistic
KNOWledgE gRaphs
(ProLingKNOWER)**

Profiling Linguistic Knowledge Graphs

Blerina Spahiu Renzo Alva Principe Andrea Maurino
 University of Milan-Bicocca University of Milan-Bicocca University of Milan-Bicocca
 Milan, Italy Milan, Italy Milan, Italy

{blerina.spahiu | renzo.alvaprincipe |
 andrea.maurino} @unimib.it

Abstract

Recently the number of approaches that model and interconnect linguistic data as knowledge graphs has experienced outstanding growth. However, despite the increasing availability of applications that manage such data, little attention has been given to their structural features. In this paper, we propose specific metrics to describe the structural features of knowledge graphs. Such metrics are evaluated on linguistic data and our findings provide a basis for a more efficient understanding of linguistic data.

1 Introduction

Language resources such as dictionaries, terminologies, corpora, etc., are adopting Semantic Web technologies to make their discovery, reuse and integration easy (Cimiano et al., 2020). The Linked Data (LD) paradigm materialises Semantic Web by enabling data belonging to different topics (Spahiu et al., 2019) to be interconnected within a data-to-data cloud¹. The linguistics community has taken advantages of the potential of the LD and has developed the Linguistic Linked Open Data (LLOD) cloud² for improving the usability and the discovery of language and linguistic resources.

In this vein, knowledge is represented into graphs using nodes and arcs. Such knowledge is stored and represented in RDF format³. The nodes represent entities while arcs represent relations among entities. Entities can have a relation of the form `rdf:type` denoting their types. The sets of possible types and relations are organized into schemas or ontologies, which define the meaning of the terms

used in the knowledge graph through logical axioms.

KGs are often large and continuously evolving. As an example we can mention LOD cloud with more than 1,301 data sets as of March 2022. This huge adoption of KGs into applications, is due to the fact that, with respect to relational models, KGs represent a flexible data model (e.g., Google’s Knowledge Graph, Facebook’s Graph API, Wikidata, etc.) where numerous editors are engaged in content creation, where the schema is ever changing, where data are incomplete, and where the connectivity of resources plays a key role. As the number of approaches that model linguistic data as knowledge graphs is increasing rapidly (Cimiano et al., 2020), understanding their structure remains a fundamental step for their reuse. For example, before using a dataset one could be curious of How types are related to each other? or How many triples are used to describe entities?. In such a scenario, users want to know some structural features of these datasets, but this information is not completely covered in the state-of-the-art tools and approaches.

Even though the use of KGs in different applications is a matter of fact, it has a cost. When a user needs to use a KG for his/her use case, several are the challenges to be faced: (1) No prior knowledge about the data, (2) Missing schema or underspecification, (3) Lack of compliance with respect to the ontology, (4) Scalability challenges of large-scale RDF processing.

Such challenges might be addressed by knowledge graph profiling tools and approaches. Profiling approaches provide insights about the data in form of summaries, statistics or both (Spahiu et al., 2023). Being able to access and explore the profile of a

¹<https://lod-cloud.net/>

²<http://linguistic-lod.org/llod-cloud>

³<https://www.w3.org/RDF/>

KG, a user can formulate and optimize queries, understand how graphs evolve and change, as well as enable data-management operations, such as compression, indexing, integration, enrichment and so forth. ABSTAT⁴ is a data profiling tool proposed to mitigate some of the above challenges and help users understanding the content of a dataset effortlessly.

In this paper we make the following contributions: (i) enrich the profile produced by ABSTAT with 24 new statistics; (ii) provide a list of applications where such statistics are useful; (iii) provide an empirical analysis of the structural features of linguistics datasets, and (iv) provide a short discussion of such features. The paper is structured as follows: Section 2 discusses approaches and tools used to profile KGs. In Section 3 we provide a brief description of ABSTAT profiles and provide the list of the new statistics added to such tool. Section 4 provides the analysis and findings by applying the enriched profile to LLOD datasets. The discussion analysis is described in Section 5 while conclusions and future work end the paper in Section 7.

2 Related Work

RDF graph profiling has been intensively studied, and various approaches and techniques have been proposed to provide a concise and meaningful representation of an RDF KG. There are different recent surveys that discuss some of the approaches to profile knowledge graphs such as (Čebirić et al., 2019), (Zneika et al., 2019) and (Song et al., 2018). In a recent work (Spahiu et al., 2023) we have reviewed and categorise profiling approaches. However, in this work, we focus only on approaches that aim to produce profiles that quantitatively represent the content of the graph and provide an empirical analysis of the structural features of KGs.

ExpLOD (Khatchadourian and Consens, 2010) is used to summarize a dataset based on a mechanism that combines text labels and bisimulation contractions. It considers four RDF usages that describe interactions between data and metadata, such as class and predicate instantiating, and class and predicate usage on which it creates RDF graphs.

It provides also statistics about the number of equivalent entities connected using the owl:sameAs predicate to describe the inter-linking between datasets. The ExpLOD summaries are extracted using SPARQL queries or algorithms such as partition refinement.

RDFStats generates statistics for datasets behind SPARQL endpoint and RDF documents (Langegger and Woss, 2009). These statistics include the number of anonymous subjects and different types of histograms; URIHistogram for URI subject and histograms for each property and the associated range(s). It also uses methods to fetch the total number of instances for a given class, or a set of classes and methods to obtain the URIs of instances.

LODStats is a profiling tool that can be used to obtain 32 different statistical criteria for RDF datasets (Auer et al., 2012). These statistics describe the dataset and its schema and include statistics about the number of triples, triples with blank nodes, labeled subjects, number of owl:sameAs links, class and property usage, class hierarchy depth, cardinalities etc. These statistics are then represented using Vocabulary of Interlinked Datasets (VOID) and Data Cube Vocabulary⁵.

Sansa is a graph processing tool that provides a unified framework for several applications such as link prediction, knowledge base completion, querying, and reasoning (Jabeen et al., 2020). It computes several RDF statistics (such as the number of triples, RDF terms, properties per entity, and usage of vocabularies across datasets), and applies quality assessment in a distributed manner.

The approach most similar to ABSTAT is Loupe (Mihindukulasooriya et al., 2015). Loupe extracts types, properties and namespaces, along with a rich set of statistics about their use within the dataset. It offers a triple inspection functionality, which provides information about triple patterns that appear in the dataset and their frequency. Triple patterns have the form <subjectType, property, objectType>. Differently from ABSTAT, Loupe does not adapt a minimalization approach thus, Loupe's profiles contain much

⁴<http://abstat.disco.unimib.it/>

⁵<http://www.w3.org/TR/vocab-data-cube/>

more triple patterns and are not as concise as ABSTAT profiles.

3 Profile Description

ABSTAT is a data profiling framework aiming to help users understanding the content of big datasets by exploring its semantic profile (Spahiu et al., 2023). It takes as input a data set and an ontology (used by the data set) and returns a semantic profile (Fig.1). Thanks to the highly distributed architecture, ABSTAT is able to profile very big KGs (Alva Principe et al., 2021). The semantic profile produced by ABSTAT consists of a summary of patterns and several statistics (Fig. 1). The informative units of ABSTAT's summaries are Abstract Knowledge Patterns (AKPs), named simply patterns in the following, which have the form (subjectType, pred, objectType). Patterns represent the occurrence of triples <sub, pred, obj> in the data, such that subjectType is the most specific type of the subject and objectType is the most specific type of the object (Spahiu et al., 2016). Despite patterns, ABSTAT extracts also some statistics as the occurrence of types, predicates, patterns and cardinality descriptors (Fig. 1).

Even though ABSTAT profiles provide valuable information about the content of the dataset, it still misses some basic information that could help users in gaining a fast overview of some characteristics that these datasets have.

Below we enumerate the list of new statistics that are added to the semantic profile produced by ABSTAT:

- # triples: This statistic computes the number of triples in an RDF dataset.
- # entities: This statistic computes the number of entities in an RDF dataset.
- # triples per entity (min, max, average): This statistic calculates the minimum, average and the maximum number of triples used to describe an entity.
- # internal and external concepts: This statistic computes the number of concepts that are considered to be internal of the

dataset (defined in the pay-level domain) and external concepts (not defined in the pay-level domain)⁶.

- # internal and external properties: This statistic computes the number of properties that are considered to be internal of the dataset (defined in the pay-level domain) and external properties (not defined in the pay-level domain).
- # blank nodes as subject and # blank nodes as object: This statistic counts the number of blank nodes that occur at the subject and at the object position of a triple.
- In and out degree: This statistic counts the number of links coming from the other datasets (in-degree) and the number of links going from the dataset to others (out-degree). The in-degree calculates the number of triples of the form (subject inPLD, predicate, object notPLD) while the out-degree counts the number of triples of the form (subject notD, predicate, object inLD).
- # owl:sameAs triples: This statistic counts the number of triples that use (and those that do not use) the predicate owl:sameAs.
- # rdfs:label triples: This statistic counts the number of triples that use the predicate rdfs:label.
- The list of typed and untyped literals: This statistic gives the list of typed and untyped literals used in a dataset.
- The average length of untyped literals: This statistics calculates the average length of the untyped literals.
- # of datatypes and their frequency: This statistics provides the number and the frequency of use for each datatype used in a dataset.

⁶The pay-level domain is defined as the part of a domain name, which can typically be registered by companies, organisations, or private end user (Gotttron et al., 2015)

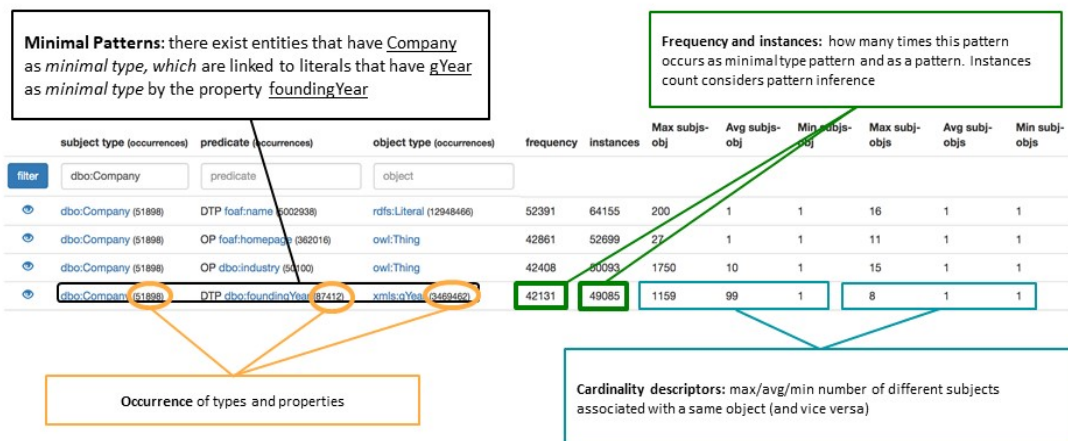


Figure 1: ABSTAT profile for a sample of DBpedia dataset.

- The list and the occurrence of the used languages: This statistic enumerates the list with the occurrence of each language used in the dataset.
- The list and the occurrence of the used vocabularies: This statistic enumerates the list with the occurrence of each vocabulary used in the dataset.

All the above statistics are implemented as API calls from the interface of ABSTAT tool.

4 Experiments

In this section we provide an analysis of the structural features by applying the above statistics in linguistics datasets from the Linguistic Linked Data Cloud.

4.1 Linguistics Datasets

The experiments were run using all the datasets from the Linguistic Linked Open Data Cloud. There are in total 136 datasets belonging to the linguistic domain in the LOD cloud. However, only 72 of them do provide a URL for the dump (di Buono et al., 2022). During the inspection of the availability of the dump it was possible to download and process the dump for only 48 datasets, while for the other (i) either the URL was not available anymore, or (ii) the dump was available but the dataset had many syntactic errors, or (iii) they were not in RDF.

4.2 Empirical Analysis

In this section we analyse the results for each of the above statistics applied to our datasets

corpus.

triples, # entities, and # triples per entity: From all the datasets from the LLOD cloud the biggest dataset is iate with 74, 023, 248 triples and 20, 726, 310 entities while the smallest datasets with respect to the number of triples is lemonbuy with 961 triples and apertium-rdf-en-es is the smallest dataset with respect to the number of entities, i.e., 2. Datasets belonging to the apertium datasets have from 2 to maximum 6 entities while 47, 445 to maximum 156, 941 triples. Thus the average number of triples for entities is greater for apertium datasets. The datasets that uses in average less triples per entity are wn-wiki-instances, srcmf, linked-hypernoms, cedict with around 1 triple per entity.

internal and external concepts: The analysis shows that only three datasets cedict, gwa-ili, iso-639-oasis use 2 internal concepts each to describe entities. As a consequence, the number of external concepts for all the datasets is greater. The dataset with the highest number of external concepts is linked-hypernoms with 361. Finally, wn-wiki-instances dataset has 0 internal and 0 external concepts.

internal and external properties: Similar analysis for the concepts is present for the number of internal and external properties. Only 5 datasets use internal properties to describe resources, i.e cedict (4), iso-639-oasis (3), lexvo (13), saldo-rdf (2), and word-net (26). All the rest 43 datasets have 0 internal properties but borrow them from exter-

nal vocabularies. The dataset with the highest number of external properties is *getty-aat* with 196 properties. Around 77% of the datasets have less than 10 properties.

blank nodes as subject and # blank nodes as object: The analysis about the use of blank nodes shows that only *lemonbuy* uses blank nodes in the subject position while 52% of datasets use blank nodes in the object position. The dataset with the highest number of blank nodes is *cedict* (554367) and *wordnet* (423986).

In and out degree: Datasets in the LLOD are more generally connected from inside to outside, meaning that the object of their triples reside in other datasets. In fact, only 12,5% of the datasets have 0 outgoing links, while 62% have 0 incoming links. *iate* dataset has the highest number of outgoing links with 16, 881, 770 links while *saldo-rdf* plays the role of a central hub with 320, 059 incoming links. Fig. 2 shows the distribution of the number of outgoing and incoming links for each dataset in the LLOD.

owl:sameAs triples: Regarding the type of outgoing and incoming links we further analyse the use of owl:sameAs predicate. The distribution of the number of such triples within the LLOD is shown in Fig. 3. The datasets with the highest number of owl:sameAs triples is *iate*, which also had the highest number of outgoing links. Around 46% of the datasets have less than 3 sameAs links, while less than 10% have more than 100, 000 sameAs links.

rdfs:label triples: We analysed the use of the predicate rdfs:label by the entities of LLOD. Around 77% of the datasets have less than 10 triples with the predicate rdfs:label. 4 datasets have more than 100, 000 rdfs:label triples, i.e., *basque-eurowordnet-lemon-lexicon-3-0* (134, 748), *lexvo* (146, 530), *catalan-eurowordnet-lemon-lexicon-3-0* (213, 787), and *sli_galnet_rdf* (723, 348) triples.

typed and untyped literals: The graph in Fig. 4 shows the distribution of typed and untyped literals. As from the graph *iate* has most of typed (7, 803, 650) and untyped (12, 922, 660) literals. 11 datasets do not have any untyped literals.

The average length of untyped literals: Top three datasets that have in average the longest untyped literals are *news-100-nif-ner-corpus* (70), *gwa-ili* (62), and *reuters-128-nif-ner-corpus* (60).

The list and the occurrence of datatypes: The most used datatype in the LLOD is <http://www.w3.org/2001/XMLSchema#integer> (8, 710, 881), followed by <http://www.w3.org/2001/XMLSchema#date> (37347) and <http://www.w3.org/2001/XMLSchema#dateTime> (36428). The less used datatype instead is <http://www.w3.org/2001/XMLSchema#boolean> (2).

The list and the occurrence of the used languages: There are 176 languages used to tag literals in the LLOD datasets. The dataset with most languages is *lexvo* with 175 different languages. Around 90% of the datasets have less than five languages. The most used language is English (36), Swedish (6), and French (5).

The list and the occurrence of the used vocabularies: The analysis shows that the dataset that uses most vocabularies to describe its data is *lexvo* (626). The distribution of the number of vocabularies per dataset is given in Fig. 5. The most used vocabularies among LLOD datasets are *rdf* (48), *rdfs* (37), and *owl*.

5 Discussion

In this work, we have analysed structural features of Linguistic LOD datasets. All datasets show a skewed structure with respect to the number of internal and external concepts and properties. In fact, almost all the datasets had more external concepts and properties. Complementing the previous finding, our evaluation also revealed that most datasets are extensively typed (more than 99% of datasets have typed entities). Regarding the in & out degree, most of the datasets had more outgoing links. In fact, most of the datasets make use of the owl:sameAs predicate. However, our findings are not in line with what is being described in the LLOD website⁷. This is for two reasons: (i) we consider the dump of the datasets having the topic linguistic in the

⁷<https://linguistic-lod.org/>

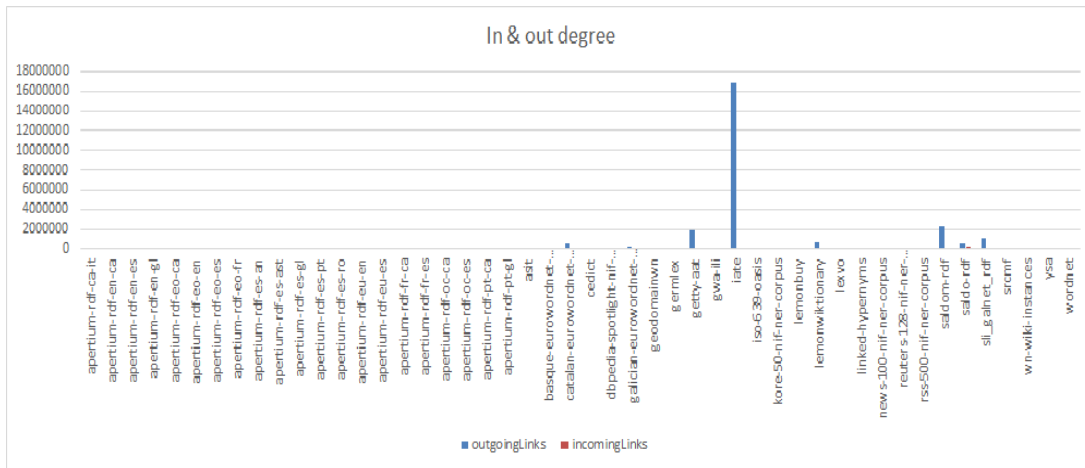


Figure 2: In & out degree for LLOD datasets.

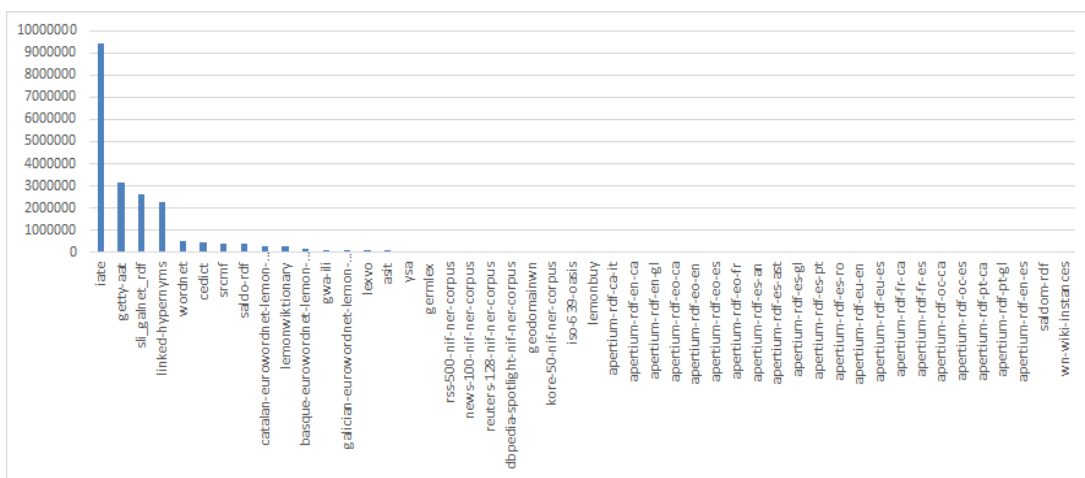


Figure 3: Distribution of number of owl:sameAs triples per LLOD datasets.

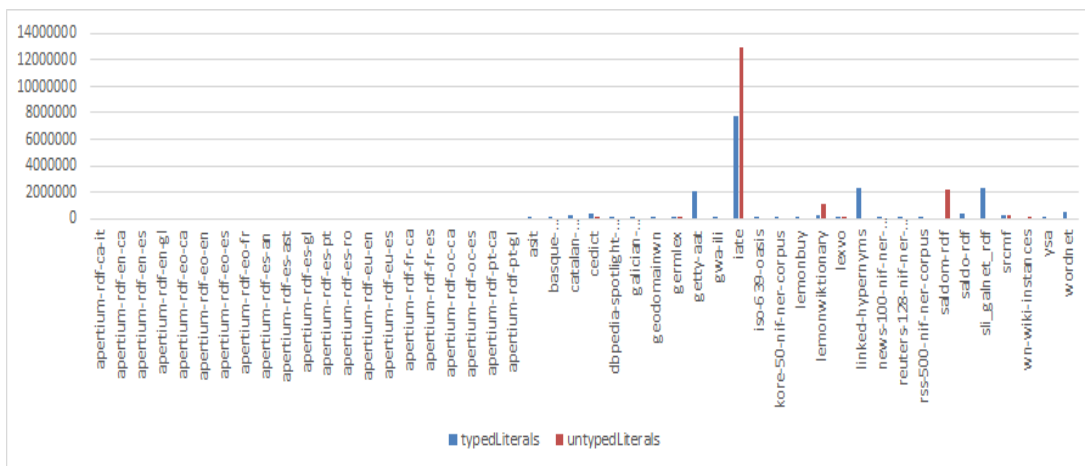


Figure 4: Distribution of typed and untyped literals per LLOD datasets.

metadata of the LOD cloud, and (ii) the version of the LLOD datasets might be different.

We observed that rdfs:label predicate is not often used as three-quarters of the datasets use

it within less than 10 triples. Also, the distribution of typed and untyped literals is skewed. While most of the smallest datasets (with respect to the number of triples) do use typed

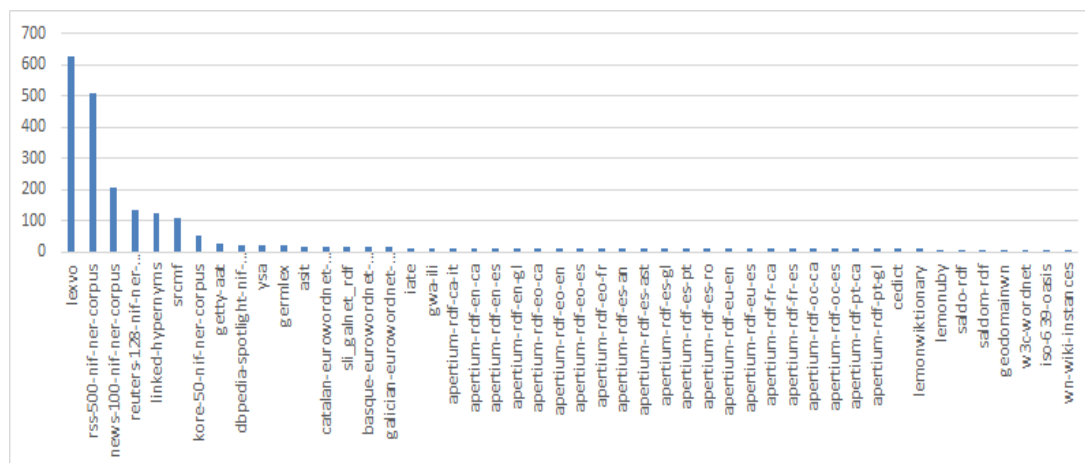


Figure 5: Number of vocabularies per LLOD datasets.

literals, for the biggest ones the number of untyped literals is greater than the typed ones.

The most frequent used language within LLOD dataset is English (99% of the datasets). Moreover, lexvo is the dataset with the highest number of languages (175) out of 176 of the languages in total. Regarding the used vocabularies, rdf remains the most used vocabulary by most of the datasets in the LLOD.

6 Analysis of key statistics and their application significance

In this section, we group the above statistics in regard to their application.

Entity Summarization: For this application scenario, the statistics in Table 1 provide (i) a quantitative understanding of the size and density of the knowledge graph, allowing for efficient summarization techniques; (ii) help identifying the source and coverage of concepts and properties used in the graph, aiding in accurate and comprehensive entity summarization; and (iii) providing information about entity equivalence and human-readable labels, enabling improved entity summarization and labeling.

Recommendation Systems: The statistics useful for recommendation systems (i) identify the level of information available for each entity, enabling more informed and personalized recommendations, (ii) analyse the connectivity of the dataset with external sources helping in incorporating relevant information from external sources for more accurate recommen-

dations, and (iii) assists in identifying equivalent entities, which can enhance recommendation algorithms by considering similar or related items.

Question Answering: For this downstream application these statistics provide (i) a sense of the knowledge graph's size and coverage, aiding in understanding the scope and potential for answering a wide range of questions; (ii) identify the level of detail available for each entity, assisting in generating comprehensive and informative answers, (iii) provide human-readable labels for entities, improving the clarity and understandability of question answering results.

Information Extraction: The statistics for this application offer (i) insights into the overall scope and coverage of the knowledge graph, helping in identifying relevant entities and relationships for extraction tasks, and (ii) assist in identifying instances where entities are represented as blank nodes, allowing for appropriate handling during information extraction processes.

Link Prediction: Link prediction is supported by (i) providing information about the richness of entity descriptions, aiding in more accurate link prediction by considering entities with more detailed representations, (ii) analyzing the connectivity of the dataset with external sources helps in predicting links between the knowledge graph and external entities; and (iii) in identifying equivalent entities, supporting link prediction across different datasets or ontologies.

Table 1: Application-specific metric

	Entity Summarisation	Recommendation Systems	Question Answering	Information Extraction	Link Prediction	Anomaly Detection	Semantic Search	Data Integration and Fusion
# triples	x		x	x			x	x
# entities	x		x	x			x	x
# triples per entity	x	x	x		x	x		x
# internal and external concepts	x							x
# internal and external properties	x							x
# blank nodes as subject				x		x		
# blank nodes as object				x		x		
in and out degree		x			x			
# owl:sameAs triples	x	x			x			x
# rdfs:label triples	x		x				x	
list of typed and untyped literals							x	
average length of untyped literals							x	
# of datatypes and their frequency						x		
list and the occurrence of the used languages							x	
list and the occurrence of the used vocabularies							x	

Anomaly Detection: The statistics for this application scenario (i) identify entities with abnormal numbers of triples, aiding in anomaly detection by flagging entities with unusual representations or relationships, and (ii) assist in identifying instances where blank nodes are involved in triples, which can be indicative of potential anomalies or incomplete information.

Semantic Search: These statistics for Semantic Search offer: (i) they indicate the knowledge graph's size and coverage, ensuring comprehensive and accurate semantic search results; (ii) they provide human-readable labels for entities, enhancing the relevance and presentation of search results; (iii) they include textual information linked to entities, thereby improving the retrieval of relevant results.

Data Integration and Fusion: Such statistics help understanding the size and scope of the knowledge graph, supporting data integration efforts by assessing the compatibility and overlap with external datasets and assist in identifying concepts and properties shared.

7 Conclusion and Future Work

In this paper we present a first preliminary analysis of structural features of LLOD datasets. We extend the profile built by ABSTAT tool with 24 new statistics in order to have a more detailed view of the content of RDF datasets. Such statistics have been applied to datasets that belong to the linguistics domain of the LOD datasets. We were not able to manage all the datasets belonging to

this domain as for many we were not able to find the dump or it had syntactic errors. However, we provide an empirical analysis of the content for 48 datasets.

Currently we are extending the profile with some fine-grained statistics. As future work we plan to integrate all statistics as API calls in the ABSTAT profile. Moreover, we plan to build an interactive interface where users can make more insightful analysis by cross-checking some of the statistics provided by ABSTAT.

References

- Renzo Arturo Alva Principe, Andrea Maurino, Matteo Palmonari, Michele Ciavotta, and Blerina Spahiu. 2021. Abstat-hd: a scalable tool for profiling very large knowledge graphs. *The VLDB Journal*, pages 1–26.
- Sören Auer, Jan Demter, Michael Martin, and Jens Lehmann. 2012. Lodstats—an extensible framework for high-performance dataset analytics. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 353–362. Springer.
- Šejla Čebirić, François Goasdoué, Harimos Kondylakis, Dimitris Kotzinos, Ioana Manolescu, Georgia Troullinou, and Mussab Zneika. 2019. Summarizing semantic graphs: a survey. *The VLDB Journal*, 28(3):295–327.
- Philipp Cimiano, Christian Chiarcos, John P McCrae, and Jorge Gracia. 2020. *Linguistic Linked Data*. Springer.
- Maria Pia di Buono, Hugo Gonçalo Oliveira, Verginica Barbu Mititelu, Blerina Spahiu, and

- Gennaro Nolano. 2022. Paving the way for enriched metadata of linguistic linked data. *Semantic Web*, 13(6):1133–1157.
- Thomas Gottron, Malte Knauf, and Ansgar Scherp. 2015. Analysis of schema structures in the linked open data graph based on unique subject uris, pay-level domains, and vocabulary usage. *Distributed and Parallel Databases*, 33(4):515–553.
- Hajira Jabeen, Damien Graux, and Gezim Sejdiu. 2020. Scalable knowledge graph processing using sansa. In *Knowledge Graphs and Big Data Processing*, pages 105–121. Springer.
1. Shahan Khatchadourian and Mariano P Consens. 2010. Explod: summary-based exploration of interlinking and rdf usage in the linked open data cloud. In *Extended Semantic Web Conference*, pages 272–287. Springer.
- Andreas Langegger and Wolfram Woss. 2009. Rdfstats-an extensible rdf statistics generator and library. In *2009 20th International Workshop on Database and Expert Systems Application*, pages 79–83. IEEE.
- Nandana Mihindukulasooriya, Maña Poveda-Villabón, Raúl García-Castro, and Asunción Gómez-Pérez. 2015. Loupe-an online tool for inspecting datasets in the linked data cloud. In *International Semantic Web Conference (Posters & Demos)*.
- Qi Song, Yinghui Wu, Peng Lin, Luna Xin Dong, and Hui Sun. 2018. Mining summaries for knowledge graph search. *IEEE Transactions on Knowledge and Data Engineering*, 30(10):1887–1900.
- Blerina Spahiu, Andrea Maurino, and Robert Meusel. 2019. Topic profiling benchmarks in the linked open data cloud: Issues and lessons learned. *Semantic Web*, 10(2):329–348.
- Blerina Spahiu, Matteo Palmonari, Renzo Alva Principe, and Anisa Rula. 2023. Understanding the structure of knowledge graphs with abstat profiles. Submitted to *Semantic Web Journal*.
- Blerina Spahiu, Riccardo Porrini, Matteo Palmonari, Anisa Rula, and Andrea Maurino. 2016. Abstat: ontology-driven linked data summaries with pattern minimalization. In *European Semantic Web Conference*, pages 381–395. Springer.
- Mussab Zneika, Dan Vodislav, and Dimitris Kotzinos. 2019. Quality metrics for rdf graph summarization. *Semantic Web*, (Preprint):1–30.

Pruning and re-ranking the frequent patterns in knowledge graph profiling using machine learning

Gollam Rabby

L3S, Leibniz
University Hannover,
Hanover, Germany
and
VSE, Prague, Czechia
gollam.rabby@L3S.de

Farhana Keya

TIB Leibniz ICST,
Hanover, Germany
and
VSE, Prague, Czechia
keya@tib.eu

Vojtěch Svátek

VSE, Prague, Czechia
svatek@vse.cz

Blerina Spahiu

University of
Milano-Bicocca,
Milan, Italy
blerina.spahiu@unimib.it

Abstract

Sets of frequent schema-level patterns characterizing a given knowledge graph (KG) represent a central output of profiling tools such as ABSTAT, as they could provide a quick overview of the coverage of the KG and its adequacy for various tasks. However, the number of patterns may be huge. The most frequent ones are often not useful for semantically characterizing the KG since they feature generic (OWL, SKOS, etc.) classes and even XML data types. We hypothesize that the pattern profile suitability for a ‘rapid skimming’ scenario might be improved by applying pattern post-processing, namely, their pruning and/or re-ranking. In this paper, we investigate, for this purpose, different machine learning (ML) methods trained on manually labelled examples (whole namespaces or individual IRIs of entities). Random Forest, Decision Tree and Multi-layer Perceptron Classifiers get higher accuracy than others.

1 Introduction

Because of the high number and large size of knowledge graphs (KGs), which makes it difficult to rapidly identify the KG suitable for a particular application, KG *profiling* was recently introduced as a means of quantifying the structure and contents of KGs to judge their suitability for particular applications. Of the many quantitative and qualitative characteristics that can describe a KG, the schema-level pattern of the form $\langle \text{subjectType}, \text{pred}, \text{objectType} \rangle$ as an abstract representation of the KG instances is particularly interesting from the point of view of knowledge engineering. Profiling tools based on schema patterns, such as ABSTAT (Spahiu et al., 2016) or Loupe (Mihindukulasooriya et al., 2015), give the user specific insights into frequent paths interconnecting entities at the instance level while remaining relatively concise. The outcome depends

on the ontology employed and the degree of explicit typing of entities. The internals of these tools consist of sophisticated graph-theoretic methods, and some rely on massive parallelization of the computation. However, the results in their generic form may not always fit every kind of usage. The scenario we have in mind is that of *rapid skimming through multiple KGs* to identify those having adequate coverage of some topic/s (contrasting to a scenario requiring detailed scrutiny of a dataset’s schema). For this, the output of a state-of-the-art tool such as ABSTAT (even a ‘minimal,’ non-redundant set) still contains too many patterns that are ‘boring’ concerning such skimming.

In our previous work (Rabby et al., 2022) we directly applied a handful of manually-written heuristics in order to (further) prune as well as re-rank the output of ABSTAT. The current paper extends this previous attempts by exploring, for the same purpose, various machine learning (ML) methods which have been trained on manually labeled examples.

2 ABSTAT

ABSTAT is a scalable profiling tool that aims to support users in exploring and understanding large RDF KGs. Given a KG in the form of a dataset and an ontology (optional), ABSTAT computes a profile comprising a summary of the dataset content and statistics. A summary is a set of data-driven ontology patterns in the form $\langle \text{subjectType}, \text{pred}, \text{objectType} \rangle$, which represent the occurrence of the triples $\langle \text{subj}, \text{pred}, \text{obj} \rangle$ in the dataset. Minimalization is applied on types and properties; that is, subjectType is a minimal type for subj (i.e., no type for subj is in subsumption relation with subjectType), objectType is a minimal type of the obj and subj is linked to obj through pred or any other super-property of pred , at this moment defining a clear distinction between patterns (a redundant pat-

Table 1: The distribution of the categories with frequency.

Category	Frequency
Remove	216
Put to the bottom	179
None	306

Table 2: Accuracy, Macro, and Weighted average for the different machine learning methods; RF = Random Forest; LinearSVC = Linear Support Vector Classifier; LR = Logistic Regression; MultinomialNB = Multinomial Naive Bayes; KNeighbors = K-Nearest Neighbors; SVC = Support Vector Classifier; DT = Decision Tree; MLP = Multi-layer Perceptron Classifier; AdaBoost = Adaptive Boosting Classifier.

ML methods	Accuracy	Macro avg	Weighted avg
RF	0.49	0.45	0.49
LinearSVC	0.48	0.45	0.45
LR	0.48	0.45	0.45
MultinomialNB	0.48	0.44	0.44
KNeighbors	0.43	0.38	0.38
DT	0.49	0.46	0.46
MLP	0.49	0.46	0.45

tern set) and minimal patterns. We will henceforth refer to minimal patterns as patterns. In addition, statistics such as the frequency of how many assertions in the dataset are represented by each pattern are also extracted. (Spahiu et al., 2016) describes the details of this KG profiling tool. The pruning effect of minimization becomes more effective when at the same time, ontologies encode a rich type hierarchy, and entities are primarily associated with many types (e.g., DBpedia). However, since ABSTAT is designed to summarize assertions in the KG while maintaining full coverage of them, it could be that a KG featuring many entities without a type and with a poor (absent) type hierarchy, fed to ABSTAT, leads to a summary with some pattern which may not be informative to the user because of its high generality.

3 Methods

The motivation for post-processing is to suppress the patterns that contain overly general namespaces or individual schema IRIs, so that, ideally, only patterns expressing ontological relationships properly characterizing the KG are left (thus also reducing the overall size of the pattern set) or at least prioritized in the list.

Input data To create the input dataset for manual labelling, we generated a list of frequent KGs patterns produced by ABSTAT (as stored in its

database), and collected the IRIs of all entities appearing in them. This became a basis for a table to be used by human annotators, which contained 700 randomly picked entities. Three annotators (from among the paper authors) eventually labelled about 400-500 of them each, using a set of three labels: “None”, “Put to the bottom”, and “Remove”. A single label for each IRI was obtained by majority vote. The frequency count of the ultimate values is in Table 1.

Entity representation The Term Frequency and Inverse Document Frequency (TF-IDF) is one of the most popular text representation methods, widely employed in numerous previous studies. To construct the TF-IDF input data table, our experiment used the unigrams and bigrams extracted from the (parsed) entity IRI.

Machine learning methods We used the random forest (Breiman, 2001), linear support vector classifier (Suthaharan and Suthaharan, 2016), logistic regression (LaValley, 2008), multinomial naive bayes (Xu et al., 2017), K-Nearest neighbors (Peterson, 2009), decision tree (Safavian and Landgrebe, 1991) and multi-layer perceptron classifier (Ramchoun et al., 2016) implementation from the scikit-learn library, with hyperparameter optimization (see Table 3). We also utilized the k-fold cross-validation from the scikit-learn. It provides cross-validation with grid search hyperparameter

Table 3: Overview of input Parameter grid (Optimal configurations are bold).

Machine learning algorithm	Parameter grid
Random Forest	'n_estimators': [100, 200 , 300], 'max_depth': [2 , 5, 10], 'min_samples_split': [2, 5, 10], 'min_samples_leaf': [1, 2 , 4]
Linear Support Vector Machine	'C': [0.1 , 1, 10], 'loss': ['hinge', ' squared_hinge '], 'max_iter': [1000 , 2000, 3000]
Logistic Regression	'C': [0.1, 1 , 10], 'solver': [' liblinear ', 'saga'], 'max_iter': [100 , 200, 300]
MultinomialNB	'alpha': [0.1, 1 , 10], 'fit_prior': [True , False]
KNeighbors	'n_neighbors': [3, 5 , 7], 'weights': ['uniform', ' distance '], 'algorithm': [' auto ', 'ball_tree', 'kd_tree', 'brute']
Decision Tree	'criterion': [' gini ', 'entropy'], 'max_depth': [None , 5, 10, 15], 'min_samples_split': [2 , 5, 10], 'min_samples_leaf': [1 , 2, 4], 'max_features': [' auto ', 'sqrt', 'log2']
MLP	'hidden_layer_sizes': [(10), (50), (100)], 'activation': ['relu', ' tanh '], 'solver': [' adam ', 'sgd'], 'alpha': [0.0001 , 0.001, 0.01], 'learning_rate': ['constant', ' adaptive ']

optimization via the GridSearchCV¹ classes.

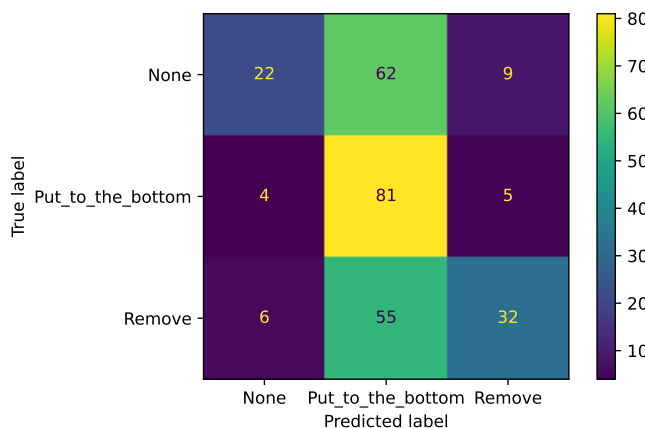


Figure 1: Confusion matrix for the Random Forest model.

4 Results and Discussion

For the ML methods, We used 70% training data and 30% test data by random sampling. We also observed that the dataset was imbalanced (cf. Table 1). To overcome the imbalance issue, we utilized the oversampling method (Chawla et al., 2002). The overall accuracy was used to evaluate the ML

¹scikit-learn-GridSearchCV

methods, but we also computed the per-class accuracy. Table 2 shows the Accuracy, Macro average, and Weighted average of the different ML methods for testing data. From Table 2, the random forest method outperforms with 0.49 accuracy, like the decision tree and multi-layer perceptron method. The linear support vector method, logistic regression, and multinomial naive bayes methods also achieved similar performance with 0.48 accuracy. The confusion matrix (in Fig. 1) also assesses the performance of the random forest method for this experiment. It concisely represents the model's predictions, enabling a detailed analysis of each class's classification accuracy and error rates.

We also processed all the KGs by ABSTAT; since we worked with the public web application, which has a maximum KG upload limit of 10 GB, this reduced the number of KGs. More precisely, the KGs used to analyze the post-processing effect comes from different domains (such as linguistics, COVID-19, etc.) are listed in Table 3. We observe that KGs are very heterogeneous; for instance, there are KGs that barely or do not at all provide types for entities.

Once profiles are computed, ABSTAT returns a set of patterns. Then we applied customizable heuristic post-processing relying on the best ML method (from Table 2). For each ML method, the

Table 4: Patterns before and after post-processing with ML vs. manual patterns (linguistic and COVID-19 KGs).

KG name	Before	Post-processing with ML	Post-processing with manual patterns
basque-eurowordnet-lemon-lexicon-3.0	74	32	47
catalan-eurowordnet-lemon-lexicon-3.0	78	32	47
dbpedia-spotlight-nif-ner-corpu	52	5	37
apertium-rdf-ca-it	15	2	2
wordnet	39	35	36
wn-wiki-instances	4	0	0
asit-data	67	27	52
Reuters-128	21	1	15
lemonwiktionary	19	0	0
apertium-rdf-fr-ca	15	2	0
SimpleEntries	4752	2533	4445
news-100-nif-ner-corpus	21	1	15
drugbank	1408	13	13
pro-sars2	12	0	0
COKG-19-Schema	7	0	0
cord19-akg	108	55	55

post-processing tool provides the options “None”, “Put to the bottom” and “Remove”, and applies them to the results. For example, Table 4 presents the pattern frequency difference for the KGs upon application of the “Remove” option with the random forest ML method and manual post-processing. The difference is tiny for some KGs, such as WordNet, Drugbank, etc. In contrast, it is quite significant for most others, outliers being SimpleEntries or Asit-data, with much larger reduction obtained using ML than using the manual method. We primarily aimed to reduce the number of patterns in this study; the option “Put to the bottom” is also offered by the ML-based post-processing tool since even patterns containing generic concepts and datatypes can be interesting, particularly for the subsequent detailed scrutiny of a chosen dataset. After familiarizing with the essential nature of a KG, the user may wish to study even such ‘de-prioritized’ patterns at the bottom of the list.

From Table 4, we can say that, for most of the KGs, with the ML and manual methods, ABSTAT pattern post-processing has a huge impact. Also, ML and manual methods of post-processing have significant differences. The top patterns before and after post-processing are available from an auxiliary page ².

²[ABSTAT-patterns-post-processing-with-ML](#)

The post-processing is even more significant for the ML-based approaches than the manual approach, although the number of KGs is too small to make ultimate conclusions. Also, we observed that the dataset that we utilized for the ML methods has a higher effect on (1) KGs with a very low percentage of typing assertions as ABSTAT by default assigns `owl:Thing` as the type for un-typed entities and (2) KGs with a majority of data type relational assertions as many of the elements in the dataset.

5 Conclusions and future work

The experiment suggests that simple heuristics leading to the suppression of patterns containing generic concepts or datatypes might improve the output of state-of-art profiling tools with different ML methods in the context of rapid skimming of multiple KGs.

The present method of training dataset construction primarily relied on manual labeling of the *individual entities* (complemented by whole namespaces, whose pruning is primarily relevant for meta-level vocabularies such as RDF, OWL, or SKOS). However, we are aware that the interestingness of a pattern may be estimated more precisely based on whole pattern triples. We also plan to apply manual labeling at the pattern level. However, the much

larger combinatorial space to be covered will require a significantly increased labor force, possibly recruited via a crowd-sourcing platform.

While the experiment was carried out via a separate ML-based post-processing tool, we will explore how a similar functionality could be achieved within ABSTAT without compromising its current user experience or risking inadequate information loss. Additionally, the dataset utilized by the different ML methods was small; we could also consider enriching the dataset in the future. Also, the generic concepts that occur in many KGs could be eliminated by applying a threshold value on the *inverse KG frequency* (analogous to the common IDF metric).

Acknowledgments

The research was supported by CHIST-ERA within the CIMPLE project (CHIST-ERA-19-XAI-003), and by Nexus Linguarum (COST Action CA18209).

References

- Leo Breiman. 2001. Random forests. *Machine learning*, 45:5–32.
- Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.
- Michael P LaValley. 2008. Logistic regression. *Circulation*, 117(18):2395–2399.
- Nandana Mihindukulasooriya, María Poveda-Villalón, Raúl García-Castro, and Asunción Gómez-Pérez. 2015. Loupe-an online tool for inspecting datasets in the linked data cloud. *ISWC (Posters & Demos)*, 1:1.
- Leif E Peterson. 2009. K-nearest neighbor. *Scholarpedia*, 4(2):1883.
- Gollam Rabby, Farhana Keya, Vojtech Svatek, and Renzo Arturo Alva Principe. 2022. [Effect of heuristic post-processing on knowledge graph profile patterns: cross-domain study](#). In *ProLingKnower 2022*. Published on Zenodo.
- Hassan Ramchoun, Youssef Ghanou, Mohamed Ettaouil, and Mohammed Amine Janati Idrissi. 2016. Multilayer perceptron: Architecture optimization and training.
- S Rasoul Safavian and David Landgrebe. 1991. A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, 21(3):660–674.
- Blerina Spahiu, Riccardo Porrini, Matteo Palmonari, Anisa Rula, and Andrea Maurino. 2016. Abstat: ontology-driven linked data summaries with pattern minimalization. In *The Semantic Web: ESWC 2016 Satellite Events, Heraklion, Crete, Greece, May 29–June 2, 2016, Revised Selected Papers 13*, pages 381–395. Springer.
- Shan Suthaharan and Shan Suthaharan. 2016. Support vector machine. *Machine learning models and algorithms for big data classification: thinking with examples for effective learning*, pages 207–235.
- Shuo Xu, Yan Li, and Zheng Wang. 2017. Bayesian multinomial naïve bayes classifier to text classification. In *Advanced Multimedia and Ubiquitous Engineering: MUE/FutureTech 2017 11*, pages 347–352. Springer.

**Sentiment Analysis and
Linguistic Linked Data
(SALLD)**

Sentiment analysis with emojis: a model for Brazilian Portuguese

Vinícius Moitinho da Silva Santos

Raquel Meister Ko Freitag

Hector Julian Tejada Herrera

Ayla Santana Florêncio

Pedro Paulo Oliveira Barros Souza

Túlio Sousa de Gois

Federal University of Sergipe

Abstract

Sentiment analysis in multimodal texts that include emojis is a complex task because of a lack of tools and cultural specificities. There are some options available (VADER or the emoticons library), but most fail to perform this analysis accurately and efficiently in languages other than English, such as Brazilian Portuguese. This study presents a model based on the sum of polarities to contribute to the improvement of sentiment analysis in different languages. A set of the 100 most used emojis in 2021 by Unicode (Daniel, 2021) was judged by a group ($n = 13$) into three categories: positive, negative, and neutral. Based on the agreement results, a sentiment analysis model using Python was run, which consisted of summing the polarities of the emojis. Two training databases from Twitter: one in Brazilian Portuguese about the Brazilian elections of 2019 ($n = 61,590$) and another in English about the 2022 World Cup ($n = 22,525$). After the filter to exclude tweets without emojis from the agreement test was applied, a dataset with 511 tweets in Brazilian Portuguese and 2,531 tweets in English was used. A sentiment analysis model was run with the datasets to classify the sentiments based on the sum of polarities developed in the previous stage. The results were compared with those of VADER (Hutto, C.J., & Gilbert, 2014), a natural language processing tool that has been validated by linguists and data scientists for performing similar tasks. The results show that the new model for Brazilian Portuguese has slightly lower performance. However, for English, the new model had an accuracy of 51% compared to VADER's accuracy of 41%. This suggests that the new model may be a useful tool for sentiment analysis in English texts containing emojis. Improvements in Brazilian Portuguese are required to broaden the accuracy of sentiment analysis in texts that include emojis. In addition, it is necessary to expand the range of emojis covered by the model and perform classification using machine-learning techniques,

which may further improve the accuracy of the model. This study was developed following Open Science standards, with data and code available to the scientific community for enhanced transparency and reproducibility, while also promoting the digital inclusion of nonhegemonic languages such as Brazilian Portuguese.

1 Introduction

The area of natural language processing has considered only the textual linguistic clue, without any support for embodied resources that make up the expressivity of human language (Bühler, 2011), like facial gestures, body gestures and prosodic signals, which make up, together with the linguistic clue, the human interaction. The linguistic system has textual resources of expressiveness, such as punctuation marks (exclamation marks, quotation marks, italics, and bold, for example), but they are not always enough to express the demand for meaning related to the emotional state of users, especially in written interaction situations. It is in this context that multimodal resources emulate the expressive dimension of natural human language, with its embodied resources. This makes that in situations of written interaction, such as in social networks and microblogs like Twitter, the use of visual resources such as emoticons, emojis and memes have been recurrent in the construction of meaning, especially feelings.

One of the most widely used features, not only on Twitter but also in other social networks and instant messaging applications, are emojis (Bai et al., 2019). An emoji is classified as a pictogram or ideogram, an image that conveys a message. When it comes to communicative expressiveness, emojis are often associated with representing the emotional state of their users (Alexandrino, 2016). The search for cues of emotional states in written texts has been the focus of sentiment analysis, which determines the polarity of texts based on values associated with each word. Going beyond the lin-

guistic clue, the use of emojis can also assist in the classification of texts according to their polarity (Cavalcante, 2017), especially in situations of irony where the linguistic clue does not fully express the user's sentiment. Another important aspect to consider is that emotions and embodied resources are sensitive to cultural context (Tejada et al., 2022) and are not universal; therefore, the interpretations and meanings attributed to emojis can vary, requiring specific libraries for each language.

This paper presents the procedures for a sentiment analysis that, considering the expressive nature of emojis in social networks, includes emojis in the polarity classification, specifically, the construction of a lexicon dictionary composed of emojis and their respective polarities, validated on a dataset of Brazilian Portuguese, a language still underrepresented in terms of technologies for natural language processing.

This work aims to develop and validate a lexical dictionary to identify the polarity of emojis, aiming for the implementation of a sentiment analysis model, the *EmojiMapper*. Through the presence of emojis in texts, the tool will be able to assign a polarity to the sentence based on the balance of the polarities of the present emojis. Moreover, the model will be validated through tests using previously analyzed datasets and comparing the results with another sentiment analysis model that has support for emojis, using VADER.

The paper is divided as follows: section 2 presents the theoretic foundation on sentiment analysis, its approaches and the VADER analyzer; section 3 deals with the methodology applied for the development of the proposed model, elucidating tools, techniques and used databases; section 4 presents the obtained results; and finally, section 5 deals with the authors' conclusions, based on the results, and presents possible future work.

2 Sentiment Analysis

Sentiment analysis (SA) is a process that seeks to identify and categorize, using computational methods, the emotions, opinions, and attitudes people express through text (Medhat et al., 2014). Considered a type of text classification, SA is an important part of Natural Language Processing, a field of study resulting from the intersection between linguistics and computation that mainly deals with the linguistic interaction between human and machine (Devika et al., 2016).

SA involves the lexicon-based approach and the machine learning-based approach (Bonta and Janardhan, 2019). For the inclusion of emojis in SA, the lexicon-based approach, with the VADER tool, was the starting point.

The lexicon-based approach uses the classification of each lexicon item for sentiment to describe the polarity of a textual content, which can be positive, negative, or neutral. The classification of items can be dictionary-based or corpus-based (Sadia et al., 2018).

The construction of the lexicon starts by compiling the words of interest and assigning their respective polarities. In the case of lexicon dictionaries, the construction of the list is initially performed manually, with the collection and classification of the objects of interest, creating a dictionary-like structure containing the object (word or symbol) and their polarity (Bonta and Janardhan, 2019). Unlike the corpus-based approach, the dictionary does not consider the context of the selected objects. However, the dictionary-based list allows the selection of specific terms from a field, while the more comprehensive corpus-based approach considers a large volume of data in different contexts, and may lose precision.

Starting from the lexicon-based approach, it is possible to find some well described and tested tools in the literature, as discussed in Bonta et al. (2019). However, considering the model proposed in this paper, the tool that has parameters able to be compared and tested is VADER (Hutto and Gilbert, 2014).

VADER (Valence Aware Dictionary for Sentimental Reasoning) is a model built for sentiment analysis that uses quantitative and qualitative methods, combining a list of lexicon attributes, and syntactic and grammatical rules (Hutto and Gilbert, 2014). Unlike other tools, VADER can also assign polarities to emojis and demonstrates good performance in texts originating from social networks (Bonta and Janardhan, 2019).

3 Methodology

3.1 Lexicon dictionary

The construction of the lexicon dictionary used in the proposed model started with the selection of the 100 most used emojis in 2021 according to Unicode. For the classification of the selected data, a concordance test was performed in which expert judges ($n = 13$) rated their perception of

polarity for each individual symbol, which could assume three different values presented in Table 1. After the individual categorization, a table was constructed containing the emoji identifications and their respective polarities established based on the majority choice.

After the construction of the lexical dictionary, a coding step was performed in Python language to develop the application. To build the tool, the following functions were implemented. In the tool, functions were implemented to do the cleaning of datasets and individual texts, classify an input set based on the polarity of the emojis present in the text, and validate the result. Validation occurs through a routine that calculates the accuracy of the model against the test data.

3.2 Database

Once the lexical dictionary was completed and integrated into the python script, two datasets consisting of tweets previously classified based on the polarity presented in each text were selected. The data was obtained from the Kaggle platform and used to test and validate the EmojiMapper tool.

The first dataset selected was 'Twitter in Portuguese - Elections 2019¹,' which initially contained texts from 61.590 tweets in Brazilian Portuguese, focusing on the 2019 elections as the central theme. The second selected database was 'FIFA World Cup 2022 Tweets²,' which contained 22.525 tweets in English, with the central theme being the 2022 World Cup. A different language was chosen to observe how the model performed on datasets with diverse natures.

Both databases were filtered using an internal function of EmojiMapper called 'cleanData.' This function takes the database to be filtered as a parameter and returns a new dataset suitable for the model application. The filtered datasets only contain tweets that have one or more of the 100 emojis mapped in the lexical dictionary step. After filtering, two sub-databases were obtained: one derived from the first dataset, containing 511 tweets (Dataset 1), and another derived from the second dataset, containing 2,531 tweets (Dataset 2)."

¹<https://www.kaggle.com/datasets/adilmar/twitter-pt>

²<https://www.kaggle.com/datasets/tirendazacademy/fifa-world-cup-2022-tweets?resource=download>

Classification	Start of range	End of range
Negativo	-1	-0.05
Neutro	-0.05	0.05
Positivo	0.05	1

Table 1: Rating ranges

3.3 Experiment

After filtering the data, it was possible to apply it to the model and evaluate its accuracy on the test sets. To establish a comparison parameter with another similar tool, we utilized VADER (Hutto and Gilbert, 2014). The classification of each database was performed using EmojiMapper's internal function called 'classify.' This function takes the filtered dataset as a parameter and returns a list containing the predicted responses, classified as positive, negative, or neutral. The same procedure was carried out using VADER, and the data classification thresholds for this tool are shown in Table 1. After the classification of the data by both models, a performance metric analysis was conducted.

4 Results

After applying the model to the test sets, metrics were obtained to compare the effectiveness of the tools. Table 2 presents the results obtained from the experiment. It can be observed that EmojiMapper had lower accuracy than VADER for Dataset 1, while the opposite was true for Dataset 2, favoring EmojiMapper. A possible cause for the discrepancy between the metrics may be related to the implementation of each tool. Unlike EmojiMapper, VADER does not directly assign polarity values to emojis. Instead, it employs a methodology to describe the emoji and assigns polarity to the descriptive terms.

Considering this hypothesis, it is possible to explain the phenomenon of VADER's better performance in Dataset 1, as it consists of texts related to politics, and the nature of this dataset is ironic, this can be verified by analyzing the *provaIronia.csv* dataset present in the tool repository. By examining the emoji description, VADER obtains a well-defined context of the message, while EmojiMapper tends to interpret the symbol literally. However, it is worth noting that in datasets without a predominance of irony, EmojiMapper performs better. This can be attributed to EmojiMapper considering the literal meaning of the emoji and its strong correlation with the text, making it an indicator of the

Dataset	EmojiMapper	VADER
Dataset 1	44%	46%
Dataset 2	51%	41%

Table 2: Model Accuracy

message's polarity.

The model used in this study is available at <https://github.com/vmoitinhoss/Emojimapper> and can be accessed freely, adhering to the principles of open science.

5 Conclusion

Based on the results obtained in the experiment, it can be concluded that EmojiMapper demonstrates itself as a viable and effective alternative for performing sentiment analysis on datasets without an ironic nature. It may even outperform a validated and prestigious tool. However, it is important to note that this work is still a proof of concept, as its scope of application is limited to texts that contain one or more of the 100 emojis present in the model. Moreover, an improvement is urgently needed to better deal with the irony phenomenon, considering the relationship between joint emojis.

For future research, it would be interesting to explore the feasibility of machine learning methods for weighting the importance of emojis based on the nature of the data. This approach could help overcome the performance issues encountered in datasets with an ironic nature.

References

- Ana Débora Oliveira Alexandrino. 2016. Análise de redes sociais aplicada a tweets sobre séries de tv.
- Qiao Bai, Qibin Dan, Zhonghai Mu, and Meina Yang. 2019. A systematic review of emoji: Current research and future perspectives. *Frontiers in psychology*, 10:2221.
- Vandana Bonta and N. K. N. Janardhan. 2019. A comprehensive study on lexicon based approaches for sentiment analysis. *Asian Journal of Computer Science and Technology*, 8(S2):1–6.
- Karl Bühler. 2011. Theory of language. the representational function of language, translated by donald fraser goodwin in collaboration with achim eschbach. *Amsterdam/Philadelphia, Benjamins*.
- Pedro Emanuel Cunha Cavalcante. 2017. Um dataset para análise de sentimentos na língua portuguesa.
- M. D. Devika, C. Sunitha, and A. Ganesh. 2016. Sentiment analysis: a comparative study on different approaches. *Procedia Computer Science*, 87:44–49.
- Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, pages 216–225.
- Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113.
- Ayesha Sadia, Faiza Khan, and Fahad Bashir. 2018. An overview of lexicon-based approach for sentiment analysis. In *2018 3rd International Electrical Engineering Conference (IEEC 2018)*, pages 1–6.
- Julian Tejada, Raquel Meister Ko Freitag, Bruno Felipe Marques Pinheiro, Paloma Batista Cardoso, Victor Rene Andrade Souza, and Lucas Santos Silva. 2022. Building and validation of a set of facial expression images to detect emotions: a transcultural study. *Psychological Research*, 86(6):1996–2006.

SLIWC, Morality, NarrOnt and Senpy Annotations: four vocabularies to fight radicalization

J. Fernando Sánchez-Rada and Oscar Araque and Guillermo García-Grao
and Carlos Á. Iglesias

{jf.sanchez,o.araque,g.ggrao,carlosangel.iglesias}@upm.es

Department of Telematic Engineering Systems

Intelligent Systems Group (GSI)

ETSI de Telecomunicación, Universidad Politécnica de Madrid

Abstract

This paper describes the vocabularies used in PARTICIPATION, a Horizon2020-funded project aimed at preventing extremism, radicalization, and polarization. To fully take advantage of Linked Data, all data in the project need to be expressed in a semantic format, and all annotation services should be accessible through a semantic API. Most of the data can be expressed by extensively leveraging common vocabularies in the Linguistic Linked Data sphere. However, certain key concepts were not present in any of the popular vocabularies, such as ideologies, morality, and narratives. Some types of analysis also required the use of resources aligned with Linguistic Inquiry and Word Count (LIWC) software. As a result, four vocabularies were developed: Senpy Annotations, SLIWC, Morality, and NarrOnt. Senpy Annotations is a vocabulary designed to represent any kind of annotation in the context of NLP services and resources. SLIWC is a vocabulary and SKOS taxonomy that aims to represent LIWC dimensions. The NarrOnt (Narrative Ontology) vocabulary models the concepts of a narrative and an ideology linked to a piece of content. Lastly, morality is a vocabulary for expressing annotations that follow the Moral Foundation Theory (MFT). These vocabularies have been designed and published using Linked Data principles and best practices. Most importantly, they follow an orthogonal design, integrate well with existing vocabularies, and only describe specific parts of a domain. We believe that the usefulness of these vocabularies will extend beyond the scope of this specific project.

1 Introduction

This work stems from efforts to semantically annotate resources and services in the context of PARTICIPATION, a project aimed at detecting and preventing extremism, radicalization, and polarization. According to previous work, the definition of formats and schemas followed a Linked Data

approach to take advantage of all efforts of the Natural Language Processing (NLP) community both in the definition of specific vocabularies and in the integration of different vocabularies for new types of analysis and domains. However, the radicalization domain requires the use of techniques and resources that have not been fully incorporated into the Linguistic Linked Data sphere yet. More specifically, we identified the need to express the domain of ideologies, morality, and narratives, as well as resources aligned with the Linguistic Inquiry and Word Count (LIWC) software.

As a result, we have developed four vocabularies: Senpy Annotations, SLIWC, Morality, and NarrOnt. These vocabularies have been designed and published using Linked Data principles and best practices. Therefore, they follow an orthogonal design, integrate well with existing vocabularies, and describe specific domains. As a consequence, we believe they will prove to be useful beyond the context of this specific project.

2 The Linked Data approach

Part of the work in the project involves several types of data processing and visualization of social media content. This includes several sources, such as microblogging platforms, news sites, and social news aggregators. The majority of the processing involves cleaning, filtering, and automatic annotation. However, the specific processes are varied and constantly evolving to deal with the dynamic nature of online social networks and the multidisciplinary nature of the work.

In order to seamlessly deal with multiple sources of information and provide different types of annotation, all data captured from social media is converted to a common semantic format. All other processes then enrich this data by adding semantic annotations to it. Using a common format allows each process to consume data from multiple sources, regardless of its origin. Modelling

each annotation process as an independent additive process allows future growth. Both of these features could be achieved without Linked Data by modelling data as documents and defining each document property separately. On the other hand, using a Linked Data approach is a better alternative for two main reasons. First, the use of existing work reduces the development and modelling effort. There is already a set of well-known formats, protocols, and libraries, most of which rely on web standards, as well as multiple quality vocabularies to express concepts in most domains. Secondly, properly reusing these works translates into interoperability and compatibility with other projects. Lastly, but most importantly, a Linked Data approach results in data that can be understood not only by humans but also by machines. As proof of the last two points, semantically-annotated data could be easily exposed from an endpoint capable of responding to meaningful queries, such as “where was #example hashtag twitted from on January 1st?”.

One of the downsides of a Linked Data approach is that many vocabularies may be needed to model the different types of data in the platform. Although this increases interoperability, it requires understanding them well. The following is an overview of all of the existing vocabularies used to model the data in the project:

- Semantically-Interlinked Online Communities (SIOC) (Breslin et al., 2006)¹ The SIOC Core Ontology provides the main concepts and properties required to describe information from online communities (e.g., message boards, wikis, weblogs, etc.) on the Semantic Web.
- Schema.org (Guha et al., 2016)² Provides schemas for structured data on the Internet, on web pages, in email messages, and beyond.
- Dublin Core Metadata Initiative (DCMI) (Initiative et al., 2012)³. Provides a model for structured metadata to support resource discovery.
- Marl (Westerski et al., 2011)⁴. Marl is a standardized data schema designed to annotate

¹<http://rdfs.org/sioc/spec/>

²<https://schema.org/>

³<http://purl.org/dc/terms/>

⁴<https://www.gsi.upm.es/ontologies/marl/>

and describe subjective opinions expressed on the Web or in particular Information Systems.

- DBpedia (Auer et al., 2007)⁵. DBpedia is a community project that extracts structured, multilingual knowledge from Wikipedia and makes it freely available on the Web using Semantic Web and Linked Data technologies. In the context of this project, DBpedia serves both as a vocabulary to express properties and, most importantly, as a source of URIs to attach to people, entities, and other encyclopedic knowledge.
- NLP Interchange Format (NIF) (Hellmann et al., 2013)⁶. NIF is an Resource Description Framework (RDF)/Web Ontology Language (OWL)-based format that aims to achieve interoperability between NLP tools, language resources and annotations.
- Onyx (Sánchez-Rada and Iglesias, 2016)⁷. Onyx is a standardized data schema designed to annotate and describe the emotions expressed by user-generated content on the Web or in particular Information Systems.

3 Vocabularies

The vocabularies in the previous section were insufficient to model all the types of annotation necessary for this project. Instead of creating a single vocabulary with all the missing elements, these missing pieces have been separated into smaller individual vocabularies to foster re-usability. To encourage the use of different vocabularies in real-life scenarios, the vocabularies have been grouped under a common umbrella of PARTICIPATION ontologies. They are accompanied by web documentation describing their usage⁸.

When designing a vocabulary, it is often necessary to reach a balance between expressiveness and simplicity. More general vocabularies tend to make use of additional nodes, which translates into more nodes in the knowledge graph. This is usually not a problem, other than having the side effect of making queries slightly more complex

⁵<https://www.dbpedia.org/>

⁶<https://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core/nif-core.html>

⁷<https://www.gsi.upm.es/ontologies/onyx/>

⁸<https://www.gsi.upm.es/ontologies/participation>

and verbose or nested. But this project imposes additional constraints that make such complexity more difficult. One of the main constraint is use of common formats such as JSON-LD. This makes it so that regular document stores can be used to save commonly accessed data, and annotation services can serve their results in a more developer-friendly format. When translating a knowledge graph to a JSON-LD document (a tree), there are certain degrees of freedom. This is done by design to allow for the same data to be represented using different schemas. Nonetheless, a deep graph structure will translate into a deeply nested document. A common design principle for the vocabularies presented is that the complete annotations (see Section 4) remain reasonably shallow.

3.1 Senpy annotations

As explained in Section 2, the semantic model for text representation has been based on earlier work (Sánchez-Rada et al., 2020). Therefore, it heavily employs the NIF 1.0 (Hellmann et al., 2013) vocabulary and adds annotations through external vocabularies such as Marl and Onyx. Past experience has shown that some aspects of these vocabularies related to how were not limited to each specific domain (e.g., emotion annotation) and could be applied to other NLP tasks such as those involved in this paper. A better strategy would be to express these common parts in its own separate vocabulary.

Hence, a decision was made to design a very simple and modular vocabulary for the sole purpose of expressing annotations in text. This new vocabulary, called Senpy annotations, follows a structure similar to that of the newer versions of NIF. But, in contrast with NIF, this vocabulary can be easily adapted to provide a better mapping in the JSON-LD representation.

The vocabulary revolves around the concept of an annotation (`sa:Annotation`). The `sa:Annotation` class is designed to be used to annotate specific entries, as will be shown in Section 4. Any entity (e.g., a tweet, a lexicon entry) can be tagged with an annotation through the `sa:hasAnnotation` property. To differentiate between annotations to a single element (e.g., in a lexicon) and an annotation that applies to a larger piece of text (e.g., the count of words in a sentence), there is a special type of annotation, `sa:AggregatedAnnotation`.

An `sa:AggregatedAnnotation` may specify both how many elements were used in the aggregation (`sa:count`), as well as the ratio of these elements to the total (`sa:ratio`). These classes can be specialized (subclassed) by specific vocabularies for annotation. As an example of this, another vocabulary in this project (which we will explain below) extends Senpy annotations to include the categories in Moral Foundation Theory. In documents, the actual annotations are an aggregate of the individual words/lemmas. Hence, corpora annotations should use the `sa:AggregatedAnnotation`, which also allows quantifying the frequency or ratio of appearance within the text.

3.2 SLIWC

The way in which the Linguistic Inquiry and Word Count (LIWC) (Pennebaker, 2011) program works is fairly simple. Basically, it reads a given text and counts the percentage of words that reflect different emotions, thinking styles, social concerns, and even parts of speech.

An important part of the LIWC project is the LIWC dictionaries. The importance and popularity of LIWC have led other researchers to adopt their annotation conventions and to use the same format to produce dictionaries that are compatible with LIWC programs.

In the Participation project, we have produced a semantic version of the LIWC annotation schema⁹. It reuses the Senpy Annotations ontology to represent the general concepts used in LIWC annotation (e.g., dimensions, categories, word-level dimensions, document-level dimensions, etc.). Then, it uses these concepts to provide elements specific to the LIWC dictionaries, such as specific categories and their hierarchical relation to one another. These categories have been modelled both as an ontology (i.e., classes) and as a SKOS taxonomy so that the hierarchical structure can be exploited independently of the ontological relations.

Using SLIWC to annotate is very simple. To add information about the LIWC category or dimension that is being annotated in a piece of text, an annotation uses the `sa:hasCategory` property, which links to a specific instance in the SKOS taxonomy. The same procedure works both for annotating lexical entries and word-

⁹<https://www.gsi.upm.es/ontologies/participation/sliwc>

level annotations (*Annotation*) as well as for annotating at a more general document-level (*AggregatedAnnotation*). A simplified example of SLIWC annotations is illustrated in Figure 1.

3.3 Morality

The popularity of LIWC has led to several LIWC-like dictionaries in the wild, such as the Moral Foundations Dictionary (Graham et al., 2009)¹⁰, which includes new annotations on morality. The theory proposes that several innate and universally available psychological systems are the foundations of intuitive ethics (Graham et al., 2013). Each culture then constructs virtues, narratives, and institutions on top of these foundations, thereby creating the unique moralities we see around the world and conflicting within nations, too. The main foundations according to this theory are *care/harm*, *fairness/cheating*, *loyalty/betrayal*, *authority/subversion* and *sanctity/degradation*.

In order to use annotations for morality both in resources (dictionaries) and in the results of analyses, we have developed an extension of the Senpy Annotations ontology that includes the concepts defined in the Moral Foundations Dictionary. In particular, it provides a class for moral annotations and categories for each of the extremes in each of the dimensions/foundations. Moreover, each category is linked to its foundation (e.g., *Harm*, *InGroup*) and the relationship of the category to the foundation (*Virtue*, *Vice*). An example of a simple annotation of a tweet can be seen in Figure 2.

3.4 Narrative

The concept of narrative in the NLP community and in the humanities, social, and cognitive sciences is related but generally not synchronized (Piper et al., 2021). However, it is undeniable that recent work on detecting narrative (and counter-narrative) in texts is helping fight extremism and disinformation (Network, 2015; Upal, 2015).

Narrative Ontology (NarrOnt) is a pragmatic model of the ideologies and narratives present in user-generated content, especially on social media. The ontology provides the *Annotation* concept, which directly subclasses *sa:Annotation*. Narratives are represented

with the *Narrative* class. Several narratives are included in the ontology, such as *ProReligion*, *CounterSeparatism*, etc. An example annotation of the narrative in a Tweet is illustrated in Figure 3.

4 Use case

The set of vocabularies in this work has been evaluated in two ways. First, we apply them in different scenarios using real excerpts of data extracted from social networks. The following two sections distill this process using placeholder data, with the main purpose of exemplifying the use of these vocabularies in a more realistic scenario where multiple annotations are needed. The examples will cover two distinct use cases separately: annotating corpora (i.e., set of Tweets with different labels) and annotating lexicons (i.e., dictionaries).

The second means of evaluation for these ontologies is their use in the project: to enable the creation of four different morality and narrative detection services; to automatically annotate more than 1,2 million tweets and 100,000 comments on Reddit using multiple services (including morality and narrative); and to power multiple dashboards for the exploration of radicalism in English, Italian, and Spanish, using the enriched data; to power advanced queries for advanced project partners, using SPARQL.

4.1 Annotation of a corpus of microblogging posts

The annotation of microblogging posts followed a model similar to TweetsKB (Fafalios et al., 2018), a public RDF corpus of anonymized data for a large collection of annotated tweets. As most of the annotated corpora in the Participation project and that of TweetsKB were limited to Twitter, we will refer to microblogging posts as tweets. Nevertheless, the model can be easily applied to any similar platform, such as Mastodon or BlueSky.

In TweetsKB, the information retrieved from a tweet is represented by the *sioc:Post* class. The SIOC Core Ontology, Schema.org and DCMI provide properties and attributes for most of the relevant fields in a tweet, such as the *soic:content* attribute for the text, *soic:has_creator* for the author user, *schema:inLanguage* for the language on which it is written, *schema:mentions* for its hashtags, *dc:created* for the creation date, and

¹⁰<https://moralfoundations.org/other-materials/>

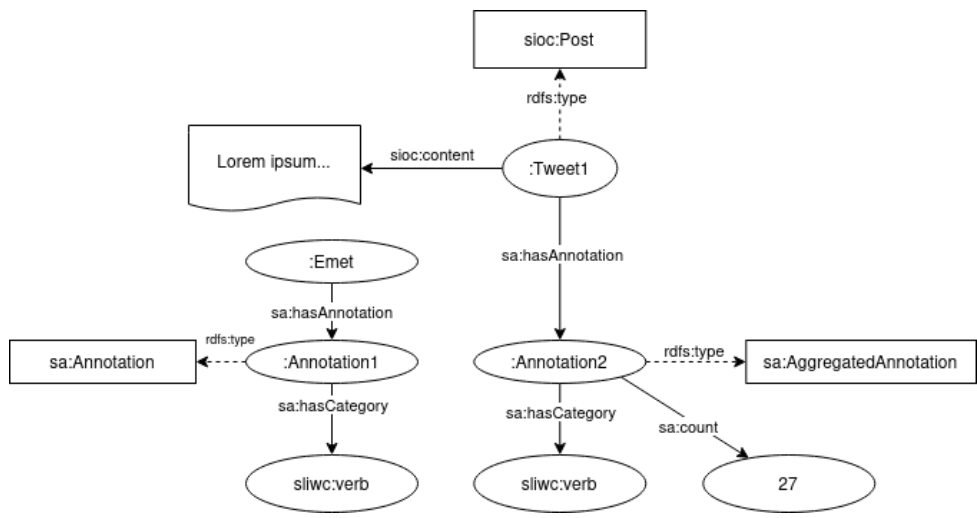


Figure 1: Example of LIWC-aligned annotations of a Tweet.

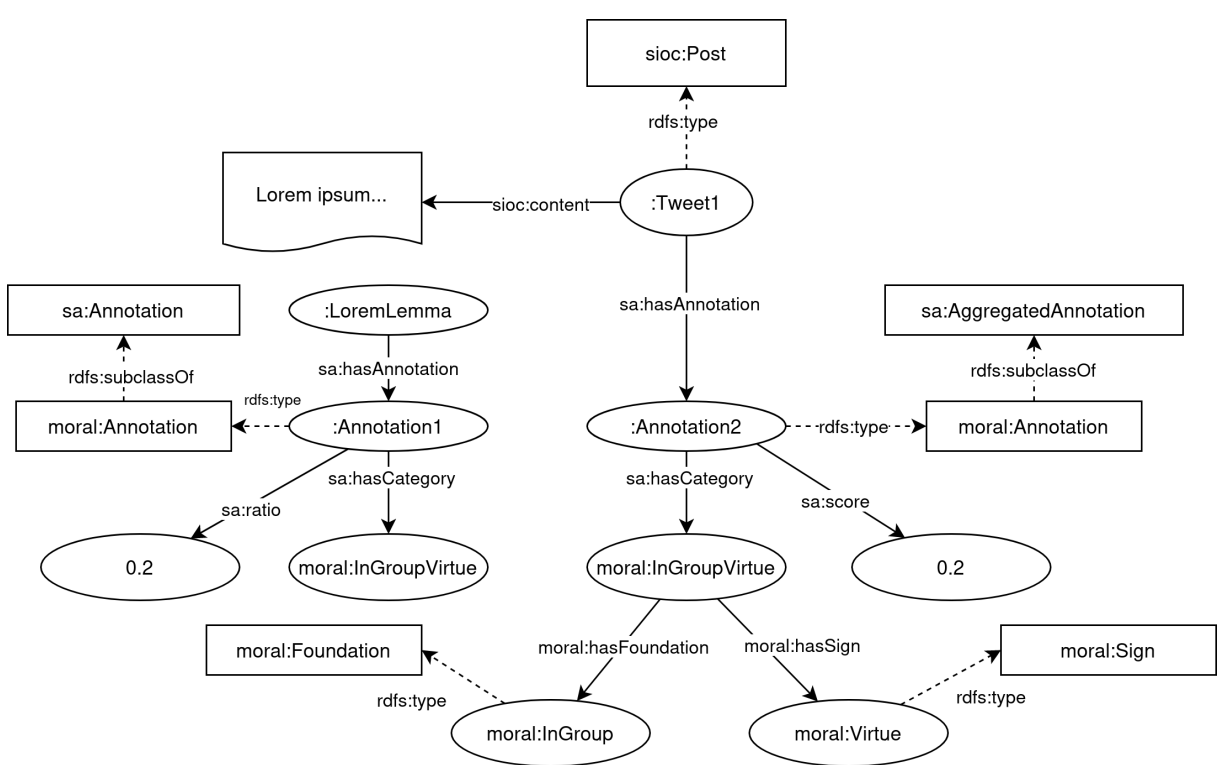


Figure 2: Example of annotation of morality (MFT) in a tweet and in an LIWC-aligned lexicon entry using the Morality ontology.

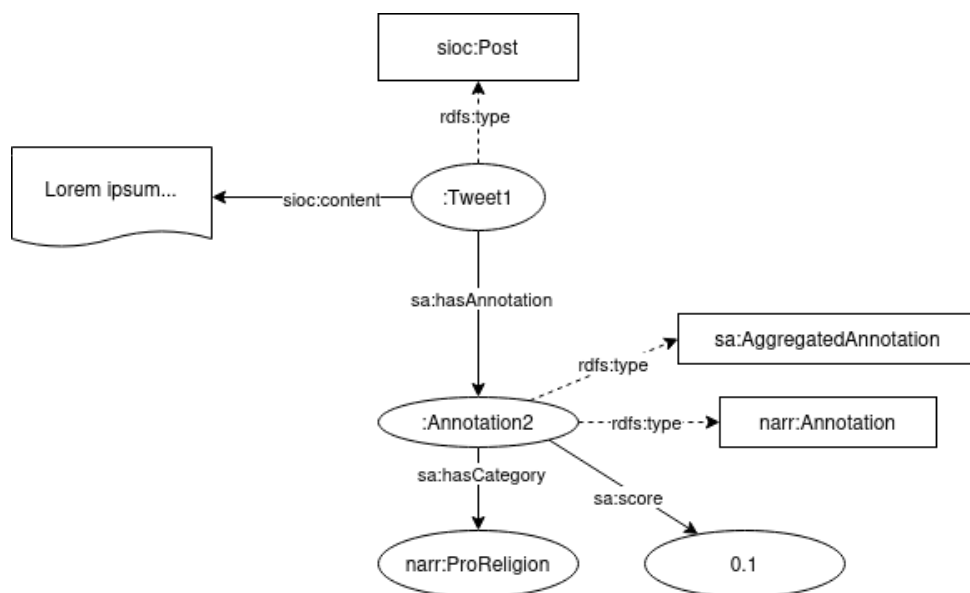


Figure 3: Example of annotation of the narrative in a tweet and a lexicon entry using the Narrative ontology.

`schema:locationCreated` for the location it was posted from.

Lastly, tweets can also be annotated with emotion labels, which are represented using the Onyx ontology. An element is annotated with emotions through the `onyx:EmotionSet` class and the `onyx:hasEmotionSet`.

An `onyx:EmotionSet` is comprised of one or more emotions, defined as `onyx:Emotion`, where the properties `onyx:hasEmotionCategory` and `onyx:hasEmotionIntensity` represent the type of emotion and value, respectively. Finally, the `nif:isString` property from the NIF ontology is used to provide compatibility with other NLP services.

A complete example of the annotation of a tweet can be observed in Figure 1.

It is important to note that the above example can be trivially translated into a mostly flat tree structure, making it ideal for representation as a JSON-LD document.

4.2 Annotating a lexicon

The annotation of a lexicon is very similar to that of a tweet. In this case, the difference is that lexical entries are represented using the lemon ontology. An example annotation of a lexicon can be observed in Figure 2.

4.3 Semantic queries

The data in the project is available to experts through an instance of Fuseki, allowing them to perform semantic queries through SPARQL Protocol and RDF Query Language (SPARQL).

For instance, it is possible to write a query that returns the narrative of every tweet that contains words from a specific Linguistic Inquiry and Word Count (LIWC) category, as well as the ratio at which that category appears. Figure 4 shows such a query, with the LIWC category of Death. An excerpt of the results returned by Fuseki can be observed in Figure 5.

Figure 4: Example SPARQL query that fetches the ratio of the LIWC Death annotation for each narrative.

Listing 1: Example of annotation of a corpus entry

```

@prefix sa: <http://www.gsi.upm.es/ontologies/participation/senpy/ns
↪ #> .
@prefix sliwc: <http://www.gsi.upm.es/ontologies/participation/sliwc/
↪ ns#> .
@prefix narr: <http://www.gsi.upm.es/ontologies/participation/
↪ narrative/ns#> .
@prefix moral: <http://www.gsi.upm.es/ontologies/participation/moral/
↪ ns#> .

:Tweet1 a sioc:Post ;
  sa:hasAnnotation [
    a sa:AggregatedAnnotation ;
    a narr:Annotation ;
    sa:hasCategory narr:ProReligion ;
    sa:ratio 0.1 .
  ] ;
  sa:hasAnnotation [
    a sa:AggregatedAnnotation ;
    sa:hasCategory moral:IngroupVirtue ;
    sa:ratio 0.1 ;
  ] ;
  sa:hasAnnotation [
    a sa:AggregatedAnnotation ;
    sa:hasCategory sliwc:Filler ;
    sa:ratio 0.34 ;
    sa:count 23 .
  ] ;
  sa:hasAnnotation [
    a sa:AggregatedAnnotation ;
    sa:hasCategory sliwc:Adverb ;
    sa:ratio 0.15 ;
    sa:count 11 .
  ] .

```


Listing 2: Example of lexicon annotation

```

@prefix sa: <http://www.gsi.upm.es/ontologies/participation/senpy/ns
↳ #> .
@prefix sliwc: <http://www.gsi.upm.es/ontologies/participation/sliwc/
↳ ns#> .
@prefix moral: <http://www.gsi.upm.es/ontologies/participation/
↳ morality/ns#> .

_:compassion a lemon:Lexicalentry;
  lemon:sense [
    lemon:reference wn:synset-fear-noun-1;
    sa:hasAnnotation [
      a sa:Annotation, moral:Annotation ;
      sliwc:hasCategory moral:IngroupVirtue .
    ] .
  ] ;
  sliwc:hasAnnotation [
    a sa:Annotation, moral:Annotation ;
    sliwc:hasCategory moral:IngroupVirtue .
  ] ;
  lexinfo:partDfSpeech lexinfo:noun .

```

	ratio	ideology
1	"4.166667e-02^^xsd:double	narr:Far_right
2	"7.142857e-02^^xsd:double	narr:Far_right
3	"5.405405e-02^^xsd:double	narr:Far_right
4	"4.761905e-02^^xsd:double	narr:Far_right
5	"2.222222e-02^^xsd:double	narr:Far_right
6	"2.564103e-02^^xsd:double	narr:Far_right

Figure 5: Part of the triples returned by the query from Figure 4.

Another example, displayed in Figure 6, demonstrates how to get the text of all tweets from a specified narrative, specifically pro far-right. This query also requests the moral categories present in the text and their ratios. It also orders the results by ascending date. Figure 7 shows a fragment of the result from that query.

5 Conclusions and future work

This work shows a successful use case of semantically annotating resources using a mixture of existing vocabularies and ad-hoc vocabularies for niche or otherwise unexplored domains. In particular, four vocabularies have been presented, which can be used independently or in conjunction. When analyzed in isolation, these vocabularies are rather simple by design. But their true power lies in their composition and orthogonal design, which

```

11 SELECT ?text ?category ?ratio
12 WHERE {
13   ?subject sa:hasAnnotation [
14     a narr:Annotation ;
15     sa:hasCategory [ a narr:ProFar_right ] ;
16   ] ;
17   dc:created ?date ;
18   nif:isString ?text ;
19   sa:hasAnnotation [
20     a moral:Annotation ;
21     sa:hasCategory ?category ;
22     sa:ratio ?ratio
23   ]
24 }
25 ORDER BY ASC(?date)

```

Figure 6: Example SPARQL query that fetches the text of every Pro far-right tweet and their moral values.

is a testament to the power of the Linked Data approach. Although these vocabularies have been conceived with the main use case of fighting radicalism in the PARTICIPATION project, they have also been designed with extensibility, composability, and reusability in mind. We hope that this work will inspire other researchers to use these vocabularies, extend them, and share their results with the community.

	text	category	ratio
1	"<user> <user> <hashtag> happynewyear<number> to my christ believing brother mike pompeo, as indeed to vp <allcaps> <user> and to the greatest president in living history <hashtag> maga <allcaps> <allcaps> <user> also to all americans and my fellow english patriots and believers in the truth! god bless you! <url>"	moral:IngroupVirtue	"4.347826e-02"^^xsd:double
2	"make the start-up nation a vacci-nation! <hashtag> vaccination <hashtag> better <hashtag> maga"	moral:IngroupVirtue	"2.222222e-01"^^xsd:double
3	""remember, comrades, your resolution must never falter. no argument must lead you astray" george <allcaps> orwell <allcaps> - animal farm. <hashtag> lockhimup <hashtag> trumptapes <hashtag> trumptreason <hashtag> maga <allcaps> <allcaps> <hashtag> gop <allcaps> <allcaps> <hashtag> trumpmarch <hashtag> trumpsupporters <url>"	moral:IngroupVirtue	"3.448276e-02"^^xsd:double
4	"in <hashtag> hannover sprach <hashtag> javidkistel am wochenende von nanopartikeln in <hashtag> corona-impfstoffen: „mit denen können sie euch steuern.“ unter den teilnehmern bei <hashtag> h<number> waren zudem sympathisanten der „<hashtag> querdenken“-bewegung und „<hashtag> qa <allcaps>non" <hashtag> dud <allcaps><number> <hashtag> impfgegner <url>"	moral:HarmVirtue	"2.777778e-02"^^xsd:double

Figure 7: Part of the response from the query in Figure 6.

Acknowledgements

This work is part of the PARTICIPATION project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 962547.

References

- Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The Semantic Web*, pages 722–735. Springer.
- John G Breslin, Stefan Decker, Andreas Harth, and Uldis Bojars. 2006. Sioc: an approach to connect web-based communities. *International Journal of Web Based Communities*, 2(2):133–142.
- Pavlos Fafalios, Vasileios Iosifidis, Eirini Ntoutsis, and Stefan Dietze. 2018. Tweetskb: A public and large-scale rdf corpus of annotated tweets. In *European Semantic Web Conference*, pages 177–190. Springer.
- Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, volume 47, pages 55–130. Elsevier.
- Jesse Graham, Jonathan Haidt, and Brian A Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5):1029.
- Ramanathan V Guha, Dan Brickley, and Steve Macbeth. 2016. Schema. org: Evolution of structured data on the web. *Communications of the ACM*, 59(2):44–51.
- Sebastian Hellmann, Jens Lehmann, Sören Auer, and Martin Brümmer. 2013. Integrating NLP using linked data. In *International Semantic Web Conference*, pages 98–113. Springer.
- Dublin Core Metadata Initiative et al. 2012. Dublin core metadata element set, version 1.1.
- Radicalisation Awareness Network. 2015. Counter narratives and alternative narratives. *Ran Issue Paper*.
- James W Pennebaker. 2011. Using computer analyses to identify language style and aggressive intent: The secret life of function words. *Dynamics of Asymmetric Conflict*, 4(2):92–102.
- Andrew Piper, Richard Jean So, and David Bamman. 2021. Narrative theory for computational narrative understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 298–311.

- J. Fernando Sánchez-Rada, Oscar Araque, and Carlos A. Iglesias. 2020. [Senpy: A framework for semantic sentiment and emotion analysis services](#). *Knowledge-Based Systems*, 190:105193.
- J Fernando Sánchez-Rada and Carlos A Iglesias. 2016. Onyx: A linked data approach to emotion representation. *Information Processing & Management*, 52(1):99–114.
- Afzal Upal. 2015. Alternative narratives for preventing the radicalization of Muslim youth. *Journal for Deradicalization*, (2):138–162.
- Adam Westerski, Carlos Angel Iglesias, and Fernando Tapia Rico. 2011. Linked opinions: Describing sentiments on the structured web of data. In *SDoW@ ISWC*.

Czech Offensive Language: Testing a Simplified Offensive Language Taxonomy

Olga Dontcheva-Navrátilová

Renata Povolná

Masaryk University, Brno, Czech Republic

navratilova@ped.muni.cz

povolna@ped.muni.cz

Abstract

This contribution presents the results of an annotation campaign carried out on a Czech Corpus of Offensive Language (CCOL) compiled for the purposes of this study. The annotation was based on a Simplified Offensive Language (SOL) Taxonomy (Lewandowska-Tomaszczyk 2022) which has been proposed as part of the research work undertaken within COST Action NexusLinguarum WG 4.1.1. The aim of the study is to test the applicability of the SOL taxonomy to the Czech language, to identify the level of inter-rater agreement for all categories of the taxonomy and to compare the results to an earlier annotation campaign on English Offensive Language within the same research project. The findings of this study hope to support the application of the suggested SOL taxonomy as an ontology for effective detection and encoding of offensive language in Linguistic Linked Open Data (LLOD).

1 Introduction

Online newspaper and social media platforms have created virtual places where people can exchange opinions and views not limited by space constraints. Apart from speeding up the process of production, consumption and sharing information, these platforms have led to the emergence of huge amounts of data and the surge of offensive language (Kennedy et al. 2017, Casselli et al. 2020). As a result, there is a need for the development of methods for the automatic detection of offensive language applicable in LLOD.

In agreement with Lewandowska-Tomaszczyk et al. (forthcoming) offensive language is understood as hurtful, derogatory or obscene utterances produced by one person (or a group of people) to another or to a group of persons (see also Wiegand et al. 2021) with the intention to cause offence or insult. Offensive language, sometimes called abusive or toxic language, or hate speech, refers to the use of explicit language means representing verbal attacks towards individuals or groups of individuals. This paper does not consider visual means although they are natural part of social media platforms and their role in creating offensiveness is generally recognized (see e.g. Lewandowska-Tomaszczyk et al. 2021).

2 Offensive language taxonomy

Several attempts have been made to create an effective offensive language taxonomy (e.g. Basile et al. 2019, Liu et al. 2019, Fortuna et al. 2021, Kogilavani et al. 2021). The taxonomies suggested by Lewandowska-Tomaszczyk et al. (2022, 2023) developed within COST Action NexusLinguarum draws on Zampieri et al.'s (2019) three-level categorisation of offensive language, in which level one discriminates between offensive and non-offensive posts, level two identifies the offensive type

(targeted vs non-targeted insult/offence) and the third level identifies the target of offence, i.e. individual, group or other. Within the SOL taxonomy approach (Lewandowska-Tomaszczyk 2022), an additional sub-level is added to the target of offence specifying whether the target is absent or present as an interaction participant. In addition, a specific level focusing on the type of lexical items is introduced differentiating between vulgar and non-vulgar expressions. The offensive type is split into four kinds of speech acts, i.e. *hate*, *insult*, *discredit* and *threat*. The offence is further specified in terms of the specific property of the target that is aimed at (e.g. *ageism*, *ideologism*, *ableism*, *racism*, *sexism*). Finally, the taxonomy considers implicit types of offence expressed via figurative means, labelled aspects, namely *exaggeration*, *irony*, *metaphor*, *rhetorical question*, *simile* or *other*.

3 Data and annotation

The Czech Corpus of Offensive Language (CCOL) comprises 400 comments, each consisting of one to three adjacent utterances, extracted from online discussions in ten Czech national newspapers and news platforms, such as SeznamZpravy, Idnes.cz, Forum24, Novinky.cz, HlídacíPes, published in the period January-February 2023. The corpus is sampled to represent discussions on a variety of topics, including home and foreign news, home and foreign politics, sport, celebrities, crime, finance, travelling, weather and health. The corpus was annotated by two annotators who are linguists and share a similar social background, age, and profession. In order to test whether the L1 of the annotator is an important variable, the L1 of one of the annotators taking part in the

1. Offensive	Yes No
2. Target 1	Group Ind. Wrt. Gr./Gr. Wrt. Ind. [by reference to group stereotypes] Individual Non-targeted
3. Target 2	Absent Present
4. Vulgar	No Yes
5. Speech act	Hate speech (referring to group stereotypes) Insult (not referring to group stereotypes) Discredit (e.g. lying-cheating, immorality, unfairness) Threat (inducing fear)
6. Aspect (specific property of the target aimed at)	Ageism Homophobic Ideologism Other Physical/mental disabilities (ableism) Prophane (religion) Racist Sexist Social class (classism) Xenophobic
7. Category of figurative language (implicit offence)	Exaggeration Irony Metaphor Other Rhetorical question Simile
Table 1: Simplified offensive language taxonomy	

campaign was Czech and the other had a different L1 but had been living and working in Czechia for 30 years. Prior to annotating the corpus, the two annotators carried several training sessions, in which they discussed the offensive language taxonomy, practiced annotating samples, compared their results and resolved disagreements.

The CCOL was annotated with the assistance of INCEpTION tool (<https://github.com/inception-project/inception>), a semantic *annotation* platform, and classified according to the SOL

Taxonomy (Lewandowska-Tomaszczyk et al. 2021) proposed as part of the research work undertaken within COST Action NexusLinguarum WG 4.1.1, summarised in Table 1. The annotation campaign took place in the period February-March 2023.

4 Results

Annotator agreement was measured according to the Cohen's Kappa measure; drawing on Landis and Koch (1997) and Sim and Wright (2005), the strength of agreement for the kappa coefficient was established on the scale: ≤ 0 =poor, .01–.20=slight, .21–.40=fair, .41–.60=moderate, .61–.80=substantial, and .81–1=almost perfect.

The Cohen's Kappa results for inter-rater agreement summarised in Table 2 show that the annotator agreement is high. More specifically, it is almost perfect for the categories Target 1 (0.89), Target 2 (0.93) and Vulgar (0.85), and substantial for the Offensive type categories (0.74 for both Insult and Discredit); the slight agreement for the *threat* category may be explained by its very low occurrence in the annotations. During the curation campaign, it was revealed that in terms of target, most of the comments in the CCOL aimed at individuals and groups, while non-targeted comments were rare (e.g. *A Hitler dělal to, co teď Russáci* [And Hitler did what the Russians are doing now], CZ-OL-131). There were some ambiguous cases, where even in the case of Czech, which discriminates T/V forms, it was impossible to decide whether the target is a group, or an individual addressed by the V-form. Occasional disagreements in the Vulgar category seem to reflect metaphorical uses of lexical items (e.g. *Člověče, vytáhněte si hlavu*

z řitního otvoru a možná to pochopíte [Man, pull your head out of your asshole, and maybe you'll understand.], CZ-OL-292). The differences in the offensive type identification concerned the perceived intensity of offence categorised as *threat* (e.g. *Už tam zůstaň na věčné časy, šmejde* [Stay there for eternity, scum.], CZ-OL-22).

As to Aspects of offensive language, or properties of the target, and categories of implicit realisations (categories of figurative language), interrater agreement differs at the three sub-levels: there is substantial agreement at the first level of Aspect 05 and Category 06, i.e. 0.70 and 0.61 respectively,

Annotation type	Agreement
Target 1 – Individual/group	0.89
Target 2 – present/absent	0.93
Vulgar	0.85
Offensive type – hate speech/insult	0.74
Offensive type discredit	0.74
Offensive type threat	0.11
Aspect 05	0.70
Aspect 05a	0.52
Aspect 05b	0
Category 06	0.61
Category 06a	0.53
Category 06b	0

Table 2: Inter-rater agreement for the Czech Offensive Language Corpus

but only moderate agreement at the second level Aspect 05a (0.52) and Category 06a (0.53); the value 0 for the third level (Aspect 05b and Category 06b) reflects the very low occurrence of simultaneous selection of more than three categories per instance of offensive language. When coding Aspects of offensive language, or properties of the target, the annotators were expected to select up to three properties available in the set (*ageism, homophobic, ideologism, albeism, profane,*

racist, social class, xenophobic and other) and mark them as Aspect 05, 05a and 05b. The annotators were instructed to select the most salient property as Aspect 05, but no further guidance was provided for assigning properties to the individual sub-types. Similarly, the instructions concerning the identification of the three sub-categories (06, 06a and 06b) of implicit realisations (categories of figurative language, i.e. *exaggeration, irony, metaphor, simile, rhetorical question, irony and other*) did not explain how the individual categories of figurative language should be assigned to the sub-types.

Out of the properties of the target, the most frequently appearing in the CCOL were *ideologism* (e.g. *České ošetrovatelství katastrofa, vládo, už se konečně proberte* [Czech healthcare is a disaster, government, wake up already], CZ-OL-355), *albeism* (physical/mental) (see the example of metaphor below), and *sexism* (e.g. *některé ženy by neměly mít peníze, aspoň by nedělaly krávovinu* [some women shouldn't have money, at least they wouldn't do shit], CZ-OL-19). As to the categories of figurativeness, *metaphor* (e.g. *Lituji pana premiéra, že se musí až do poslední chvíle scházet s tou vypitou troskou...* [I pity the Prime Minister for having to meet with that drunken wreck until the last minute.], CZ-OL-323), *simile* (e.g. *Pokud se nechovají jako ruská šovinistická prasata, tak s nimi nemají sebemenší problém* [As long as they don't act like Russian chauvinist pigs, they don't have the slightest problem], CZ-OL-127), *irony* and *rhetorical question* appear to be most prominent. The curation campaign showed that the lower level of agreement is most likely

affected by the absence of specific instructions concerning the order in which the individual properties of the target and categories of figurative language should be marked during the annotation process. In the absence of such instructions, the annotators ranked the properties and categories differently, for instance, annotator 1 classified *metaphor* as category 06a and *irony* as category 06b, while annotator 2 had *metaphor* as category 06b and *irony* as category 06a. The same concerns the properties of the target, where the annotators often listed the same properties, but in a different order. This suggests that the annotation scheme is robust, but should include a hierarchy of potential realisation of categories, in order to improve inter-rator agreement. In addition, some divergencies in the annotation of the two annotators are caused by differences in the splitting of a particular document into several consecutive parts, for instance, one annotator has identified as offensive a single expression, and the other has marked as offensive a whole clause, or one annotator has analysed a complex sentence as consisting of two clauses realising two speech acts of offense, while the other has marked the whole sentence as one speech act of offence. This could also be resolved during the training campaign by specific instructions on selection criteria.

Overall, in the case of the CCOL, the use of the SOL (Lewandowska-Tomaszczyk et al. 2022) has yielded a considerably higher degree of inter-rater agreement in comparison with annotation campaigns using a more elaborate taxonomy of offensive language, such as the English Offensive Language Corpus annotation performed earlier within COST Action NexusLinguarum

(Lewandowska-Tomaszczyk et al. 2023), where interrater agreement was fair (for Offensive type 0.32 and for Aspect between 0.29 and 0.18 for the individual sub-categories). Apart from the simplification of the taxonomy, this considerably higher degree of inter-rater agreement seems to stem from the careful selection of the data included in the corpus, the extensive training campaign and the similarity in the professional and social background of the two annotators. The CCOL campaign also indicates that inter-rater agreement is not strongly affected by the L1 factor (one of the annotator's L1 was different from Czech), as what seems of primary importance is the knowledge of the cultural and social context, in which offensive language is used. A comparison of this annotation campaign with the earlier campaign using the extended offensive language taxonomy on English offensive language (Lewandowska-Tomaszczyk et al. Forthcoming) suggests that the substantially lower inter-rater agreement (moderate and fair agreement) achieved in the English offensive language campaign may be attributed, apart from the random selection of data and short training campaign, to the choice of annotators, who, not only were speakers of various L1s different from English, but also lived in various non-English speaking contexts failing to provide them with shared cultural and social knowledge for the analysis of the English data.

5 Conclusions

This study tested the applicability of the SOL taxonomy to the Czech language, seeking to identify the level of inter-rater agreement for all categories of the taxonomy in CCOL. The results showed that the SOL

taxonomy can be successfully applied to the Czech language and that the level of inter-rater agreement was generally high. This suggests that the taxonomy is applicable as an ontology for detection and encoding of offensive language in Linguistic Linked Open Data (LLOD).

Acknowledgement

We would like to thank Slavko Žitnik, University of Ljubljana, Slovenia, for the preparation of the INCEpTION tool with the assistance of which we have annotated our data.

References

- Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F. M., Rosso, P. & Sanguinetti, M. 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In: *Proceedings of the 13th International Workshop on Semantic Evaluation. Association for Computational Linguistics*, 54-63.
- Caselli, T., Basile, V., Mitrovic, J., Kartoziya, I. & Granitzerz, M. (2020) I feel offended, don't be abusive! Implicit/explicit messages in offensive and abusive Language. In Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), 6193-6202.
- Fortuna, P., Soler-Company, J. & Wanner, L. (2021) How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management* 58(3), 102524.
- INCEpTION Annotation platform <https://inception-project.github.io/>
- Kennedy, G., McCollough, A., Dixon, E., Bastidas, A., Ryan, J., Loo, C. & Sahay, S. (2017) Technology solutions to combat online harassment. *Proceedings of the First Workshop on Abusive Language Online*. 73-77.
- Kogilavani, S. V., Malliga, S., Jaiabinaya, K. R., Malini, M. & Manisha Kokila, M. (2021) Characterization and mechanical properties of offensive language taxonomy and detection techniques. *Materials Today: Proceedings*.

- Landis J. R. & Koch G. G. (1997) The measurement of observer agreement for categorical data. *Biometrics* 33, 159-174.
- Lewandowska-Tomaszczyk, B. (2022) A simplified taxonomy of offensive language (SOL) for computational applications. *Konin Language Studies* 10 (3). 213-227.
- Lewandowska-Tomaszczyk, B., Žitnik, S., Bączkowska, A., Liebeskind, C., Mitrović, J. & Valunaite Oleškevičienė, G. (2021) Lod-connected offensive language ontology and tagset enrichment. In: Carvalho, R. & Rocha Souza, R. (eds) *Proceedings of the workshops and tutorials held at LDK 2021 co-located with the 3rd Language, Data and Knowledge Conference*. CEUR Workshop Proceedings. 135-150.
- Lewandowska-Tomaszczyk, B., Žitnik, S., Liebeskind, C., Valunaite Oleske-vicienė, G., Bączkowska, A., Wilson, P. A., Trojszczak, M., Brač, I., Filipić, L., Ostroški Anić, A., Dontcheva-Navratilova, O., Borowiak, A., Despot, K. & Mitrović, J. (accepted) Annotation scheme and evaluation: The case of OFFENSIVE language. *Rasprave*.
- Lewandowska-Tomaszczyk, B., Bączkowska, A., Liebeskind, C., Valunaite Oleskevičienė, G. & Žitnik, S. (2023) An integrated explicit and implicit offensive language taxonomy. *Lodz Papers in Pragmatics* 23.1.
- Liu, P., Li, W. & Zou, L. (2019). nlpUP at SemEval-2019 Task 6: Transfer learning for offensive language detection using bidirectional transformers. In: May, J., Shutova, E., Herbelot, A., Zhu, X., Apidianaki, M. & Mohammad, S. M. (eds) *Proceedings of the 13th international workshop on semantic evaluation*. Association for Computational Linguistics. 87-91.
- Sim, J. & Wright, C. C. (2005) The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. *Physical Therapy* 85 (3): 257-268. Doi:10.1093/ptj/85.3.257
- Wiegand, M., Ruppenhofer, J. & Eder, E. (2021) Implicitly Abusive Language – What does it actually look like and why are we not getting there? In: Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T. & Zhou, Y. (eds) *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics. Stroudsburg. 576-587.
- Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N. & Kumar, R. (2019) SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In: *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis, Minnesota, USA: Association for Computational Linguistics. 75-86.

Terminology in the Era of Linguistic Data Science (TermTrends)

Football terminology: compilation and transformation into OntoLex-Lemon resource

Jelena Lazarević

University of Belgrade Faculty
of Philology
Serbia
jelazarevic1@gmail.com

Ranka Stanković and

Mihailo Škorić and Biljana Rujević

University of Belgrade
Faculty of Mining and Geology
Serbia
(ranka.stankovic|mihailo.skoric
|biljana.rujevic)@rgf.bg.ac.rs

Abstract

The purpose of this article is to present the ongoing project which is the compilation of the first digital Football dictionary in the Serbian language, as well as to demonstrate the application of OntoLex and associated modules. The OntoLex-FrAC module for a football-specific dictionary includes information about frequency, attestation, and corpus usage. In this case, a domain-specific corpus was created by the name of SrFudKo, containing news articles about football in Serbian. Multi-word terms were automatically extracted from the Serbian corpus, then manually evaluated and classified as either sport or football-related. An inflection lexicon was produced and transformed into the OntoLex-Lemon format, Frequency information from the extraction phase was assigned to the entries. Finally, a few lexical entries were linked with the attestations from the corpus.

1 Introduction

This paper will use the expression "the language of football", as a reference to the terminology and specialized expressions used relating to football. We are aware that this is not a language in a traditional sense, but rather a specific type of jargon belonging to the domain of sporting terminology. Said terminology includes terms such as goal, corner, throw-in, offside, etc. These concepts are essential for understanding and communication about football. Here are some of the terms used, related to football:¹

- **Goal:** *fundamental scoring event in football, that occurs when a player successfully kicks the ball crosses the goal line of the opposing team, typically resulting in one point being awarded to the team that scored.*

¹The definitions are adapted from UEFA dictionary <https://www.uefa.com/insideuefa/dictionary/> and Wikipedia's Glossary of association football terms https://en.wikipedia.org/wiki/Glossary_of_association_football_terms

- **Corner:** *restart of play that occurs when the attacking team plays the ball from the corner of the field towards the opposing team's goal.*
- **Offside:** *position where a player plays the ball more advanced than the position of the last player on the opposing team.*
- **Foul:** *break of the rules of the game by a player making contact with an opponent.*

In addition to the use of said terminology and specialized expressions, there is a specific writing style, which emphasizes important events and moments during the match. Moreover, journalists often use technical terms and match analysis, in order to explain tactical decisions and player performances. They also rely on statistical data and analysis for the sake of adding depth and context to their reports.

To recapitulate, news articles about football use a specific language, crafted to provide accurate and informative reports about the sport. Shown here are examples of characteristic multi-worded expressions often used in articles about football:

- **Potentially dangerous situation:** *situation where the ball is near the goal and there is a strong possibility that the opposing team may score.*
- **Effective play:** *team's use of sound tactics and strategies, resulting in positive outcomes.*
- **Strong play:** *style of play in which a team employs physicality, often utilizing powerful kicks and high jumps, to gain an advantage over their opponents.*
- **Best chance:** *situation in which a player has a favorable opportunity to score a goal, often resulting from a good position or a well-placed pass.*

There are currently no digital terminological dictionaries that cover this area in the Serbian language, which is the main motivation for creating a Serbian lexicon of football terms and expressions. There is a traditional, analog dictionary (Miha-jlović, 2003) that covers four languages: Serbian, English, French, and Spanish. It is mentioned in the review of Serbian-Spanish dictionaries (Pejovic, 2021), but its structure does not meet the requirements of contemporary lexicography: lexical entry contains only translation equivalents. The number of terms and their selection are subjective, while the development of the dictionary was not corpus-driven.

De Oliveira Chishman et al. (2015) discussed the relevance of the Sketch Engine software (Kilgarriff et al., 2014) to build Field-Football Expressions Dictionary², a trilingual terminological resource based on the notion of the frame and on linguistic corpora. They described the analysis procedures to identify polysemic words and collocations in the corpus. Its building process involved, amongst other stages, the compilation of three comparable corpora for Spanish, Portuguese, and English.

Bergh and Ohlander (2019) have shown that, over the past hundred years, football vocabulary has become more *mainstream*, while some football terms have formed a strong presence in the minds of fans. Thus, the language of football remains in a state of constant flux, responding to the developments in and around the game. They conclude that due to its status and large media coverage of the “people’s game”, the English general purpose dictionaries are recognizing more of this footballing vocabulary as part of the general language.

The language of sport has always been a field of rich specialized linguistic communication (Liponski, 2009). Within sports, football is an especially important element of communication (Penn, 2016), due to the fact that in general human communication, football represents a significant topic. Communication about sports is primarily carried out by the media in constant contact with their target group of readers – sports fans. The language of sports, especially in Europe, is primarily the language of football, which has therefore turned into a public discourse accessible to all (Bergh and Ohlander, 2012).

The language of sports and therefore of sports journalism differs from other forms of expression.

Compared to literary language, there are differences in the degree of formality of expression and the style of presenting information. The use of collocations and idioms is present in the media coverage, which makes the articles seem much closer to the readers.

In his research, Čudomirović (2014) analyzed how the media constructed the national identity of the Serbian National Team during the 2010 World Cup matches. The corpus used for analysis included 35 reports from daily newspapers in Serbia. His findings showed that the press constructed the Serbian national identity as both highly homogeneous and self-focused, with an emphasis on achieving and maintaining unity within the nation.

There are numerous examples of research in the field of football language worldwide. However, the most interesting is *Kicktionary*, a multilingual (German – English – French) electronic dictionary of the football language, that includes 1926 football terms, of which 599 are in English, 792 in German and 535 in French (Schmidt, 2009). The terms are structured into a hierarchy of scenarios and frameworks, which further include multiple concepts. Each word is illustrated with one or more example sentences from the authentic: written or spoken football language.

The main goal of the *Kicktionary* project was to explore how the linguistic theories of lexical semantics, as well as corpus linguistic methods, hypertext technologies, and computational language-processing techniques, can help to create a lexical resource – better than, or at least different from, traditional analog dictionaries. Although primarily intended for humans, *Kicktionary* has also been used to create models for automatic text markup. Specifically, an adapted version of the frame semantic parsing model *LOME* was used to automatically label texts with frames and semantic roles according to the *Kicktionary* lexical resource (Minnema, 2021).

Inspired by numerous works, our research question is the following: Is it possible to semi-automatically generate a list of terms and football expressions for the Serbian language?

The Section 2 Materials and methods will firstly present the dataset, i.e., the corpus of texts used for the research, the usage and dictionary microstructure, the methods of automatic extraction of terms and manual evaluation criteria, followed by a short

²<http://dicionariofield.com.br/langselect>

outline of the OntoLex-Lemon³ core model (McCrae et al., 2017), a widely used vocabulary for modeling machine-readable dictionaries on the Semantic Web and as Linguistic Linked Open Data (LLOD), with extension Morph⁴ (Klimek et al., 2019; Chiarcos et al., 2022c) and OntoLex-FrAC module (Chiarcos et al., 2022a, 2020).

The Section 3 is dedicated to the results, where the typical examples for several observed syntactic groups will be shown. The Section 4 is dedicated to the examples of lexical entries published in the form of linked data following the OntoLex-Lemon and OntoLex-FrAC specifications. Ultimately, this study offers conclusive considerations and directions for further research.

2 Materials and Methods

2.1 FudKo Corpus

The *srFudKo* corpus is comprised of articles about football in the Serbian language. These articles are gathered from five Serbian digital news sites: *B92*, *Blic*, *Mondo*, *Politika*, and *Sport Klub*. The articles were automatically retrieved through various web scraping techniques, following the harmonization of the gathered structure, and the text was cleansed. Articles shorter than 3000 characters, sentences in other languages, and tables containing only numerical results were eliminated. The article titles were also analyzed, resulting in the removal of 130 duplicate articles detected by their titles. They were then manually examined and removed.

The corpus was prepared as a collection of XML files, in which articles are marked with the following structural labels: `<data>` - the basic elements of each document, `<post>` - published article, `<date>` - article date, `<title>` - article title, and `<p>` - paragraph or text of the article. XML files are organized by year and by the portals from which they were downloaded, so 11,117 articles are distributed across 37 files.

Regarding the distribution across portals, *Politika* is the most represented with 3257 articles. They are followed by *Mondo* with 2639 articles, *B92* with 2514 articles, *Sport klub* with 1937 articles, and *Blic* with 770 articles (Table 1). The articles taken from the *Politika* website cover a long period from 2006 to 2021, making this the largest partition. *Sport Klub* covered the years 2017 to 2021, while *Mondo* covered the years 2013 to

Portal	Period	Number of	
		Articles	Words
Politika	2006-2021	3257	3.1M
Mondo	2013-2021	2639	2.8M
B92	2013-2021	2514	1.9M
Sport klub	2017-2021	1937	1.6M
Blic	2020-2021	770	0.7M

Table 1: Articles distribution across portals

2021. The *B92* website was downloaded from 2017 to 2021, and *Blic* was parsed for only two years: 2020 and 2021, making this partition the smallest.

The corpus was tagged with part-of-speech and lemma using a tagger: *SrpKor4Tagging-TreeTagger* for the Serbian language⁵ (Stanković et al., 2020; Stanković et al., 2022) integrated into the *TXM tool* (Heiden, 2010). The tagger was trained on the manually annotated corpus *SrpKor4Tagging*⁶, which combines literary one-third and administrative two-thirds texts in Serbian.

The corpus was annotated with two sets of part-of-speech tags: *Universal POS* and *SrpLemKor* (a set created based on the traditional, descriptive grammar of the Serbian language), and lemmatized, containing 342,803 tokens. The lemmatization is based on electronic morphological dictionaries for Serbian (Krstev, 2008; Vitas and Krstev, 2012), specifically on the derived distribution intended for tagging *SrpMD4Tagging*⁷ (Serbian Morphological Dictionaries for Tagging).

The TXM platform has proven to be very successful for corpus analysis, frequency distributions, and visual presentation. After filtering articles and cleaning the text, the *srFudKo* corpus contains 10,100,553 tokens, of which 8,618,426 are words, and the remainder consists of punctuation marks.

2.2 Dictionary Usage and Microstructure

A sports dictionary of football can be useful for various individuals involved in the sport. They include players, coaches, referees, commentators, journalists, and fans who wish to enhance their understanding and communication in the realm of

⁵<https://live.european-language-grid.eu/catalogue/ld/9296>

⁶<https://live.european-language-grid.eu/catalogue/corpus/9295>

⁷<https://live.european-language-grid.eu/catalogue/lcr/9294>

³<https://www.w3.org/2016/05/ontolex/>

⁴<https://www.w3.org/community/ontolex/wiki/Morphology>

football. This dictionary will also be used in NLP (Natural Language Processing) applications related to the football domain.

Football players, both amateur and professional, can benefit from a sport dictionary of football, helping enhance their understanding of technical terms, rules, positions, tactics, and strategies used in the game. Thus it can help them communicate effectively with their teammates and coaches. This is especially helpful in the case of foreign players that do not speak the native language of their teammates. Football coaches can also utilize this type of dictionary to reinforce their knowledge of the game and stay updated in the latest terminology. It can assist them in explaining concepts to players, designing training sessions, and developing game plans.

Referees and officials responsible for enforcing football rules can use this football dictionary to ensure a comprehensive understanding of the terms used in the game. This helps them make accurate decisions and maintain consistency during matches. Commentators and analysts who provide match commentary or analysis can utilize a football dictionary to expand their vocabulary and improve their understanding of the game. It allows them to deliver more informative and engaging commentary to viewers. Football journalists and writers can reference a specialized sporting dictionary of football to ensure accuracy in their match reports, using appropriate terminology while discussing player profiles, match analysis, or tactical elements.

Football fans who wish to deepen their knowledge of the sport can benefit from a football dictionary, which enables them to understand better match broadcasts, articles, discussions, and conversations related to the sport. It also enhances their overall enjoyment and engagement with the game.

The microstructure of this football dictionary will include a range of information related to lemma (base word), inflected forms, examples or attestations, frequencies, multi-word expressions, and collocations. Here's a breakdown of each component:

- The lemma represents the base, canonical form, and serves as the entry point in the dictionary. For example, in the football domain a lemma could be *gol* (goal) or *udarac* (kick).
- The inflected forms of a lemma are important for Serbian as a highly inflected language. For instance, variations of the lemma

udarac could include *udarca*, *udarcu*, *udarci*, *udarcima*, *udarce*, etc.

- The illustrative examples or attestations showcase the usage of the lemma in different contexts. These examples demonstrate how the word is used in football-related sentences or phrases.
- The multiword expressions, including fixed phrases, idioms, or collocations specific to the football domain will be included and related to its component words.
- Word usage frequency indicates how common or uncommon a word is within the football domain. Frequencies will be represented through numerical values, both in domain-specific football corpus and in the general-purpose Corpus of the contemporary Serbian language *SrpKor2013* (Utvić, 2011; Vitas and Krstev, 2012), as illustrated through the examples in the Section 4.
- The term collocation refers to words that frequently occur together with a specific lemma. In a dictionary focused on football, collocations can highlight common word combinations or phrases that involve the main lemma.

The current focus is based on a monolingual dictionary. However, future research will include term translation equivalents in the target language. These would also provide corresponding phrases or idioms in the other language, allowing users to understand football-related expressions in both languages. It is important to state that definitions are not part of the initial phase but are planned for the following phase. This is due to the fact that the initial phase is focused on automatic procedures that are already developed. For the definition extraction in Serbian, initial results are presented in (Stanković et al., 2021). However, the solution requires improvement and adaptation for this particular case of use.

Including multi-word expressions and their bilingual equivalents will enhance the dictionary's coverage of idiomatic and context-specific language usage in the football domain, helping users grasp the nuances and intricacies of the language related to the sport.

The outlined micro-structure of the football dictionary aims to provide comprehensive information

about the lemma, its variations, usage examples, frequency of occurrence, and common word combinations, allowing users to better understand and utilize football-related vocabulary.

2.3 Terminology Extraction Approach

The process of football terminology extraction from the corpus *srFudKo* included the following steps:

1. automatic extraction of candidates,
2. manual evaluation and classification,
3. import to lexical database,
4. export to other formats (DELA⁸ for Unitex⁹, RDF, etc.).

The statistical measure *Keyness* is used in the step of terminology extraction, for identifying terms that are significantly more frequent in the football corpus *srFudKo*, compared to the Corpus of contemporary Serbian *SrpKor2013* (Utvić, 2011; Vitas and Krstev, 2012). The relevance and specificity of a term within a football domain are calculated through the ratio of term frequency in the corpus *srFudKo*, as the target corpus, compared to its frequency in *SrpKor2013*, as the reference corpus. The terms with a high keyness score are considered to be highly relevant, distinct to the football domain, and thus can be used as potential candidates for the terminology lexicon. The keyness function was applied to single-word lemma and multi-words extracted with various syntactic patterns (Krstev et al., 2015).

Multi-word candidates are extracted from texts in their various inflected forms using lexical resources and local grammars developed for Serbian (Krstev et al., 2015) with patterns explained in Section 3. The lemmatization of extracted multi-word candidates, that is, their linking to one normalized form is of low importance for the English language. However, in terms of highly-inflected languages, such as Serbian and other Slavic languages, this task can hardly be avoided, as each nominal multi-word unit (MWU)¹⁰ can have many inflected forms (from five to ten or even more) and

⁸Dictionnaires électroniques du LADL - Laboratoire d'Automatique Documentaire et Linguistique

⁹<https://unitexgramlab.org>

¹⁰We use the term *multi-word unit* as a general term for MWE, collocation, multi-word term, or multi-word named entity

many of these forms (but usually not all) can, in general, be extracted from a corpus (Krstev et al., 2015).

The hybrid system called *Srp-TE* (Stanković et al., 2016) was used, which relies on the application of syntactic patterns and electronic Morphological dictionaries for the Serbian language *SrpMD* (Krstev, 2008) that contain both single and multi-word units, covering general lexicon, proper names, toponyms, encyclopedic knowledge, and terminology from numerous domains.

Class names correspond to FSTs (Finite-state transducers) used for the inflection of MWUs belonging to that class. For example, MWUs are composed of an adjective (A) followed by a noun (N), which concord in gender, number, case, and animacy, belong to the AXN class. The letter X represents a component that remains unchanged when the MWU inflects. It can also denote a separator, like a space or a hyphen. The number preceding X indicates how many of these parts there are in the MWU, with 2X representing two uninflected components, one of which is a separator, 4X representing four components, two of which are separators, etc.

The most frequent syntactic structures, for example AXN, 2XN, N2X, N4X, AXN2X, NXN, AXAXN, N6X, AXN4X, 2XAXN, AXN6X, N8X, are implemented. In the Section 3 explanations are given, with examples for the most productive syntactic structures.

2.4 Ontolex-lemon and OntoLex-FrAC

The use of the OntoLex-Lemon is increasing in terms of lexical resources in the web of data. The lexical entries (single and multi-words) from the football domain, extracted by the approach described in the Section 2 are represented using the OntoLex-Lemon.

The morphological dictionary of multi-word units was produced using a multipurpose tool (Stanković et al., 2011), then transformed with a custom application, following the OntoLex specifications, and published examples (Chiarcos et al., 2022b). The grammatical information, morpho-syntactic features about word forms were given by tag properties in accordance with the *LexInfo vocabulary*¹¹.

The Section 4.2 presents the use of the OntoLex core module and the module for Frequency, Attes-

¹¹<https://lexinfo.net/>

tations, and Corpus-Based Information (OntoLex-FrAC) (Chiarcos et al., 2022a). The information found in the corpora, such as attestations and frequency information of tokens (forms) and lemmas (lexical entries) that are automatically derived from corpora, are introduced following the OntoLex-FrAC specifications.

3 Terminology Extraction Results

The terminology extraction in this research study relies upon the results of previous research, both for building and using a terminology system, which includes data, application, and user-interface layers, covering different data and software technologies. The rule-based automatic multi-word term extraction and lemmatization are first used in the domain of library-information terminology (Krstev et al., 2015; Stanković et al., 2016). This data-driven approach was used for raw material terminology (Kitanović et al., 2021), and corpus-based bilingual terminology extraction in the power engineering domain (Ivanović et al., 2022).

The conversion of electronic dictionaries from a file system to a lexical database *LeXimirka*, based on the Lemon model has resulted in a robust system, that not only manages electronic dictionaries but also incorporates a connection with corpora, including results of systems for automatic - single and multi-word terminology extraction (Stanković et al., 2018; Lazić and Škorić, 2020).

Figure 1 presents a web form with lexical entry *utakmica* ('match, sports competition') several parts:

1. inflected forms with grammatical information,
2. concordances in the selected corpus, in this case *srFudKo*,
3. frequencies of inflected forms in the selected corpus for lexical entry of syntactic patterns, in this case, adjective-noun A (N), where the noun is the current lexical entry,
4. lemma frequencies, where in case of syntactic patterns, all components are lemmatized,
5. links to multi-word lexical entries in *LeXimirka* where current entry is one component.

Before the extraction procedure was conducted as part of this research study, the football domain

was not specifically processed. However, the electronic morphological dictionary already had a number of terms related to the sporting domain. Using the marker `DOM=Sport`, a total of 185 simple words and 240 multi-word units were marked, belonging to the domain of sport. The semantic marker `+Sport` denoting sporting disciplines was assigned to four simple words and 13 multi-word units. After processing the football domain corpus *SrFudKo* in the Serbian language, some additional entries were prepared. A new marker `DOM=Fudbal` was introduced for the football domain. The list of candidates already in the morphological dictionary was extracted using the keyness function and a new marker was assigned, based on annotations from two independent evaluators and a supervisor that resolved differences. The first author was one of the evaluators, and she has nearly a decade-long experience in sports journalism, reporting primarily on football and creating football-themed articles in multiple languages, which allows her to offer her practical expertise to the academic realm. The second evaluator is a dedicated enthusiast of football.

As for the nouns, a total of 915 nouns that are characteristic of football and sporting articles were marked, while an additional 219 nouns were marked as belonging to the football domain (e.g. *gol, fudbal, fudbaler, poluvreme, golman, mreža, penal (goal, football, football player, halftime, goalkeeper, net, penalty)*). When it comes to verbs, there are 196 sports and 5 specific football terms (e.g. *predriblati, uklizati, uštopovati, proklizati (to feint, to tackle, to intercept, to slide tackle)*).

Presented here are some of the most productive patterns:

- AXN – an adjective followed by a noun; the adjective and the noun have to concord in all four grammatical categories; e.g. *bela tačka, crveni karton, fudbalski klub, (penalty mark, red card, football club)*,
- N2X – a noun followed by a non-inflecting word, usually a noun in the genitive or in the instrumental case; e.g. *OFK Beograd, het-trik, FS Srbija, plej-aut, (OFK Belgrade, hat-trick, FS Serbia, play-out)*,
- N4X – a noun followed by two words that do not inflect in the MWU: 1) A noun followed by a prepositional phrase; e.g. *uzbuđenje pred golom, centaršut iz kornera, (excitement in*

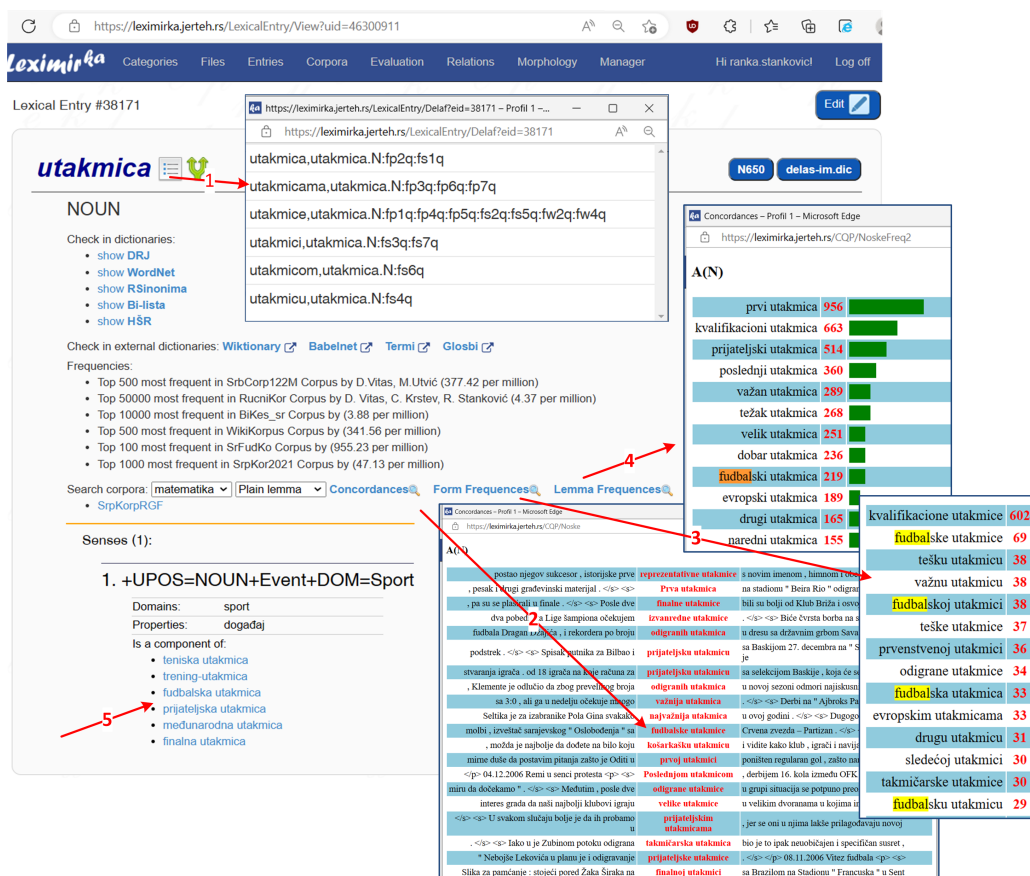


Figure 1: Panel from LeXimirka lexical resources management system

- front of the goal, corner kick); 2) A noun followed by two adjectives/nouns in the genitive case or instrumental case; e.g. *ivica kaznenog prostora, utakmica visokog rizika*, (edge of the penalty area, high-risk match (a match with potential for violence or disturbances)),
- AXN2X – a noun preceded by an adjective concurring in the gender, number, case and animateness and followed by a word that does not inflect in the MWU, usually a noun in the genitive or instrumental case; e.g. *grupna faza fudbala, prvo kolo kvalifikacija, evropska kuća fudbala*, (group stage of the league, first qualifying round, the Union of European Football Associations (UEFA)),
 - AXAXN – a noun preceded by two adjectives, concurring in gender, number, case and animateness; e.g. *zimski prelazni rok, Svet-sko fudbalsko prvenstvo, centralni vezni igrač*, (winter transfer window, FIFA World Cup, central midfielder),
 - N6X - a noun followed by three words that do not inflect in the MWU: *učešće u ligi šampi-ona, udarac sa ivice šesnaesterca, borba na sredini terena*, (participation in the Champions League, shot from the edge of the penalty area, a battle in the middle of the field),
 - AXN4X – a noun preceded by an adjective concurring in the gender, number, case and animateness, followed by two words that do not inflect in the MWU or by two adjectives/nouns in the genitive case or in the instrumental case: *Svet-sko prvenstvo u fudbalu, prvo mesto na tabeli, žuti karton zbog simuliranja*, (FIFA World Cup, first place on the table, yellow card for simulation),
 - 2XAXN - an adjective followed by a noun concurring all four grammatical categories, preceded by a word that does not inflect in the MWU; *FK Crvena zvezda, crveno-beli dres, crno-beli tabor*, (FC Red Star, the Red and White jersey, the Black and White side),
 - N8X - a noun followed by four words that do not inflect in the MWU: *udarac sa ivice kaznenog prostora, bod u borbi za opstanak*,

(shot from the edge of the penalty area, point in the fight for survival).

4 FudLe: Linked Data Lexicon

4.1 OntoLex Core Part of FudLe

We illustrate the conversion of electronic dictionary entries with the term *fudbalska utakmica* (eng. football match), which is a competition between two football teams. In Serbian Morphological E-Dictionary (SrpMD) of Compounds (Krstev and Vitas, 2009) in the form of DELAC (Savary et al., 2007) the original dictionary entry is:

```
fudbalska (fudbalski.A2:aefs1g)
utakmica (utakmica.N650:fs1q),
NC_AXN+DOM=Sport+Comp
```

The finite state transducer (FST) NC_AXN generates the inflected forms for the morphological e-dictionaries of compound words, where NC stands for Noun compound and AXN depicts adjective-noun compound, where the adjective concurs with the noun in its grammatical number, gender, case, and animacy. For the components that the FST inflects, it requires information about lemma (*fudbalski* and *utakmica*), the FST (A2 and N650) for simple component word and values for grammatical features (aefs1g and fs1q).

The grammatical features are: *a* - positive degree, *e* - form both definite and indefinite, *f* - feminine grammatical gender, *s* - singular number, *l* - nominative case, *g* - no consequence for animacy, *q* - inanimate. Most of the grammatical features are easily mapped to *Lexinfo* but the dilemma for their mapping was *lexinfo:otherAnimacy* adequate for *g* - no consequence for animacy and for the forms that are both definite and indefinite.

Here, we assume that the term *fudbalska utakmica* is a multi-word expression, since it is in the SMD and it can be found in terminological dictionaries. However, it can be treated as a collocate as well. By using the *OntoLex-Lemon* vocabulary, we can declare that it is a (lexicalized) MWU with its specific meaning.

A part of the *LeXimirka* MS SQL Server database's data model, is shown in Figure 2, which displays tables for lexical entries and inflected forms, as well as components for multi-word units. Grammatical information is linked to the inflected forms through data categories and their values. The system is provided with metadata related to linked information between data categories in the Serbian

morphological dictionaries and the *Lexinfo* vocabulary.

The following listing presents an example of a multi-word unit, where the name: *le_fudbalska_utakmica_220902* is composed of prefix *le* that stands for *LexicalEntry*, term *fudbalska_utakmica* and primary key *220902* from the table *LexicalEntry* from database *LeXimirka*. Similarly, prefix *cm* denote entries from the table *Component* and prefix *fm* denote entries from the table *Form*.

```
:le_fudbalska_utakmica_220902
  a ontolex:LexicalEntry,
    ontolex:MultiwordExpression;
  ontolex:canonicalForm
    [ontolex:writtenRep
      "fudbalska utakmica"@sr];
  lexinfo:partOfSpeech lexinfo:noun;
  ontolex:sense
    [ontolex:reference <https://
      dbpedia.org/ontology/FootballMatch>];
  decomp:constituent :cm_fudbalska_20258;
  decomp:constituent :cm_utakmica_20259;
  rdf:_1 :le_fudbalski_78369; # lexical
  rdf:_2 :le_utakmica_38171. # entries

# component of canonical form
:cm_fudbalska_20258 a decomp:Component;
decomp:correspondsTo :le_fudbalski_78369;
morph:grammaticalMeaning
  [lexinfo:degree lexinfo:positive;
  lexinfo:gender lexinfo:feminine;
  lexinfo:number lexinfo:singular;
  lexinfo:case lexinfo:nominative;
  lexinfo:lexinfo:inanimate].
...
```

The inflected forms of single and multi-word units in morphological dictionaries are followed by a set of data category values. The majority of inflected forms have ambiguous grammatical interpretations. The following example presents typical instances of single and multi-word units - *fudbalska utakmica*.

```
# inflected forms for adjective
fudbalska:aefs1g:aefs5g:aemw2g:aemw4g...
fudbalskoj:aefs3g:aefs7g
fudbalskim:aefp3g:aefp6g:aefp7g:aemp3g...
...
# inflected forms for noun
utakmica:fp3q:fp6q:fp7q
utakmici:fs3q:fs7q
utakmicama:fs6q
...
# multiword inflected forms
fudbalska utakmica:fs1q
fudbalskoj utakmici:fs3q:fs7q
fudbalskim utakmicama:fp3q:fp6q:fp7q
...
```

The following example presents the first lexical entry - the adjective component *fudbalski* with a

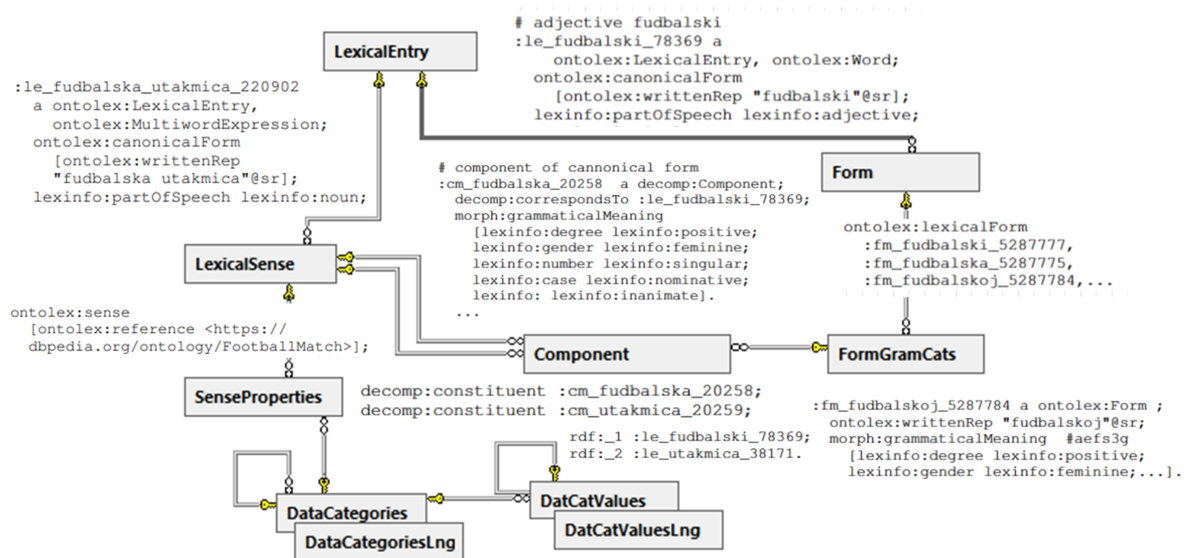


Figure 2: MS SQL Server database diagram with tables related to lexical entries and inflected form

sample of the inflected forms, accompanied by its grammatical information. It is followed by a lexical entry *utakmica* as the second component in its inflected form *utakmici*.

```

# adjective fudbalski
:le_fudbalski_78369 a
  ontolex:LexicalEntry, ontolex:Word;
  ontolex:canonicalForm
  [ontolex:writtenRep "fudbalski"@sr];
  lexinfo:partOfSpeech lexinfo:adjective;
  ontolex:lexicalForm
  :fm_fudbalski_5287777,
  :fm_fudbalska_5287775,
  :fm_fudbalskoj_5287784, ...
:fm_fudbalskoj_5287784 a ontolex:Form ;
  ontolex:writtenRep "fudbalskoj"@sr;
  morph:grammaticalMeaning #aefs3g
  [lexinfo:degree lexinfo:positive;
  lexinfo:gender lexinfo:feminine;...].
  ...
# noun utakmica
:le_utakmica_38171 a
  ontolex:LexicalEntry, ontolex:Word;
  ontolex:canonicalForm
  [ontolex:writtenRep "utakmica"@sr];
  lexinfo:partOfSpeech lexinfo:noun;
  ontolex:lexicalForm
  :fm_utakmica_4569852,
  :fm_utakmice_4569854,
  :fm_utakmici_4569855, ...
:fm_utakmici_4569855 a ontolex:Form ;
  ontolex:writtenRep "utakmici"@sr;
  morph:grammaticalMeaning #fs3q
  [lexinfo:gender lexinfo:feminine;
  lexinfo:number lexinfo:singular;...].
  ...
  
```

The examples of inflected forms *a* in multi-word lexical entry *fudbalska utakmica* and form *fudbalskoj utakmici* is given with its grammatical infor-

mation:

```

:le_fudbalska_utakmica_220902
  ontolex:lexicalForm
  :fm_fudbalska_utakmica_2309936,
  :fm_fudbalske_utakmice_2309938.
  :fm_fudbalskoj_utakmici_2309942,
  ...
# inflected forms
:fm_fudbalskoj_utakmici_2309942
  a ontolex:Form;
  ontolex:writtenRep
  "fudbalskoj utakmici"@sr;
  morph:grammaticalMeaning
  [lexinfo:gender lexinfo:feminine;
  lexinfo:number lexinfo:singular;
  lexinfo:case lexinfo:acusative;
  lexinfo:animacy lexinfo:inanimate];
  morph:grammaticalMeaning
  [lexinfo:gender lexinfo:feminine;
  lexinfo:number lexinfo:singular;
  lexinfo:case lexinfo:locative;
  lexinfo:animacy lexinfo:inanimate].
  ...
  
```

4.2 OntoLex-FrAC Part of FudLe

The OntoLex Module for Frequency, Attestation, and Corpus Information (FrAC) is still under development and in this paper, we are relying on a Draft Community Group Report.¹² Due to the potential changes in the FrAC model, the modeling examples presented may be subject to modifications in future development.

The auxiliary class `:SrFudKo` is defined to provide convenient handling and shorter notation. Currently, the version of the corpus *srFudKo* published

¹²<https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/index.md> accessed 7.8.2023

in the *noSketch engine* (Kilgarriff et al., 2014) instance is managed by the Society for Language Resources and Technologies - JeRTeH, is linked.¹³

We introduce specialized sub-classes for the two frequency types: `:SrFudKo_token_freq`, for inflected forms frequency and `:SrFudKo_lemma_freq` for a total of all inflected form-frequencies of a lexical entry. Just to mention that in this case: the "token" can be also a multi-word unit. This represents a more compact encoding, as the data does not have to be repeated for each individual observable.

```
# football corpus
:SrFudKo a owl:Class;
  rdfs:subClassOf [a owl:Restriction;
    owl:onProperty frac:observedIn ;
    owl:hasValue <https://noske.jerteh.rs/
#dashboard?corpname=FudKo>] .
:SrFudKo_token_freq rdfs:subClassOf
  frac:Frequency, :SrFudKo,
  [a owl:Restriction;
    owl:onProperty dct:description;
    owl:hasValue "token frequency"].

# general language corpus
:SrpKor2021 a owl:Class;
  rdfs:subClassOf [a owl:Restriction;
    owl:onProperty frac:observedIn ;
    owl:hasValue <https://noske.jerteh.rs/
#dashboard?corpname=SrpKor2021>] .
:SrpKor2021_token_freq rdfs:subClassOf
  frac:Frequency, :SrpKor2021,
  [a owl:Restriction;
    owl:onProperty dct:description;
    owl:hasValue "token frequency"].

...
```

Let us notice that absolute and relative (per million) frequencies from several corpora are available in the lexical database for (simple) words. Figure 1 shows that the information about the frequency class: top 100, 500, 1000, etc. is available as well. It can be seen that the lemma *utakmica* is in the top 100 most frequent lemmas in the SrFudKo corpus with a relative frequency of 955.23 per million and in the top 1000 most frequent in SrpKor2021 corpus with a relative frequency of 47.13 per million. The absolute frequencies for the inflected form (token) *utakmici* and lexical entry (lemma) *utakmica* are encoded as follows:

```
# inflected form frequency
:fm_utakmici_4569855 frac:frequency
  [a :SrFudKo_token_freq;
  rdf:value "3739"].
:fm_utakmici_4569855 frac:frequency
  [a :SrpKor2021_token_freq;
  rdf:value "23055"].

# lemma frequency
```

¹³<https://jerteh.rs/>

```
:le_utakmica_38171
  [a :SrFudKo_token_freq;
  rdf:value "29479" ] .
:le_utakmica_38171
  [a :SrpKor2021_token_freq;
  rdf:value "138573" ] .
```

In terms of multi-word units, absolute frequencies are retrieved using the CQL (Corpus Query Language) expressions, while relative frequencies are calculated by dividing the headword frequency.

The dilemma in terms of frequencies was related to the multi-word expressions frequency: whether or not the same property should be used `SrFudKo_token_freq` or it should be introduced the `SrFudKo_mwe_freq`. The possible solution may be the following:

```
:SrFudKo_mwe_freq rdfs:subClassOf
  frac:Frequency, :SrFudKo,
  [owl:Restriction;
    owl:onProperty dct:description;
    owl:hasValue "mwe frequency"].
```

Furthermore, the frequencies are given for the multi-word inflected forms *fudbalskoj utakmici* and the multi-word lexical entry *fudbalska utakmica*.

```
# mwe inflected form frequency
:fm_fudbalskoj_utakmici_2309942
  frac:frequency
  [a :SrFudKo_mwe_freq ;
  rdf:value "38" ] ;
  frac:head :fm_utakmici_4569855 .
:fm_fudbalskoj_utakmici_2309942
  frac:frequency
  [a :SrpKor2021_mwe_freq ;
  rdf:value "495" ] ;
  frac:head :fm_utakmici_4569855 .

# mwe lemma frequency
:le_fudbalska_utakmica_220902
  frac:frequency
  [a :SrFudKo_mwe_freq;
  rdf:value "219"];
  frac:head
  :le_utakmica_38171 .
:le_fudbalska_utakmica_220902
  frac:frequency
  [a :SrpKor2021_mwe_freq;
  rdf:value "2749"];
  frac:head
  :le_utakmica_38171 .
```

The attestation example "*Odavno na Banovom brdu nije bilo toliko gledalaca na jednoj fudbalskoj utakmici.*", translated to English: "*It has been a long time since there were so many spectators at one football match at Banovo Brdo*" is encoded by using properties `frac:attestation` and `frac:quotation`. It has been added manually, but automatizing the process is expected in the future:

```
# single word inflected form attestation
:fm_utakmice_4569854 [
```

```

frac:quotation "Gledao sam sve
utakmice tih timova .";
frac:observedIn :SrfudKo].
:fm_utakmice_4569854 [
frac:quotation "Mi u ovoj vašoj
utakmici, u vašoj trgovini,
nećemo da učestvujemo ..";
frac:observedIn :SrpKor2021].

# multiword inflected form attestation
:fm_fudbalskoj_utakmici_2309942
frac:attestation [
frac:quotation "Odavno na Banovom
brdu nije bilo toliko gledalaca
na jednoj fudbalskoj utakmici.";
frac:observedIn :SrfudKo].
:fm_fudbalskoj_utakmici_2309942
frac:attestation [
frac:quotation "Gospodo, ponašajte se
pristojno, nije ovo fudbalska utakmica
, ovo je parlament Srbije .";
frac:observedIn :SrpKor2021].

```

5 Conclusion

The initial results of the ongoing activity in the creation of the Serbian language Football dictionary are presented, fully proving that it is possible to semi-automatically generate lists of terms and football expressions for the Serbian language. The corpus-driven approach is complemented by manual evaluation and classification of term entries. Current activities include 1) refining the produced data set through additional semantic annotation inspired by the *Kicktionary* (Schmidt, 2009) project, 2) automatic morphological inflection, which is followed by manual evaluation of the morphological classes for all new multi-word units, 3) refining the exporting procedures from the *LeXimirka* database to the *ttl*, 4) the automatic selection of good corpus examples, 5) including footballing terms' derivation and variation, and ultimately 6) word embedding integration. We will follow the initiatives related to the improvement of terminology modules for Ontolex and improve our resources according to new specifications.

Acknowledgements

This paper is partially supported by the COST Action NexusLinguarum – “European network for Web-centred linguistic data science” (CA18209), supported by COST (European Cooperation in Science and Technology).

References

Gunnar Bergh and Sölve Ohlander. 2012. Free kicks, dribblers and wags. exploring the language of “the

people’s game”. *Moderna språk*, 106(1):11–46.

Gunnar Bergh and Sölve Ohlander. 2019. A hundred years of football english: A dictionary study on the relationship of a special language to general language. *Alicante Journal of English Studies / Revista Alicantina de Estudios Ingleses*, 32:15–43.

Christian Chiarcos, Elena-Simona Apostol, Besim Kabashi, and Ciprian-Octavian Truică. 2022a. **Modelling frequency, attestation, and corpus-based information with OntoLex-FrAC**. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4018–4027, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Christian Chiarcos, Christian Fäth, and Maxim Ionov. 2022b. **Unifying morphology resources with ontomorph. a case study in german**. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4842–4850.

Christian Chiarcos, Katerina Gkirtzou, Fahad Khan, Penny Labropoulou, Marco Passarotti, and Matteo Pellegrini. 2022c. Computational morphology with ontomorph. In *Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference*, pages 78–86.

Christian Chiarcos, Maxim Ionov, Jesse de Does, Katrien Depuydt, Fahad Khan, Sander Stolk, Thierry Declerck, and John Philip McCrae. 2020. Modelling frequency and attestations for ontomorph. In *Proceedings of the 2020 Globalex Workshop on Linked Lexicography*, pages 1–9.

Rove Luiza De Oliveira Chishman, Aline Nardes dos Santos, Diego Spader de Souza, and João Gabriel Padilha. 2015. The relevance of the sketch engine software to build field-football expressions dictionary. *Revista de Estudos da Linguagem*, 23(3):769–796.

Serge Heiden. 2010. The txm platform: Building open-source textual analysis software compatible with the *tei* encoding scheme. In *24th Pacific Asia conference on language, information and computation*, volume 2–3, pages 389–398. Institute for Digital Enhancement of Cognitive Development, Waseda University.

Tanja Ivanović, Ranka Stanković, Branislava Šandrih Todorović, and Cvetana Krstev. 2022. **Corpus-based bilingual terminology extraction in the power engineering domain**. *Terminology*, 28:2.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. **The Sketch Engine: ten years on**. *Lexicography*, 1(1):7–36.

Olivera Kitanović, Ranka Stanković, Aleksandra Tomašević, Mihailo Škorić, Ivan Babić, and Ljiljana Kolonja. 2021. A data driven approach for raw material terminology. *Applied Sciences*, 11(7):2892.

- Bettina Klimek, John P McCrae, Julia Bosque-Gil, Maxim Ionov, James K Tauber, and Christian Chiarcos. 2019. Challenges for the representation of morphology in ontology lexicons. *Proceedings of eLex*.
- Cvetana Krstev. 2008. *Processing of Serbian. Automata, texts and electronic dictionaries*. Faculty of Philology of the University of Belgrade.
- Cvetana Krstev, Ranka Stanković, Ivan Obradović, and Biljana Lazić. 2015. Terminology acquisition and description using lexical resources and local grammars. In *Proceedings of the 11th Conference on Terminology and Artificial Intelligence, Granada, Spain, 2015*.
- Cvetana Krstev and Duško Vitas. 2009. An effective method for developing a comprehensive morphological e-dictionary of compounds. In *Proceedings of Lexis and Grammar Conference, Bergen*, pages 204–212.
- Biljana Lazić and Mihailo Škorić. 2020. From dela based dictionary to leximirka lexical database. *Infotheca*.
- Wojciech Liponski. 2009. "hey, ref! go, milk the canaries!" on the distinctiveness of the language of sport. *Studies in Physical Culture and Tourism*, 16(1).
- John P McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The Ontolex-Lemon model: development and applications. In *Proceedings of eLex 2017 conference*, pages 19–21.
- Aleksandar Mihajlović. 2003. *Fudbalski rečnik / Football Dictionary / Dictionnaire du football / Diccionario del fútbol*. A. Mihajlović, Belgrade.
- Gosse Minnema. 2021. Kicktionary-lome: a domain-specific multilingual frame semantic parsing model for football language. *arXiv preprint arXiv:2108.05575*.
- Andjelka Pejovic. 2021. Logros lexicográficos del hispanismo serbio y el croata. *Revista de Lexicografía*, 26:113–130.
- Roger Penn. 2016. Football talk: sociological reflections on the dialectics of language and football. *European Journal for Sport and Society*, 13(2):154–166.
- Agata Savary, Cvetana Krstev, and Duško Vitas. 2007. Inflectional non compositionality and variation of compounds in french, polish and serbian, and their automatic processing. *Bulag-Bulletin de Linguistique Appliquée et Générale*, 32:73–94.
- Thomas Schmidt. 2009. The kicktionary—a multilingual lexical resource of football language. In *Multilingual FrameNets in computational lexicography: methods and applications*, pages 101–132. de Gruyter.
- Ranka Stanković, Cvetana Krstev, Biljana Lazić, and Mihailo Škorić. 2018. Electronic dictionaries-from file system to lemon based lexical database. In *Proceedings of the 11th International Conference on Language Resources and Evaluation-W23 6th Workshop on Linked Data in Linguistics: Towards Linguistic Data Science (LDL-2018), LREC 2018, Miyazaki, Japan, May 7-12, 2018*.
- Ranka Stanković, Cvetana Krstev, Ivan Obradović, Biljana Lazić, and Aleksandra Trtovac. 2016. Rule-based automatic multi-word term extraction and lemmatization. In *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016, Portorož, Slovenia, 23–28 May 2016*.
- Ranka Stanković, Cvetana Krstev, Mihailo Škorić, Rada Stijović, and Nebojša Vasiljević. 2021. Towards automatic definition extraction for serbian. *Proceedings of the XIX EURALEX Congress of the European Association for Lexicography: Lexicography for Inclusion (Volume 2). 7-9 September (virtual)*, pages 695–704.
- Ranka Stanković, Ivan Obradović, Cvetana Krstev, and Duško Vitas. 2011. Production of morphological dictionaries of multi-word units using a multipurpose tool. In *Proceedings of the Computational Linguistics-Applications Conference, October 2011, Jachranka, Poland*, pages 77–84.
- Ranka Stanković, Mihailo Škorić, and Branislava Šandrih Todorović. 2022. Parallel bidirectionally pre-trained taggers as feature generators. *Applied Sciences*, 12(10):5028.
- Ranka Stanković, Branislava Šandrih, Cvetana Krstev, Miloš Utvić, and Mihailo Škorić. 2020. [Machine learning and deep neural network-based lemmatization and morphosyntactic tagging for serbian](#). In *Proc. of The 12th LREC*, pages 3947–3955, Marseille, France. European Language Resources Association.
- Miloš Utvić. 2011. Annotating the corpus of contemporary serbian. In *Proceedings of the INFOtheca '12 Conference*, pages 36–47.
- Duško Vitas and Cvetana Krstev. 2012. Processing of Corpora of Serbian Using Electronic Dictionaries. *Prace Filologiczne*, XVIII:279–292.
- Jovan Čudomirović. 2014. Mobilizacija publike: Izveštavanje dnevnih novina u srbiji o nastupima fudbalske reprezentacije. *Zbornik Matice srpske za filologiju i lingvistiku*, 57(2):143–159.

The Importance of Being Interoperable: Theoretical and Practical Implications in Converting TBX to OntoLex-Lemon

Andrea Bellandi

Cnr-Istituto di Linguistica Computazionale
“Antonio Zampolli”, Italy
andrea.bellandi@ilc.cnr.it

Silvia Piccini

Cnr-Istituto di Linguistica Computazionale
“Antonio Zampolli”, Italy
silvia.piccini@ilc.cnr.it

Giorgio Maria Di Nunzio

Dip. di Ingegneria dell’Informazione
Università di Padova, Italy
giorgiomaria.dinunzio@unipd.it

Federica Vezzani

Dip. di Studi Linguistici e Letterari
Università di Padova, Italy
federica.vezzani@unipd.it

Abstract

This paper introduces a methodology, design, and implementation of an interactive converter for transforming terminological data from the TermBase eXchange (TBX) format to the OntoLex-Lemon model. The paper highlights the differences between the two models, emphasizing their different technologies and data structures.

The proposed software architecture implements the conversion process through three main phases: analysis, filtering, and assembling. The analysis phase includes parsing the TBX file and generating an intermediate representation stored in a SQLite database. The filtering phase allows users to query and filter the data on the basis of their specific requirements. Finally, the assembling phase builds the OntoLex-Lemon lexicon by processing the filtered data and serializing it as RDF triples.

The converter aims to enable end users to actively participate in the conversion process, particularly in complex decision-making steps dealing with term variation, polysemy, and sense-concept relations.

1 Introduction

In the last decade Linked Data (LD) has been confirmed as one of the promising approaches for representing and connecting research data and metadata (Frey and Hellmann, 2021). In the context of linguistic resources, Linguistic Linked Open Data (LLOD) is a paradigm that promotes the publication and interlinking of resources such as lexicons, corpora, and terminologies. LLOD allows for a standardized way to access data, enabling researchers to explore, analyze, and utilize linguistic data for various language-related applications (Cimiano et al., 2020). Among the various data models, the OntoLex-Lemon model has gained

popularity as the de-facto standard for representing lexical data using the Resource Description Framework (RDF) to express the information on the Semantic Web as LD (McCrae et al., 2017). However, there are specific cases where some types of linguistic resources have their own standard formats. This is the case of terminological resources encoded according to the TermBase eXchange (TBX) ISO standard 30042¹ – an XML-based family of terminology exchange formats compliant with the Terminological Markup Framework (TMF - ISO 16642:2017)². TBX, as well as other LD approaches, ensures consistency and interoperability by establishing a common structure and vocabulary for describing terminology across different systems and applications.

A number of methods and approaches, like for example the TBX2RDF conversion system (Cimiano et al., 2015; Montiel-Ponsoda et al., 2015), have been proposed to convert terminological data from the XML-based TermBase eXchange (TBX) format to OntoLex-Lemon, enabling their integration into the linguistic Linked Data ecosystem. Guidelines for a virtualization approach known as Term-à-LLOD have been developed to facilitate this conversion process (di Buono et al., 2020). In addition, there have been recent efforts to enhance OntoLex-Lemon with a dedicated module for representing terminology information³.

Our proposal focuses on the mismatches between the two representations (one terminographical the other lexicographical) that, in order to be tackled and solved, require a necessary intervention of the user. In fact, these mismatches call into ques-

¹<https://www.iso.org/standard/62510.html>

²<https://www.iso.org/standard/56063.html>

³<https://www.w3.org/community/ontolex/wiki/Terminology>

tion theoretical aspects that have been neglected by the previous works and that instead require active decisions by the scholars interested in converting their own data. In particular, the theoretical aspects related to this work have been discussed in a seminal paper (Piccini et al., 2023) and have been taken up, inspiring the preliminary design and implementation of such tool (Bellandi et al., 2023). We report here a brief summary of the considerations presented by (Bellandi et al., 2023):

- **lexicographical vs. terminological view.** A purely terminological vision (TBX) is transformed into a lexicographic standpoint (Ontolex-Lemon), where the conceptual dimension is not longer central and, conversely, sense acquires a crucial role.
- **ontology reuse.** The LD paradigm strongly encourages the reuse of existing vocabularies. According to this principle, the converter should make it possible to decide which data categories to use.
- **deductive rules.** The structure of the TBX file has some implicit relations among terms that get lost in the conversion from TBX to OntoLex-Lemon. The most important one is the information about synonymy among terms.
- **knowledge extraction.** In some cases the terminographer does not have a specific data category available in the TBX file to describe a particular behavior of the term. In such cases he/she can simply use the «note» field to store that information.
- **enriching the TBX.** After the knowledge extraction from unstructured notes, we can enrich the original TBX as well as its OntoLex-Lemon counterpart with the new extracted information.

In this paper, we focus on the methodology, design, and implementation of the interactive converter that will allow terminologists to actively participate in the conversion process. In particular, we describe the conversion steps that require the user to make decisions about aspects such as variation, polysemy, and sense-concept relations.

2 How do TBX and *lemon* Differ

In this section, we briefly summarize the differences between TBX and OntoLex-lemon.

A basic key difference between the two models lies in their underlying technologies: TBX utilizes XML as its representation language, while Ontolex-Lemon is based on RDF and leverages the semantic capabilities of the Semantic Web. This distinction influences the way data is structured and the interoperability possibilities with other linked data resources. However, it is important to recognize that converting TBX to LD involves more than a shift from an XML-based to an RDF-based structure; it requires theoretical reflection and consideration of the conceptual and organizational differences between the two models (Piccini et al., 2022). In fact, the organizational differences are also reflected by the aim of the two models: TBX primarily emphasizes the exchange and management of terminological resources, ensuring consistency and interoperability among terminologists and language professionals. In contrast, Ontolex-Lemon is specifically tailored for representing lexical data, aiming to capture detailed linguistic information and to enable semantic integration with other RDF datasets.

The objective of this paper is therefore to examine the prerequisites of a converter capable of processing the latest editions of TBX and Ontolex-Lemon. The analysis will particularly concentrate on the theoretical consequences that arise from the shift from a structure centered on concepts (TBX) to one centered on senses (Ontolex-Lemon).

3 Towards a TBX to *lemon* Converter

Given the different nature of the two models, we propose to create an interactive and configurable converter that can indulge the theoretical vision of the user who carry out the conversion, whether they are terminologists, translators, or lexicographers. In light of this, converting a TBX resource to Ontolex-Lemon should require a dedicated software architecture as depicted in Figure 1. The latter translates a TBX source into RDF triples, going interactively through three main phases: i) *analysis*, ii) *filtering*, and iii) *assembling*.

3.1 Phase 1: Analysis

Concerning the first phase, the parser component is in charge of analyzing the XML input file, potentially written in different TBX public dialects

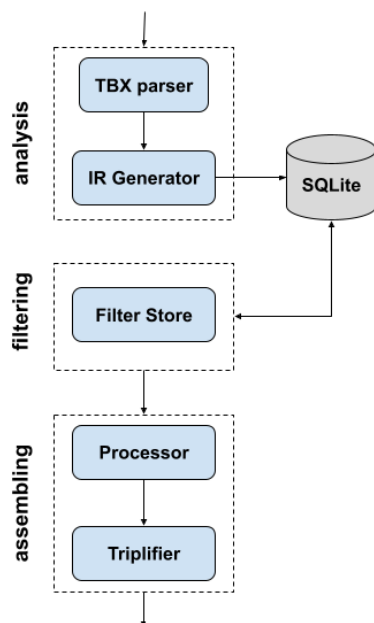


Figure 1: The architecture of the three phases of the converter from a TBX to an Ontolex-lemon representation.

(core, min, basic), and is aimed at producing an intermediate representation (IR) of the information contained. IR represents a partial conversion of the TBX elements such as concepts, terms, and languages, in a series of RDF triples, without making any assumption on the final output.

3.2 Phase 2: Filtering

IR is stored in a SQLite database, together with some metadata (for example transaction types, creation dates, subject fields), allowing the *filtering* phase to implement fast and feasible querying for user-specific filters to select and eventually enrich the data.

3.3 Phase 3: Assembling

Starting from the filtered data, the third phase constructs the Ontolex-Lemon lexicon by processing the languages, the concepts, and the terms (the Processor component in Figure 1), and serializes them as RDF triples according to the Ontolex-Lemon data model (the triplifier component in Figure 1).

The Processor is the crucial component of the software architecture because it is in charge of taking into account the desiderata of the user who makes the conversion. It potentially can be composed of a pipeline of processors that implements those desiderata starting from the IR data, for example:

- bypassing the Ontolex-Lemon Lexical Sense class and linking lexical entries directly to the designated concepts,
- linking the terms denoting the same concept across different languages by means of the translation property,
- creating polysemous entries in Ontolex-Lemon in those cases in which the terms designate different concepts but are characterized by the same orthographic form and share the same etymology,
- creating relationships of synonymy between terms designating the same concept in a given language.

Currently, the software prototype performs a conversion process based on the default behavior. The following section is devoted to presenting a simple example of default conversion.

4 A Conversion Example

The hierarchical structure of a TBX file is basically the following:

- a set of concept entries (tag <conceptEntry>),
- within each concept entry, a set of language sections (tag <langSec>),
- for each language section, a set of terms that designate the concept for that language (tag <termSec>).

Figure 2 depicts a fragment of an example of a TBX-basic terminological database with one concept. In particular,

- the fragment, reports a concept called *c1*, related to the e-mobility field,
- and two language sections, for English and French, with their respective definitions for that concept.
- There are two terms for concept *c1* in English, neighborhood "car vehicle" and "NEV", while one in French, "véhicule de proximité". For each term, some kind of information is specified, such as morphology, term type, and administrative status.

```

<conceptEntry id="c1">
  <min:subjectField>e-mobility</min:subjectField>
  <langSec xml:lang="en">
    <descripGrp>
      <basic:definition>A battery-electric car that is
        capable of traveling at a maximum speed of 25 miles
        per hour (mph) and has a maximum loaded weight
        of 3,000 lbs.
      </basic:definition>
    </descripGrp>
    <termSec>
      <term>neighborhood electric vehicle</term>
      <basic:termType>fullForm</basic:termType>
      <min:partOfSpeech>noun</min:partOfSpeech>
      <basic:grammaticalGender>masculine
      </basic:grammaticalGender>
      <min:administrativeStatus>preferredTerm-admn-sts
      </min:administrativeStatus>
    </termSec>
    <termSec>
      <term>NEV</term>
      <basic:termType>acronym</basic:termType>
      <min:partOfSpeech>noun</min:partOfSpeech>
      <min:administrativeStatus>admittedTerm-admn-sts
      </min:administrativeStatus>
    </termSec>
  </langSec>
  <langSec xml:lang="fr">
    <descripGrp>
      <basic:definition>Véhicule à deux places,
        activé par un moteur électrique à courant
        continu alimenté par des batteries au plomb
        rechargeables à partir d'une prise de courant
        résidentielle de 110 volts.
      </basic:definition>
    </descripGrp>
    <termSec>
      <term>véhicule de proximité</term>
      <basic:termType>fullForm</basic:termType>
      <min:partOfSpeech>noun</min:partOfSpeech>
      <min:administrativeStatus>admittedTerm-admn-sts
      </min:administrativeStatus>
    </termSec>
  </langSec>
</conceptEntry>

```

Figure 2: A TBX-basic dialect example.

Our converter performs the conversion and the result is reported in Figure 3. RDF triples are encoded in turtle syntax, and they are grouped according to the TBX entities they correspond to.

Concerning the <conceptEntry> entity, concepts are converted by means of the SKOS ontology, according to [Reineke and Romary \(2019\)](#). All the subject fields correspond to SKOS concept schemes, while concepts are mapped to SKOS concepts. The membership of concepts to their subject fields is formalized through the SKOS *inScheme* relationship. The SKOS *definition* property of a concept represents the definition of that concept provided by the TBX resource, whether the definition is given at the concept level or at the language level. Figure 2 reports an example related to the latter case. A definition of the concept in each language is formalized as Figure 3 shows. Other TBX data categories, such as note, source, and cross reference, are mapped to SKOS *note*, Dublin core *source*, and RDF *seeAlso* properties, respectively.

Concerning the <langSec> entity, the related *lemon* lexica are created. Referring to the example in Figure 2, both English and French lexica are defined as in the second group of triples in Fig-

concepts

```

:c1 a skos:Concept ;
  skos:prefLabel "c1"@en ;
  skos:inScheme :sbjf_1 ;
  skos:definition "A battery .."@en ;
  skos:definition "Véichule à deux .."@fr .

:sbjf_1 a skos:ConceptScheme ;
  skos:prefLabel "e-mobility"@en .

```

languages

```

:lexEN a lime:Lexicon ;
  dct:language "en" ;
  lime:entry :t1, :t2 .

:lexFR a lime:Lexicon ;
  dct:language "fr" ;
  lime:entry :t3 .

```

terms

```

:t1 a ontolex:LexicalEntry ;
  lexinfo:partOfSpeech lexinfo:noun ;
  lexinfo:termType lexinfo:fullForm ;
  ontolex:canonicalForm [
    lexinfo:gender lexinfo:masculine ;
    ontolex:writtenRep "neighborhood
      electric veichle"@en .
  ] ;
  ontolex:sense :t1_sense .

:t1_sense a ontolex:LexicalSense ;
  lexinfo:normativeAuthorization
  lexinfo:preferredTerm .

:t2 a ontolex:LexicalEntry ;
  lexinfo:partOfSpeech lexinfo:noun ;
  lexinfo:termType lexinfo:acronym ;
  ontolex:canonicalForm [
    ontolex:writtenRep "NEV"@en .
  ] ;
  ontolex:sense :t2_sense .

:t2_sense a ontolex:LexicalSense ;
  lexinfo:normativeAuthorization
  lexinfo:admittedTerm .

:t2 lexinfo:acronymFor :t1 .

:t3 a ontolex:LexicalEntry ;
  lexinfo:partOfSpeech lexinfo:noun ;
  ontolex:canonicalForm [
    ontolex:writtenRep "véichule de proximité"@fr .
  ] ;
  ontolex:sense :t3_sense .

:t3_sense a ontolex:LexicalSense ;
  lexinfo:normativeAuthorization
  lexinfo:admittedTerm .

```

Figure 3: The converted data in *lemon* is serialized by means of the turtle syntax.

ure 3. Furthermore, the terms of each language are defined as entries of the suitable lexicon. If the definition contained in the <langSec> had had a source or/and an external reference, we would have used the reification mechanism⁴ in order to represent the source and the reference of the concept definition, by means of Dublin core *source*, and RDF *seeAlso* properties, respectively.

Finally, terms contained in the <termSec> entity are represented as lexical entries in the Ontolex-Lemon model. Each term is mapped to a Lexical Entry element, without specifying its particular type (word or multi-word), and it is represented as a canonical form of that lexical entry. According to the "*semantics by reference*" paradigm of Ontolex-Lemon, the meaning of a lexical entry is

⁴The reification is a mechanism allowing to write RDF triples about RDF triples. In our case, we could specify both the source and the link of concept definitions.

specified by referring to the created SKOS concept that represents its meaning. The default conversion process creates a lexical sense for each lexical entry and links it to the suitable concept by means of the *reference* property. Since the model does not contain a complete collection of linguistic categories, it relies on Lexinfo vocabulary⁵. As a consequence, morphological information, such as part of speech, gender, and number is associated with the forms, while usage context, term type, and administrative status are associated with the senses, according to the Lexinfo schema.

5 Conclusion and Future Work

In this paper, we have presented the current work on the definition of a methodology for the conversion of terminological data as well as the design and implementation of an interactive converter from TBX to Ontolex-Lemon. Despite the already available tools for this type of conversion, we believe that transforming TBX data to Ontolex-Lemon can be more challenging than just carefully mapping and transforming all of the (meta)data of the different elements from one model to another. In fact, the two different frameworks (TBX concept-oriented and Ontolex-Lemon sense-centered) necessitate a deep understanding of both models and the ability to reconcile the differences in their structures and semantics during the conversion process.

The current prototype of the conversion tool allows the user to explore and analyze the structure (what data categories are available) and the statistics (how many concepts, languages, and terms) of the TBX file. In addition, the user can also make some choices about the mapping and identification of TBX concepts into SKOS concepts across different languages and from TBX terms to Ontolex-lemon lexical concepts. As future work, we are currently working on parameterizing the default behavior on some steps such as:

- make explicit the choice of the use of Ontolex-lemon senses (or not);
- make explicit the decision of the management of synonymy and the equivalents across multiple languages;
- extrapolate information from TBX textual data categories (for example the element

<note>) that can be mapped into Ontolex-lemon properties.

6 Acknowledgment

This work has been carried out in the framework of agreement between Consiglio Nazionale delle Ricerche – Istituto di Linguistica Computazionale and RUT Foundation. This work is also part of the initiatives carried out by the Center for Studies in Computational Terminology (CENTRICO) of the University of Padua and in the research directions of the Italian Common Language Resources and Technology Infrastructure CLARIN-IT.

References

- Andrea Bellandi, Giorgio Maria Di Nunzio, Silvia Piccini, and Federica Vezzani. 2023. *From TBX to Ontolex Lemon: Issues and Desiderata*. In *Proceedings of the 2nd International Conference on Multilingual Digital Terminology Today (MDTT 2023)*, volume 3427 of *CEUR Workshop Proceedings*, Lisbon, Portugal. CEUR. ISSN: 1613-0073.
- Philipp Cimiano, Christian Chiacros, John P. McCrae, and Jorge Gracia. 2020. *Linguistic Linked Open Data Cloud*, pages 29–41. Springer International Publishing, Cham.
- Philipp Cimiano, John P. McCrae, Víctor Rodríguez-Doncel, Tatiana Gornostay, Asunción Gómez-Pérez, Benjamin Siemoneit, and Andis Lagzdins. 2015. *Linked terminologies: applying linked data principles to terminological resources*. In *Proceedings of the eLex 2015 Conference*, pages 504–517.
- Maria Pia di Buono, Philipp Cimiano, Mohammad Fazleh Elahi, and Frank Grimm. 2020. *Terme-à-LLOD: Simplifying the Conversion and Hosting of Terminological Resources as Linked Data*. In *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)*, pages 28–35, Marseille, France. European Language Resources Association.
- Johannes Frey and Sebastian Hellmann. 2021. *FAIR Linked Data - Towards a Linked Data Backbone for Users and Machines*. In *Companion Proceedings of the Web Conference 2021, WWW '21*, pages 431–435, New York, NY, USA. Association for Computing Machinery.
- John McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. *The OntoLex-Lemon Model: Development and Applications*. In *Electronic lexicography in the 21st century: Lexicography from scratch. Proceedings of eLex 2017*, pages 587–597, Brno. Lexical Computing CZ s.r.o. <http://en.wikipedia.org/wiki/Galway>; <https://en.wikipedia.org/wiki/Leiden>.

⁵LexInfo is an ontology that provides data categories for the *lemon* model. Please, see <https://lexinfo.net/>

Elena Montiel-Ponsoda, Julia Bosque-Gil, Jorge Gracia, Guadalupe Aguado de Cea, and Daniel Vila-Suero. 2015. Towards the Integration of Multilingual Terminologies: an Example of a Linked Data Prototype. In *Terminology and Artificial Intelligence (TAI)*, pages 205–206.

Silvia Piccini, Federica Vezzani, and Andrea Bellandi. 2022. [Entre TBX et Ontolex-Lemon : Quelles Nouvelles Perspectives en Terminologie?](#) (poster). In *Proceedings of the 1st International Conference on Multilingual Digital Terminology Today*, volume 3161 of *CEUR Workshop Proceedings*, Padua, Italy. CEUR. ISSN: 1613-0073.

Silvia Piccini, Federica Vezzani, and Andrea Bellandi. 2023. [TBX and ‘Lemon’: What perspectives in terminology?](#) *Digital Scholarship in the Humanities*, 38(Supplement_1):i61–i72.

Detlef Reineke and Laurent Romary. 2019. [Bridging the gap between SKOS and TBX.](#) *edition - Die Fachzeitschrift für Terminologie*, 19(2). Publisher: Deutscher Terminologie-Tag e.V. (DTT).

Formalizing Translation Equivalence and Lexico-Semantic Relations Among Terms in a Bilingual Terminological Resource

Giulia Speranza, Maria Pia di Buono and Johanna Monti

University of Naples "L'Orientale"

{gsperanza, mpdibuono, jmonti}@unior.it

Abstract

In this paper we investigate the feasibility of applying the Semantic Web formalisms, in particular the OntoLex-Lemon model, to represent bilingual terminological resources, both from a conceptual and a lexico-semantic point of view. As a proof of concept for our study we select a bilingual Italian-English terminological resource in the specialized domain of archaeology, in order to identify possible modelling solutions as well as potential challenges.

1 Introduction

Recent years have witnessed a significant increase in the conversion and development of lexical resources into RDF, following the Linguistic Linked Open Data (LLOD) principles¹. Indeed, there is a growing recognition of the importance of the interoperability, reuse and accessibility of data, also in the field of language resources Khan et al. (2022). The employment of Semantic Web formalisms, such as the OntoLex-Lemon model, allows the enrichment of linguistic and terminological resources with structured semantic information, making them easily integrated with other semantic resources, such as ontologies, linked datasets and semantic knowledge bases, thus preventing the so-called data-silos. The rich semantic information that can be easily represented in a resources by means of the Semantic Web formalisms is also beneficial in many applicative scenarios where Natural Language Processing (NLP) is concerned.

Among several formalisms that have been proposed for the formalization of such resources, the OntoLex-Lemon model allows to represent in detail the meaning of terms, the semantic relationships between them, and other related linguistic information, enabling a complete and accurate representation of the entries in terminological resources.

¹https://en.wikipedia.org/wiki/Linguistic_Linked_Open_Data

Furthermore, the OntoLex-Lemon model is flexible and easily extendable, offering several representation possibilities to meet different formalization needs.

These achievements are also due to the efforts, experiments, and proposals of a community of researchers and scholars of the W3C Ontology-Lexica Community Group² and the Nexus Linguarum COST Action³, who collaborate on the systematization of models and modules that continue to evolve in order to meet the needs of the LLOD community.

The LLOD principles are being applied to the formalization of several types of resources. Indeed, the analysis carried out by di Buono et al. (2022) about the existing resources and their metadata used to represent them within the LOD Cloud and AnnoHub, which resulted in the creation of METASHARE Enriched LLD (MELLD)⁴, a new enriched metadata resource, show that out of the 666 total LLOD resources, 315 are Corpora, 303 are Lexicons and Dictionaries and only 30 are catalogued as Terminologies, Thesauri and Knowledge Bases.

Furthermore, the comprehensive survey by Gro-mann et al. (forthcoming) sheds light on the linguistic description levels represented in the LLOD resources available and reports several studies focused on the description of the Translation and Terminology level.

Finally, for the description and representation of the terminologies some proposals are also emerging and being discussed such as the TermLex (Martín-Chozas and Declerck, 2022), an extension module for the OntoLex-Lemon model.

In order to contribute to the discussion we investigate the feasibility of applying the Semantic

²<https://www.w3.org/community/ontolex/>

³<https://nexuslinguarum.eu/>

⁴<https://github.com/unior-nlp-research-group/melld>

Web formalisms, in particular the OntoLex-Lemon model, to represent bilingual terminological resources, both from a conceptual and a lexical point of view. As a proof of concept for our study we select a bilingual Italian-English terminological resource in the specialized domain of archaeology, in order to identify possible modelling solutions as well as potential challenges.

2 Case Study

As case study for our modelling experiment we select a bilingual Italian-English terminological resource (TR) in the specialized domain of archaeology.

The TR has been created by means of a semi-automatic extraction process based on appositional constructions from a parallel domain corpus (Speranza et al., 2021, 2022). The TR is composed of 300 terms in each language in the form of single and multi-word units (MWUs) terms.

Furthermore, by means of the terminology extraction methodology previously applied to create the TR, we were also able to enrich it with other information such as Part of Speech (PoS), terminological variants, examples of terms in the context of a sentence and reformulations of technical terms. The inclusion of lay reformulations of technical terms, retrieved hinging on appositional constructions structures, can be a useful information in a TR to be employed for the simplification and exemplification of technicalisms in different communicative scenarios involving experts and non-experts.

Starting from our case study our representation needs concern the following information:

- **Terminological entry:** Single and multi-word terms, Syntactic and grammatical information (PoS, gender and number), Context (Example sentence)
- **Lexico-semantic relations:** diaphasic and synonymous variants and taxonomical and translation equivalence relations

3 Modelling Strategy

In order to formalize the TR according to the Linked Open Data principles applied to Linguistics (Cimiano et al., 2020), we choose to adopt the OntoLex-Lemon core model, including some of its specific modules (see table 1), such as the Variation and Translation Module (`vartrans`), the

Decomposition Module (`decomp`) as well as the `LexInfo`.

Furthermore, since we also need to represent the Conceptual level of the entries we use the `Skos Models`.

Prefix	Namespaces
ontolex	http://www.w3.org/ns/lemon/ontolex#
vartrans	http://www.w3.org/ns/lemon/vartrans#
decomp	http://www.w3.org/ns/lemon/decomp#
lexinfo	http://www.lexinfo.net/ontology/2.0/lexinfo#
skos	http://www.w3.org/2004/02/skos#

Table 1: Models and modules' prefixes and namespaces

In particular, we use the Ontolex-Lemon core model to formalize the terminological entries and we use the `LexInfo` Model as the Data Category Ontology for the representation of grammatical information about the terms.

Furthermore, the `decomp` module is used for representing the internal structure of MWUs terms, since in our TR many MWUs are endocentric MWUs which present a fixed head, which is usually post-modified by prepositional phrases or through adjectival post-modification as in *anfora a piramide*, *anfora da trasporto*, *anfora punica*.

In addition, we use `skos` for reporting an example of sentence containing the term, which can also be useful for the user of a TR.

Then, we use the `vartrans` module for representing both the monolingual lexico-semantic relations in Italian or English and the translation equivalence relations between the two languages. Indeed, the `vartrans` module has been developed to record "lexico-semantic relations across entries in the same or different languages" (Montiel-Ponsoda et al., 2015). In addition, translation relations in Ontolex-Lemon are intended as a special type of lexico-semantic variation (Bosque-Gil et al., 2015) or a special case of a sense relation (McCrae et al., 2017).

Finally, in order to provide for each terminological entry in the resource a conceptual scheme, we use the SKOS Core Vocabulary⁵. SKOS is in fact used for expressing the basic structure of concept schemes i.e., thesauri, taxonomies, terminologies, glossaries and other types of controlled vocabulary.

⁵<https://www.w3.org/TR/2009/NOTE-skos-primer-20090818/>

3.1 Conceptual Level

Following the Ontolex-Lemon module Specifications⁶, SKOS and Ontolex-Lemon can be used in conjunction to provide more detailed information about the "labels". As a consequence, by means of the `skos:concept` property we choose to link each lexical entry to the conceptual schema proposed in the Italian *Istituto Centrale per il Catalogo e la Documentazione* (ICCD) Thesaurus of Archaeological Finds in the SKOS version. The ICCD's Thesaurus is indeed organized according to a hierarchical classification which provides general categories (macro-categories) and specific categories (sub-categories) to conceptually organize the archaeological terms.

For example, the archaeological find *amuleto* (amulet) is a term listed under the macro-category (I° Level) *Strumenti, Utensili e Oggetti d'Uso* (Tools); more precisely belonging to the sub-category (II° level) *Amuleti e Oggetti per uso cerimoniale, magico e votivo* (Magic and votive supplies) (Di Buono, 2015).

In the SKOS version of the ICCD's Thesaurus Felicetti et al. (2013) converted the 10 macro-categories of the taxonomic hierarchy of the ICCD's Thesaurus into different corresponding URIs distinguished by different identifiers from 001 to 010, representing different macro-categories (i.e., *Abbigliamento e Ornamenti personali* (Clothing and Accessories) (001), *Arredi* (Furnishing) (002), *Edilizia* (Building) (003), etc.), linked by means of the `skos:hasTopConcept` property.

In such a way, the Italian lexical entry *anfora da trasporto* can be connected to the conceptual level by means of the `ontolex:sense` property and the lexical sense can point to the `skos:Concept` by means of the `ontolex:reference` property, thus reusing previously set URIs to uniquely identify the concepts in our TR (see figure 1).

Linking each lexical entry to an ontology entity in the CIDOC Conceptual Reference Model (CRM) (Doerr, 2003), which is the reference ontology for Cultural Heritage domain, even if the OntoLex-Lemon module easily allows this operation by means of the `ontolex:denotes` property, would only provide us with a single class for linking our terms in the archaeological domain, namely E22 Human-Made Object, since all of our terms conceptually belong to the class of objects made by humans (Human Made Objects).

⁶<https://www.w3.org/2016/05/ontolex/>

```

:anfora_da_trasporto_lex a
  ontolex:lexicalEntry, ontolex:
  MultiWord;
  ontolex:sense
    :anfora_da_trasporto_sense

:anfora_da_trasporto_sense a
  ontolex:LexicalSense;
  ontolex:reference
    <https://dati.beniculturali.it/lodview/vocabularies/reperti_archeologici/def/009.005.000.005.002>

    <https://dati.beniculturali.it/lodview/vocabularies/reperti_archeologici/def/009.005.000.005.002> a
      ontolex:LexicalConcept;
      skos:inScheme
        <https://dati.beniculturali.it/lodview/vocabularies/reperti_archeologici/def/>;

```

Figure 1: RDF serialization of the conceptual level of the term *anfora da piramide*

3.2 Terminological Entry Level

In order to test the representation of the grammatical information of the terms, we report in Figure 2 the formalization of the Italian lexical entry *anfora da trasporto*.

By means of the Ontolex-Lemon core model we are able to represent different information such as the type of forms a lexical entry can have: a canonical form (*anfora da trasporto*) and another form (*anfоре da trasporto*). With LexInfo we can further specify some grammatical and syntactic information such as the number (singular and plural), the gender (masculine or feminine) and the PoS about the term.

The decomposition of the MWU terms is realized resorting to the property `decomp:constituent` that relates a lexical entry to its components, as in figure 3.

Moreover, by means of the property `decomp:correspondsTo` we are also able to link the single components of the MWU to the corresponding lexical entries, enabling, as a consequence, the further specification of the linguistic information connected with the lexical entries. Finally, in order to specify the order of the components, it is possible to use the RDF properties `rdf:_1`, `rdf:_2`, etc.

In addition, in our TR we also provide for each entry an example sentence containing the term extracted from the parallel corpus. We formalize this information resorting to the `skos` module which

```

:anfora_da_trasporto_lex a
  ontolex:lexicalEntry, ontolex:
  MultiWord;
  ontolex: canonicalForm
    :form_anfora_da_trasporto_sn;
  ontolex: otherForm
    :form_anfore_da_trasporto_pl;

:form_anfora_da_trasporto_sn a
  ontolex:Form;
  ontolex:writtenRep
    "anfora da trasporto"@it;
  lexinfo:partOfSpeech lexinfo:noun;
  lexinfo:gender lexinfo:feminine;
  lexinfo:number lexinfo:singular.

:form_anfore_da_trasporto_pl a
  ontolex:Form;
  ontolex:writtenRep
    "anfora da trasportoe"@it;
  lexinfo:partOfSpeech lexinfo:noun;
  lexinfo:gender lexinfo:feminine;
  lexinfo:number lexinfo:plural.

```

Figure 2: RDF serialization of the term *anfora da trasporto*

offers the possibility to use the `skos:example` property, as in the figure 4 but it could also be represented resorting to the OntoLex module for Frequency, Attestations, and Corpus-Based Information (OntoLex-FrAC) (Chiarcos et al., 2022), as example sentences are, in our case, corpus attestations.

3.3 Lexico-semantic Relations

3.3.1 Diaphasic variations

As far as the monolingual terminological variation in each language is concerned, the OntoLex-Lemon model Specifications include the diatopic, diaphasic, diachronic, diastratic and dimensional variants as examples of terminological relations.

In our TR, we mainly need to represent the diaphasic relations, especially when Latin or Greek origin terms coexist with the target language variants and are employed in different communicative registers, namely in different communicative situations (Montiel-Ponsoda et al., 2013). In this case, both terminological variants share the same conceptual meaning by pointing to the same `Skos:concept`, while changing their respective surface forms. Therefore, by means of the class `vartrans:TerminologicalVariants` and the property `vartrans:category:diaphasic` we are able to frame this kind of terminological relation between functional variants

```

:anfora_da_trasporto_lex a
  ontolex:LexicalEntry;
  decomp:constituent :anfora_component;
  rdf:_1 :anfora_component;
  decomp:constituent :da_component;
  rdf:_2 :da_component;
  decomp:constituent
    :trasporto_component;
  rdf:_3 :trasporto_component;

:anfora_component a decomp:Component;
  decomp:correspondsTo :anfora_lex.
:da_component a decomp:Component;
  decomp:correspondsTo :da_lex;
:trasporto_component a decomp:Component;
  decomp:correspondsTo :trasporto_lex.

```

Figure 3: RDF serialization of the MWU decomposition of the term *anfora da trasporto*

```

:anfora_da_trasporto_sense a
  ontolex:LexicalSense;
  ontolex: reference
  <https://dati.beniculturali.it/lodview/vocabularies/reperti_archeologici/def/009.005.000.005.002>

  <https://dati.beniculturali.it/lodview/vocabularies/reperti_archeologici/def/009.005.000.005.002> a
    ontolex:LexicalConcept;
  skos:inScheme
  <https://dati.beniculturali.it/lodview/vocabularies/reperti_archeologici/def/>;
  skos:example
  rdf:value "Significativa anche la
  presenza di un'anfora da
  trasporto di produzione greca"
  @it

```

Figure 4: RDF serialization of the context sentence example for the term *anfora da trasporto*

as in the example of the term *foculo* and its Latin origin variant *foculum* in Figure 5.

3.3.2 Taxonomic relations

In the modelling phase we are also confronted with the need of representing the semantic relation of hypernymy/hyponymy, which can be represented with the `vartrans` module in combination with the `LexInfo` categories (`LexInfo:hypernym` or `LexInfo:hyponym`). We use the property `vartrans:senseRelation`, which connects together two lexical entries' senses and allows the declaration of the `category:hypernym` and the indication of the relation direction from the source to the target term. In Figure 6 we report the example of the formalization of the relation

```

:foculo_lex a ontolex:LexicalEntry;
ontolex:lexicalForm :foculo_form;
ontolex:sense :foculo_sense.
:foculo_form ontolex:writtenRep
  "foculo"@it.
:foculo_sense ontolex:reference
<https://dati.beniculturali.it/lodview/vocabularies/reperti_archeologici/def/009.000.000.011>

:foculum_lex a ontolex:LexicalEntry;
ontolex:lexicalForm :foculum_form
;
ontolex:sense :foculum_sense.
:foculum_form ontolex:writtenRep
  "foculum"@it .
:foculum_sense ontolex:reference
<https://dati.beniculturali.it/lodview/vocabularies/reperti_archeologici/def/009.000.000.011>

:foculo_foculum_relation a
vartrans:TerminologicalRelation;
vartrans:source :foculo_sense;
vartrans:target :foculum_sense;
vartrans:category :diaphasic.

```

Figure 5: RDF serialization of the diaphasic terminological relation between the entries *foculo* and *foculum*

between the term *rython* which is a hyponym and *coppa* (cup) which is its hypernym, namely a more generic term.

3.3.3 Synonymous reformulations

By means of the methodology applied to extract bilingual terms from the parallel corpus which is based on a special kind of linguistic constructions between brackets named appositional constructions, we were able to retrieve from our parallel corpus terms and their exemplifications or simplifications in Italian (a) and English (b). Technically speaking, we were able to retrieve *anchors* and *supplements*, which are the two elements composing the appositional construction (Huddleston and Pullum, 2005) as in the example (1).

- (1) a. **rhyton** (una coppa a forma di corno)
 b. **rhyton** (a horn-shaped cup)

In a terminological resource it could be useful to also include this kind of synonymous reformulation of technical terms.

In this specific case, the `skos:definition` property is not taken into consideration since what we need to formalize is not a canonical definition as intended by the ISO 1087:2019⁷: "Representation

⁷<https://www.iso.org/obp/ui/#iso:std:>

```

:rython_lex a ontolex:LexicalEntry;
ontolex:sense :rython_sense;
ontolex:canonicalForm
  :rython_form.
:rython_sense ontolex:reference
<https://dati.beniculturali.it/lodview/vocabularies/reperti_archeologici/def/009.005.000.046>
:rython_form ontolex:writtenRep
  "rython"@it .

:coppa_lex a ontolex:LexicalEntry ;
ontolex:sense :coppa_sense ;
ontolex:canonicalForm :coppa_form.
:coppa_sense ontolex:reference
<https://dati.beniculturali.it/lodview/vocabularies/reperti_archeologici/def/009.005.000.046>
:coppa_form ontolex:writtenRep
  "coppa"@it .

:senseRelation a vartrans:SenseRelation;
vartrans:source :coppa_sense;
vartrans:target :rython_sense;
vartrans:category
  lexinfo:hypernym.

```

Figure 6: RDF serialization of the hypernymic relation between the terms *rython* and *coppa*

of a concept (3.2.7) by an expression that describes it and differentiates it from related concepts" which normally are much more complex and articulated (Magris, 1998).

This kind of reformulation could be intended as a very short descriptive definition of the term in plain language with the aim of simplify and explain the technical concept. From this point of view, they can not obviously include a fine-grained and nuanced level of definition.

3.3.4 Translation equivalence relations

Finally, since we need to formalize a bilingual TR, among the several possibilities provided in the OntoLex-Lemon model Specifications, we choose to represent equivalent translations by means of the `vartrans:Translation` class and the properties `vartrans:source` and `vartrans:target`, which also enable the explicit indication of the translation direction. The two lexical entries in the two languages (Italian and English) can be connected to the conceptual level by means of the `ontolex:sense` property, pointing to the `skos:Concept`. Since the two entries in the two languages share the same concept, they can be linked together in a relationship

`iso:1087:ed-2:v1:en`

of translation equivalence at sense level by means of the `vartrans` module, by even specifying the translation direction from the Italian source (*anfora da trasporto*) to the English target (transport amphora) (see figure 7).

```

:anfora_da_trasporto_sense
  ontolex:reference
  <https://dati.beniculturali.it/lodview/vocabularies/reperti_archeologici/def/009.005.000.005.002>
:transport_amphora_sense
  ontolex:reference
  <https://dati.beniculturali.it/lodview/vocabularies/reperti_archeologici/def/009.005.000.005.002>

:trans a vartrans:Translation ;
  vartrans:source
    :anfora_da_trasporto_sense ;
  vartrans:target
    :transport_amphora_sense .

```

Figure 7: RDF serialization of relation of translation equivalence between the lexical entry *anfora da trasporto* and *transport amphora*

4 Conclusions and Future Works

In this paper we tried to formalize a bilingual terminological resource in Italian and English using the vocabularies offered by the Semantic Web Formalisms.

OntoLex-Lemon model with its modules in conjunction with LexInfo and SKOS resulted to be detailed and flexible enough for covering all the representation needs of our specific TR both from the monolingual and the bilingual point of view.

During the modelling phase we were, nevertheless, confronted with the challenge of representing special kinds of synonymous reformulations extracted from the corpus that we wanted to include in the TR. Possible modelling solutions are offered by the Lexinfo category `synonym` which "Indicates the the terms have the same meaning lexicographically"⁸ or by the Lexinfo category `gloss`, which according to the TEI is "A phrase or word used to provide a gloss or definition for some other word or phrase."⁹ Nonetheless, these options might be limiting from one perspective, since they do not account for the actual status of linguistic reformulations of terminology in plain language.

⁸<https://lexinfo.net/index.html>

⁹<https://tei-c.org/release/doc/tei-p5-doc/it/html/examples-gloss.html>

Future works might therefore be needed to meet specific necessities related to particular representations as long as further information about terms such as reformulations of technical terms or very short descriptive definitions are needed to be included and addressed in a TR more directly.

Finally, in terms of applicability, terminological resource formalized with OntoLex-Model can also be easily converted in other formats which are also widely employed for the representation, storing and sharing of terminological resources, such as the TBX, which can be used in CAT-Tools for translation purposes.

References

- Julia Bosque-Gil, Jorge Gracia, Guadalupe Aguado-de Cea, and Elena Montiel-Ponsoda. 2015. Applying the ontolox model to a multilingual terminological resource. In *European Semantic Web Conference*, pages 283–294. Springer.
- C. Chiarcos, Elena Simona Apostol, Besim Kabashi, and Ciprian-Octavian Truică. 2022. Modelling frequency, attestation, and corpus-based information with ontolox-frac. In *International Conference on Computational Linguistics*.
- Philipp Cimiano, Christian Chiarcos, John P McCrae, and Jorge Gracia. 2020. Linguistic linked open data cloud. In *Linguistic Linked Data*, pages 29–41. Springer.
- Maria Pia Di Buono. 2015. Information extraction for ontology population tasks. an application to the italian archaeological domain. *International Journal of Computer Science: Theories and Applications*, 3(2):40–50.
- Maria Pia di Buono, Hugo Gonçalo Oliveira, Verginica Barbu Mititelu, Blerina Spahiu, and Gennaro Nolano. 2022. Paving the way for enriched metadata of linguistic linked data. *Semantic Web*, vol. 13(6):1133–1157.
- Martin Doerr. 2003. The cidoc conceptual reference module: an ontological approach to semantic interoperability of metadata. *AI magazine*, 24(3):75–75.
- Achille Felicetti, Tiziana Scarselli, Maria Letizia Mancinelli, and Franco Niccolucci. 2013. Mapping iccd archaeological data to cidoc-crm: the ra schema. In *CRMEX@ TPDFL*, pages 11–22.
- Dagmar Gromann, Elena-Simona Apostol, Christian Chiarcos, Marco Cremaschi, Jorge Gracia, Katerina Gkirtzou, Chaya Liebeskind, Verginica Mititelu, Liudmila Mockiene, Michael Rosner, et al. forthcoming. Multilinguality and llod: A survey across linguistic description levels. *Semantic Web*.

- Rodnry Huddleston and Geqffrry Pullum. 2005. The cambridge grammar of the english language. *Zeitschrift für Anglistik und Amerikanistik*, 53(2):193–194.
- Anas Fahad Khan, Christian Chiarcos, Thierry Declerck, Daniela Gifu, Elena Gonzalez-Blanco Garcia, Jorge Gracia, Maxim Ionov, Penny Labropoulou, Francesco Mambrini, John P McCrae, et al. 2022. When linguistics meets web technologies. recent advances in modelling linguistic linked open data. *Semantic Web Journal*.
- Marella Magris. 1998. La definizione in terminologia e nella traduzione specialistica. *EUT-Edizioni Università di Trieste*.
- Patricia Martín-Chozas and Thierry Declerck. 2022. Representing multilingual terminologies with ontolx-lemon. In *1st International Conference on "Multilingual digital terminology today. Design, representation formats and management systems"*, June 16 – 17, Padova, Italy.
- John P McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The ontolx-lemon model: development and applications. In *Proceedings of eLex 2017 conference*, pages 19–21.
- Elena Montiel-Ponsoda, Julia Bosque-Gil, Jorge Gracia, Guadalupe Aguado de Cea, and Daniel Vila-Suero. 2015. Towards the integration of multilingual terminologies: an example of a linked data prototype. In *TIA*, pages 205–206.
- Elena Montiel-Ponsoda, John P McCrae, Guadalupe Aguado de Cea, and Jorge Gracia del Río. 2013. Multilingual variation in the context of linked data. In *Proceedings of 10th International Conference on Terminology and Artificial Intelligence (TIA'13)*.
- Giulia Speranza, Maria Pia Di Buono, and Johanna Monti. 2021. Terms and appositions: What unstructured texts tell us. In *International Conference on Automatic Processing of Natural-Language Electronic Texts with NooJ*, pages 219–230. Springer.
- Giulia Speranza, Maria Pia Di Buono, and Johanna Monti. 2022. Tailoring terminological resources to the users' needs: a corpus-based study on appositive constructions in italian and english. In *CEUR Workshop Proceedings: 1st International Conference on "Multilingual Digital Terminology Today. Design, representation formats and management systems"*, 16 - 17 June 2022, Padua, Italy.

Domain-Specific Keyword Extraction using BERT

Jill Sammet

Kiel University
Kiel, Germany
jillsammet@web.de

Ralf Krestel

ZBW - Leibniz Information Centre
for Economics & Kiel University
Kiel, Germany
rkr@informatik.uni-kiel.de

Abstract

Maintaining domain-specific thesauri is a costly endeavor. Terms might get added, removed, or merged over time to reflect new trends and keep the thesaurus consistent. This work is done by domain experts following pre-defined rules. Instead of curating the thesaurus manually, we investigate the use of language models to automatically propose novel terms to be added. To this end, we present an approach for keyword extraction from titles and abstracts of domain-specific documents. We report results on fine-tuned BERT models and compare them with different baselines. We further show that our proposed approach outperforms others in various evaluation scenarios.

1 Introduction

The Thesaurus for Economics (STW) is the world-wide largest bilingual vocabulary used for representing and researching economics-related content. It consists of almost 6000 subject headings and more than 20.000 additional entry terms, both available in English and German. It broadly covers topics from the economics domain and other related fields (Kempf and Neubert, 2016). Numerous organizations, libraries, and institutions use the STW for subject indexing and research, e.g., the German Institute for Economic Research.¹ The thesaurus is provided by the Leibniz Information Centre for Economics (ZBW), a large information service provider with the worldwide largest stock of economics literature.² The thesaurus is currently maintained manually by a small team of domain experts. They are responsible for deciding whether new terms should be added to the thesaurus, removed, or merged, as well as for finding relationships between those terms. The thesaurus relies on term suggestions from users. To alleviate the task

of finding and selecting novel relevant terms manually, we propose a data-driven, automatic way to suggest novel terms for the thesaurus by automatically extracting keywords from domain-specific publications. This approach can not only be used for keyword suggestions for the STW, but also for finding terms for indexing of document collections. We investigate three pre-trained BERT models that are fine-tuned for the task of token classification with the goal to extract domain-specific keywords, which in turn can be filtered to find new suggestions for the thesaurus.

2 Related Work

In recent years, various BERT models have been proposed for the task of keyword and key phrase extraction: Lim et al. (2020) proposed an approach of using two pre-trained BERT models, namely BERT and SciBERT, and fine-tuned them on a task similar to named entity recognition. The former model is pre-trained on the English Wikipedia and the BookCorpus with 3.3B tokens (Devlin et al., 2018) and the latter on the Semantic Scholar Corpus with 3.1B tokens (Beltagy et al., 2019). For the fine-tuning, each token was assigned to a label, marking either the beginning, middle or end of a key phrase. The models have been evaluated on three different datasets: KDD, WWW and Inspec.³ KDD consists of abstracts of papers from the ACM conferences on Knowledge Discovery and Data Mining (KDD). WWW consists of abstracts from the World Wide Web Conference (WWW). Both KDD and WWW only include publications between 2004-2014, with 715 and 1330 documents respectively (Gollapalli and Caragea, 2014). Inspec consists of 2000 abstracts of scientific Computer Science journals between 1998 and 2002 (Hulth, 2003). Their reported results show that while their BERT model did not attain state-of-the-art results

¹<https://www.zbw.eu/de/stw-info/anwendungen/>

²<https://www.zbw.eu/en/about-zbw>

³<https://github.com/LIAAD/KeywordExtractor-Datasets>

as the maximum performance of their model differs from the state-of-the-art between 0.08 - 5.2%, their SciBERT model overtook the state-of-the-art in all of their datasets with a 3.92 - 8.57% improvement. Qian et al. (2021) proposed a BERT-based approach for extracting keywords from scientific texts. In their work, BERT is used to extract key sentences from abstracts of papers from the Wanfang database.⁴ by dividing abstracts into a set of sentences. For each sentence, BERT is then used to find other sentences with high semantic similarity to the sentence in question. These extracted sentences are then ranked by their similarity and eventually a set of sentences is extracted to further retrieve keywords from. The keyword extraction itself is done by a combination of term frequency-inverse document frequency (TF-IDF) weighting, latent Dirichlet allocation (LDA), and TextRank. The model was evaluated using precision, recall and F1-scores. The results showed an improvement of 1.5% in the F1-score compared to the approach without prior sentence extraction with BERT.

Borisov et al. (2021) also used BERT for keyword extraction by fine-tuning BERT for the task of named entity recognition. They labeled three datasets with a 1 if the word is a keyword, and with a 0 if it is not a keyword. They used two separate datasets, one based on articles from news pages, and one derived from the Qulac datasets for IR-keywords (Aliannejadi et al., 2019). They used two categories for the evaluation of the model: test dataset accuracy and human evaluation. For evaluating the test dataset accuracy they measure precision and recall, as well as the average correct tag identification (ACTI), which tests the overall quality of the assigned tags, e.g., if a word is correctly tagged as a keyword or not, and the correct per response fill (CpRF), which captures the ratio of fully and partially correct predictions. For the human evaluation, a team of human annotators scores each keyword on a score from 1 to 5. The BERT model showed promising results with a precision of 0.86 and a recall of 0.88. The ATCI score measured 0.97, implying that most of the tags have been correctly assigned. The CpRF score of 0.76 implies that two third of the terms have been correctly predicted. The human evaluation score was 3.96, indicating high quality keywords.

In 2022 BERT has been used for domain-specific keyword extraction in combination with an addi-

tional Bi-LSTM layer for a sequence labeling task (Pezzo, 2022) fine-tuned on statistics-related textbooks. BERT is used to generate the contextualized word embeddings for the input, which are then fed into a Bi-LSTM layer that helps with the classification of the tokens. Each token is assigned the label "0" if it is predicted as a keyword and the label "1" if not. The difference to the previously presented methods is that this approach is unsupervised, meaning the model has not been trained on labeled texts but on unlabeled texts. The results of the model showed that it performed better than other commonly used keyword extraction methods such as KeyBERT, TextRank, LDA, TF-IDF or TopicRank by a large margin. The model's F1-score was 59.10, whereas the highest F1-score of the compared models was 43.78, obtained by TopicRank.

3 Dataset

In this work, a dataset derived from ECONIS, an online catalogue that contains titles and abstracts from economics literature maintained by ZBW - Leibniz Information Centre for Economics from various economic domains, is used.⁵ From the ECONIS dataset, the title, abstract, and metadata of scientific publications are extracted. The full-text body is not used to minimize the complexity of the approaches. The chosen metadata contains the publication year and language of the document. Additionally, three sets of indexing terms are assigned to the publications: assigned by its authors, specialists, and the STW each. Specialists are people from ZBW, that are responsible for subject indexing of documents. They are also responsible for the STW indexing labels, but for that category only terms from the thesaurus can be considered. The dataset is further reduced to publications published between 2009–2021. These restrictions lead to a dataset with 575K entries.

4 Methods

Our approach consists of two steps. First, we fine-tune a BERT model and use it to classify tokens as keyword candidates. Second, we filter the obtained candidates based on frequency and trend.

⁴<http://www.wanfangdata.com>, accessed 07.07.2023

⁵<https://www.econbiz.de/Record/datenbank-econis-online-katalog-der-zbw/10001514790>, accessed 18.11.2022

4.1 Extraction Process

To extract domain-specific keywords from documents, three BERT models are fine-tuned for the task of token classification. The first model is SciBERT (Beltagy et al., 2019), which is pre-trained on the semantic scholar corpus. The second model, FinBERT, is pre-trained on financial-communication texts, namely the three financial corpora, *corporate reports 10-K & 10-Q*, *earnings call transcripts* and *analyst reports* (Huang et al., 2022). The third model considered is DistilBERT, which is the lighter version of the original *BertBase*. It is trained on Wikipedia and a book corpus (Sanh et al., 2020).

To train the models for the downstream task, a labeled dataset is needed. Binary labels are applied to the terms in the documents of the dataset. "1" implies a word is a keyword or part of a key phrase and "0" that the term is not a keyword or part of a key phrase. The labels are assigned to the word based on whether they belong to a term in the STW. Thus, the words of the term "tax consultancy" are each assigned the label "1", however, if the term "consultancy" occurs alone, it is assigned a "0", as it is not an entry in the STW on its own. To fine-tune and then evaluate the models, the dataset needs to be split into training and test set. A subset of the STW terms is randomly sampled and the documents containing any of those terms are assigned to the test set. This ensures that hold-out STW terms have not been seen during fine-tuning. Hereby it can be evaluated how many of these terms that the model has not seen during fine-tuning are predicted as keywords during the evaluation. This subset of terms is referred to as the *control set* and it amounts to 970 terms from which 457 are descriptors and 513 non-descriptors. Descriptors describe the preferred term used for a concept. Non-descriptors describe the same concept, but are secondary terms, e.g., synonyms. The test set thus contains 131K documents and the training set for fine-tuning 443K documents. Each BERT-model variant is fine-tuned for 3 epochs using the training set. The batch size of each model is 32, as recommended by the authors of BERT and the input token length is 512 tokens, the maximal input size for BERT-models (Devlin et al., 2018). The learning rate for fine-tuning is set to $5e - 5$.

4.2 Filtering Process

To be able to suggest new terms for a thesaurus, the extracted keywords from the given documents need to be further filtered, because not every extracted keyword is a valuable addition to the STW. The filtering process consists of multiple steps. First, from the pool of extracted keywords, terms are removed that are already part of the STW as well as duplicated terms. This includes singular and plural forms of STW terms.

In the next step, adjectives denoting affiliations to a country are removed, e.g. *French social reform* becomes *social reform*. The adjective makes the term too specific for it to be a relevant term for the STW, considering that the thesaurus needs to be as general as possible. After removing the adjectives, it is verified again whether these terms now belong to an existing entry of the STW, and removed if they do.

The next filter ensures the relevance and frequency of the keyword candidate. Two types of filtering methods are introduced: the frequency filter and the trend filter. The frequency filter considers the frequency of a keyword. If its frequency reaches a threshold, the term is selected as a potential keyword candidate. For the evaluation, a threshold of 300 was chosen. This threshold has been set empirically by analyzing the frequency of existing STW terms during the given time period in the ECONIS dataset. The second filtering method is the trend filter. It selects keywords based on whether their usage has increased in the last three years (between 2019–2021), compared to their frequency in 2009–2018. For this, the average frequencies of those time spans are compared. If the latter average frequency of the term has increased, it is considered as a keyword candidate. Both cases are considered as some terms might not have a high frequency overall, as they have not or barely been mentioned in the literature, but have had a strong increase in recent years, e.g., *Coronavirus* has had a strong increase in recent years for obvious reasons. These terms are just as important as words that are frequent in the literature overall. In the least step of the filtering process, the keyword candidates are standardized to a uniform format, e.g., all candidates are singularized with a capitalized first letter, e.g., *social reforms* becomes *Social reform*.

5 Evaluation

The performance of the proposed models is compared to three common keyword extraction methods: TF-IDF (Luhn, 1957), TextRank (Mihalcea and Tarau, 2004), and KeyBERT (Grootendorst, 2020).

5.1 Term Suggestion

First, each method is evaluated on how effectively it recognizes terms from the control set, thus from the subset of terms that the models have not seen in the fine-tuning phase. Table 1 shows the performance of the methods based on the number of found descriptors (D) and non-descriptors (ND) from the control set in the test set. Besides splitting up the set into descriptors and non-descriptors, each entry for a concept is considered, thus an entry is considered as found by the model if either the descriptor or any of the non-descriptors for this entry are found. An important note to make is that TF-IDF has been given an advantage for this evaluation: because TF-IDF only extracts unigrams from texts but a lot of the terms from the control set and the STW are n-grams, the 10 extracted keywords have been concatenated to one large sequence of terms for each document. It is then evaluated if each subterm of an n-gram occurs in this sequence, if it is the case, then the term is considered as found. If only a subset of words of the term has been found, then the term is not considered as being found. In practice, it would not be known what terms are expected to be found, thus every combination of the extracted terms would have to be considered.

Beginning with the results for the descriptors, TF-IDF has in fact found 100% of the descriptors with its given advantage. Aside from that, DistilBERT performed the best by finding about 84% of the descriptors in the control set. This leaves a 20% margin compared to the next best method, which is TextRank. However, the two remaining fine-tuned models SciBERT and FinBERT show worse results than DistilBERT and TextRank. The results for the non-descriptors show that this time TF-IDF only finds about 13% of the non-descriptors, thus performing the worst out of all evaluated methods. Again, DistilBERT shows the best performance by finding 61% of the non-descriptor terms, which shows a 20% increase compared to the results of TextRank once again. Hence, counting an STW entry as found if either the descriptor or any of the non-descriptors are found, TF-IDF results in

Table 1: Percentage of found terms from the control set in the test set

	D	ND	Both
DistilBERT	83.6%	61.0%	90.8%
SciBERT	55.6%	27.3%	68.1%
FinBERT	50.5%	24.0%	60.8%
KeyBERT	35.4%	24.8%	46.0%
TextRank	63.7%	41.5%	76.2%
TF-IDF	100.0%	12.7%	100.0%

finding 100% of the entries, due to its performance on the descriptors. DistilBERT extracts terms for nearly 91% of the entries from the control set, given its performance on both the descriptors and the non-descriptors. This shows that the DistilBERT model works well in finding new and domain-specific keywords from documents. However, SciBERT and FinBERT do not show promising results.

Besides the performance on the control set, it is also interesting how the extracted keywords compare to the labels assigned to the documents in the dataset, thus how many of the STW terms have been extracted as keywords by the methods. Therefore, precision, recall, and F1-scores are calculated for every method. Precision describes how many of the retrieved keywords are marked as keywords in the labeled dataset, while recall determines how many of the overall keywords have been retrieved (Roelleke, 2013). Table 2 shows these results when considering terms that have been retrieved only partially, as each term has its own label. With these measures, it can be evaluated how well the proposed models and other keyword extraction techniques can recognize the terms that are part of the STW. Based on these values, all proposed BERT models outperform the baseline methods by a large margin. SciBERT, FinBERT and DistilBERT have each resulted in precision and recall values higher than 94%. These values are very high, which is likely due to the fact that these models have been trained on documents containing a large amount of STW terms. Hence they are much more likely to extract these terms as keywords. The other methods lack the domain-expertise as they have not been trained on the same data. Aside from these models, TF-IDF (Luhn, 1957) performed the best from the baseline methods, but it only reached values of up to 44%, thus resulting in a large margin compared to the fine-tuned BERT models. This shows the advantage of training a keyword extraction model

Table 2: Comparison of the extracted keywords with the labeled test set

	Precision	Recall	F1
DistilBERT	0.97	0.99	0.98
SciBERT	0.97	0.97	0.97
FinBERT	0.94	0.94	0.94
KeyBERT	0.28	0.22	0.25
TextRank	0.43	0.33	0.38
TF-IDF	0.44	0.41	0.42

on a domain-related dataset, as it is familiar with terms that it has seen during pre-training.

5.2 Manual Evaluation

To suggest new terms for the STW, the extracted keywords and key phrases have been run through the filtering process, filtering out terms that are already part of the STW and then applying either a frequency (FF) or trend filter (TF). The threshold of the frequency filter is set to 300. Then for each keyword extraction method and filter type, 100 terms have been randomly selected from the pool of keywords. Each of these sampled keywords is then presented to an expert from the STW team for evaluation. Based on the performance of the three proposed BERT models in the prior experiments, DistilBERT is selected to be further evaluated manually together with the baseline methods. All of the terms, in total 800, are then combined into one randomly sorted list and are presented to the STW team member along with the frequency of the suggested term. For each term, the STW member then labels the keyword with "1", if he/she thinks that the term has the potential to be added to the STW as either a descriptor or a non-descriptor, and label "0" if it is not a fitting word for the STW.

Table 3 shows the precision results of the manual evaluation for each filtering type. For the keywords that have been selected based on their frequency, the baseline methods TextRank and TF-IDF did not perform well. TF-IDF actually performed the worst on both filter types, having only 17 frequency-based keywords selected as potential keywords and 31 terms for the time filter (out of 100). TextRank performed slightly better than TF-IDF but worse than the other methods. While for KeyBERT 44 out of 100 terms have been marked as potential keyword candidates, DistilBERT found even more, resulting in 51% of the suggested terms being potential keywords

Table 3: Precision after frequency filtering (FF) and trend filtering (TF)

	FF	TF	Overall
DistilBERT	51%	59%	55%
TextRank	22%	42%	32%
TF-IDF	17%	31%	24%
KeyBERT	44%	36%	40%

for the STW. The DistilBERT model performs even better for trend-filtered keywords. 59 of the 100 selected terms qualify as potential keywords for the STW. The model outperforms the baseline methods by a large margin of 17%. The second-best performance shows TextRank, which still only suggested 42 potential keywords. The table also shows the overall percentage of terms that can be considered as potential candidates for the STW. The results show that the DistilBERT model suggests the best keyword candidates for the STW. More than 55% of the suggested terms qualify as potential candidates for addition to the thesaurus. Compared to the baseline methods, our model showed an increased performance of 15%.

These results also show that for 3 out of 4 applied keyword extraction methods, the trend filtering resulted in more potential keywords than the frequency filter.

5.3 Document Indexing

Next up, we evaluate whether the extracted keywords from the different methods can be used to index documents. Based on the performance of the proposed models on their ability to extract a significant portion of the STW terms, they might be able to produce indexing terms for documents directly. Thus, we analysed how many of the extracted keywords correspond to indexing terms from any of the three label sets: STW labels, author labels, and specialist labels, as described in Section 3. While only for a small portion of the dataset these indexing terms are provided, it can at least be evaluated whether the models extract these existing terms. Hence it would be even more useful if this model predicts the labels well enough to be used for automating the labeling of documents. Unfortunately, only around 126K of the 575K entries of the entire dataset are indexed with any of the terms from the three index labeling sets, resulting in only around 22%. For the test set, only 1.3% of the documents

Table 4: Available indexing labels in the test set

Indexing Set	Available Labels
STW	3345
Author	435
Specialist	36

Table 5: Percentage of extracted keywords corresponding to document labels

	STW	Author	Specialist
DistilBERT	91.3%	34.9%	27.8%
SciBERT	85.0%	33.1%	0.0%
FinBERT	75.0%	32.0%	19.4%
KeyBERT	48.5%	21.4%	25.0%
TextRank	25.2%	11.3%	11.1%
TF-IDF	29.6%	12.6%	8.3%

contain any indexing terms in the metadata. Table 4 lists the labels in the test set for the different indexing sets.

Table 5 shows the number of labels that have been correctly predicted by the keyword extraction methods. Overall, in each of the label categories, our DistilBERT model performed the best by finding the largest number of labels each. For the STW Labels, the DistilBERT model correctly predicted approximately 91% of the given labels. For the baseline methods, KeyBERT performed the best, but only extracted around 48% of the labels. The results are similar for the author labels: While the DistilBERT model only predicts around 35% of the labels this time, it still performed better than the baseline methods, from which KeyBERT performs the best again with 21% of found labels. For the specialist labels, only 36 labels were available in the test set. While DistilBERT performs the best again by predicting 28% of the labels, it did not perform better by a large margin compared to the other methods this time, as the performance of KeyBERT comes close with 25%. Following these results, our DistilBERT model performs the best in finding labels for documents. Especially in the case of the STW labels our model may be useful, as these results suggest that it finds the correct words in documents. Considering the fact that only a small amount of texts have any labels available, it might be worth using this model to suggest indexing terms for documents.

6 Discussion

Analyzing the keywords extracted by either of the methods together with comments from the domain expert, some common errors from the methods can be identified. One of the occurring problems relates to the part-of-speech of the extracted keywords. The STW only accepts entries of nouns, not verbs or adjectives, which have been commonly extracted by all of the methods. This can be improved by implementing an additional part-of-speech filter in the filtering process to only consider nouns as candidates for the STW. A similar problem occurs with the extraction of proper names and corporation names. These are terms that are not considered for the STW, but at this point, the proposed model does not recognize them and thus also not remove these terms from the candidate pool. The results in the previous section suggest that the fine-tuned DistilBERT model can be used to label documents with indexing terms from the STW. Given the fact that all three of the proposed models are fine-tuned the same way, it can be presumed that the increased performance of BERT relates to the pre-trained model itself, thus the corpus of the DistilBERT model appears to create the best-fitting model for this use case. This is supported by the fact that SciBERT as well as FinBERT in multiple cases did not know a token, thus labeling them with the as [UNK]. However, since only a small part of the test set had been labeled at all, the experiment should also be carried out on a larger set of indexed documents, e.g., the complete dataset. Furthermore, the methods predict more keywords for a document than the number of indexing terms available for each document. Therefore it would be beneficial to rank candidates from a document and only suggest the most important ones. For future work, a way of building an actual term hierarchy could be considered, making use of hierarchical connections among thesaurus terms. While first experiments on clustering terms did not show promising results, finding a way to not only grouping terms but also determining the descriptor terms would be helpful.

7 Conclusion

In this work, the three pre-trained BERT models DistilBERT, SciBERT, and FinBERT were fine-tuned for the task of token classification with the goal of domain-specific keyword extraction. Their performance has been compared to three baseline methods used for keyword extraction, namely TF-

IDF, TextRank and KeyBERT. The results showed that DistilBERT performed the best overall, as it was able to extract domain-specific keywords reliably, but also to suggest more potential new terms for the Thesaurus for Economics (STW) compared to the other methods. This suggests that fine-tuning a model on domain-related documents does indeed help in retrieving domain-specific terms compared to not fine-tuned methods. In future research, the filtering process could be further optimized to achieve higher precision by limiting the number of suggested terms.

References

- Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. 2019. [Asking clarifying questions in open-domain information-seeking conversations](#). In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.
- Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. [Scibert: A pretrained language model for scientific text](#). In *EMNLP*. Association for Computational Linguistics.
- Oleg Borisov, Mohammad Aliannejadi, and Fabio Crestani. 2021. [Keyword extraction for improved document retrieval in conversational search](#).
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristin Toutanova. 2018. [Bert: Pre-training of deep bidirectional transformers for language understanding](#).
- Sujatha Das Gollapalli and Cornelia Caragea. 2014. [Extracting keyphrases from research papers using citation networks](#). In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, AAAI'14*, page 1629–1635. AAAI Press.
- Maarten Grootendorst. 2020. [Keybert: Minimal keyword extraction with bert](#).
- Allen H. Huang, Hui Wang, and Yi Yang. 2022. [Finbert: A large language model for extracting information from financial text](#). *Contemporary Accounting Research*.
- Anette Hulth. 2003. [Improved automatic keyword extraction given more linguistic knowledge](#). In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, EMNLP '03*, page 216–223, USA. Association for Computational Linguistics.
- Andreas Oskar Kempf and Joachim Neubert. 2016. [The role of thesauri in an open web: A case study of the stw thesaurus for economics](#). *Knowledge Organization*, 43:160–173.
- Yeonsoo Lim, Deokjin Seo, and Yuchul Jung. 2020. [Fine-tuning bert models for keyphrase extraction in scientific articles](#). *Journal of Advanced Information Technology and Convergence*, 10(1):45–56.
- H. P Luhn. 1957. [A statistical approach to mechanized encoding and searching of literary information](#). *IBM Journal of Research and Development*, 1:309–317.
- Rada Mihalcea and Paul Tarau. 2004. [TextRank: Bringing order into text](#). In *EMNLP*, pages 404–411. Association for Computational Linguistics.
- Lorenzo Pezzo. 2022. [Keyed alike: Towards versatile domain-specific keyword extraction with bert](#). Master thesis. Utrecht University.
- Yili Qian, Chaochao Jia, and Yimei Liu. 2021. [Bert-based text keyword extraction](#). *Journal of Physics: Conference Series*, 1992(4):042077.
- Thomas Roelleke. 2013. [Information retrieval models: Foundations and relationships](#). *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 5.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. [Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter](#).

Extracting the Agent-Patient Relation From Corpus With Word Sketches

Antonio San Martín and Catherine Trekker Juan Carlos Díaz-Bautista
 University of Quebec in Trois-Rivières Autonomous Mexico State University
 Trois-Rivières, Canada Toluca, Mexico
 {sanmarti,trekkers}@uqtr.ca jdiazb002@alumno.uaemex.mx

Abstract

Word sketches are a powerful function of Sketch Engine that automatically summarizes the most common usage patterns of a search word in a corpus. While they have proven to be a valuable tool for collocational analysis in both general and specialized language, their potential for the extraction of terminological knowledge is yet to be fully realized. To address this, we introduce a novel semantic sketch grammar designed to extract the agent-patient relation, an important yet understudied relation. This paper presents the various stages of developing the rules that compose this sketch grammar as well as the evaluation of their precision. The errors identified during the evaluation process are also analyzed to guide future improvements. The sketch grammar is available online so that any user can apply it to their own corpora in Sketch Engine.

1 Introduction

Word sketches (WSs) are a powerful function of corpus analysis tool Sketch Engine (<https://www.sketchengine.eu/>) (Kilgarriff et al., 2014) that automatically summarizes the most common usage patterns of a search word in a corpus. A WS is composed of columns listing the words that are related (most often syntactically) to the search word in the corpus. This includes, for instance, the verbs having the search word as subject or object, or the words modified by the search word (Figure 1). WSs have proven valuable for collocational analysis in both general and specialized language, as they enable the easy identification of a word’s combinatorial behavior.

verbs with "research" as subject	verbs with "research" as object	nouns modified by "research"
be 757,580 ...	do 592,104 ...	project 456,192 ...
have 343,294 ...	conduct 374,631 ...	interest 207,959 ...
show 208,067 ...	be 159,713 ...	paper 190,150 ...
focus 119,616 ...	support 101,061 ...	team 187,967 ...

Figure 1: Three WS columns of the search word *research* in the enTenTen21 corpus

However, the default WS in Sketch Engine is not adapted to the extraction of terminological knowledge. For this reason, the EcoLexicon Semantic Sketch Grammar (ESSG) (León-Araúz et al., 2016; León-Araúz and San Martín, 2018; San Martín et al., 2022) expanded WS functionality to enable the identification of some of the most common relations used in Terminology and Ontology Engineering with new WS columns (generic-specific, part-whole, cause, function, and location) in English and French (Figure 2).

"soybean" is a type of...	"oxygen" is a part of...	"hurricane" is the cause of...
crop 19 ...	atmosphere 23 ...	damage 24 ...
vegetable 6 ...	molecule 21 ...	pressure 23 ...
plant 5 ...	compound 18 ...	erosion 23 ...
oil 4 ...	water 16 ...	storm 20 ...

Figure 2: Semantic WS columns generated with the ESSG in the EcoLexicon English Corpus (León-Araúz and San Martín, 2018)

This paper presents the first version of a novel semantic sketch grammar designed to extract the agent-patient relation in the form of WSs. An example of this relation is the one between *mechanic* and *tire* in “...the mechanic inflated the tires...”, “...mechanics mount tires...” and “...the tires were balanced by a mechanic...”. In all three examples, *mechanic* is the agent of the action that affects *tire*, which is the patient (*mechanic* affects *tire*)¹.

The agent-patient is a valuable relation for the extraction and representation of terminological knowledge because the organization of specialized domains is shaped by the interaction between different agents and patients (Faber, 2015). Despite its importance, it is an understudied relation, and terminologists and ontologists currently lack a straightforward way of extracting it from corpora. Our proposal seeks to bridge this gap by providing

¹Inspired by the “affects” relation in EcoLexicon (León-Araúz and Faber, 2010), a terminological knowledge base on the environment, we will use the verb *affect* to represent the agent-patient relation in a triplet.

a solution for extracting this semantic relation in the form of WSs. By facilitating the analysis of the interplay of agents and patients within specialized domains, this tool can contribute to both practical terminological and ontological work and academic research.

The remaining sections of this paper are structured as follows. Section 2 describes the process of WS generation. In Section 3, we present our definition of agent, patient, and the agent-patient relation. Section 4 introduces the methods and materials employed in developing the new agent-patient sketch grammar. Sections 5 and 6 outline the two main development phases. The evaluation results are discussed in Section 7. Finally, Section 8 gives the conclusions derived from this research and outlines future work.

2 Word Sketch Generation

WS generation in Sketch Engine is based on the matching of patterns encoded as rules expressed in CQL language (Jakubíček et al., 2010). A CQL rule is composed of tokens in the form of attributes (part-of-speech tag, lemma, word form, etc.) and values combined with regular expressions. For example, the rule `[tag="J.*"] [tag="N.*"] [lemma="management"]` matches concordances containing the lemma *management* preceded by a noun and an adjective (e.g., “natural resource management”, “effective risk management”, and “cold chain management”).

Within a CQL rule intended for WSs, the position of the words to be extracted as the WS results are identified. For instance, the rule `1: [tag="J.*"] [tag="J.*"]? 2: [tag="N.*"]` enables the extraction of an adjective (1:) that is followed by another optional adjective and a noun (2:). It also allows the inverse: the extraction of a noun (2:) preceded by an optional adjective, which itself is preceded by another adjective (1:). In this case, Sketch Engine identifies matches of the rule (a noun preceded by one or two adjectives) in the corpus, and subsequently extracts the left-most adjective and the noun from each matched concordance. However, a significant limitation of WSs is that results are restricted to single words.

For WS generation, the CQL rules designed to identify the same relation are grouped into a gramrel (for “grammatical relation”). Each gramrel can produce one or more WS columns (normally one

relation and its reverse). The collection of gramrels that generate a WS is referred to as a sketch grammar. For instance, the gramrel included in Sketch Engine’s default sketch grammar that identifies the relation between the object of a sentence and its verb generates two WS columns (“objects of “X”” and its reverse “verbs with “X” as object”) by means of three rules (Figure 3). The first rule identifies the object-verb relation in the active voice and the other two in the passive voice (one without the verb *to be* and the other with it).

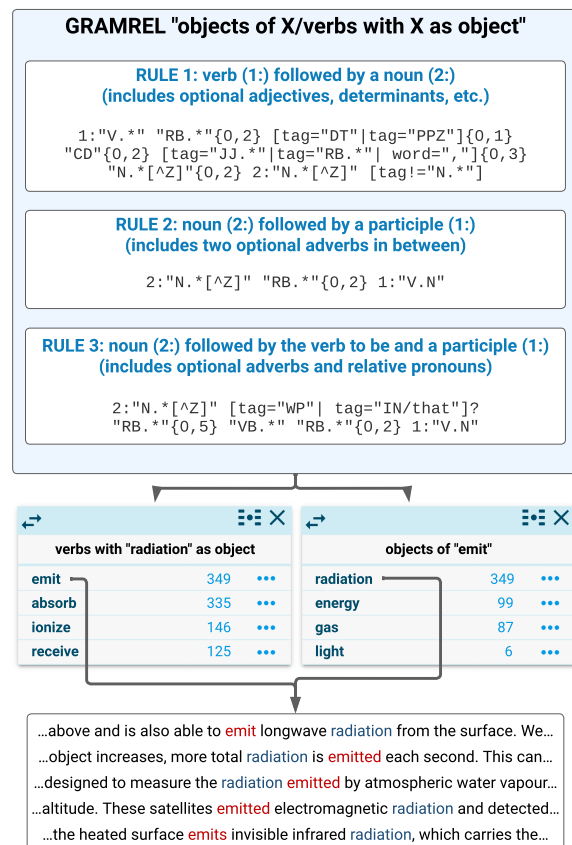


Figure 3: The "objects of “X”/verbs with “X” as object” gramrel in the default English sketch grammar with an example from the EcoLexicon English Corpus

While the default sketch grammar is mainly based on syntactic relations, the ESSG extracts semantic relations by means of knowledge patterns, i.e., lexico-syntactic patterns that match contexts where a specific semantic relation is conveyed (Meyer, 2001). For instance, the knowledge pattern “X and other Y” (e.g., “...theophylline and other bronchodilators...”) conveys a generic-specific relation (*theophylline* is-a *bronchodilator*).

While our new agent-patient sketch grammar extracts a semantic relation, our methodology does

not rely on knowledge patterns². Instead, our starting point is the syntactic relation between the nouns functioning as subject and object in the same sentence. This is based on the premise that the subject typically functions as the agent and the object as a patient. Even though the subject-object relation does not always correspond to an agent-patient semantic relation (and vice versa), the results of a pilot study confirmed the feasibility of this approach (San Martín and Trekker, 2021).

3 Defining the Agent-Patient Relation

We define the agent-patient relation as one in which one participant in the action (the agent) affects another participant (the patient) in some way. In this sense, we adopt the notions of agent and patient in a broad sense, aligning with Dowty's (1991) macroroles of proto-agent and proto-patient, or Van Valin's (2004) actor and undergoer. This implies that our definition of agent also encompasses other semantic roles that affect another participant in the action such as effector, actor, instrument, and others. Similarly, our interpretation of patient is inclusive of roles that other authors might label not only as patient but also as theme, referent, goal, beneficiary, result, etc. As a result, according to our definition, agents and patients can be nouns that refer to any type of concept including concrete and abstract entities, processes, states, and attributes.

The extent to which an agent's action must impact a patient in order to establish the existence of an agent-patient relation is not clear-cut. Whereas "...the researcher vaccinated the rats..." is indisputably agentive and "...the researcher imagined colorful rats...", non-agentive, there are many borderline cases, such as "...the researcher possesses rats..." or "...the researcher exhibits the rats..."

To better delimit the agent-patient relation for the creation and subsequent evaluation of CQL rules, we used a pre-existing list of verb senses to determine which ones are to be considered agentive and which are not. We chose that of Faber and Mairal Usón (1999), which classifies the English verb lexicon into 13 verb sense categories (such as existence, movement, and position), which are further subdivided into 389 subcategories.

We labeled each verb sense in the list as agentive, non-agentive, or intransitive, based on their nature. Given the fuzziness of the agent-patient

relation, there were unavoidably subjective choices. Most verb senses were deemed either agentive or intransitive. Agentive subcategories include, among others, all causative senses, which means that our definition of the agent-patient relation subsumes the causal relation. Intransitive subcategories are those involving a single argument.

The non-agentive subcategories included those verb senses overlapping with the part-whole and location relations. Additionally, other subcategories that were considered non-agentive include, among others, those expressing perception, cognition, feeling, and speech. Some possession verb senses were also considered non-agentive, such as those expressing basic possession (*have*, *possess*, *own*). However, when the agent carries out an action to possess something (*take*, *get*, *obtain*) or there is a transfer of possession (*give*, *provide*, *exchange*), the verb senses are considered agentive. The final classification of verb senses is available at <http://doi.org/10.5281/zenodo.8121939>³.

As will be seen later, verbs that most frequently activate intransitive or non-agentive senses were filtered out in the CQL rules.

4 Materials and Methods

The development of a new sketch grammar is based on the encoding of CQL rules and their subsequent enhancement based on the evaluation of the matching concordances in a given corpus (León-Araúz et al., 2016). For this agent-patient sketch grammar (consisting of a single gramrel)⁴, we used the Elsevier OA CC-BY Corpus (Kershaw and Koeling, 2020), which is composed of 40,000 open-access articles in English published between 2014 and 2020 in Elsevier journals. The corpus in its version available in Sketch Engine contains 187,615,459 words and 232,511,611 tokens. It covers a wide variety of domains (e.g., Medicine, Computer Science, Social Sciences, Economics, Arts, etc.). This ensures that the sketch grammar is domain-independent.

³In this URL, the final sketch grammar can also be found, as well as all the lists of verbs and phrases used to build the CQL rules that are mentioned later in the paper.

⁴In San Martín and Trekker (2021), we created a preliminary version of this gramrel. The one presented in this study partly follows the same methodology, but with numerous improvements and modifications. These differences cannot be discussed because of space restrictions.

²However, some of the CQL rules, as will be seen below, could be considered knowledge patterns.

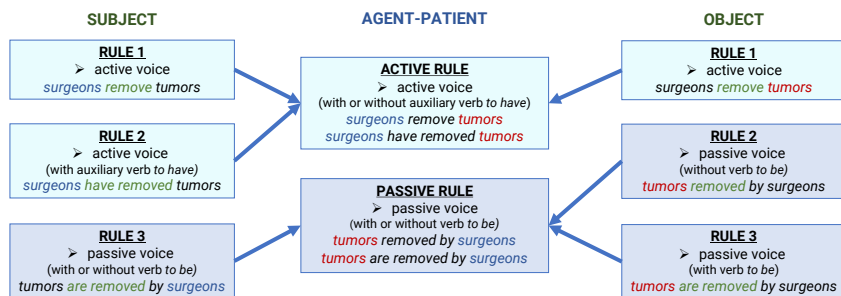


Figure 4: Generation of the simple version of the agent-patient rules

Our initial step was to generate a simple version of the gramrel by integrating the two default gramrels "objects of "X"/verbs with "X" as object" (object gramrel) and "subjects of "X"/verbs with "X" as subject" (subject gramrel) (Figure 4). The active-voice rules were combined into a new rule ('active-simple'), while the passive ones were also consolidated into another one ('passive-simple').

We then proceeded to the subject-object enhancement, which consisted of enriching and refining the simple version to improve its precision and recall with respect to the extraction of the subject-object relation. This was followed by the agent-patient enhancement, aimed at improving its capacity to extract the agent-patient relation.

Throughout both enhancement phases, minor and major evaluations were carried out, with the authors of the paper acting as evaluators. All evaluations were collaboratively reviewed and agreed upon, aimed at iteratively refining the rules, determining whether 20 random concordances extracted with the evaluated rule conveyed the subject-object relation or the agent-patient relation (depending on the enhancement phase). For a concordance to be considered valid, the rule also had to correctly identify the nouns functioning as subject and object (or agent and patient) within the concordance.

The count of valid concordances was used to estimate precision and determine whether the evaluated modifications should be retained. When the results were inconclusive, additional sets of 20 concordances were evaluated. Recall was prioritized over precision since users ultimately access the results of the gramrel through WSs, where the potentially most relevant results (with higher frequency) are at the top of the WS column.

In this paper, we only present the results of the major evaluations which involved the assessment of 250 random concordances and were reserved for definitive versions of the rules.

5 Subject-Object Enhancement

For the subject-object enhancement phase, the rules resulting from combining the subject and object gramrels ('active-simple' and 'passive-simple') were enriched and refined to increase recall without compromising precision. Each enrichment was subject to a minor evaluation. These enhancements included, among others, the addition of optional modal and auxiliary verbs, the possibility of more than one main verb, optional gerunds and participles where adjectives were already possible, an optional comma before the optional relative pronoun as well as some minor adjustments to avoid noise (for instance, excluding the presence of *so* before the optional relative pronoun to avoid noise created by the occurrence of *so that*).

Both versions of the rules were subject to a major evaluation. For a concordance to be considered valid, there needs to be a subject-object relation between the identified nouns, and both of them need to be the head of their noun phrase.

The evaluation results (Figure 5) indicate that the simple and enhanced versions yield comparable subject-object precision. However, the enhanced active rule extracts 53.74% more concordances, and the enhanced passive rule extracts 31.86% more concordances than their simple counterparts.

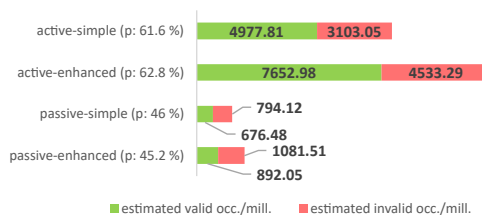


Figure 5: Precision and occurrences per million tokens of the simple and enhanced rules

6 Agent-Patient Enhancement

Since the two enhanced rules provided a precision comparable to the simple ones but with higher recall, the agent-patient enhancement was performed on these two rules. However, before proceeding, an evaluation of the agent-patient precision of the same concordances was performed to establish a reliable baseline.

Evaluators answered the following question for each concordance: “Does the identified agent have an effect on the identified patient?”. When the concordance was not considered valid, the error or errors at cause were noted. Although an agent-patient relationship is established in the concordance, if the correct agent and patient are not identified, the concordance is considered invalid. The list of errors and their distribution in this evaluation and the subsequent ones are reproduced and explained in section 7.2.

According to the results of the evaluation (Figure 6), ‘active-enhanced’ has an agent-patient precision of 31.2% and ‘passive-enhanced’, 38.4%. Both values are significantly lower than their subject-object precision. This indicates that solely focusing on improving subject-object precision is insufficient for effectively capturing the agent-patient relation. Consequently, we proceeded to the agent-patient enhancement, which was divided into three stages described in the remainder of this section.

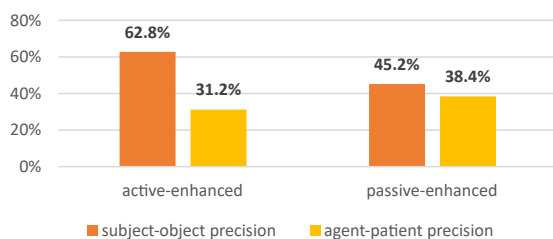


Figure 6: Evaluation results of ‘active-enhanced’ and ‘passive-enhanced’

6.1 First Stage

This first stage, aimed at improving precision⁵, consisted of creating a version of the rules where verbs that do not convey the agent-patient relation are excluded. To compile a list of non-agentive verbs, we first extracted the 1000 most frequent verbs in the Elsevier corpus as well as the 1000 most frequent verbs in the same corpus occurring within our ac-

⁵Henceforth, precision is understood specifically as agent-patient precision.

tive and passive enhanced rules. The elimination of duplicates produced a list of 1083 verbs, which was reduced to 1054 verbs after the consolidation of spelling variants and lemmatization errors.

Each verb was subjected to a minor evaluation in which its presence was forced in the active and passive rules. The purpose of the evaluation was to determine whether the verb more frequently activates agentive or non-agentive verb senses, based on our classification of verb senses.

Since verbs can have both agentive and non-agentive senses because of polysemy, verbs with non-agentive senses in 75% or more of the concordances were classified as non-agentive. As a result, a total of 275 non-agentive verbs (e.g., *say*, *define*, *display*...) were identified, as well as 693 agentive verbs (e.g., *convert*, *target*, *structure*...).

We also identified intransitive and inverting verbs. Intransitive verbs produce noise because they cannot instantiate an agent-patient relation. An intransitive verb is one that in 75% or more of the concordances was found to be intransitive. A total of 76 intransitive verbs were thus identified (e.g., *exist*, *go*, *live*...).

As for inverting verbs, they are verbs in which the subject functions as the patient and the object as the agent. For instance, *undergo* in “...women undergo an outpatient hysteroscopy...” (hysteroscopy affects woman). We identified 10 inverting verbs (e.g., *experience*, *resist*, *tolerate*...).

With the final list of verbs, we created four variants of the rules. The first two rules (‘active-exc’ and ‘passive-exc’) exclude non-agentive, intransitive, and inverting verbs⁶. Conversely, the other two rules (‘active-inv’ and ‘passive-inv’) only permit inverting verbs and reverse the order in which the agent and the patient appear.

6.2 Second Stage

The second stage, aimed at improving recall, consisted of the creation of a version of the active rule that allows certain prepositional verbs⁷ that convey an agent-patient relation (e.g., *lead to*, *contribute to*, *aim at*, *help in*). A version of the passive rule that permits certain verbs followed by prepositions other than *by* was also created (e.g., *attribute to*, *expose to*, *filter through*).

⁶The gerund verb forms *using* in ‘active-exc’ and *facing* in ‘active-inv’ were excluded too because they generated excessive noise.

⁷By prepositional verbs, we also mean particle verbs.

For the active rule ('active-prep'), we initially allowed the optional presence of a preposition or a particle after the main verb. However, the evaluation of the concordances of 26 prepositions and particles in that position showed that this approach created a significant amount of noise. Nonetheless, this evaluation allowed us to identify 148 prepositional verbs that could potentially be agentive.

After an individual evaluation of each one, the list was reduced to 107 agentive prepositional verbs (e.g., *act on*, *contribute to* or *deal with*⁸). This permitted the creation of the rule 'active-prep'. Also identified were 16 inverting prepositional verbs (e.g., *suffer from*, *depend on* or *result from*), resulting in the rule 'active-prep-inv'.

Some examples of valid concordances from these two rules include "...Government can contribute to realising a circular economy..." (*government* affects *economy*) and "...mice reacted to fear conditioning stimuli..." (*stimulus* affects *mouse*).

Using this method and by means of minor iterative evaluations, we identified three verbs that can appear in passive voice without a by-phrase but which are followed by a prepositional phrase with agentive meaning: *attributed to*, *exposed to* and *filtered through*. The rule 'passive-prep' forces their presence.

Some examples of valid concordances retrieved with this rule include "...Supernatants were filtered through a 0.45 μ m membrane..." (*membrane* affects *supernatant*) and "...sorption could therefore be attributed to the sludge..." (*sludge* affects *sorption*).

6.3 Third Stage

Finally, the third stage, also aimed at improving recall, consisted of developing a version of the active rule that allows verb phrases expressing an agent-patient relation (e.g., *to have impact/effect/influence on*, *to play a role in*, *to make a contribution to*...). Additionally, we created a version of the passive rule where *by* is replaced by expressions such as *using*, *by means of*, *with the help of*, etc. (e.g., "...rules are instituted with the help of a dietician...").

In the case of verb phrases, the patient is not the object of the sentence but rather the head of the prepositional phrase that follows. For instance, in "competition has a sizeable negative impact on pupil wellbeing", *wellbeing* serves as the patient

⁸The gerund of *deal* (i.e., *dealing*) was excluded from the rule because, unlike other tenses, it mostly had a non-agentive sense.

despite not being the object. Considering this, we developed a version of the active rule ('active-phrases') that forces the presence of agentive verb phrases such as *play a role in*, *have effect on* or *make use of* and retrieves as patient the head of the prepositional phrase that follows.

Each verb phrase was individually evaluated to ensure a minimum precision level of 50%. An example of valid concordances extracted with this rule are "...Mitochondria play key roles in mammalian apoptosis..." (*mitochondrion* affects *apoptosis*) and "...Imports have large positive effects on firm productivity..." (*import* affects *productivity*).

Additionally, we created a passive rule ('passive-not-by') where the by-phrase is replaced by expressions referring to an instrument or a means such as *using*, *by means of*, and other variants. Each of the expressions in the rule was evaluated to determine whether they provided at least 50% precision. An example of some valid concordances extracted with this rule are "...The pycnometer was calibrated using a standard calibration ball..." (*ball* affects *pycnometer*) and "...sequences can be folded by addition of metal ions..." (*ion* affects *sequence*).

7 Evaluation Results

7.1 Precision

Figure 7 presents the results of the evaluation of each of the rules that make up the new agent-patient gramrel. The figure also includes the number of valid matches that each rule is estimated to retrieve from the Elsevier corpus (expressed as occurrences per million tokens). This estimate was calculated by applying the precision percentage to the total number of matches retrieved by each rule.

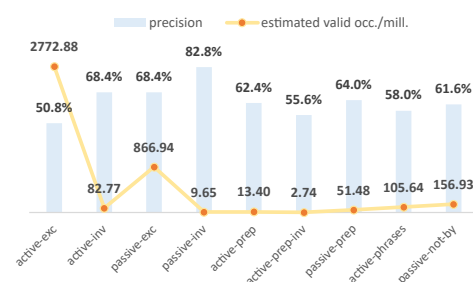


Figure 7: Evaluation results per rule

With an overall precision of 54.9%, the new gramrel significantly outperforms the baseline (32.2%) (Figure 8). Each individual rule also surpasses the baseline in precision. However, the total count of valid occurrences per million tokens retrieved by the gramrel is slightly lower than the baseline, although the number of invalid matches (i.e., noise) is nearly three times lower.

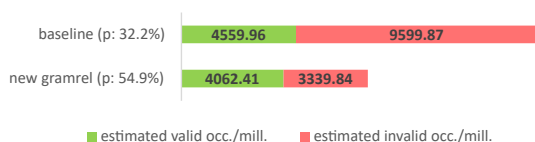


Figure 8: Precision and occurrences per million tokens of the baseline and the new gramrel

Nearly 90% of the valid occurrences recovered by the new gramrel are attributed to two rules: ‘active-exc’ and ‘passive-exc’, which capture the subject-object relation but block selected verbs. Passive rules also exhibit more precision than active rules because of their inherent restrictiveness. Unlike the flexibility in verb tense allowed by active rules, passive rules need the presence of a past participle, which mitigates potential noise.

It is worth noting that whereas assessing rule precision through random concordances is useful during the development process, only the analysis of the resulting WS can validate the usefulness of the sketch grammar. Terms unlikely to be queried by a user through the WS function (due to their irrelevance in terminological analysis or because they do not engage in agent-patient relations) are identified as potential agents or patients in these random concordances. Consequently, random concordances tend to be noisier than those associated with genuine WS queries made by terminologists or ontologists. Moreover, WSs show the most frequent results at the top, which tend to be linked to a higher number of valid concordances.

Since this agent-patient sketch grammar is still in development and WS evaluation is a labor-intensive task, the resulting WSs will only be evaluated when the final version is completed.

7.2 Types of Errors

The following six types of errors were identified during the evaluation:

1. *Non-agentive*: The relation between the two nouns is not agent-patient because the verb

sense is non-agentive (e.g., “...results indicate a temperature increase...”). Evaluators referred to the verb sense classification to determine the agentivity of the verb sense within each concordance. The *non-agentive* error also includes the cases in which the agent was erroneously retrieved as a patient and vice versa. For example, in “...drivers experiencing more fatigue...”, the correct relation is “*fatigue* affects *driver*” and the inverse would be considered an error under this category.

2. *Not head*: The retrieved noun is not the head of the grammatical subject or object. This can be caused by multiword terms, prepositional phrases, relative clauses, etc. For instance, in “...The discharge of untreated or partially treated domestic wastewater to the aquatic environment severely threatens public health...”, *environment* was mistakenly detected as the agent instead of *discharge*.

When the agent or patient is a noun phrase, it may be unclear which is the most semantically significant noun. To ensure objectivity, we followed a strict syntactic criterion with a short list of exceptions such as *group of*, *part of*, etc., where it was determined that the correct noun is not the head. For instance, in “...A number of researchers have used salt...”, although *researchers* is not the head, it was considered a valid concordance.

3. *Not noun*: A noun that is not the subject or object is retrieved because the subject or object is not a noun phrase, but rather a clause or a pronoun (e.g., “...Understanding how meteorology impacts the seasonality of Lyme disease case occurrence can aid in targeting limited prevention resources...”). This type of error also includes cases where an incorrect noun is retrieved as agent because the subject is not explicit in the sentence (e.g., “...Accelerometers are glued to the surface of the plate using hot glue...”).
4. *POS tagging*: Due to a POS tagging error, an incorrect agent-patient relation is retrieved. For instance, the concordance “...the total number generated matches the distribution of the dwelling stock...” was incorrectly retrieved because *matches* was tagged as a noun instead of a verb.

5. *Not by-phrase*: For passive rules, the noun that follows the preposition *by* is not the logical subject. For instance, in “...This enables dry commodities to be marketed by weight...”, weight is not the passive logical subject, but the head of an adverbial. Nonetheless, in those cases in which the adverbial headed by *by* introduces an instrument or a means, they were considered valid. For instance, in “...the tissue had already been stabilised by fixation...”, although fixation is not the logical subject, the concordance was considered valid (*fixation* affects *tissue*).
6. *Segmentation*: An invalid agent-patient relation is retrieved due to a segmentation error (e.g., “...and to extract B. Exponentially growing cells were...”).

Figure 9 illustrates the distribution of error types per rule. Since a single concordance can contain more than one type of error, the count of errors may not match the number of invalid concordances (out of 250 evaluated concordances per rule).

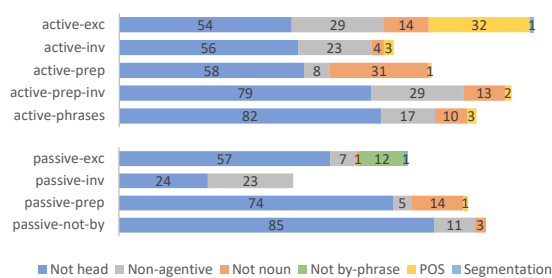


Figure 9: Distribution of error types per rule

The *not head* error accounts for over half of the errors in all rules. This error is a byproduct of the fact that Ws can only extract one-word results.

The way our rules select which noun to identify as agent or patient is inherited from how it is done in Sketch Engine’s default sketch grammar. Before the verb, the rules capture the rightmost noun and, after the verb, the rightmost noun before any non-noun token. This approach yields precise results in the absence of prepositional phrases (e.g., “...energy suppliers use wastewater heat to produce...”).

However, the presence of prepositional phrases before the verb is the cause of a considerable amount of noise (e.g., “...Hydrodynamics in bubble columns strongly influence mass transfer...”). In fact, the difference in the number of *not head* errors between rules can be primarily attributed

to the varying frequency of prepositional phrases occurring before the verb in each rule.

As for the *POS tagging* error, it is significantly more prevalent in the ‘active-exc’ rule because of the POS tagger’s difficulty in distinguishing between past tense verbs and past participles (e.g., “...there is growing evidence that increased production and productivity can lead...”) as well as present participles and nouns (e.g., “...solar absorption cooling system...”).

In ‘active-prep’, we found more *not noun* errors than in other rules because some of the prepositional verbs included in the rule have a greater tendency to have a clause as subject, notably *lead to* and *contribute to* (e.g., “...Increasing the amount of rutile phase compared to that of the anatase phase led to decrease the photodegradation...”).

Finally, the *not by-phrase* error is exclusive to ‘passive-exc’ and ‘passive-inv’ because the other passive rules do not match concordances with by-phrases. However, in ‘passive-inv’, we did not find this error because the inverting verbs allowed by this rule do not normally induce this error.

7.3 Avenues of Improvement

The evaluation of the rules has underscored the priorities to be addressed for the development of the final version of the sketch grammar.

The fact that most concordances retrieved by the gramrel are extracted by the ‘active-exc’ and ‘passive-exc’ rules suggests that future improvement efforts should focus on increasing the precision of these two rules. One way to accomplish this would be to limit the retrieval as a patient of the object of common verb phrases. For instance, the rule ‘active-exc’ currently retrieves non-agentive concordances such as “...30% of cycling takes place in roads...” or “...data may shed light on HBP dysfunction...”. These noisy concordances could be excluded by not allowing *place* and *light* as patient when their respective verbs are *take* and *shed*.

Still another possibility is the expansion of our list of non-agentive, intransitive, and inverting verbs, which are specifically excluded in ‘active-exc’ and ‘passive-exc’.

Finally, considering that the *not head* error accounts for over half of all errors across all rules, it could be productive to examine how different types of multiword terms in the agent or patient position, as well as the presence of prepositional phrases, can be accounted for in the rules.

8 Conclusions and Future Work

In this paper, we have presented the development of an innovative sketch grammar that enables users to extract the agent-patient relation from any English user-owned corpus in Sketch Engine. The current version of the agent-patient sketch grammar can be downloaded at <http://doi.org/10.5281/zenodo.8121939>, where instructions on how to use it with their own corpora in Sketch Engine are also found.

Figure 10 shows a sample of the resulting agent-patient WS columns for the term *farmer* when the sketch grammar is applied to an 8-million-word specialized corpus on agriculture. Some of the concordances that are accessible via the WS are also reproduced.

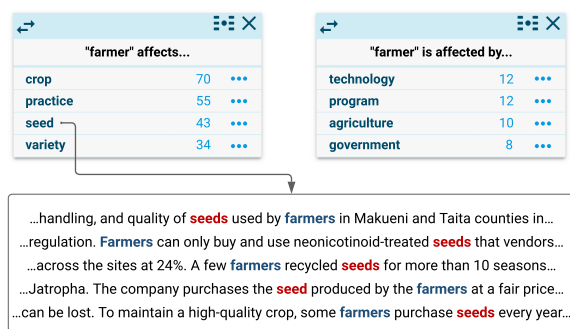


Figure 10: Agent-patient WS columns of *farmer* in an agricultural corpus

The current agent-patient sketch grammar, though currently functional, is still under development and will undergo future enhancements to increase both precision and recall, including those previously mentioned in this paper. As with the current version, subsequent iterations will be made freely accessible online.

The agent-patient sketch grammar can greatly benefit terminologists and ontologists since it facilitates access to one aspect that reflects how specialized domains are structured that was previously very time-consuming to extract. Beyond its practical applications, this sketch grammar is a valuable research tool. We plan to use it in future studies to further explore the agent-patient relation in specialized domains.

Acknowledgements

This research was carried out as part of projects 2020-NP-267503 funded by Quebec's Society and Culture Research Fund, 430-2023-0248 funded

by the Social Sciences and Humanities Research Council of Canada, PID2020-118369GBI00 funded by the Spanish Ministry of Science and Innovation, and A-HUM-600-UGR20 funded by the Regional Government of Andalusia.

References

- David Dowty. 1991. *Thematic Proto-Roles and Argument Selection*. *Language*, 67(3):547–619.
- Pamela Faber. 2015. *Frames as a Framework for Terminology*. In H. J. Kockaert and F. Steurs, editors, *Handbook of Terminology, volume 1*, pages 13–33. John Benjamins, Amsterdam.
- Pamela Faber and Ricardo Mairal Usón. 1999. *Constructing a Lexicon of English Verbs*. Mouton de Gruyter, Berlin.
- Miloš Jakubíček, Adam Kilgarriff, Diana McCarthy, and Pavel Rychlý. 2010. *Fast syntactic searching in very large corpora for many languages*. In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, pages 741–747, Tohoku University, Sendai, Japan. Institute of Digital Enhancement of Cognitive Processing, Waseda University.
- Daniel Kershaw and Rob Koeling. 2020. *Elsevier OA CC-BY Corpus*.
- Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. *The Sketch Engine: ten years on*. *Lexicography*, 1(1):7–36.
- Pilar León-Araúz and Pamela Faber. 2010. *Natural and contextual constraints for domain-specific relations*. In *The Workshop Semantic Relations, Theory and Applications*, pages 12–17, Valletta.
- Pilar León-Araúz and Antonio San Martín. 2018. *The EcoLexicon Semantic Sketch Grammar: from Knowledge Patterns to Word Sketches*. In *Proceedings of the LREC 2018 Workshop "Globalex 2018 – Lexicography & WordNets"*, pages 94–99, Miyazaki. Globalex.
- Pilar León-Araúz, Antonio San Martín, and Pamela Faber. 2016. *Pattern-based word sketches for the extraction of semantic relations*. In *Proceedings of the 5th International Workshop on Computational Terminology (Computerm2016)*, pages 73–82, Osaka, Japan. The COLING 2016 Organizing Committee.
- Ingrid Meyer. 2001. *Extracting knowledge-rich contexts for terminography - A conceptual and methodological framework*. In *Recent Advances in Computational Terminology*, pages 279–302. John Benjamins, Amsterdam.

Antonio San Martín and Catherine Trekker. 2021. Adapting word sketches for specialized knowledge extraction. In *Proceedings of the 14th International Conference of the Asian Association for Lexicography (ASIALEX)*, pages 64–87, Jakarta. ASIALEX.

Antonio San Martín, Catherine Trekker, and Pilar León-Araúz. 2022. Repérage automatisé de l’hyponymie dans des corpus spécialisés en français à l’aide de Sketch Engine. *Terminology*, 28(2):264–298.

Robert D. Van Valin. 2004. Semantic macroroles in Role and Reference Grammar. In R. Kailuweit and M. Hummel, editors, *Semantische Rollen*, pages 62–82. Narr, Tübingen.

Index of Authors

Index of Authors

A

Apostol, Elena-Simona 340, 410
 Araque, Oscar 617
 Arcan, Mihael 134, 374
 Arfon, Elin 306
 Armaselu, Florentina 340, 410
 Arroyo, David 514

B

Baczkowska, Anna 434
 Bajčetić, Lenka 49
 Bal, Bal Krishna 328
 Ballier, Nicolas 281
 Banerjee, Shubhanker 246
 Basile, Pierpaolo 86
 Baumann, Andreas 288
 Bausch, Nicole 288
 Bellandi, Andrea 646
 Benson, Juliane 288
 Bernad, Jorge 147
 Bigeard, Sam 364
 Blaschke, Theresa 274
 Bloos, Sarah 288
 Brasoveanu, Adrian M. P. 294
 Brate, Ryan 97
 Buitelaar, Paul 134, 374

C

Callus, Dorians 364

Camacho, Jose Manuel 514
 Can, Fazli 549
 Carvalho, Sara 30
 Cassotti, Pierluigi 86
 Cecchini, Flavio Massimiliano 74
 Chakravarthi, Bharathi Raja 246
 Chen, Pin-Er 455
 Chiarcos, Christian 166, 180, 340, 434
 Chou, Hsin-Yu 455
 Cimiano, Philipp 207
 Comito, Carmela 559
 Cordeiro, João 470

D

Damova, Mariana 434, 440
 Das, Debopam 449
 Declerck, Thierry 49, 364
 Delort, Clara 504
 Deng, Delin 232
 Dereza, Oksana 109
 Di Buono, Maria Pia 347, 538, 659
 Di Nunzio, Giorgio Maria 646
 Di Pierro, Davide 86
 Díaz-Bautista, Juan Carlos 666
 Dinarelli, Marco 154
 Dontcheva-Navrátilová, Olga 627
 Doyle, Adrian 109
 Drozd, Agata 482
 Dužij, Maxim 392

E

Eckelt, Klaus 534
 Egg, Markus 449
 Elahi, Mohammad Fazleh 207
 El-Haj, Mo 220, 262, 306
 Ell, Basil 193, 207
 Elmer, Christina 520
 Eroglu, Erdem Ege 549

F

Ferilli, Stefano 86
 Florêncio, Ayla Santana 613
 Flossdorf, Jonathan 520
 Fraefel, Andreas 294
 Fransen, Theodorus 109
 Freitag, Raquel 613
 Frincu, Marc 226
 Frincu, Simina 226
 Frontini, Francesca 316

G

Gaillat, Thomas 281
 Galassi, Andrea 559
 Garabík, Radovan 402
 Garcia, Maria González 334
 García-Grao, Guillermo 617
 Geleta, Raisa Romanov 532
 Gifu, Daniela 410
 Gillis-Webber, Frances 37
 Giordano, Luca 538

Góis, Túlio Sousa 613
 Gracia, Jorge 4, 147
 Groth, Paul 256
 Guardiola, Patricia 591
 Gugliotta, Elisa 154, 579

H

Haliloglu, Dilruba Sultan 549
 Hammouda, Tymaa 306
 Holter, Ole Magnus 193
 Hornig, Nico 520
 Hsieh, Shu-Kai 455

I

Iglesias, Carlos Á. 617
 Ionov, Maxim 385
 Ivasiuk, Bogdan 559

J

Jablonzay, Nikoletta 288
 Jarrar, Mustafa 306
 Jentsch, Carsten 520
 Jo, Eunkyong 504

K

Keya, Farhana 607
 Khairova, Nina 559
 Khallaf, Nouran 306
 Khan, Anas Fahad 30, 86, 316, 340, 410
 Kirchmair, Thomas 288
 Kitanović, Olivera 180

Klenner, Manfred 122

Knez, Timotej 322

Knight, Dawn 306

Krestel, Ralf 659

L

Lacy, Anna 591

Lang, Christian 239

Lazarević, Jelena 634

Leal, António 470

Li, Jen-Yu 281

Liebeskind, Chaya 340, 410, 434, 466

Little, Suzanne 134

Lo Scudo, Fabrizio 559

M

Mallart, Cyriel 281

Mallia, Michele 579

Marongiu, Paola 86

Matthews, Benjamin 364

Maurino, Andrea 598

Maynard, Diana 33

McCrae, John P. 4, 109, 246

McGillivray, Barbara 86, 340, 410

Mishev, Kostadin 434

Mititelu, Verginica 347

Monti, Johanna 659

Moreno-Schneider, Julian 334

Morris, Jonathan 306

Müller, Henrik 520

Mulligan, Bret 591

Mündges, Stephan 520

O

Ogrodniczuk, Maciej 482

Oliveira, Hugo Gonçalo 347, 358

Olsen, Sussi 364

Ostroški Anić, Ana 399, 402

Ozcelik, Oguzhan 549

P

Pais, Sebastião 470

Panasci, Livia 579

Pannach, Franziska 274

Parada-Cabaleiro, Emilia 532

Pasricha, Nivranshu 374

Passarotti, Marco Carlo 74

Paterson, Hugh 591

Paudel, Shishir 328

Pedonese, Giulia 74

Penteliuc, Marius E. 226

Perez-Miguel, Luis 514

Piccini, Silvia 646

Pitarch, Lucia 147

Polat, Fina 256

Potter, Andrew 493

Povolná, Renata 627

Principe, Renzo Alva 598

R

Rabby, Gollam 607
 Rackevičienė, Sigita 402
 Rahnenführer, Jörg 520
 Rani, Priya 109
 Ransmayr, Jutta 34
 Rayson, Paul 262, 306
 Rehm, Georg 334
 Reyes, Juan-Francisco 428
 Rieger, Jonas 520
 Rodrigues, Ricardo 358
 Romary, Laurent 316
 Rosner, Mike 385
 Rujević, Biljana 634

S

Sammet, Jill 659
 San Martín, Antonio 666
 Sánchez-Rada, J. Fernando 617
 Santos, Vinícius Moitinho da Silva 613
 Schedl, Marcus 532
 Selmistraitis, Linas 402
 Sérasset, Gilles 49, 417
 Shrestha, Dhiraj 328
 Silvano, Purificação 434, 470
 Simpkin, Andrew 281
 Sitter, Emilie 288
 Škorić, Mihailo 634
 Souza, Pedro Paulo Oliveira Barros 613
 Spahiu, Blerina 399, 347, 598, 607
 Speranza, Giulia 659

Stanković, Ranka 180, 634
 Stearns, Bernardo 109, 281
 Strossa, Petr 392
 Suryawanshi, Shardul 134
 Süsstrunk, Norman 294
 Svátek, Vojtěch 392, 607

T

Takanami, Ryutaro 220
 Takebayashi, Kohei 566
 Tejada, Julian 613
 Tiddi, Ilaria 256
 Tittel, Sabine 61
 Tomaszewska, Aleksandra 482
 Tono, Yukio 566
 Trajanov, Dimitar 434
 Trekker, Catherine 666
 Truica, Ciprian-Octavian 340, 410, 434
 Tseng, Yu-Hsiang 455
 Tu, Ngoc Duyen Tanja 239

U

Utka, Andrius 340, 402, 410
 Utvić, Miloš 180

V

Valūnaitė-Oleškevičienė, Giedrė 340, 347, 402, 410, 434, 466
 Van den Bosch, Antal 97
 Van Erp, Marieke 97
 Venant, Rémi 281
 Verborgh, Ruben 34

Vezzani, Federica 646

Vossen, Piek 256

Vuth, Nakanyseth 417

W

Wang, Po-Ya Angela 455

Weichselbraun, Albert 294

Xuereb, Loran Ripard 364

Y

Yenicesu, Arda Sarp 549

Yildirim, Onur 549

Z

Zeidler, Laura 239

Ziembicki, Daniel 482

Žitnik, Slavko 322

Zmandar, Nadhem 262

Zurowski, Sebastian 482