

LREC-COLING 2024

**The 9th Workshop on Linked Data in Linguistics:
Resources, Applications, Best Practices
(LDL-2024) @LREC-COLING-2024**

Workshop Proceedings

Editors

John P. McCrae, Christian Chiarcos, Katerina Gkirtzou,
Maxim Ionov, Fahad Khan, Patricia Martín-Chozas and
Elena Montiel-Ponsoda

25 May, 2024
Torino, Italia

**The 9th Workshop on Linked Data in Linguistics:
Resources, Applications, Best Practices
(LDL-2024) @LREC-COLING-2024**

Copyright ELRA Language Resources Association (ELRA), 2024
These proceedings are licensed under a Creative Commons
Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

ISBN 978-2-493814-38-8
ISSN 2951-2093 (COLING); 2522-2686 (LREC)

Jointly organized by the ELRA Language Resources Association
and the International Committee on Computational Linguistics

Preface by the Program Chairs

The Linked Data in Linguistics (LDL) workshop series has established itself as the premier venue for discussing the application of Semantic Web technologies to the fields of linguistics, digital lexicography, and digital humanities (DH).

While recent years have witnessed a steady growth in the adoption of the technology in these areas, its uptake in other relevant domains, most notably in the case of natural language processing (NLP), continues to lag behind. This year, aside from embracing the full bandwidth of applications of LLOD technologies and the closely related area of knowledge graphs in linguistics, we welcome contributions addressing the application of LLOD technologies to NLP applications, as well as those dealing with emerging hot topics of future bridges between structured (linguistic) knowledge and neural methods.

In addition, this year's edition of the workshop will be a venue for in-depth discussions on community standards and best practices, and, above all, those related to the work of the W3C community groups OntoLex,¹ LD4LT² and BPMLOD.³ To this end, it will include featured talks on the latest achievements, developments, and perspectives of these W3C Community Groups.

This year, we received a total of 19 submissions and accepted 8 of these papers for oral presentation and a further 7 of these papers for poster presentation. The papers covered a wide range of topics related to the application of linked data to linguistics. Several papers covered issues related to lexicography, especially those using the OntoLex-lemon module and it was applied to many languages including Latin, Bosnian, Croatian, Serbian and Proto-Indo-European. Further, we had papers examining lexicons for Portuguese borrowings in Asian languages and the Babylonian Talmud. In addition, several papers looked at extensions to the OntoLex model and challenges in lexicographic modelling including morphological description and diachronic analysis.

Several papers have also looked into the wider applications of linguistic linked data and have highlighted specific challenges in applications to the digital humanities. This includes works looking at challenges of museum cataloguing, scholarly information extraction and linking challenges in the humanities. Further, papers looked at the challenge of interoperability around lexicons through new platforms and online services. Finally, a new challenge was the use of linked data to connect lexicons with corpora and several papers tackled this challenge, by proposing new formats for the representation and linking of corpora as well as by extensions to the OntoLex-lemon model to enable such linking.

In addition to this rich array of papers, the feature talks for the three W3C community groups will be given by Penny Labrapoulou on the progress of the LD4LT group and Katerina Gkirtzou on the BPMLOD group. The OntoLex group meeting will focus on the developments of the model and will be led by Christian Chiarcos.

Overall, this is an exciting programme, reflecting the diversity of the research area and the many exciting research directions for linked data and its application to linguistic questions.

This workshop is organised in the scope of COST Action CA18209 NexusLinguarum,⁴ supported by COST (European Cooperation in Science and Technology).

LDL 2024 Organisation Committee

¹Ontology-Lexica Community Group, <https://www.w3.org/community/ontolex/>

²Linked Data in Language Technology Community Group, <https://www.w3.org/community/ld4lt/>

³Best Practices in Multilingual Linked Open Data, <https://www.w3.org/community/bpmlod/>

⁴<https://nexuslinguarum.eu/>

Organising Committee

Christian Chiarcos (University of Augsburg, Germany)
Katerina Gkirtzou (Athena Research Center, Greece)
Maxim Ionov (University of Cologne, Germany)
Fahad Khan (Consiglio Nazionale delle Ricerche, Italy)
John P. McCrae (University of Galway, Ireland)
Elena Montiel Ponsoda (Universidad Politécnica de Madrid, Spain)
Patricia Martín Chozas (Universidad Politécnica de Madrid, Spain)

Program Committee

Sina Ahmadi (George Mason University, USA)
Verginica Barbu Mititelu (Research Institute for Artificial Intelligence of the Romanian Academy, Romania)
Paul Buitelaar (Insight, Ireland)
Sara Carvalho (University of Aveiro, Portugal)
Rute Costa (NOVA FCSH/NOVA CLUNL, Portugal)
Milan Dojchinovski (Czech Technical University, Czech Republic)
Agata Filipowska (Uniwersytet Ekonomiczny w Poznaniu, Poland)
Francesca Frontini (CNR-ILC, Italy)
Frances Gillis Webber (University of Cape Town, South Africa)
Voula Giouli (Athena Research Center, Greece)
Dagmar Gromann (University of Vienna, Austria)
Yoshihiko Hayashi (Waseda University, Japan)
Alik Kirillovich (ex. Higher School of Economics, Russia)
Penny Labropoulou (Athena Research Center, Greece)
Chaya Liebeskind (Jerusalem College of Technology, Israel)
David Lindemann (University of the Basque Country, Spain)
Francesco Mambrini (Università Cattolica del Sacro Cuore, Italy)
Monica Monachini (CNR-ILC, Italy)
Steven Moran (University of Neuchâtel, Switzerland)
Diego Moussallem (Paderborn University, Germany)
Roberto Navigli ("La Sapienza" Università di Roma, Italy)
Petya Osenova (IICT-BAS, Bulgaria)
Ana Ostroški Anić (Institute of Croatian Language and Linguistics, Croatia)
Giulia Pedonese (CNR-ILC, Italy)
Sigita Rackevičienė (Mykolas Romeris University, Lithuania)
Felix Sasaki (SAP, Germany)
Andrea Schalley (Karlstad University, Sweden)
Gilles Sérasset (University Grenoble Alpes, France)
Milena Slavcheva (IICT-BAS, Bulgaria)
Blerina Spahiu (Bicocca University, Italy)
Ranka Stanković (University of Belgrade, Serbia)
Armando Stellato (University of Rome, Italy)
Federica Vezzani (University of Padua, Italy)

Table of Contents

<i>LLODIA: A Linguistic Linked Open Data Model for Diachronic Analysis</i> Florentina Armaselu, Chaya Liebeskind, Paola Marongiu, Barbara McGillivray, Giedre Valunaite Oleskeviciene, Elena-Simona Apostol, Ciprian-Octavian Truica and Daniela Gifu	1
<i>Cross-Lingual Ontology Matching using Structural and Semantic Similarity</i> Shubhanker Banerjee, Bharathi Raja Chakravarthi and John Philip McCrae	11
<i>Querying the Lexicon der indogermanischen Verben in the LiLa Knowledge Base: Two Use Cases</i> Valeria Irene Boano, Marco Passarotti and Riccardo Ginevra	22
<i>Defining an Ontology for Museum Critical Cataloguing Terminology Guidelines</i> Erin Canning	32
<i>The MOLOR Lemma Bank: a New LLOD Resource for Old Irish</i> Theodorus Fransen, Cormac Anderson, Sacha Beniamine and Marco Passarotti	37
<i>CHAMUÇA: Towards a Linked Data Language Resource of Portuguese Borrowings in Asian Languages</i> Fahad Khan, Ana Salgado, Isuri Anuradha, Rute Costa, Chamila Liyanage, John P. McCrae, Atul Kumar Ojha, Priya Rani and Francesca Frontini	44
<i>LODinG: Linked Open Data in the Humanities</i> Jacek Kudera, Claudia Bamberg, Thomas Burch, Folke Gernert, Maria Hinzmann, Susanne Kabatnik, Claudine Moulin, Benjamin Raue, Achim Rettinger, Jörg Röpke, Ralf Schenkel, Kristin Shi-Kupfer, Doris Schirra, Christof Schöch and Joëlle Weis	49
<i>DigitAnt: a platform for creating, linking and exploiting LOD lexica with heterogeneous resources</i> Michele Mallia, Michela Bandini, Andrea Bellandi, Francesca Murano, Silvia Piccini, Luca Rigobianco, Alessandro Tommasi, Cesare Zavattari, Mariarosaria Zinzi and Valeria Quochi	55
<i>Teanga Data Model for Linked Corpora</i> John P. McCrae, Priya Rani, Adrian Doyle and bernardo stearns	66
<i>The Services of the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin</i> Marco Passarotti, Francesco Mambrini and Giovanni Moretti	75
<i>An Annotated Dataset for Transformer-based Scholarly Information Extraction and Linguistic Linked Data Generation</i> Vayianos Pertsas, Marialena Kasapaki and Panos Constantopoulos	84
<i>Linguistic LOD for Interoperable Morphological Description</i> Michael Rosner and Maxim Ionov	94
<i>Modeling linking between text and lexicon with OntoLex-Lemon: a case study of computational terminology for the Babylonian Talmud</i> Flavia Sciolette	103

OntoLex Publication Made Easy: A Dataset of Verbal Aspectual Pairs for Bosnian, Croatian and Serbian

Ranka Stanković, Maxim Ionov, Medina Bajtarević and Lorena Ninčević 108

Towards Semantic Interoperability: Parallel Corpora as Linked Data Incorporating Named Entity Linking

Ranka Stanković, Milica Ikonić Nešić, Olja Perisic, Mihailo Škorić and Olivera Kitanović 115

LLODIA: A Linguistic Linked Open Data Model for Diachronic Analysis

**Florentina Armaselu, Chaya Liebeskind, Paola Marongiu,
Barbara McGillivray, Giedre Valunaite Oleskeviciene,
Elena-Simona Apostol, Ciprian-Octavian Truică, Daniela Gifu**

University of Luxembourg, Jerusalem College of Technology, Université de Neuchâtel,
King's College London, Mykolas Romeris University,
National University of Science and Technology Politehnica Bucharest, Romanian Academy - Iasi Branch
florentina.armaselu@uni.lu, liebchaya@gmail.com, paola.marongiu@unine.ch,
barbara.mcgillivray@kcl.ac.uk, gvalunaite@mr.uni.eu,
{elena.apostol, ciprian.truica}@upb.ro, daniela.gifu@uaic.ro

Abstract

This article proposes a linguistic linked open data model for diachronic analysis (LLODIA) that combines data derived from diachronic analysis of multilingual corpora with dictionary-based evidence. A humanities use case was devised as a proof of concept that includes examples in five languages (French, Hebrew, Latin, Lithuanian, and Romanian) related to various meanings of the term *revolution* considered at different time intervals. The examples were compiled through diachronic word embedding and dictionary alignment.

Keywords: linguistic linked open data, diachronic analysis, multilingual word embeddings

1. Introduction

In this article, we propose a model and dataset that bring together two areas of research often considered separately, linguistic linked open data (LLOD) and diachronic word embedding. The goal is to address the question of how to approach semantic change detection and modelling by combining algorithmic processing with the expressive power of the Semantic Web formalism (Khan et al., 2022). While our model and proof of concept were intended to represent the meaning of words based on corpus and dictionary evidence, they also served as a testbed for our ideas and a way of encoding through structured forms not only the analysis results but also our own understanding of how words and concepts evolve across language, time and space. The model called LLODIA (linguistic linked open data for diachronic analysis) elaborates on existing vocabularies and methods, such as OntoLex-Lemon, OntoLex-FrAC and the “perdurantist” approach (McCrae et al., 2017; Chiarcos et al., 2022a,b; Welty et al., 2006), and creates wrappers and bridges between concepts and resources previously not linked within the diachronic analysis context.

We started from the assumption that embedding results from semantic change analysis need to be assessed in a unified view against a reference background. For this purpose, we included in our modelling both information resulting from corpus processing and comparison with dictionary attestations. Our tests mainly consisted of static word embedding, gensim word2vec (Mikolov et al.,

2013; Rehurek and Sojka, 2010) and fastText (Bojanowski et al., 2017), applied to our corpora in five languages (French, Hebrew, Latin, Lithuanian and Romanian). Experiments with contextual word embedding implementations such as AllenNLP (Gardner et al., 2018) and ELMo (Peters et al., 2018) have been applied so far to the Romanian corpus (Truică et al., 2023).

This paper focuses on the design of the LLODIA model and proof of concept. Section 2 presents the methodology devised for the different corpora in our dataset to build the model and the steps in the construction of the model itself. Section 3 explains in more detail the main LLODIA classes and properties and how the word embedding results have been modelled using them. In Sections 4 and 5, we discuss modelling examples and queries to illustrate the usage of the model. Section 6 synthesises our findings and presents some hypotheses for future work.

2. Methodology

Our method consisted of integrating diachronic word embedding results into LLOD modelling and including dictionary- and corpus-based evidence that referred to word meanings observed or attested at certain time points and intervals.

2.1. Diachronic Word Embedding

The French dataset contained a selection of about 6.4 million tokens from the National Library of Lux-

embourg Open Data monograph collection,¹ with a time span from 1690 to 1918. We cut the corpus into 6 time slices that were chosen based on events and periods related to the history of Luxembourg and the rules and policies regarding the use of the three languages (French, German, Luxembourgish) in the Grand Duchy.² These elements were considered to have an impact on the evolution of language and word meanings. The corpus was lemmatised and stopwords were removed. We applied gensim word2vec (100-dimension vectors, 5-word context window) to each time slice and cosine similarity measures to compute lists of neighbours for words belonging to topics such as socio-political, cultural, and historical. The word “révolution” was chosen for LLOD modelling since the different meanings detected and its potential for cross-language analysis were considered relevant to the study. The lexicographic resources used as references were the CNRTL’s lexical portal³ and Wiktionary.⁴ The former offered rich attestation and etymological information about the analysed term. The latter provided multilingual information regarding etymology and translation in the five languages and English that we used as a pivot.

The Hebrew dataset comprised 76,710 articles, approximately 100 million word tokens sourced from the Responsa Project⁵, spanning from the 11th century to the 21st century. The corpus was divided into four time periods, namely the 11th century until the end of the 15th century, the 16th century, the 17th through the 19th centuries, and the 20th century until the present day. These time periods were selected based on the historical development of halakhic (Jewish religious laws) rulings (Liebeskind and Liebeskind, 2020). These advancements were deemed to influence the evolution of language and the meanings of words. The Hebrew Responsa data set underwent minimal pre-processing before being used with gensim word2vec. The word2vec model used 100-dimensional vectors and a context window of 5 words. Due to the underwhelming performance of modern Hebrew POS taggers on the Responsa dataset (Liebeskind et al., 2012), the pre-processing step only involved tokenizing the text based on white spaces. The lexical resources utilized were Wiktionary and Milog⁶. The latter provided an additional meaning of the explored word

¹Bibliothèque nationale du Luxembourg (BnL) Open Data MONOGRAPH TEXT-PACK: <https://data.bn1.lu/data/historical-newspapers/>.

²For instance, the invasion of Napoleonic troops (1795), the Congress of Vienna (1815), the Royal Decree (1834) stating the official languages, etc.

³<https://www.cnrtl.fr/portail/>.

⁴<https://www.wiktionary.org/>.

⁵<https://www.responsa.co.il/>.

⁶<https://milog.co.il/>.

that is present in the dataset but was not included in Wiktionary.

For the experiments on Latin we used LatinISE (McGillivray and Kilgarriff, 2013), a 13-million token corpus of Latin texts spanning from the 4th century BCE to the 21st century CE. We worked on the lemmatised version of the corpus. We trained a fast-Text model (Bojanowski et al., 2017) on LatinISE with 100 dimensions, a context window of 5, and a minimum frequency count of 5. We used the *Dictionary of Medieval Latin from British Sources* DMLBS (Ashdowne, 2016), accessed via the Logeion platform⁷ to build a sense inventory for *revolutio*, and the LatinISE corpus to retrieve the sense attestations.

For the modelling experiments in Lithuanian, we used Sliekkas (Gelumbeckaitė et al., 2012) where the representation of the original spelling is transliterated into modern Lithuanian, followed by linguistic and morphological annotations. The lemmatised text was used for modelling from a freely accessible, annotated corpus (ca. 350,000 words) including 16th century religious literature and works by the Lithuanian national poet Kristijonas Donelaitis (1714–1780). Also for the sense attestations we used Lietuvių kalbos žodynas⁸ and to identify the etymology, we referred to LIETUVIUZODYNAS.lt.⁹

To detect semantic change in Romanian, a low resource language, Truică et al. (2023) used two static word embedding techniques on the RoDICA corpus.¹⁰ The experimental results showed that Word2Vec Skip-Gram with negative sampling and Orthogonal Procrustes (SGNS-OP) and Word2Vec Skip-Gram negative sampling and Word Injection (SGNS-WI) perform well in detecting semantic change on small datasets, while contextual word embeddings such as ELMo work better on larger datasets and are not suited for languages where collecting a large dataset can be a problem. Previously, Gifu (2016a,b) used RoDICA corpus to analyse topics over time and diachronic similarity between cognate languages by statistical analysis of word distribution over epochs. For Romanian, the RoDICA corpus did not contain any relevant occurrence of the showcase word “revoluție” (eng. revolution), thus the modelling using LLODIA only focuses on dictionary data from the online Explanatory Dictionary of the Romanian Language – DEXonline¹¹. DEXonline acts as a lexical resource that offers information regarding the etymology and the different meanings of the target word.

⁷<https://logeion.uchicago.edu/>

⁸<http://www.lkz.lt/>.

⁹<https://www.lietuviuzodynas.lt/terminai>.

¹⁰<http://lsplr.iit.academiaromana-is.ro/resources/detail/7/>

¹¹<https://dexonline.ro/>

2.2. LLOD Modelling

The LLOD modelling included three main phases. Given the potential of generative AI (GenAI) and large language models (LLMs) to produce outputs in various tasks, such as math problem solving, coding and creative writing, based on step by step prompting (Wei et al., 2023; Chen et al., 2023), a series of prompts have been designed in the early stage to model in RDF/XML a set of examples based on the French word embedding results and dictionary consultation. The aim was to assist the team with RDF/XML modelling when expert assistance was not available. Tests with several GenAI agents were performed and after considering preliminary results, ChatGPT (OpenAI, 2023; Bubeck et al., 2023) and Microsoft Copilot (Ortiz, 2023) were selected for this task.

The prompts in the first phase included several categories. For instance, asking the agents general questions about RDF/XML syntax, class and property generation (Copilot), or to extract examples from an OntoLex-FrAC article (Chiarcos et al., 2022a) and express them into RDF/XML (ChatGPT-4). The RDF/XML format was chosen since XML was more familiar to the members of the team from the humanities area and having less experience with the Turtle language. Another category contained instructions for RDF/XML encoding of (1) resources (corpus, dictionaries), citations and related metadata (title, creator, publisher, publication date, time span), (2) embedding results (vectors, frequency counts, neighbour lists), and sense discrimination and dictionary alignments derived from the French use case on the term *révolution*. The goal was to create templates that could be used for the modelling examples in the other languages of the project.

In the second phase, the results of these conversations were analysed and compared with existing LLOD vocabularies, knowledge repositories and models, such as Dublin Core, DBpedia, ontolex, frac, lexicog, lexinfo, vartrans, lemonEty.¹² Then, the observations based on the French examples were generalised taking into account the broader LLOD context to define the classes and properties of the LLODIA model. Oxygen XML Editor¹³ and Protégé¹⁴ were used for creating, editing and validating the classes, properties and instances of the OWL-based implementation

¹²<http://www.w3.org/ns/lemon/ontolex#>,
<http://www.w3.org/ns/lemon/frac#>,
<http://www.w3.org/ns/lemon/lexicog#>, <http://www.lexinfo.net/ontology/2.0/lexinfo#>,
<http://www.w3.org/ns/lemon/vartrans#>,
<http://lari-datasets.ilc.cnr.it/lemonEty#>.

¹³<https://www.oxygenxml.com/>.

¹⁴<https://protege.stanford.edu/>

of the model.

Once the ontology and the first examples for French were created, validated and tested using the two editors, in the third phase, the model was enriched with examples in the other languages included in the study, and further refined based on observations and exchanges derived from the encoding of the various cases and their particularities. The following section describes in more detail the main characteristics of the proposed model.

3. LLODIA model

The main class of the LLODIA model is `LexicalRecord`, a wrapper around an `ontolex:Form`, which contains temporal information on when certain linguistic events about the form were observed. For this purpose a time interval was devised using the `dct:Period`¹⁵ class was devised, including `dct:start` and `dct:end` properties, to be associated with the record. `LexicalRecord` was conceived as a subclass of `frac:Observable` referring to entities about which a series of corpus- and dictionary-based observations can be documented.

Figure 1 shows the connections of the class `LexicalRecord` to other classes. For instance, the invertible LLODIA properties `form`, `timeSlice`, `lexicalConcept`, and `isRecordOf` link a record with a form, the time interval in which a series of observations were performed, a lexical concept and a lexical chronicle (collection of lexical records). As shown in the figure, the chronicle contains 9 record instances, including “`r_révolution_1`” about the French form *révolution*, its frequency observed in the time slice 1690-1794, and an associated `frac:FixedSizeVector` resulting from applying static word embedding to that corpus segment.

Listing 1: Lexical concept related to a lexical record.

```
<ontolex:LexicalConcept rdf:about="
  lc_révolution_1">
  <ontolex:reference rdf:resource="
    c_bnlm_fra"/>
  <frac:embedding rdf:resource="
    neighb_révolution_1"/>
  <frac:attestation rdf:resource="
    ca_révolution_1"/>
  <ontolex:lexicalizedSense rdf:
    resource="
    d_plex_fra_révolution_n_I.B.2"/>
</ontolex:LexicalConcept>
```

Further information about the form was encoded by means of the class `ontolex:LexicalConcept` associated with the lexical record. We considered that lists of

¹⁵<http://purl.org/dc/terms/>.

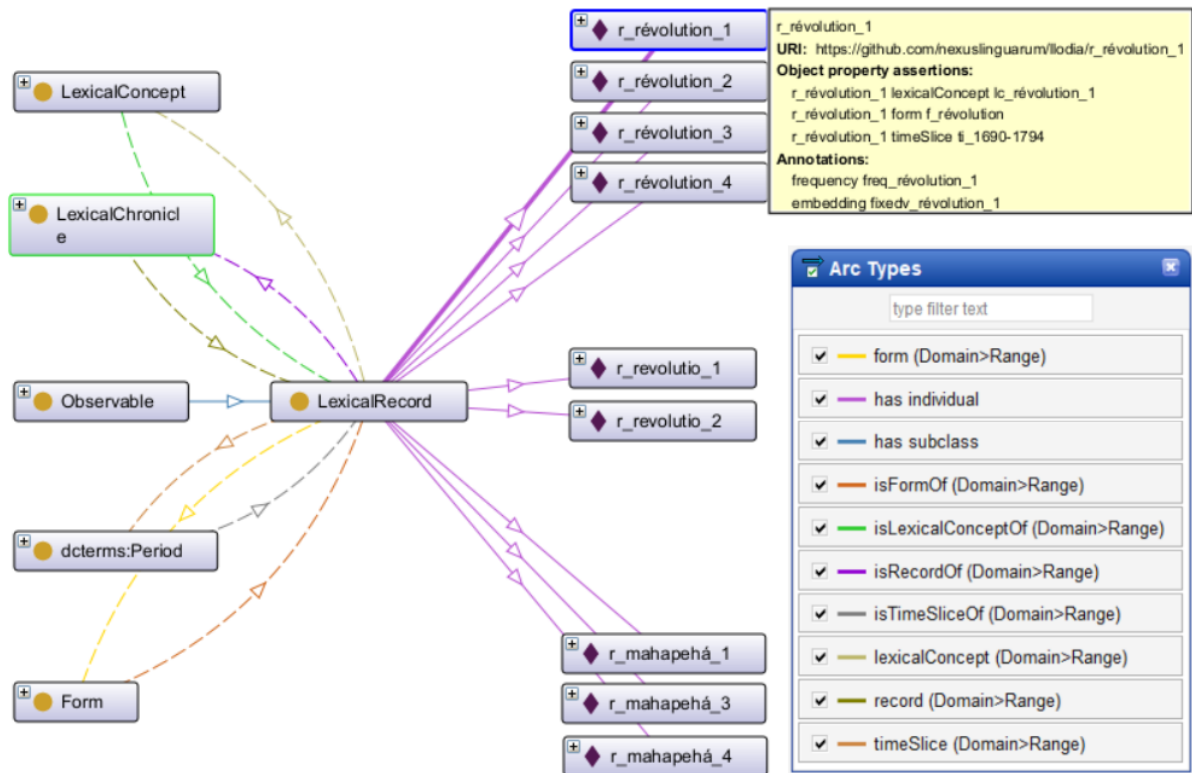


Figure 1: LexicalRecord and arc connections to classes (dashed) and individuals (solid) in Protégé.

neighbours and attestations from the corpus can capture certain aspects of the meaning of the word corresponding to the observed time period. Listings 1 and 2 describe the lexical concept associated with the “r_révolution_1” record, which refers to a list of neighbours computed through cosine similarity, a corpus attestation and a link to a lexical sense provided by a reference dictionary. Therefore, a `LexicalConcept` encompasses corpus-based evidence (neighbours, vector/embedding-related information, citation), while the related `LexicalSense` encapsulates dictionary-based evidence (sense, its domain and meaning explanation, attestation date). Thus, a record that documents the usage of a form observed in a certain time interval and corpus is indirectly connected through concepts to one or more senses in a reference lexicographical resource.

Listing 2: Corpus attestation of a lexical concept.

```
<frac:Attestation rdf:about="
  ca_révolution_1">
  <ontolex:reference rdf:resource="
    c_bnlm_fra"/>
  < dct:date>1789</dct:date>
  <frac:citation>
    <cito:Citation rdf:about="
      cc_révolution_1">
      <dct:title>L'art de conduire et
        régler les pendules ...
```

```
</dct:title>
<dct:creator>F. Rosset</dct:
  creator>
<dct:publisher>Chez la Veuve de
  J. B. Kleber ...
</dct:publisher>
<dbo:country rdf:resource="http
  ://dbpedia.org/resource/
  Luxembourg"/> ...
<rdf:value rdf:datatype="xsd:
  string">La roue de longue
  tige ou grande moyene fait
  une révolution par heure ...
</rdf:value>
<rdfs:comment>p. 13</rdfs:
  comment>
<dct:source rdf:resource="https
  ://viewer.eluxemburgensia.lu
  /ark:70795/dqgfr3/pages/17/
  articles/DTL612"/>
</cito:Citation>
</frac:citation>
</frac:Attestation>
```

We defined two types of resources, `Corpus` and `Dictionary`, as LLODIA subclasses of `dcmitype:Collection` and `lexicog:LexicographicResource`. They were utilised to attest word forms and their meaning in a given time period and space, since information about the publishers and their location was also encoded, when available. `Corpus`

attestations were related to lexical concepts and associated neighbour list, while dictionary attestations were connected to lexical senses that were further linked to lexical entries corresponding to the observed forms (integrated as `ontolex:canonicalForm`). Translation and etymological relations across languages were encoded via `vartrans:TranslationSet` and `lemonEty:Etymology`, inspired by (Abromeit et al., 2016; Khan, 2018; Khan et al., 2020), and based on information extracted from multilingual resources such as Wiktionary or monolingual dictionaries. We considered that this type of corpus- and dictionary-based evidence allows the researcher to document and contextualise word meanings and their evolution and circulation over time and space.¹⁶

To test these assumptions, we created a set of interconnected examples for the term *revolution* in the six languages included in the study, with English as a pivot for general explanations of the sense meanings and descriptions of the process. When not enough evidence was available from the corpora, the information from the dictionaries was used instead. The following sections provide an overview of the observations encoded as a proof of concept and a series of queries on the model.

4. Multilingual Proof of Concept

The modelling task has drawn our attention to the dynamics of association between corpus and dictionary forms that express and record meaning characterisations and their usage over time and space. The following examples illustrate this aspect from the perspective of the datasets and languages considered for analysis.

4.1. French

The results of word embedding on the French corpus indicated that the term *révolution* occurred 16, 276, 97 and 82 times in four of the six time slices defined for analysis (1690-1794, 1831-1866, 1867-1889 and 1890-1918). For the neighbours intended to be included in the LLODIA encoding, we used the top 20 most similar words with *révolution* computed via cosine similarity. We devised a series of prompts for ChatGPT-4 to assist with the task of selection and alignment with dictionary senses. The agent was asked to separate the lists into sub-lists that could most likely be aligned with the senses of the word *révolution* according to the CNRTL's lexical portal. The process was iterative and in

¹⁶The model and proof of concept has been published in the Nexus Linguarum GitHub repository (Armaselu et al., 2024): <https://github.com/nexuslinguarum/lloodia/>.

subsequent steps the citations extracted from the four corpus segments were also included in the prompts. Then, the output of the GenAI agent was manually checked and the terms from the sub-lists of neighbours considered most relevant to the chosen senses were selected.

The concept and associated dictionary sense for *révolution* assigned to the first time slice of the corpus corresponded to the domain of (1) mechanics as related to the circular motion of a body around its axis. The neighbours selected to model this concept included 10 terms, such as *moyene*, *ajouter*, *chant*, *enveloppeur*, *corde*, *tige*, with similarity measures between 0.89 and 0.79, and a citation from the field of clockwork mechanics describing the movement of wheels, minute and hour hands. The attestation date of this sense in the dictionary was 1727, with a citation from a French author, while the corpus citation was dated 1789 and indicated a Luxembourgish publisher. The list selected for the second corpus segment included 6 terms, e.g., *paraboloïde*, *polaire*, *lemniscate*, with similarity values between 0.65 and 0.58, and a citation pertaining to the domain of (2) geometry and the motion of a geometric form around an axis. The dictionary and corpus attestations pointed to the years 1799 and 1844, and to a French author and respectively Belgian publisher for the corpus citation.

A similar procedure was applied for the two other time intervals. The concepts and dictionary senses for *révolution* corresponding to them were related to the domains of (3) geophysics (natural phenomena changing the physical characteristics of the Earth) and (4) politics (sudden overthrow of the political regime of a nation) for the third segment, and (5) the French Revolution for the fourth one. The lists of neighbours selected for these concepts included terms such as *écroulement*, *plutonien*, *explosion* for concept (3), *nationalité*, *avènement*, *fédératif* for (4) and *vandalisme*, *insurrection*, *insurgé* for (5), with cosine similarity values in the range 0.70 - 0.61, 0.67 - 0.60 and respectively 0.64 - 0.57. The dictionary attestation years for the corresponding senses indicated 1749, 1636 and 1789, while the corpus attestation dates that we recorded for the related concepts were 1883 for (3) and (4), and 1904 for (5). GenAI prompts were tested for French, and then for Hebrew and Lithuanian and the outputs were manually checked and compared with the results of the evaluation method called LLM-Eval (Lin and Chen, 2023) applied for these languages.

4.2. Latin

According to the DMLBS (Ashdowne, 2016), the term *revolutio* has the following (main) senses: 1. (act of) rolling back or aside 2. (act of) unrolling or opening (book) 3. act of revolving, circular movement, revolution (referred to celestial motion or to cyclical

passage of time); 4. regular and recurring succession of persons in office, rotation; 5. something that forms a circular shape, coil, spiral; 6. act of turning over 7. reflection on, consideration of, going back over a past event; 8. repetition 9. relapse. The term is etymologically derived from the verb *revolvere* which means ‘to roll back; to unroll, unwind; to revolve, return’ and is attested from the Classical era, e.g., in Cicero and Livius, although it becomes especially frequent in the Augustan period e.g., in Vergil.¹⁷ In the LatinISE corpus (McGillivray and Kilgarriff, 2013), this lemma occurs 21 times, all in Medieval and early modern texts. It occurs twice, within the same sentence, in *Problemata Heloissae cum Petri Abaelardi solutionibus* by Peter Abelard (1110) with the sense 1 (act of rolling back or aside), referred to the movement of a stone.

The remaining 19 occurrences are found in the following texts, where *revolutio* expresses sense 3 (act of revolving, circular movement, revolution, referred to celestial motion or to cyclical passage of time): *Sermones* by Peter Abelard (1110); *De luce seu de inchoatione formarum* and *De impressionibus aeris seu de prognosticatione* by Robert Grosseteste (1200); *Missale Romanum* (1570). We trained fastText embeddings on LatinISE with window size 5 and minimum frequency count 5, turning subwords off.¹⁸ The first ten closest neighbours of *revolutio* in the model (with their associated cosine similarity scores) are: *vergiliarum* ‘Pleiades’ (constellation)(0.80), *solstitialis* ‘of the summer solstice or referred to solar revolution’, (0.80), *autumnale* ‘autumnal’ (0.79), *solstitium* ‘solstice’ (0.78), *arcticum* ‘northern, arctic’ (0.77), *tricesima* ‘the thirtieth’ (0.77), *cente(n)simus* ‘the hundredth’ (0.77), *semicirculus* ‘half-circle’ (0.77), *sexdecim* ‘sixteen’ (0.76), *octobri* ‘of october’ (0.76). All the 10 closest neighbors refer to the semantic field of astronomy, time calculation, or the motion of rotation and revolution of the Earth around the sun. None of them pertains to the act of physical rolling motion i.e., the one illustrated in sense 1 in DMLBS. This is easily understandable given that this sense occurs only two times within the corpus, both in the same sentence, and therefore, the model training is affected by data sparsity.¹⁹

As it can be observed from the description of the

¹⁷The entries for *revolvere* and *revolutio* are not yet available in the most comprehensive Latin lexicographic resource, the monolingual dictionary *Thesaurus Linguae Latinae* (*Thesaurus-Kommission, 1900–*), therefore we relied on the definitions and attestations provided in other Latin dictionaries.

¹⁸Tests with a higher minimum count and wider windows (10 to 50) led to unsatisfactory results. We turned subwords off in order to avoid getting orthographically similar words among the closest neighbours.

¹⁹Extending the number of closest neighbours to 20 did not improve the results.

occurrences of *revolutio* in the corpus, senses 1 and 3 are both attested for the first time in the corpus in 1110 (in the two texts by Peter Abelard). This, combined with the limited number of occurrences of *revolutio* with sense 1, has made it impossible to achieve satisfactory results when applying fastText on the corpus divided into smaller time spans.

4.3. Hebrew

Wiktionary defines the term in Hebrew *מהפכה* (*revolution*) as having the following meanings: 1. A historical event that significantly altered the trajectory of a specific nation or the course of human civilization as a whole. This could include revolutionary events like a technological revolution, such as the advent of the printing press, or a political upheaval like the French Revolution, which resulted in the overthrow of absolute monarchy. 2. Biblical terminology: destruction. 3. Derived from 2: chaos, commotion, a state of evident disarray. The Milog dictionary proposes an additional meaning for the word (4): Full restoration, altering the current arrangement and routines. The term has occurred in three distinct periods of the Responsa corpus (1st, 3rd, and 4th), each time in varying contexts. We obtained the 30 most closely related terms to the term *מהפכה* for each of the time periods. We manually chose 10 neighboring terms, excluding non-informative words that cannot be understood without context.

By examining the chosen terms, we assigned the most prevalent sense to each time period. The first period was assigned the fourth sense, as indicated by terms such as *מהטעות* (by the mistake)(0.72), *החיסרון* (the disadvantage)(0.71), and *התועבה* (the abomination)(0.698). These were primarily utilized in a religious context. The first sense has been assigned to the third and fourth periods. The third period is characterized by words such as *וייהרגו* (and they killed us)(0.71), *לאונסה* (to rape her)(0.66) and *שהונות* (that the prostitution)(0.65), which convey the themes of war and tragedy. This aligns with the historical periods of the French corpus, as it reflects the pogroms that Jews experienced during this period. The fourth period is characterized by neighboring words that are prominent in the context of medical and industrial revolutions, such as *החייאה* (resuscitation)(0.65), *ממכונות* (from machines)(0.646) and *פאטולוגיה* (pathology)(0.645).

It is important to observe that word2vec, as a non-contextualized approach, primarily provides terms that commonly occur in similar contexts as the given word. However, frequently, these contexts may not necessarily indicate the right sense of the word, even when used in the most prominent context. Moreover, on certain occasions, the word itself may be used in a manner that is outdated, conveying a meaning that is not explicitly defined in

the dictionary. For instance, a sentence extracted from the fourth period states: המכונה בעצמה כובסת הכביסה במה שהחשמל מהפכה וע"י זה נעשה הכביסה במכונה הכביסה (The machine itself washes the laundry as the electricity **turns** it and by this the washing is done in the machine by itself and not by a person). The context of this sentence is certainly related to an industrial revolution. However, the word מהפכה means turn which is not a direct sense of the word in the dictionary and is kind of archaic way to express the act of "turning" (הפיכה).

4.4. Lithuanian

For the modelling experiments related to the etymology of *revolution*, in Lithuanian we used the attestation of the dictionary LIETUVIUZODYNAS.It which shows that *revoliucija* comes from Latin *revolutio*. Another dictionary, Lietuvių kalbos žodynas, identifies that the word was first mentioned in Lithuanian texts in the 19th century. Relying on the dictionary the word has two meanings: 1. *staigus prievartinis politinės valdžios nuvertimas, sukeliantis esminius visuomenės pakitimus (a sudden, forcible overthrow of political power, causing fundamental changes in society)*; 2. *kokybinis raidos pakeitimas (qualitative change of development)*.

4.5. Romanian

For the Romanian language, we used the DEXonline digital dictionary to determine the etymology and the different meanings of the word *revoluție* (en. revolution). According to this dictionary, the etymology of the word *revoluție* comes from three terms, i.e., the Latin term *revolutio*, the French term *révolution*, and the German term *Revolution*. The term *revoluție* has the following main senses: 1) a fundamental change in the values, political institutions, social structure, leaders, and ideologies of a society (in the philosophy field); 2) revolt, uprising; 3) a radical change or transformation in a certain field; 4) a continuous periodic motion of a body following a closed curve; 5) the rotational motion of a body around a fixed straight line (geometry); 6) the motion of a body that travels a fixed curve (physics); and 7) the geological change of the Earth's crust.

5. Queries

Once the conception of our model was stabilised and examples in all five languages were produced, we wanted to check the functionality of the LLODIA model through queries. For this purpose, we have chosen Vocbench²⁰ that included a SPARQL query

²⁰<https://vocbench.uniroma2.it/>.

editor.

Our intention was to test whether temporal aspects can be included in the queries to allow for time-based comparison across languages. Listing 3 illustrates how lexical records corresponding to a certain time interval can be retrieved from the model. In this case, four records, one from the Hebrew, and the other from the French dataset were retrieved.

Listing 3: Lexical record by time slice (Vocbench SPARQL).

```
PREFIX ...
SELECT DISTINCT ?lex_record ?t_start ?
t_end WHERE {
  ?lrecord rdf:type lldia:
    LexicalRecord .
  ...
  ?tslice rdf:type dct:Period .
  ?lrecord lldia:timeSlice ?tslice .
  ?tslice dct:start ?t_start .
  ?tslice dct:end ?t_end .
FILTER (?t_start >= "1600-01-01" && ?
t_end <= "1900-12-31")
Results count: 4
lex_record t_start t_end
"r_mahapehá_3" "1601-01-01" "1900-12-31"
"r_révolution_1" "1690-01-01"
"1794-12-31"
"r_révolution_2" "1831-01-01"
"1866-12-31"
"r_révolution_3" "1867-01-01"
"1889-12-31"
```

Another element that seemed relevant to us in the context of diachronic analysis was the retrieval of attestation dates and places, to get an idea about when and where certain pieces of knowledge were produced. Listing 4 displays two dictionary and two corpus attestations, with their respective dates and place of publication for citations of the terms *revolutio* and *revolution* in Latin, Lithuanian, French and Hebrew and two time intervals.

Listing 4: Dictionary and corpus attestation by date and publisher place (Vocbench SPARQL).

```
PREFIX ...
SELECT DISTINCT ?attestation ?att_date ?
pub_place WHERE {
  ?att rdf:type frac:Attestation .
  ...
  ?att dct:date ?att_date .
  ?cit rdf:type cito:Citation .
  ?att frac:citation ?cit .
  ?cit dbo:country ?pl .
  ...
FILTER ((?att_date >= "1150" && ?
att_date <= "1180") || (?att_date >=
"1890" && ?att_date <= "1920"))
Results count: 4
attestation att_date pub_place
"da_revolutio_2" "1157" "England"
```

```
"da_revoliucija_n_1" "1894" "Lithuania"
"ca_révolution_4" "1904" "Luxembourg"
"ca_mahapehá_4_2" "1917" "Israel"
```

The query from listing 5 explores the possibility of finding similar domains across different languages, in which the various meanings of the retrieved terms were observed. The results display the domains of mechanics and astronomy and corresponding dictionary senses and their explanations in English for French, Latin and Romanian.

Listing 5: Sense by subject (Vocbench SPARQL).

```
PREFIX ...
SELECT DISTINCT ?lex_sense ?subj ?expl
WHERE {
  ?ls rdf:type ontolex:LexicalSense .
  ?ls dct:subject ?ls_subj .
  ?ls rdfs:comment ?expl.
  ...
FILTER ((?subj = "Mechanics" || ?subj =
"Astronomy") && LANG(?expl)="eng")
Results count: 3
lex_sense subj expl
"d_plex_fra_révolution_n_I.B.2" "
Mechanics" "Circular motion of a
body around its axis."@eng
"d_dmlbs_lat_revolutio_n_3.bc" "
Astronomy" "Act of revolving,
circular movement, revolution (w.
ref. to celestial motion and to
cyclical passage of time)."@eng
"d_dex_ron_revoluție_n_3" "Mechanics" "
Circular motion of a body around its
axis."@eng
```

Translation relations can also be interrogated as illustrated in listing 6 that provides the translation of the French word *révolution* in English, Hebrew, Lithuanian and Romanian.

Listing 6: Translation (Vocbench SPARQL).

```
PREFIX ...
SELECT DISTINCT ?source ?target WHERE {
  ?trans_set rdf:type vartrans:
TranslationSet .
  ?trans_set vartrans:source ?s_form.
  ?trans_set vartrans:target ?t_form.
  ?s_form rdf:value ?source .
  ?t_form rdf:value ?target .
FILTER (LANG(?source) = "fra")}
Results count: 4
source target
"révolution"@fra "revolution"@eng
"révolution"@fra "מהפכה mahapehá"@heb
"révolution"@fra "revoliucija"@lit
"révolution"@fra "revoluție"@ron
```

Finally, listing 7 presents a query about the etymons of the various forms stored in the model. The results show the common Latin root *revolutio* for *revolution* in French, Lithuanian and Romanian, the etymon of this root in Latin, and a different origin

for Hebrew. Additional etymons are displayed for Romanian, the French form *révolution* and German *Revolution*. Etymological chains can be inferred, e.g., between the French, Lithuanian and Romanian forms, and the Latin *revolutio* and its etymon *revolvō*. It should be noted that both the translation and etymological relations were defined at the level of forms but other approaches, considering for instance connections at the sense level or complex etymological relations, can be imagined as well. These aspects are currently under study.

Listing 7: Etymology (Vocbench SPARQL).

```
PREFIX ...
SELECT DISTINCT ?form ?etymon
WHERE {
  ?frm rdf:type ontolex:Form .
  ?etm rdf:type lemonEty:Etymology .
  ?etym rdf:type ontolex:Form .
  ?frm lemonEty:etymology ?etm .
  ?etm llodia:etymon ?etym .
  ?frm rdf:value ?form .
  ?etym rdf:value ?etymon .
}
Results count: 7
form etymon
"révolution"@fra "revolutio"@lat
"מהפכה mahapehá"@heb "הפך hapah"@heb
"revoliucija"@lit "revolutio"@lat
"revolutio"@lat "revolvō"@lat
"revoluție"@ron "révolution"@fra
"revoluție"@ron "revolutio"@lat
"revoluție"@ron "Revolution"@deu
```

Our assumption was that this type of model can capture some of the complexities of the linguistic phenomenon of change in meaning over time and space, and across languages. Although the proof of concept contained a limited number of examples and was affected by data sparsity in some cases, it showed that interconnections can be built between time- and space-aware representations based on multilingual and varied types of resources. The sets of neighbours and corpus citations could provide insights into the contexts where a form occurred. The senses and attached domains could enable inferences about how the corresponding meanings, recorded by reference sources and reflecting the accepted usage by the community in a certain period of time, were possibly transmitted from one language to the other, evolved independently or influenced each other across linguistic and cultural borders, or disappeared.

6. Conclusion and future work

In this article, we proposed a LLOD model for diachronic analysis (LLODIA) and a proof of concept in five languages (French, Hebrew, Latin, Lithuanian, Romanian, with English as a pivot) for the term

revolution. We argue that a combination of corpus and dictionary evidence on the evolution of word meanings and its modelling in a structured format can provide a richer basis for analysing multilingual diachronic phenomena than each part alone. For this purpose, we used word embeddings computed on diachronic corpora, reference dictionaries and existing Semantic Web vocabularies, and created new classes and properties when the elements needed for our investigation were not available.

We used a set of queries to test the capabilities of the LLODIA model to express and support inferences based on time and space dimensions and interconnections across languages. While simple translation and etymological relations at the level of forms were considered at this stage, further enquiry is intended for more complex cases that require sense-level interrelations or etymological chains.

We designed LLODIA as a small-scale model and proof of concept that may serve as a starting point for other projects that combine NLP and LLOD methods to detect and represent change of meaning over time, space and across several languages. It can also be imagined as a larger lexicographic project based on interoperability with other vocabularies and expanded as an online resource aggregator that may be enriched, queried and reasoned upon by various contributors. However, for this type of interaction a dedicated infrastructure would be needed, which is a subject matter that needs additional study and examination.

7. Acknowledgments

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.

8. Authors' contribution

F.A., sections 1, 2.1 (French), 2.2, 3, 4.1, 5 and 6; C.L., sections 2.1 (Hebrew) and 4.3; P.M. and B.M., sections 2.1 (Latin) and 4.2; G.V.O., sections 2.1 (Lithuanian) and 4.4; E.S.A. and C.O.T., sections 2.1 (Romanian) and 4.5; D.G., section 2.1 (Romanian). All the authors critically revised the final version of the manuscript.

9. Bibliographical References

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 *Proceedings of the 5th Workshop on Linked Data in Linguistics: Managing, Building and Using Linked Language Resources (LDL-2016)*, 24 May 2016, Portorož, Slovenia, pages 11 – 19.
- Florentina Armaselu, Chaya Liebeskind, Paola Marongiu, Barbara McGillivray, Giedrė Valūnaitė Oleškevičienė, Elena Simona Apostol, and Ciprian-Octavian Truică. 2024. *Linguistic Linked Open Data for Diachronic Analysis (LLODIA)*.
- R Ashdowne. 2016. Data in online version of the 'dictionary of medieval Latin from British sources' (dmlbs).
- 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.
- Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrike, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. *Sparks of artificial general intelligence: Early experiments with GPT-4*. (arXiv:2303.12712). ArXiv:2303.12712 [cs].
- Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. 2023. *Unleashing the potential of prompt engineering in large language models: a comprehensive review*. (arXiv:2310.14735). ArXiv:2310.14735 [cs].
- Christian Chiarcos, Elena-Simona Apostol, Besim Kabashi, and Ciprian-Octavian Truică. 2022a. *Modelling frequency, attestation, and corpus-based information with OntoLex-FrAC*. In *International Conference on Computational Linguistics*, pages 4018–4027.
- Christian Chiarcos, Katerina Gkirtzou, Maxim Ionov, Besim Kabashi, Fahad Khan, and Ciprian-Octavian Truică. 2022b. *Modelling collocations in OntoLex-FrAC*. In *Proceedings of Globalex Workshop on Linked Lexicography within the 13th Language Resources and Evaluation Conference*, pages 10–18.
- Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. *AllenNLP: A deep semantic natural language processing platform*. (arXiv:1803.07640). ArXiv:1803.07640 [cs].
- Jolanta Gelumbeckaitė, Mindaugas Šinkūnas, and Vytautas Zinkevičius. 2012. "senosios lietuvių kalbos tekstynas" (SLIEKKAS) - nauja diachroninio tekstyno samprata. *Darbai ir dienos*, 58:257–278.

- Daniela Gifu. 2016a. Lexical semantics in text processing. contrastive diachronic studies on Romanian language.
- Daniela Gifu. 2016b. Diachronic evaluation of newspapers language between different idioms. In *Proceedings of the IJCAI 2016 Workshop. Natural Language Processing meets Journalism*.
- Anas Khan. 2018. [Towards the representation of etymological data on the Semantic Web](#). *Information*, 9(12):304.
- 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, Ros Salvador, and Ciprian-Octavian Truică. 2022. [When linguistics meets Web technologies. Recent advances in modelling linguistic linked open data](#). *Semantic Web*.
- Fahad Khan, Laurent Romary, Ana Salgado, Jack Bowers, Mohamed Khemakhem, and Toma Tasovac. 2020. [Modelling etymology in LMF/TEI: The Grande Dicionário Houaiss da Língua Portuguesa Dictionary as a use case](#). In *LREC2020-12th Language Resources and Evaluation Conference*.
- 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.
- 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.
- Yen-Ting Lin and Yun-Nung Chen. 2023. [LLM-Eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models](#). In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, page 47–58, Toronto, Canada. Association for Computational Linguistics.
- 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 and Adam Kilgarriff. 2013. Tools for historical corpus research, and a corpus of Latin. In *New Methods in Historical Corpus Linguistics*, Tübingen. Narr.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. [Efficient estimation of word representations in vector space](#). *arXiv:1301.3781 [cs]*. ArXiv: 1301.3781.
- OpenAI. 2023. [GPT-4 Technical Report](#).
- Sabrina Ortiz. 2023. [What are Microsoft's different Copilots? Here's what they are and how you can use them](#).
- Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. [Deep contextualized word representations](#). *arXiv:1802.05365 [cs]*. ArXiv: 1802.05365.
- Radim Rehurek and Petr Sojka. 2010. [Software framework for topic modelling with large corpora](#). In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, page 45–50, Valletta, Malta. ELRA.
- Thesaurusbüro München Internationale Thesaurus-Kommission, editor. 1900–. *Thesaurus linguae latinae*. Mouton de Gruyter, Berlin.
- Ciprian-Octavian Truică, Victor Tudose, and Elena-Simona Apostol. 2023. Semantic change detection for the Romanian language. In *International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC2023)*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. [Chain-of-thought prompting elicits reasoning in large language models](#). (arXiv:2201.11903). ArXiv:2201.11903 [cs].
- Christopher Welty, Richard Fikes, and Selene Makarios. 2006. [A reusable ontology for fluents in OWL](#). In *Proceedings of the Fourth International Conference, FOIS 2006*, page 8, Baltimore, Maryland, USA.

Cross-Lingual Ontology Matching using Structural and Semantic Similarity

Shubhanker Banerjee ¹, Bharathi Raja Chakravarthi ², John Philip McCrae ¹

¹ ADAPT Centre, University of Galway

² School of Computer Science, University of Galway
shubhanker.banerjee@adaptcentre.ie

Abstract

The development of ontologies in various languages is attracting attention as the amount of multilingual data available on the web increases. Cross-lingual ontology matching facilitates interoperability amongst ontologies in different languages. Although supervised machine learning-based methods have shown good performance on ontology matching, their application to the cross-lingual setting is limited by the availability of training data. Current state-of-the-art unsupervised methods for cross-lingual ontology matching focus on lexical similarity between entities. These approaches follow a two-stage pipeline where the entities are translated into a common language using a translation service in the first step followed by computation of lexical similarity between the translations to match the entities in the second step. In this paper, we introduce a novel ontology matching method based on the fusion of structural similarity and cross-lingual semantic similarity. We carry out experiments using 3 language pairs and report substantial improvements in the performance of the lexical methods thus showing the effectiveness of our proposed approach. To the best of our knowledge, this is the first work that tackles the problem of unsupervised ontology matching in the cross-lingual setting by leveraging both structural and semantic embeddings.

Keywords: cross-lingual ontology matching, cross-lingual semantic similarity, lexical similarity

1. Introduction

An increasing amount of multilingual data on the web has led to the development of ontologies in different languages. Ontologies are used to enable the sharing of information across different systems (Davies et al., 2002; Beydoun et al., 2011; Elmhadhbi et al., 2021). Furthermore, the adoption of ontologies as databases across domains has also attracted attention (Pankowski, 2023). These applications motivate the development of tools that allow semantic interoperability of ontologies across a wide range of languages. Identifying correspondences between ontologies in different languages is called Cross-lingual Ontology Matching (CLOM) (Ibrahim et al., 2023). Cross-Lingual Ontology Matching has the potential to contribute to various areas such as ontology enrichment, peer-to-peer information sharing, and linked data. Despite these potential applications CLOM has largely been an unexplored research problem. Therefore, more efforts from the research community towards building flexible CLOM systems are needed.

In recent times, deep learning based methods have achieved good results on ontology matching (Iyer et al., 2020; Li et al., 2019b; He et al., 2022). However, these methods are dependent on large amounts of training data which are not available in cross-lingual scenarios. To tackle this challenge we present an unsupervised ontology matching approach for CLOM. The proposed approach uses a state-of-the-art text embedding model to embed the concept descriptions into low-dimensional vectors

which are then used to compute semantic similarity. Structural similarity between source and target concepts is an integral part of the proposed approach. We leverage the semantic similarity between source and target concepts to generate reference alignments. These reference alignments are used to learn structural embeddings for each concept in source and target ontologies. The semantic and structural embeddings are then used to calculate a weighted similarity to find equivalent entities in two ontologies. The main contributions of this work can be summarized as follows:

- Our experiments reveal that a weighted combination of semantic and structural similarity achieves performance gains over lexical similarity measures.
- We evaluate our method on 3 language pairs to demonstrate its extensibility.
- The proposed approach does not require manually labeled alignment data and thus is suitable for application in data-scarce scenarios.

The paper is organized as follows: Section 2 discusses the related works, Section 3 describes the methodology, the experiments are described in Section 4, the results are discussed in Section 5. The conclusion is given in Section 6. The limitations have been discussed in Section 7.

2. Related Work

Cross-lingual Ontology Matching. Traditionally, CLOM approaches involve translation of the concepts into a common language (usually English) followed by calculation of lexical similarity to identify equivalent concepts. Following this paradigm [Fu et al. \(2010\)](#) propose a CLOM approach that selects the appropriate translation from amongst multiple translations generated by their system based on synonym-based matching with the entities in the target ontology. Furthermore, to resolve conflicts in alignments their system relies on the similarity of 1-hop neighbours of the entities from source and target ontologies. The translation in [Ibrahim et al. \(2019\)](#) follows a similar approach where they select candidate translations based on similarity to target concepts. Their system outperforms the state-of-the-art systems on the Ontology Alignment Evaluation Initiative (OAEI)¹ 2018 benchmark. [Ibrahim et al. \(2020\)](#) introduced MULON, a modularized CLOM system based on lexical and semantic similarity. The alignments are computed using a combination of both similarities. They use Jaccard for lexical similarity and WordNet path-based matching for semantic similarity.

MoMatch ([Ibrahim et al., 2023](#)) is based on lexical similarity of translated entities computed using metrics such as Jaccard ([Jaccard, 1901](#)), Levenshtein ([Levenshtein et al., 1966](#)), Jaro ([Jaro, 1989](#)) and Jaro-Winkler ([Wang et al., 2017](#)). The translation is carried out using the Yandex translation API and they improve upon the performance of the state-of-the-art methods for CLOM from the OAEI 2020 benchmark. [Sharma and Jain \(2023\)](#) achieve the best results on the MultiFarm dataset ([Meilicke et al., 2012](#)) at OAEI 2023. Their method uses Levenshtein-based similarity of translated concepts and WordNet-based synonym matching to align concepts. Machine learning based methods have also been explored for CLOM; [Spohr et al. \(2011\)](#) use a small amount of manually aligned concepts to train a SVM with 20 string-based features and 22 structural features for CLOM. [Gracia and Asooja \(2013\)](#) leverage artificial neural networks (ANNs) to calculate similarity between source and target concepts using manually designed features.

Unsupervised Entity Alignment. Ontologies are graph structures that describe hierarchies between concepts within a domain ([Zhapa-Camacho and Hoehndorf, 2023](#)). Therefore, ontology matching is fundamentally similar to the task of entity alignment across knowledge graphs. Unsupervised and self-supervised methods have been proposed for aligning entities in data-scarce scenarios. [Liu et al. \(2022a\)](#) propose a self-supervised

training objective based on contrastive learning for entity alignment. To generate reference training alignments they use semantic similarity between concept descriptions from the source and target ontologies. The descriptions are encoded using LaBSE ([Feng et al., 2022](#)) and graph attention network ([Velickovic et al., 2018](#)) is used to learn structural embeddings using a self-supervised training objective based on noise-contrastive estimation ([Gutmann and Hyvärinen, 2010](#)). [Tang et al. \(2023\)](#) pose entity alignment as an optimal transport problem and report good results. In particular, they calculate fused Gromov-Wasserstein distance ([Vayer et al., 2019](#)) to minimize the distance between entities. [Mao et al. \(2021\)](#) formulate the ontology matching problem as a minimum sum assignment problem. The optimal assignments are calculated using the Hungarian ([Kuhn, 1955](#)) and Sinkhorn algorithms ([Sinkhorn, 1964](#)). Graph convolutional networks (GCN) ([Kipf and Welling, 2017](#)) have also been used to capture structural information. [Zeng et al. \(2021\)](#) use GCN to compute structural similarity between concept nodes in source and target ontologies. Textual similarity is computed using a weighted combination of Levenshtein similarity and cosine similarity of averaged word vectors. A weighted combination of structural and textual similarity is compared against a fixed threshold to align entities.

3. Methodology

We propose a framework for unsupervised cross-lingual ontology alignment. As discussed in Section 2, approaches based on lexical similarity have achieved state-of-the-art (SOTA) results on CLOM tasks. To demonstrate the effectiveness of our approach we compare it against 5 lexical similarity measures.

3.1. Task Formulation

The source and the target ontologies O_1 and O_2 respectively are inputs to the proposed CLOM system. The task of cross-lingual ontology matching is defined as finding aligned concepts between the ontologies. i.e.,

$$\phi = \{(a, b) | a \in C_1, b \in C_2, a \leftrightarrow b\},$$

where C_1 and C_2 refer to the concept sets in O_1 and O_2 , respectively, $a \leftrightarrow b$ represent alignment between source and target concepts i.e., a and b refer to the same object in the real world. In this paper, we focus on unsupervised cross-lingual ontology matching i.e., source and target concepts belong to different languages and there is no labeled alignment data available.

¹<http://oaei.ontologymatching.org/>

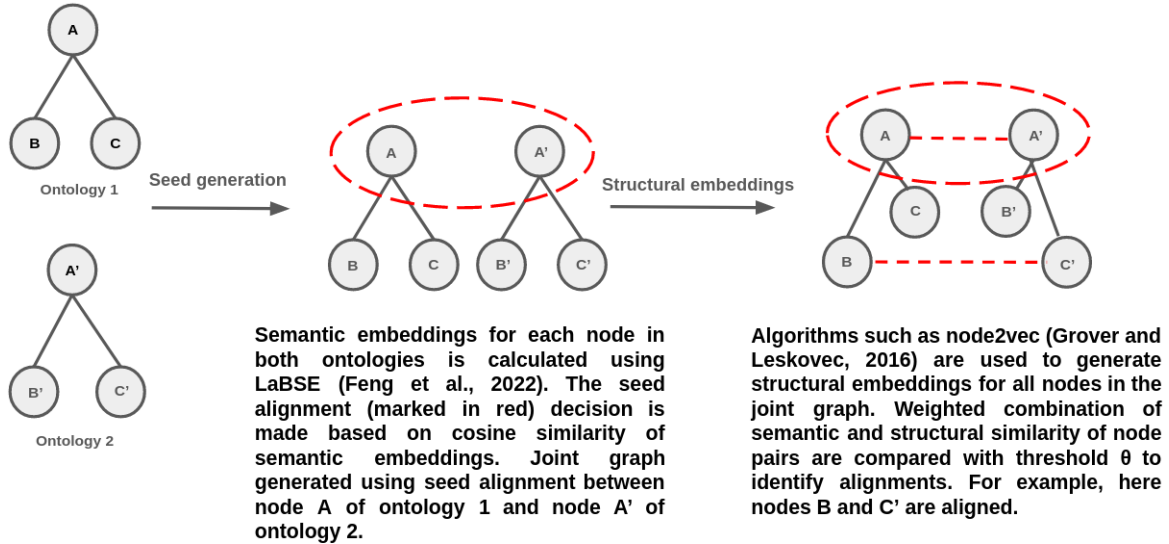


Figure 1: The source and target ontologies are inputs to our proposed CLOM framework in which we leverage both semantic and structural similarity of concepts to align the candidate nodes.

3.2. Concept Alignment

We hypothesize that aligned concepts in the source and target ontologies would have similar textual descriptions. This hypothesis postulates that the cosine similarity of text embeddings obtained from concept descriptions is positively correlated with the likelihood of the concepts being aligned/matched. Similar ideas have been explored by various supervised, semi-supervised, and self-supervised knowledge graph entity alignment approaches (Liu et al., 2022b; Wu et al., 2019; Chen et al., 2017; Tang et al., 2020). Here we leverage LaBSE (Feng et al., 2022), a multilingual model pre-trained on 109 languages for generating cross-lingual embeddings for the concepts in the source and target ontologies. Cosine similarity between the normalized embeddings is computed as a measure of semantic similarity between the corresponding concepts. Semantic similarity in the multilingual space is then leveraged for generating seed alignments between the input ontologies.

Ontologies are fundamentally graphs that represent concept hierarchies within a domain. In addition to textual descriptions, structural embeddings of the concepts in question can also constitute an important factor in determining alignment. Concept nodes with similar neighbourhoods are more likely to be aligned. We carry out experiments with various graph embedding approaches such as node2vec (Grover and Leskovec, 2016), Graph Convolutional Networks (Kipf and Welling, 2017, GCN), RGCN (Schlichtkrull et al., 2018) and TransE (Bordes et al., 2013) to learn embeddings for concept nodes. However, comparisons using embeddings learned on the two input ontologies independently are not meaningful as the embed-

dings would reside in two different vector spaces. Therefore, we leverage the seed textual alignments to consider source and target ontologies together as a graph and learn structural embeddings for all concept nodes in both ontologies. We employ two strategies for seed alignment for this task. In the first strategy, we select only those concept node pairs as alignments where the source and target concept node descriptions are semantically mutual nearest neighbours of each other. In the second strategy, we calculate the semantic similarity scores of all source and target concept pairs. The top-k most similar concept pairs are selected as seed alignments. We experiment with $k=1,3,5,7$ to quantify variation in performance as the number of seed alignments changes. To generate structural embeddings we train the node2vec model using the self-supervised loss defined by Grover and Leskovec (2016). RGCN and TransE are trained using a margin ranking loss (MR) based on negative sampling². The seed alignments are used as training data for training the GCN model using the training objective given in Equation 1

$$\mathcal{L} = \sum_{(a,b) \in S} \sum_{(a',b') \in S'} [d(a,b) + \gamma - d(a',b')]_+ \quad (1)$$

where $[\cdot]_+ = \max\{0, \cdot\}$ and (a, b) denotes a labeled concept pair from the training data. The set S' (a', b') represents negative concept pairs obtained by corrupting (a, b) using nearest neighbor sampling (Li et al., 2019a). The embeddings of the source and target concepts learned by GCN are denoted as a and b , respectively. The distance func-

²<https://pykeen.readthedocs.io/en/stable/reference/training.html>

tion measuring the distance between two embeddings is represented by $d(\cdot, \cdot)$. The hyper-parameter γ serves to separate positive samples from negative ones. Structural similarity between concept nodes is calculated using the cosine similarity of normalized structural embeddings generated by the graph embedding models.

Algorithm 1: The proposed algorithm combining semantic and structural similarity for ontology matching

Data: Source Ontology, O_1 , Target Ontology, O_2

Result: Aligned node pairs $\hat{\phi}$

- 1 **Strategy 1:** Select seed set S_1 by choosing concept node pairs (c_1, c_2) where $c_1 \in O_1$ and $c_2 \in O_2$ and descriptions of c_1 and c_2 are semantically mutual nearest neighbors of each other;
 - 2 **Strategy 2:** Calculate semantic similarity scores for all source and target concept pairs (c_1, c_2) where $c_1 \in C_1$ and $c_2 \in C_2$;
 - 3 Select the top-k most similar concept pairs as seed set S for experiments with $k = 1, 3, 5, 7$;
 - 4 Construct joint graph G_{joint} by combining O_1 and O_2 using S as reference alignments between the graphs;
 - 5 Learn structural embeddings for concept nodes in G_{joint} using one of the methods defined in Section 3.2;
 - 6 The combination of structural and semantic similarities $Sim_{Combined}$ is calculated using the Equation 2 as a measure of their alignment;
 - 7 Output the aligned concept pairs according to Equation 3;
-

	OP	DP	Concept Classes
cmt	49	10	30
confOf	13	23	39
sigkdd	17	11	50
conference	46	18	61

Table 1: Dataset statistics: The number of Object properties (OP), Data Properties (DP), and Concept classes in each ontology. For our experiments, only concept classes are considered.

Finally, as discussed above both structural and semantic similarity are positively correlated with the likelihood of alignment. Therefore, we use a weighted combination of both these measures to assign a final similarity score to a pair of concepts from the source and target ontologies as shown in

Equation 2.

$$Sim_{Combined} = \alpha \cdot Sim_{str} + (1 - \alpha) \cdot Sim_{sem} \quad (2)$$

where Sim_{str} is the structural similarity between concept nodes calculated using cosine similarity of normalized structural embeddings, Sim_{sem} is the semantic similarity of source and target concept node description calculated using the embeddings output by LaBSE. The concept pairs where $Sim_{Combined}$ is greater than a fixed threshold θ are considered to be aligned.

$$\hat{\phi} = \{(c_1, c_2) \mid c_1 \in C_1, c_2 \in C_2, Sim_{combined}(c_1, c_2) > \theta\} \quad (3)$$

where the $\hat{\phi}$ is the set of all aligned concept pairs (c_1, c_2) where C_1 is the set of all concepts in source ontology and C_2 is the set of all concepts in target ontology and θ is the fixed threshold. The algorithm for the aligning source and target concept nodes has been described in Algorithm 1.

4. Experiments

4.1. Dataset

We carry out experiments on 3 ontology pairs (cmt-confOf, conference-confOf, and conference-sigkdd) across 3 language pairs (German-English, German-French, and English-French) of the MultiFarm dataset (Meilicke et al., 2012). The MultiFarm dataset is a benchmark for multilingual ontology matching. It is used to evaluate the ability of systems to deal with ontologies in different languages. It consists of a set of 7 ontologies related to conferences. The dataset was derived by translating the OntoFarm dataset (Zamazal and Svátek, 2017) into 9 languages: Chinese, Czech, Dutch, French, German, Portuguese, Russian, Arabic and Spanish. The dataset statistics are given in Table 1.

4.2. Baselines

As discussed in Section 2, lexical string similarity measures constitute the core part of most state-of-the-art CLOM systems. Therefore, to evaluate the proposed approach we compare it to 5 lexical similarity measure commonly used in the literature, namely: Jaccard (Jaccard, 1901), Levenshtein (Levenshtein et al., 1966), Jaro (Jaro, 1989), Jaro-Winkler (Wang et al., 2017) and Tversky (Tversky, 1977). Since the baselines compute lexical similarity, we translate the source and target entities to English before using these methods. In our experiments, we have used MetaAI’s state-of-the-art NLLB model (Costa-jussà et al., 2022) to translate the source and target concepts. In particular, we use a distilled 600M parameter version of the model

	γ	Epochs	Learning rate	Walk length	# of walks	Batch size
GCN	3.0	1000	1e-5	—	—	1
node2vec	—	—	—	30	200	—
RGCN	3.0	100	—	—	—	2
TransE	3.0	100	—	—	—	2

Table 2: The hyperparameters used during training. The value of γ has been chosen based on prior work on knowledge graph entity alignment by Zeng et al. (2021). The default learning rate scheduler in pyKEEN is used for training RGCN and TransE. We set the other hyperparameters through empirical trial and error.

nllb-200-distilled-600M to limit the computational resources needed for inference.

4.3. Experimental Setup

As discussed above, to establish the effectiveness of our approach we carry out experiments on 3 ontology pairs across 3 languages. In the first step semantic similarity between source and target concepts is calculated to establish seed alignments between the ontologies. As discussed in Section 3.2 experiments are carried out with node2vec, GCN, RGCN, and TransE for the structural embeddings. The hyperparameters used during training are listed in Table 2. Furthermore, for our experiments, we empirically set the similarity threshold θ to 0.80. We carry out experiments with different values of α to ascertain the relative importance of both similarity measures for achieving good task performance.

4.4. Implementation Details

To ensure reproducibility we have used open-source libraries in our implementation. The ontologies were pre-processed using RDFlib³. We used Hugging Face⁴ to implement the translation pipeline for the baseline methods and calculate semantic similarity⁵ between the source and target concepts. GCN was implemented using Torch Geometric⁶. The node2vec algorithm was implemented using node2vec library⁷. TransE and RGCN were implemented using PyKEEN library⁸.

5. Results

Our main experimental results can be found in Table 3. It is important to note that semantic similarity

³<https://rdflib.readthedocs.io/en/stable/>

⁴<https://huggingface.co/>

⁵We used setu4993/LaBSE model from then hugging face repository to generate cross-lingual text embeddings

⁶<https://pytorch-geometric.readthedocs.io/en/latest/>

⁷<https://pypi.org/project/node2vec/>

⁸<https://pykeen.readthedocs.io/en/stable/>

using embeddings from LaBSE outperforms lexical similarity baselines in almost all cases on the F1-score, often by large margins in the range of approximately 1-40%. On the conference-sigkdd dataset node2vec-NN (NN implies node2vec with mutual nearest neighbour seed alignment strategy) has the best performance and achieves an average F1-score of 61.5% over all the language pairs. Similarly on the conference-confOf ontology pair node2vec-NN has the best performance on German-English and German-French datasets. However, on the English-French dataset, Jaro similarity outperforms all other methods. We also note that the TransE-NN based alignment approach outperforms the lexical methods in most cases but substantially lags behind node2vec-NN in all cases. The other two graph embedding methods namely GCN-NN and RGCN-NN have relatively bad performance and are outperformed by the lexical baselines in most cases. These observations indicate that using semantic similarity is a better alternative than lexical similarity for ontology matching. This result is not surprising as the semantic similarity is based on similarity of "meaning" whereas lexical similarity is based on overlap of surface forms and is dependent on the translations. Furthermore, the good performance of node2vec-NN also establishes the effectiveness of the proposed framework for ontology matching where we combine structural similarity with semantic similarity using a weighted combination. We attribute the relatively bad performance of GCN and RGCN models to the smaller size of the graph (≈ 100 nodes in source and target ontologies combined) leading to ineffective learning of node representations.

The results reported in Table 3 use $\theta = 0.80$. We recognize that fixed thresholds for alignment identification may lead to sub-optimal performance where a particular similarity threshold might not be optimal for all datasets. Higher thresholds might lead to a larger number of false negatives and a smaller threshold might lead to a larger number of false positives on different datasets. As demonstrated in Figure 2, these fluctuations might also have an impact on the overall performance.

Overall, the results indicate that incorporating

cmt-confOf			
	German-French	German-English	English-French
	Precision/Recall/F1	Precision/Recall/F1	Precision/Recall/F1
Jaro	66.6/44.4/53.3	66.6/44.4/53.3	50.0/20.0/28.5
Jaro-Winkler	44.4/57.1/50.0	66.6/75.0/70.5	50.0/55.5/52.6
Levenshtein	100.0/40.0/57.1	100.0/50.0/66.6	100.0/30.0/46.1
Jaccard	80.0/44.4/57.1	85.7/66.6/75.0	75.0/33.3/46.1
Tversky	41.6/62.5/50.0	43.7/87.5/58.3	33.3/42.8/37.5
LaBSE	83.3/50.0/62.5	83.3/55.5/66.6	66.6/60.0/63.1
LaBSE + node2vec-NN	100.0/50.0/66.6	100.0/55.5/71.4	71.4/50.0/58.8
LaBSE + GCN-NN	71.4/55.5/62.5	62.5/55.5/58.8	63.6/70.0/66.6
LaBSE + TransE-NN	100.0/50.0/66.6	100.0/50.0/66.6	100.0/50.0/66.6
LaBSE + RGCN-NN	100.0/22.2/36.3	100.0/10.0/18.1	100.0/20.0/33.3
conference-confOf			
Jaro	33.3/33.3/33.3	40.0/44.4/42.1	70.0/70.0/70.0
Jaro-Winkler	25.0/50.0/33.3	27.7/62.5/38.4	50.0/80.0/61.5
Levenshtein	66.6/18.1/28.5	75.0/27.2/39.9	83.3/45.4/58.8
Jaccard	20.0/11.1/14.2	25.0/22.2/23.5	50.0/45.4/47.6
Tversky	7.1/33.3/11.7	9.3/75.0/16.6	22.2/66.6/33.3
LaBSE	66.6/54.5/60.0	60.0/54.5/57.1	60.0/54.5/57.1
LaBSE + node2vec-NN	75.0/54.5/63.1	66.6/60.0/63.1	66.6/54.4/60.0
LaBSE + GCN-NN	50.0/60.0/54.5	46.1/60.0/52.1	44.4/72.7/55.1
LaBSE + TransE-NN	75.0/27.2/39.9	83.3/45.4/58.8	85.7/54.5/66.6
LaBSE + RGCN-NN	80.0/36.3/50.0	75.0/36.3/50.0	100.0/27.2/39.9
conference-sigkdd			
Jaro	42.8/30.0/35.2	42.8/30.0/35.2	40.0/18.1/25.0
Jaro-Winkler	25.0/40.0/30.7	29.4/50.0/37.0	33.3/27.2/30.0
Levenshtein	75.0/25.0/37.5	75.0/25.0/37.5	50.0/8.3/14.2
Jaccard	27.2/27.2/27.2	20.0/30.0/24.0	28.5/20.0/23.5
Tversky	8.5/42.8/14.2	10.5/44.4/17.0	13.3/57.1/21.6
LaBSE	55.5/41.6/47.6	42.8/54.5/47.9	60.0/50.0/54.5
LaBSE + node2vec-NN	66.6/50.0/57.1	70.0/63.6/66.6	58.3/63.6/60.8
LaBSE + GCN-NN	36.8/58.3/45.1	43.7/63.6/51.8	38.8/70.0/50.0
LaBSE + TransE-NN	60.0/25.0/35.2	80.0/33.3/47.0	66.6/33.3/44.4
LaBSE + RGCN-NN	100.0/25.0/40.0	100.0/8.3/15.3	100/16.6/28.5

Table 3: Precision, recall, and F1-scores of 5 lexical baselines compared with node2vec-NN, GCN-NN, TransE-NN, and RGCN-NN (NN indicates that mutual nearest neighbour source and target concepts are used as seed alignments between the input ontologies. This seed generation strategy is described as Strategy 1 in Algorithm 1) for $\theta = 0.80$ and $\alpha = 0.2$.

structural information improves performance as compared to only using semantic similarity. However, the performance is sensitive to the choice of embedding methods used as node2vec substantially outperforms GCN. Furthermore, these results have been reported for $\alpha=0.2$ which signifies a smaller contribution of structural similarity to the overall alignment. We discuss variation in performance of the node2vec-NN model with α in more detail in Section 5.2.

5.1. Performance vs. k

As discussed in Algorithm 1 we employ two seed generation strategies. In this section, we compare the task performance of node2vec using mutual nearest neighbour seed alignments (Strategy 1) and top-k most semantically similar seed alignments (Strategy 2). We fix $\alpha = 0.2$ and $\theta = 0.80$ for the experiments. The results are illustrated in Table 4. In general, $k = 1$ leads to bad performance. This is understandable as only 1 seed alignment between the graphs is insufficient to learn meaningful representations. As can be seen, the F1-scores

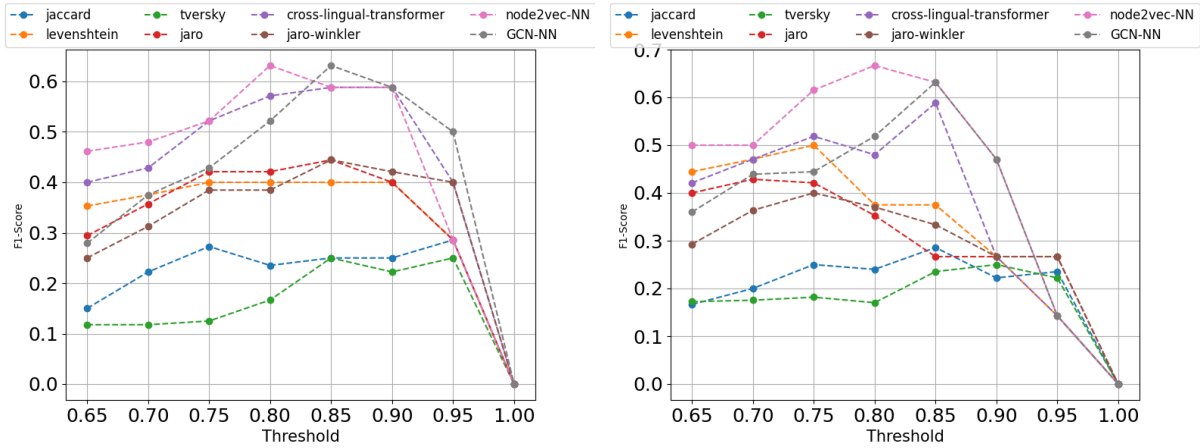


Figure 2: The variation in F1-score with change in threshold for the German-English dataset for conference-confOf pair (on the left) and the conference-sigkdd pair (on the right) with $\alpha = 0.2$.

cmt-confOf			
k	German-French	German-English	English-French
1	36.3	46.1	66.6
3	46.1	66.6	53.3
5	57.1	66.6	62.5
7	57.1	66.6	62.5
NN	66.6	71.4	58.8

conference-confOf			
k	German-French	German-English	English-French
1	28.5	58.8	28.5
3	58.8	58.8	39.9
5	55.5	66.6	47.0
7	55.5	63.1	44.4
NN	63.1	63.1	60.0

conference-sigkdd			
k	German-French	German-English	English-French
1	26.6	47.0	47.0
3	52.6	55.5	52.6
5	60.0	70.0	50.0
7	57.1	60.0	63.6
NN	57.1	66.6	60.8

Table 4: F1-scores of top-k semantically similar seed alignments where $k = 1, 3, 5, 7$ compared with mutual nearest neighbour (NN) alignments for node2vec model for $\theta = 0.80$ and $\alpha = 0.2$.

exhibit monotonic behaviour concerning the number of seed alignments in general i.e., increasing the number of alignments from 1 to 5 improves performance. However, in general $k = 7$ leads to degradation of performance as compared to $k = 5$. This can be attributed to additional noise introduced by a larger number of seed alignments. Hence, neither very low nor very high i.e., $k = 5$ is optimal for almost all datasets. In terms of the two strategies both are equally effective with nearest neighbour seed alignment outperforming $k = 5$ on 5 out of the 9 datasets.

5.2. Performance vs. α

To quantify variation in performance with changes in α we carry out experiments with varying α across different thresholds for node2vec with mutual nearest neighbour seed alignment. The results are illustrated in Figures 3, 4 and 5. As can be seen, almost all the ontology pairs and all the language pairs $\alpha = 0.2$ had the best F1-score overall. Interestingly, as the value of alpha went up the performance deteriorated with the lowest F1-scores recorded for $\alpha = 0.8$ for a given threshold. $\alpha = 0$ has good performance and for specific thresholds outperforms F1-scores achieved by using $\alpha = 0.2$. $\alpha = 0$ indicates only the

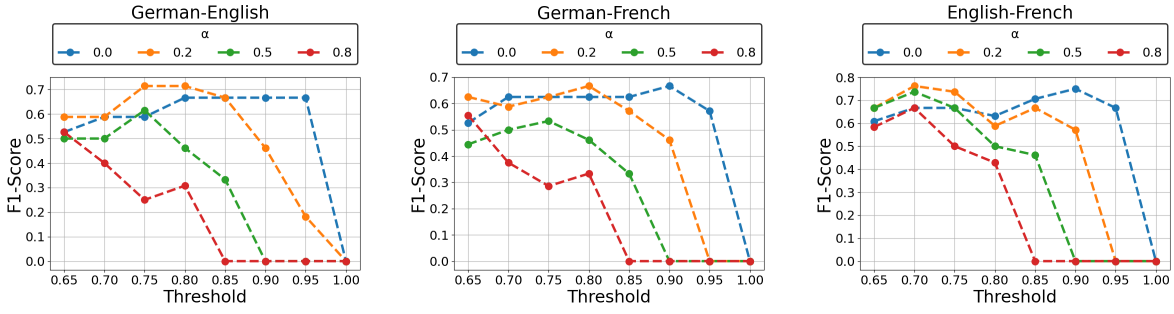


Figure 3: F1 vs. α : cmt-confOf

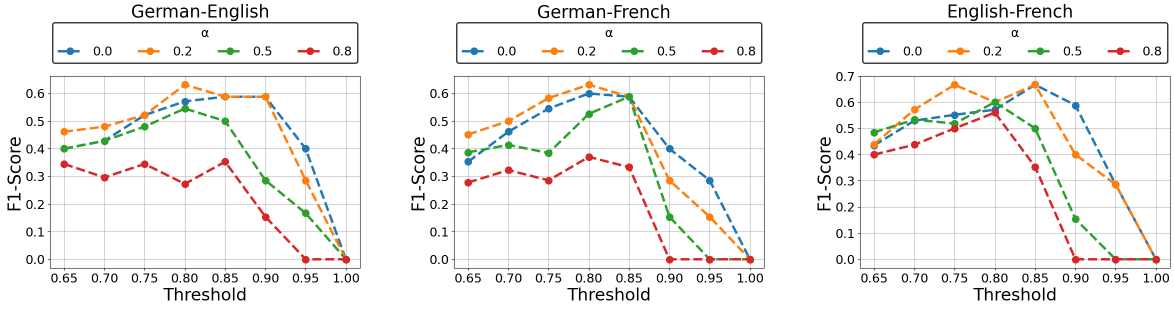


Figure 4: F1 vs. α : conference-confOf

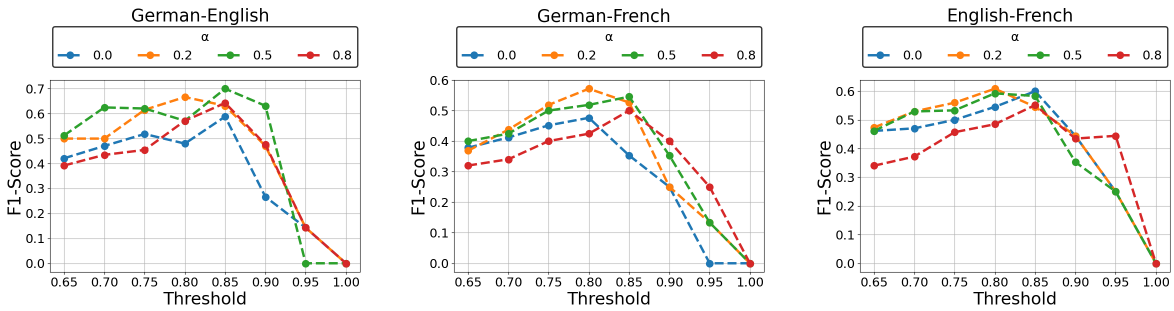


Figure 5: F1 vs. α : conference-sigkdd

semantic similarity of concept descriptions being used for ontology matching. These results suggest that while the choice is alpha is dependent on the similarity threshold being used, a value of 0.2 for α leads at threshold 0.80 leads to good results in general. Furthermore, the results demonstrate that while semantic similarity is the more important factor for ontology matching even outperforming the combined similarity for certain thresholds, the addition of structural similarity signals can lead to an improvement in task performance.

6. Conclusion

In this work, we proposed a new framework for ontology matching and evaluated it on 3 ontology pairs across 3 language pairs. The proposed framework takes into account semantic similarity between concept node descriptions in the source and target ontologies as well as the structural similarity calculated using embeddings that aggregate information

about node neighbourhood structure. We showed that our proposed system can outperform current state-of-the-art lexical similarity measures being used for CLOM. Furthermore, the results show that semantic similarity of concept node descriptions is the more important factor when aligning source and target nodes. We experiment with four structural embeddings, namely node2vec, TransE, RGCN, and GCN, and find that node2vec leads to better performance. It is also important to note that the performance of Levenshtein similarity is better than our proposed framework for German-English and English-French datasets of the cmt-confOf and conference-confOf ontology pairs respectively. Semantic similarity is used to generate seed alignments in the first stage of our approach and we explore two strategies for this purpose. Our analysis suggests that selecting top-k semantically similar concepts as seed alignments leads to better performance.

7. Limitations

Although we have shown good performance of our method for ontology matching as compared to lexical measures there are limitations worth discussing. We carry out our experiments using a fixed threshold however as discussed in Section 5, there is substantial variation in performance with changing thresholds. Choosing a threshold is associated with a trade-off between precision and recall. Manually fixing a threshold for different datasets is not optimal. Furthermore, we show that GCN is substantially outperformed by node2vec; there are more advanced alternatives such as Graph attention networks which can allow the nodes to only aggregate useful signals from their neighbours. We expect there to be an improvement in performance by using these algorithms. We show that using top-k semantically similar concepts as seed alignments is a better strategy for seed generation overall. However, the experiments do not establish an optimal value of k for all datasets. We hope to develop better seed generation strategies as a part of future work.

8. 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.

9. Bibliographical References

- Ghassan Beydoun, Graham Low, Quynh-Nhu Numi Tran, and Paul Bogg. 2011. [Development of a peer-to-peer information sharing system using ontologies](#). *Expert Syst. Appl.*, 38(8):9352–9364.
- Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. [Translating embeddings for modeling multi-relational data](#). In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2787–2795.
- Muhao Chen, Yingtao Tian, Mohan Yang, and Carlo Zaniolo. 2017. [Multilingual knowledge graph embeddings for cross-lingual knowledge alignment](#). In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 1511–1517. ijcai.org.
- Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.
- John Davies, Alistair Duke, and Audrius Stonkus. 2002. [Ontoshare: Using ontologies for knowledge sharing](#). In *Proceedings of the WWW2002 International Workshop on the Semantic Web, Hawaii, May 7, 2002*, volume 55 of *CEUR Workshop Proceedings*. CEUR-WS.org.
- Linda Elmhadi, Mohamed-Hedi Karray, Bernard Archimède, J. Neil Otte, and Barry Smith. 2021. [An ontological approach to enhancing information sharing in disaster response](#). *Information*, 12(10):432.
- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. [Language-agnostic BERT sentence embedding](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 878–891. Association for Computational Linguistics.
- Bo Fu, Rob Brennan, and Declan O’Sullivan. 2010. [Cross-lingual ontology mapping and its use on the multilingual semantic web](#). In *Proceedings of the 1st International Workshop on the Multilingual Semantic Web, Raleigh, North Carolina, USA, April 27th, 2010*, volume 571 of *CEUR Workshop Proceedings*, pages 13–20. CEUR-WS.org.
- Jorge Gracia and Kartik Asooja. 2013. Monolingual and cross-lingual ontology matching with CIDER-CL: evaluation report for OAEI 2013. In *Proceedings of the 8th International Conference on Ontology Matching-Volume 1111*, pages 109–116.
- Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864.
- Michael Gutmann and Aapo Hyvärinen. 2010. [Noise-contrastive estimation: A new estimation principle for unnormalized statistical models](#). In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010*, volume 9 of *JMLR Proceedings*, pages 297–304. JMLR.org.
- Yuan He, Jiaoyan Chen, Denvar Antonyrajah, and Ian Horrocks. 2022. [Bertmap: A bert-based ontology alignment system](#). In *Thirty-Sixth AAAI*

- Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelfth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022, pages 5684–5691. AAAI Press.
- Shimaa Ibrahim, Said Fathalla, Jens Lehmann, and Hajira Jabeen. 2020. Multilingual ontology merging using cross-lingual matching. In *2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, pages 113–120. IEEE.
- Shimaa Ibrahim, Said Fathalla, Jens Lehmann, and Hajira Jabeen. 2023. [Toward the Multilingual Semantic Web: Multilingual Ontology Matching and Assessment](#). *IEEE Access*, 11:8581–8599.
- Shimaa Ibrahim, Said Fathalla, Hamed Shariat Yazdi, Jens Lehmann, and Hajira Jabeen. 2019. From monolingual to multilingual ontologies: The role of cross-lingual ontology enrichment. In *Semantic Systems. The Power of AI and Knowledge Graphs: 15th International Conference, SEMANTICS 2019, Karlsruhe, Germany, September 9–12, 2019, Proceedings 15*, pages 215–230. Springer International Publishing.
- Vivek Iyer, Arvind Agarwal, and Harshit Kumar. 2020. [Veealign: a supervised deep learning approach to ontology alignment](#). In *Proceedings of the 15th International Workshop on Ontology Matching co-located with the 19th International Semantic Web Conference (ISWC 2020), Virtual conference (originally planned to be in Athens, Greece), November 2, 2020*, volume 2788 of *CEUR Workshop Proceedings*, pages 216–224. CEUR-WS.org.
- Paul Jaccard. 1901. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bull Soc Vaudoise Sci Nat*, 37:547–579.
- Matthew A Jaro. 1989. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association*, 84(406):414–420.
- Thomas N. Kipf and Max Welling. 2017. [Semi-supervised classification with graph convolutional networks](#). In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.
- Harold W Kuhn. 1955. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97.
- Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.
- Chengjiang Li, Yixin Cao, Lei Hou, Jiaxin Shi, Juanzi Li, and Tat-Seng Chua. 2019a. Semi-supervised Entity Alignment via Joint Knowledge Embedding Model and Cross-graph Model. 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 2723–2732.
- Weizhuo Li, Xuxiang Duan, Meng Wang, Xiaoping Zhang, and Guilin Qi. 2019b. [Multi-view embedding for biomedical ontology matching](#). In *Proceedings of the 14th International Workshop on Ontology Matching co-located with the 18th International Semantic Web Conference (ISWC 2019), Auckland, New Zealand, October 26, 2019*, volume 2536 of *CEUR Workshop Proceedings*, pages 13–24. CEUR-WS.org.
- Xiao Liu, Haoyun Hong, Xinghao Wang, Zeyi Chen, Evgeny Kharlamov, Yuxiao Dong, and Jie Tang. 2022a. [SelfKG: Self-Supervised Entity Alignment in Knowledge Graphs](#). In *Proceedings of the ACM Web Conference 2022, WWW '22*, pages 860–870, New York, NY, USA. Association for Computing Machinery.
- Xiao Liu, Haoyun Hong, Xinghao Wang, Zeyi Chen, Evgeny Kharlamov, Yuxiao Dong, and Jie Tang. 2022b. [Selfkg: Self-supervised entity alignment in knowledge graphs](#). In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, pages 860–870. ACM.
- Xin Mao, Wenting Wang, Yuanbin Wu, and Man Lan. 2021. [From alignment to assignment: Frustratingly simple unsupervised entity alignment](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 2843–2853. Association for Computational Linguistics.
- Christian Meilicke, Raúl García-Castro, Fred Freitas, Willem Robert van Hage, Elena Montiel-Ponsoda, Ryan Ribeiro de Azevedo, Heiner Stuckenschmidt, Ondřej Šváb Zamazal, Vojtěch Svátek, Andrei Taminin, Cássia Trojahn, and Shenghui Wang. 2012. [MultiFarm: A benchmark for multilingual ontology matching](#). *Journal of Web Semantics*, 15:62–68.

- Tadeusz Pankowski. 2023. Ontological databases with faceted queries. *The VLDB Journal*, 32(1):103–121.
- Michael Sejr Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. [Modeling relational data with graph convolutional networks](#). In *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*, volume 10843 of *Lecture Notes in Computer Science*, pages 593–607. Springer.
- Abhisek Sharma and Sarika Jain. 2023. [Lsmatch and Ismatch-multilingual results for OAEI 2023](#). In *Proceedings of the 18th International Workshop on Ontology Matching co-located with the 22nd International Semantic Web Conference (ISWC 2023), Athens, Greece, November 7, 2023*, volume 3591 of *CEUR Workshop Proceedings*, pages 159–163. CEUR-WS.org.
- Richard Sinkhorn. 1964. A Relationship Between Arbitrary Positive Matrices and Doubly Stochastic Matrices. *The Annals of Mathematical Statistics*, 35(2):876–879.
- Dennis Spohr, Laura Hollink, and Philipp Cimiano. 2011. A machine learning approach to multilingual and cross-lingual ontology matching. In *The Semantic Web—ISWC 2011: 10th International Semantic Web Conference, Bonn, Germany, October 23-27, 2011, Proceedings, Part I 10*, pages 665–680. Springer.
- Jianheng Tang, Kangfei Zhao, and Jia Li. 2023. [A fused gromov-wasserstein framework for unsupervised knowledge graph entity alignment](#). In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 3320–3334. Association for Computational Linguistics.
- Xiaobin Tang, Jing Zhang, Bo Chen, Yang Yang, Hong Chen, and Cuiping Li. 2020. [BERT-INT: A bert-based interaction model for knowledge graph alignment](#). In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3174–3180. ijcai.org.
- Amos Tversky. 1977. Features of similarity. *Psychological review*, 84(4):327.
- Titouan Vayer, Nicolas Courty, Romain Tavenard, Laetitia Chapel, and Rémi Flamary. 2019. [Optimal transport for structured data with application on graphs](#). In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 6275–6284. PMLR.
- Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. [Graph attention networks](#). In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.
- Yaoshu Wang, Jianbin Qin, and Wei Wang. 2017. Efficient approximate entity matching using jarowinkler distance. In *International conference on web information systems engineering*, pages 231–239. Springer.
- Yuting Wu, Xiao Liu, Yansong Feng, Zheng Wang, Rui Yan, and Dongyan Zhao. 2019. [Relation-aware entity alignment for heterogeneous knowledge graphs](#). In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5278–5284. ijcai.org.
- Ondrej Zamazal and Vladimír Svátek. 2017. The Ten-Year OntoFarm and its Fertilization within the Onto-Sphere. In *Web Semantics: Science, Services and Agents on the World Wide Web*, volume 43, pages 46–53. Elsevier.
- Weixin Zeng, Xiang Zhao, Jiuyang Tang, Xinyi Li, Minnan Luo, and Qinghua Zheng. 2021. [Towards entity alignment in the open world: An unsupervised approach](#). In *Database Systems for Advanced Applications - 26th International Conference, DASFAA 2021, Taipei, Taiwan, April 11-14, 2021, Proceedings, Part I*, volume 12681 of *Lecture Notes in Computer Science*, pages 272–289. Springer.
- Fernando Zhapa-Camacho and Robert Hoehndorf. 2023. [From axioms over graphs to vectors, and back again: Evaluating the properties of graph-based ontology embeddings](#). In *Proceedings of the 17th International Workshop on Neural-Symbolic Learning and Reasoning, La Certosa di Pontignano, Siena, Italy, July 3-5, 2023*, volume 3432 of *CEUR Workshop Proceedings*, pages 85–102. CEUR-WS.org.

Querying the *Lexicon der indogermanischen Verben* in the LiLa Knowledge Base: Two Use Cases

Valeria Irene Boano, Marco Passarotti and Riccardo Ginevra

Università Cattolica del Sacro Cuore

Milan, Italy

valeriairene.boano01@icatt.it, marco.passarotti@unicatt.it, riccardo.ginevra@unicatt.it

Abstract

This paper presents two use cases of the etymological data provided by the *Lexicon der indogermanischen Verben* (LIV) after their publication as Linked Open Data and their linking to the LiLa Knowledge Base (KB) of interoperable linguistic resources for Latin. The first part of the paper briefly describes the LiLa KB and its structure. Then, the LIV and the information it contains are introduced, followed by a short description of the ontologies and the extensions used for modelling the LIV's data and interlinking them to the LiLa ecosystem. The last section details the two use cases. The first case concerns the inflection types of the Latin verbs that reflect Proto-Indo-European stems, while the second one focusses on the Latin derivatives of the inherited stems. The results of the investigations are put in relation to current research topics in Historical Linguistics, demonstrating their relevance to the discipline.

Keywords: Linked Open Data, Latin, Indo-European

1. Introduction

In recent years, the Linked Open Data (LOD) paradigm has been increasingly applied to linguistic (meta)data to achieve their interoperability, leading to a constant growth of the Linguistic Linked Open Data Cloud.¹ Linguistic resources that are part of the cloud include textual corpora, lexicons, dictionaries and more. Among the resources interlinked in the Cloud are those published in the LiLa Knowledge Base (KB),² which contains several textual and lexical resources for the Latin language, published as LOD and linked with each other.

With regard to textual resources, LiLa includes so far more than 3,5M words from several Latin corpora, in both Classical Latin and Medieval Latin. Among them, are the corpus of Classical texts LASLA³ (CIRCSE, 2022; Fantoli et al., 2022), the *Index Thomisticus* treebank (CIRCSE, 2006-2024; Cecchini et al., 2018), which contains works of Thomas Aquinas, and *UDante* (CIRCSE, 2021b; Cecchini et al., 2020), which is a Universal Dependencies⁴ treebank for Dante Alighieri's Latin works. As for the lexical resources, LiLa currently includes, among others, the Lewis and Short Latin-English dictionary (CIRCSE, 2021a; Lewis and Short, 1879), the derivational lexicon *Word Formation Latin* (CIRCSE, 2018; Litta et al., 2019), and a resource of morphological principal parts of

Latin words, *PrinParLat* (CIRCSE, 2023b; Pellegrini, 2023).

Moreover, LiLa interlinks etymological information from two reference dictionaries: the *Etymological dictionary of Latin and other Italic Languages* (CIRCSE, 2020a; de Vaan, 2008), which focusses on Latin and other Italic languages, and the *Lexicon der indogermanischen Verben* (LIV) (CIRCSE, 2023a; Rix, ed., 2001), which features reconstructed Proto-Indo-European (PIE) verbal roots and details their developments in the attested Indo-European (IE) daughter languages, including Latin. The etymological relations between Latin words and their ancestors in PIE provided by the latter have been recently linked to LiLa (Boano et al., 2023): their integration in the KB helps to put the information contained in the LIV in relation to the one provided by other linguistic resources. To achieve this in the (recent) past different linguistic resources were consulted one at time, and their data were integrated later. Now, thanks to the interoperability among resources made possible by LiLa, this same process can be achieved automatically and it is made fully replicable.

This paper aims to show the advantages that linking the LIV to LiLa provides in approaching two research questions of Historical Linguistics. After introducing the overall architecture of the LiLa KB (Section 2) and the process performed to interlink the LIV into LiLa (Section 3), in Section 4 the paper details the two use cases, showing how the LIV's information can be queried and exploited in the LiLa KB to address the research questions concerned.

¹<https://linguistic-lod.org/>.

²<https://lila-erc.eu/>.

³https://www.lasla.uliege.be/cms/c_8508894/fr/lasla.

⁴<https://universaldependencies.org/>.

2. The LiLa Knowledge Base

The LiLa KB provides FAIR linguistic resources (Wilkinson et al., 2016) published as LOD. The syntactic interoperability between the resources of the KB is ensured by the use of the Resource Description Framework (RDF) data model (Lassila and Swick, 1998). The semantic interoperability (Ide and Pustejovsky, 2010) instead is achieved by the use of a few vocabularies widely used for the publication of linguistic resources as LOD, including the Ontolex Lemon model,⁵ the OLiA ontology (Chiarcos and Sukhareva, 2015) and the Ontolex lexicography module.⁶ The connection between the resources interlinked in the KB is achieved via the so-called Lemma Bank (LB) (CIRCSE, 2019-2024).⁷ The LB is a set of more than 200k lemmas, which was originally created from the database of the morphological analyser LEMLAT (Passarotti et al., 2017), and which is constantly extended whenever a new linguistic resource requires a new lemma to be included in the KB.

The LB constitutes LiLa’s core structure and the crossroads between all the resources part of the KB. Interoperability is achieved by linking tokens provided by textual corpora and entries in lexical resources to their corresponding lemma in the LB.

Whenever possible, lemmas, tokens and lexical entries are represented and published as LOD by means of classes and properties from the Ontolex Lemon core module. Each `ontolex:LexicalEntry`⁸ of each lexical resource is linked via the property `ontolex:canonicalForm`⁹ to the corresponding `lila:Lemma`¹⁰ in the LB. A `lila:Lemma` is a subclass of the class `ontolex:Form`,¹¹ namely a word’s citation form. The simple link established between a `lila:Lemma` and the corresponding `ontolex:LexicalEntry` ensures the interoperability between the lexical resources part of LiLa. As for the tokens of the corpora interlinked in LiLa, they are connected to the LB via the property `lila:hasLemma`.¹²

The `lila:Lemma` also carries morphological information, such as the gender and the inflection

⁵<https://www.w3.org/2016/05/ontolex/>.

⁶<https://www.w3.org/2019/09/lexicog/>.

⁷<http://lila-erc.eu/data/id/lemma/LemmaBank>.

⁸<http://www.w3.org/ns/lemon/ontolex#LexicalEntry>.

⁹<http://www.w3.org/ns/lemon/ontolex#canonicalForm>.

¹⁰<http://lila-erc.eu/ontologies/lila/Lemma>.

¹¹<http://www.w3.org/ns/lemon/ontolex#Form>.

¹²<https://lila-erc.eu/ontologies/lila/hasLemma>.

type. Some lemmas are also assigned derivational information about prefixes, suffixes and lexical bases: at the time of writing, the derivational information recorded in the LiLa LB regards Classical Latin words only, while the coverage for the Medieval Latin is significantly lower (Pellegriani et al., 2022).

The LiLa KB can be queried via a SPARQL endpoint,¹³ via a user-friendly interface¹⁴ and via an interactive search platform.¹⁵

3. The LIV and its Modelling

Etymology can be broadly defined as “the branch of linguistics which deals with determining the origin of words and the historical development of their form and meanings” (OED, s.v. *etymology*, *n.*). The LIV is the reference etymological dictionary for verbs attested in the ancient IE languages. It was curated by Helmut Rix and first published in 1998 by Reichert Verlag (Rix, ed., 1998). A second edition appeared in 2001, with the additions and corrections by Martin Kümmel and Helmut Rix (Rix, ed., 2001). This dictionary contains information regarding the PIE verb and its development in the IE languages: it details the etymology of verbs attested in IE languages by tracing them back to reconstructed PIE verbs. In particular, the LIV contains three main types of lexical items:

- Reconstructed PIE verbal roots. They constitute the entries of the dictionary, and are provided with their phonological structure and broad lexical meaning. A verbal root is the part of a word that “carries the core of the meaning, the idea of a situation, which is recognisable in all forms derived from the root” (Rix, ed., 2001, p. 5, my translation).
- Reconstructed PIE verbal stems. They consist of the verbal root processed with affixes and they encode aspectual information.
- Word forms attested in IE languages. The LIV lists word forms for several IE languages: they can be traced back to the corresponding PIE stems and are provided with their attested meaning.

As by agreement with the LIV’s publisher, only the relations established between these elements were modelled and linked to LiLa. For the modelling, we decided to use the `lemonEty` extension of the Ontolex Lemon model (Khan, 2018), which was developed precisely for representing etymological information. `lemonEty` provides three key classes:

¹³<https://lila-erc.eu/sparql/>.

¹⁴<https://lila-erc.eu/query/>.

¹⁵<https://lila-erc.eu/LiLaLisp/>.



Figure 1: The model of the LIV etymological relations, with respect to the verb *glubo*.

- `lemonEty:Etymon`:¹⁶ this class is a subclass of `ontolex:LexicalEntry` and contains all the lexical items of the source language that are introduced to explain the etymology of the target language;
- `lemonEty:Etymology`:¹⁷ this class “reifies the whole process of etymological reconstruction as scientific hypothesis” (Passarotti et al., 2020, p. 22);
- `lemonEty:EtyLink`:¹⁸ this class is used to connect linguistic items from the source language to the target language.

We modelled the PIE roots provided by the LIV (e.g. PIE **h₃emh₃-*, underlying the Latin verb *amo* ‘to love’) as instances of the class `lemonEty:Etymon`, since they are items of the source language (in this case, PIE) and are introduced “to describe the origin and his-

tory of another Lexical Entry”.¹⁹ Moreover, since `lemonEty:Etymon` is a subclass of `ontolex:LexicalEntry`, this allowed us to preserve the structure of the LIV, which treats the roots as lexical entries.

The `lemonEty:EtyLink` class was used to model the relation between a PIE stem and its corresponding Latin stem. The LIV provides in fact the Latin first-person present and first-person perfect word forms, which are traditionally used to represent all the forms derived from the present and the perfect stems. For this reason, we were able to include the Latin stems as part of the model: in particular, we reused the individuals of the class `Stem`²⁰ provided by *PrinParLat* (CIRCSE, 2023b), which is a collection of principal parts of Latin morphological paradigms already interlinked in the LiLa KB. When the Ontolex Morph module (Chiarcos et al., 2022) will be released, the PIE stems, instead, will be represented as instances of the class `morph:Morph`.²¹ this class is used to repre-

¹⁶<http://lari-datasets.ilc.cnr.it/lemonEty#Etymon..>

¹⁷<http://lari-datasets.ilc.cnr.it/lemonEty#Etymology.>

¹⁸<http://lari-datasets.ilc.cnr.it/lemonEty#EtyLink.>

¹⁹<http://lari-datasets.ilc.cnr.it/lemonEty.>

²⁰<https://lila-erc.eu/lodview/ontologies/prinparlat/Stem.>

²¹At the time of writing, the URIs provisionally point to the Morph’s GitHub page (<https://github.com/>

sent all those elements of morphological analysis which are below the word level.

Each PIE stem is also linked to the class `prinparlat:StemType`:²² new individuals were added to this class in order to include all the PIE stem types provided by the LIV. The latter are: present, aorist, perfect, causative, iterative, causative-iterative, desiderative, intensive, fientive and essive. Each of these categories expresses a specific grammatical or lexical aspect.

Both PIE and Latin stems were connected to the `lemonEty:EtymLink` via the properties `lemonEty:etySource` and `lemonEty:etyTarget`, respectively. Each Latin stem was also linked to the corresponding Latin form. For the perfect, we reused the form provided by *PrinParLat*. For the present, instead, we generated it from scratch, since *PrinParLat* supplies only the third-person present form.

Finally, the `lemonEty:Etymology` class stands as a central crossroads between all the LIV lexical items: it reifies the generic etymological relation between the Latin `ontolex:LexicalEntry` and the PIE root, while also being linked to the two etymological links.

Figure 1 shows the model applied to the case of the verb *glubo* ‘to peel’. On the left side is the lexical entry *glubo*, linked to the LiLa lemma via the property `ontolex:canonicalForm`. The two *PrinParLat* Latin stems are linked to the lexical entry via the property `vartrans:lexicalRel`.²³ The Latin stems are the starting point of two connections: one with the Latin forms, the present *glubo* and the perfect *glupsi*,²⁴ and the other with the two etymological links.²⁵ These reify

`ontolex/morph`).

²²<https://lila-erc.eu/lodview/ontologies/prinparlat/StemType>.

²³<http://www.w3.org/ns/lemon/vartrans#lexicalRel>.

²⁴The perfect form, provided by *PrinParLat* is linked via the property `morph:consistsOf` (<https://ontolex.github.io/morph/consistsOf>), while the present form, created from scratch, is linked via the property `ontolex:lexicalForm`. This does not constitute an inconsistency, rather it is a choice imposed by economic reasons: in fact, whenever a relation is already expressed by a property (in this case `consistsOf`), it is not necessary to represent it again with another one (`lexicalForm`), since this would result in redundancy.

²⁵The etymological links are connected with the source element and the target element via the properties `lemonEty:etySource` (<http://lari-datasets.ilc.cnr.it/lemonEty#etySource>) and `lemonEty:etyTarget` (<http://lari-datasets.ilc.cnr.it/lemonEty#etyTarget>), respectively.

the etymological relation between the Latin stems and the corresponding PIE stems (**g/ǵléub^h-/g/ǵlub^h-*, underlying the Latin present stem, and **g/ǵléub^h/g/ǵléub^h-s-*, underlying the Latin perfect stem), which are displayed on the right side of the picture. The PIE root (**g/ǵléub^h-*) is linked to both of them.²⁶ In the central part of the graph is the `lemonEty:Etymology` class, connected with the lexical entry, the PIE root and the two etymological links.²⁷

4. Case Studies

Thanks to the creation of a total of 385 lexical entries and to their linking to the LB, the etymological information provided by the LIV was included in LiLa. The integration of the LIV’s information into LiLa allows to put it in relation to that provided by the other resources that are part of the KB. The RDF data can be queried by means of the SPARQL query language.²⁸

Querying the LIV in LiLa allows to enhance the quality of research in the field, by providing new insights about the relations of attested Latin word forms with reconstructed PIE roots and stems. This section illustrates two case studies made possible by the interoperability of the LIV with other resources in LiLa: the first use case regards the inflection types of the Latin verbs inherited from PIE, while the second one investigates the derivatives of PIE stems in Classical Latin.

4.1. An Investigation about the Lemmas’ Inflection Types

When investigating the etymological relationship holding between Latin verbs and their ancestors in PIE, a question that emerges regards their inflection type. In particular, some inflection types seem to be more common among Latin verbs that are inherited from PIE, and less common among verbs that cannot be traced back to PIE stems (Weiss, 2020). The linking of the LIV to the LB can be effectively exploited to answer this question, as it

²⁶The property that links the PIE stems and the PIE root is `vartrans:lexicalRel`, mirroring the relationship between the lexical entry and the Latin stems.

²⁷The `lemonEty` ontology defines specific properties to link the `Etymology` to these elements, namely `lemonEty:etymology` (<http://lari-datasets.ilc.cnr.it/lemonEty#etymology>), `lemonEty:etymon` (<http://lari-datasets.ilc.cnr.it/lemonEty#etymon>) and `lemonEty:hasEtyLink` (<http://lari-datasets.ilc.cnr.it/lemonEty#hasEtyLink>).

²⁸The SPARQL queries performed to obtain the results presented in this paper can be found at <https://github.com/CIRCSE/SPARQL-queries-LIV>.

allows to quickly and easily identify the inflection type of each Latin verb listed in the LIV. More precisely, the LB records information about the inflection type of each lemma, represented using the property `lila:hasInflectionType`.²⁹ By counting the number of lemmas for each verbal inflection type, it is possible to compare the predominant inflection types of the entire LB (table 1) with those of the Latin verbs listed in the LIV (table 2).

Inflection Type Label	Number of lemmas
First conjugation	9530
Third conjugation	3398
First conjugation deponent	1019
Fourth conjugation	922
Second conjugation	823

Table 1: The inflection types of the LiLa LB.

Inflection Type Label	Number of lemmas
Third conjugation	172
Second conjugation	80
First conjugation	28
Fourth conjugation	19
Third conjugation deponent	16

Table 2: The inflection types of the Latin lexical entries in LIV.

As Tables 1 and 2 show, the distributions of the inflection types in the LB and in the LIV are very different. In the LB, the first conjugation is predominant, and only around one third of all verbs belong to the third conjugation. Among the Latin lexical entries in the LIV, however, the proportion is reversed: the number of third conjugation verbs is six times higher than that of first conjugation verbs. These data quantitatively confirm what is stated in Michael Weiss standard work, the *Outline of the historical and comparative grammar of Latin* (Weiss, 2020): “the 3rd and 4th Conjugations [...] are the main repository of present stem formations inherited from Proto-Indo-European” (p. 404).

Since the information regarding the various PIE stem types was included in the modelling of the LIV (as described in Section 3), it is possible to refine the SPARQL query, and consequently to extend the investigation, by taking this information into account. In particular, for each inflection type, it is possible to count the number of lemmas reflecting a certain PIE stem type. The *Outline of the historical and comparative grammar of Latin* (Weiss,

²⁹<http://lila-erc.eu/ontologies/lila/hasInflectionType>.

2020) gives detailed information about the sources of each Latin conjugation. For instance, PIE so-called causative-iterative and iterative stems are usually reflected in Latin by second conjugation verbs (p. 403). By restricting the results of the query to the lemmas derived from a determined stem type, it is possible to quantitatively confirm this statement.

Inflection Type Label	Number of lemmas
Second conjugation verb	5
First conjugation deponent verb	1
First conjugation verb	1

Table 3: The inflection types of the Latin reflexes of LIV causative-iterative stems.

Inflection Type Label	Number of lemmas
Second conjugation verb	13
First conjugation verb	6
Third conjugation verb	2
First conjugation deponent verb	1

Table 4: The inflection types of the Latin reflexes of LIV iterative stems.

Tables 3 and 4 show the results of the queries for the causative-iterative stems and for the iterative stems respectively. As expected, the second conjugation is predominant in both cases. These results can be considered statistically significant, since the p-value, indicating the inter-dependence between the inflection type and the PIE stem type, was calculated to be lower than 0.05. The queries performed to obtain these results are simple, but give an empirical confirmation of what is stated in the *Outline of the historical and comparative grammar of Latin*.

4.2. An Investigation about the PIE Stem Types and their Derivatives

A further research question relevant to Historical Linguistics that may be investigated with the aid of the LIV’s linking in LiLa, is whether the derivatives of Latin verbs that may be traced back to PIE stems feature specific affixes depending on their underlying PIE stem type. The derivational information that is recorded in the LiLa KB can be exploited to answer this question, too. Indeed, the lemmas of the LB are linked via the properties `lila:hasBase`,³⁰ `lila:hasPrefix`³¹

³⁰<http://lila-erc.eu/ontologies/lila/hasBase>.

³¹<http://lila-erc.eu/ontologies/lila/hasPrefix>.

and `lila:hasSuffix`³² to their derivational bases, their prefixes and their affixes, respectively. By putting this information in relation to the LIV data, it is possible to answer the question previously outlined.³³

The query counts the number of LiLa lemmas that are derived by means of a specific Latin affix and whose Latin base may be traced back to a specific PIE stem type. What emerges from the results is that some of the most frequent prefixes and suffixes in the entire LB are also the most frequent in the LIV derivatives. However, some affixes that are not in the top five ranking of the LB appear in the first five positions for the derivatives of certain PIE stems.

Prefix Label	Number of lemmas
con-	1992
e(x)-	1438
in (negation)-	1346
de-	1146
in (entering)-	1131

Table 5: The five most frequent prefixes in the Classical Latin lemmas of the LB.

Prefix Label	Number of lemmas
in (negation)-	12
prae-	10
in (entering)-	10
pro-	10
con-	7

Table 6: The five most frequent prefixes in the Classical Latin lemmas reflecting PIE desiderative stems.

Prefix Label	Number of lemmas
con-	24
ad-	21
e(x)-	19
re-	14
pro-	13

Table 7: The five most frequent prefixes in the Classical Latin lemmas reflecting PIE fientive stems.

³²<http://lila-erc.eu/ontologies/lila/hasSuffix>.

³³As described in Section 2, the derivational information recorded in LiLa currently regards only a subset of the Medieval Latin lemmas: for this reason, the results of all the queries including derivational information were restricted to Classical Latin lemmas only.

For instance, the prefix *pro-* is not among the most frequent of the LB (Table 5), but it is the fourth most frequent prefix for the derivatives of Latin lemmas reflecting PIE desiderative stems (Table 6) and in fifth position for the derivatives of Latin lemmas reflecting PIE fientive stems (Table 7). With regard to the suffixes, an interesting example of the same phenomenon is *-ment*, which is not among the top five suffixes of the LB (Table 8), but is in the fourth position for the derivatives of Latin reflexes of PIE fientive stems (Table 9). These data point to a close association between Latin affixes and specific PIE stem types. Thus, the queries performed open new perspectives about the relation between the Latin affixes and the derivatives of the PIE stems, suggesting that the semantic meaning carried by the stem influenced the choice of the affix involved in the derivational process. Indeed, each PIE stem type originally encoded a specific grammatical or lexical aspect, that is, they expressed the duration or the manner of the action (Meier-Brügger, 2003, pp. 164 ff.), as do the various prefixes and suffixes used in Latin to derive new words. This hypothesis may be further investigated since these results can be considered statistically significant: indeed, the p-value indicating the inter-dependence between the affixes involved in the derivational process and the PIE stem type underlying the lemma was calculated to be lower than 0.05.

Suffix Label	Number of lemmas
-(t)io(n)	2961
-(t)or	1837
-ari	1449
-(i)t	1381
-i	1258

Table 8: The five most frequent suffixes in the Classical Latin lemmas of the LB.

Suffix Label	Number of lemmas
-sc	63
-id	30
-ul	18
-ment	17
-(i)t	17

Table 9: The five most frequent suffixes in the Classical Latin lemmas reflecting PIE fientive stems.

To delve more into the matter, it is possible to calculate the percentage of the presence of each affix in the derivatives of Latin lemmas reflecting PIE stems compared to its total occurrences in the LB. The results show that a good part of the Classical Latin derivatives may be traced back to PIE

present stems: more precisely, with regard to both prefixes and suffixes, the percentage of derivatives that can ultimately be traced back to a PIE present stem often exceeds the threshold of 50%. As an example, table 10 shows the first five results for the suffixes involved in the derivational processes concerning Latin reflexes of PIE present stems.

Suffix label	Number of lemmas	Percentage
-(i)t	922	66,76%
-(i)es	80	57,55%
-(t)ur	127	55,22%
-or	92	55,09%
-men/min	171	50,89%

Table 10: The suffixes of Latin derivatives reflecting PIE present stems and their percentage on the total.

Suffix label	Number of lemmas	Percentage
-id	125	34,92%
-sc	103	15,37%
-(i)t	68	4,92%
-i	57	4,53%
-(t)io(n)	68	2,30%

Table 11: The suffixes of the derivatives descending from a PIE essive stem and their percentage on the total.

On the other hand, for the other PIE stem types, the situation is different: they usually cover less than 10% of the derivatives formed with a specific affix, and sometimes their percentages do not even reach the frequency threshold.³⁴ However, there is one outstanding result: the 34,92% of the LB's lemmas containing the suffix *-id* is derived from a Latin reflex of a PIE essive stem. This suffix is over two times more frequent than the second-ranked one, *-sc*, pointing to a special relation between the Latin adjectives in *-idus* and the Latin reflexes of PIE essive stems.

This special relation may be understood within the context of the so-called Caland system (Rau, 2009; Nussbaum, 1999). This PIE system consisted in a set of formal and semantic relationships between words, based on the alternation of specific affixes. The words involved were not derivatives of each other: rather “the word formation process is called recategorisation, i.e. the part of speech changes, but not the semantic content” (Balles, 2003, p. 10, my translation). The system

³⁴The frequency threshold was set on the 1% of the total occurrences of the most common affix in the Classical Latin lemmas of the LB.

has been inherited by many IE languages, including Latin. In the latter, however, it was remodelled following language-specific linguistic patterns. In particular, a Latin set of Caland formations usually features an adjective (e.g. *calidus* ‘hot’ or *liquidus* ‘fluid’), a noun (e.g. *calor*, *-ōris* ‘heat’ or *liquor*, *-ōris* ‘fluidity’), an essive verb (e.g. *caleō*, *-ēre* ‘to be hot’ or *liqueō*, *-ēre* ‘to be fluid’), an inchoative verb (e.g. *calēscō*, *-ere* ‘to become hot’ or *liquēscō*, *-ere* ‘to become fluid’) and a factitive verb (e.g. *calefaciō*, *-ere* ‘to make hot’ or *liquefaciō*, *-ere* ‘to make liquid’). These *-idus* adjectives and essive verbs, which have long been recognized as part of the Caland system in Latin, exactly correspond to the derivatives in *-id* and the Latin reflexes of PIE essive stems identified thanks to the LIV's linking to the LiLa KB. The results of the queries thus quantitatively confirm a relation which has long been noted and discussed within Historical Linguistics: this suggests that other results may also provide relevant information, which may be used to demonstrate new substantial relationships between the Latin affixes and the PIE stems.

5. Conclusion and Future Work

The linking of the LIV to the LiLa KB provides new opportunities to explore its etymological data in relation to Latin. The queries and the results shown in this paper confirm that the etymological information included in the LiLa KB can be effectively exploited to acquire new information about the relationship between Latin and PIE lexical items. The queries discussed here could not have been performed without the LIV's linking to the LiLa KB. Thus, the publication of the LIV's etymological relationships as LOD increases the research possibilities in the field, while representing an enhancement of the etymological subset of LiLa and of the LLOD Cloud.

Indeed, the queries and the results discussed in the present paper exemplify only a few of the advantages that the LIV's linking may actually provide. LiLa contains resources that supply information with regard to syntax (e.g. *Latin Vallex 2.0* (CIRCSE, 2020c; Mambrini et al., 2021), morphology (e.g. *PrinParLat* (CIRCSE, 2023b)), semantics (e.g. the Lewis and Short dictionary (CIRCSE, 2021a)) and sentiment analysis (e.g. *LatinAffectus* (CIRCSE, 2020b; Sprugnoli et al., 2020), while also providing different textual corpora, both for Classical and for Medieval Latin. All these layers of information are interoperable with each other and with the LIV. Querying their interconnected data can have a concrete impact on the academic communities of Classicists and Historical Linguists, by allowing them to carry out investigations that were not possible before.

Moreover, two future challenges can be outlined. First, the LIV does not exclusively contain etymological information with regard to Latin, but actually details the etymology of lexical items in many other IE languages: by modelling their data with the same ontologies that we used, it will be possible to enlarge the etymological network and investigate the etymological relationships between several IE languages.

Secondly, the biggest challenge not only for the LIV and LiLa, but for all the linguistic resources published as LOD, will be their integration within the world of the so-called Big Data and Large Language Models (LLMs). LLMs (such as BERT (Devlin et al., 2018) or ChatGpt (Ouyang et al., 2022)) are the future of Computational Linguistics, since they can process huge amounts of raw text, without the need to learn patterns provided by previous annotations. They can achieve very good results in several tasks, such as question answering (Jiang et al., 2021), machine translation (Lewis et al., 2020) and text generation (Li et al., 2022). In this context, the future of annotated linguistic resources is uncertain, given that they may stop being necessary altogether. However, the shift from supervised models to unsupervised machine learning methods constitutes a radical change that cannot be faced without critical thinking: if no annotation is required, the linguist's expertise and the deep analysis of the linguistic data are not required either. The challenge will thus be to preserve the original analytical component of Computational Linguistics, while taking all the benefits that LLMs can offer. In particular, this can be achieved by incorporating the linguistic resources published as LOD into LLMs. The LOD resources are stored in the form of knowledge graphs (KGs): these are able to generate interpretable results and to perform symbolic reasoning (Zhang et al., 2021), thus providing a solution for some of the limitations of LLMs (Biever, 2023). In this view, the linguistic resources published as LOD will hopefully preserve their crucial and innovative role in the discipline by establishing a fruitful relationship with the LLMs: in fact, the quality of the structured data contained in these resources can be reused to fine-tune and provide external knowledge to the LLMs (Zhang et al., 2019; Liu et al., 2021), while also being useful to analyse their results and provide interpretability (Petroni et al., 2019). This will hopefully constitute an opportunity to enhance the LLMs' performance and continue to improve the machine's capabilities with human knowledge.

6. Acknowledgements

The "LiLa - Linking Latin" 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.

7. Bibliographical References

- Irene Balles. 2003. Die lateinischen *idus*- Adjektive und das Calandsystem. *Indogermanisches Nomen. Derivation, Flexion und Ablaut. Akten der Arbeitstagung der indogermanischen Gesellschaft, Freiburg (2001)*, pages 8–29.
- Celeste Biever. 2023. Chatgpt broke the turing test-the race is on for new ways to assess ai. *Nature*, 619(7971):686–689.
- Valeria Irene Boano, Francesco Mambrini, Marco Carlo Passarotti, and Riccardo Ginevra. 2023. [Modelling and publishing the *Lexicon der indogermanischen Verben* as linked open data](#). In *Proceedings of the Ninth Italian Conference on Computational Linguistics*.
- 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. 2020. [UDante: First Steps Towards the Universal Dependencies Treebank of Dante's Latin Works](#). In *Seventh Italian Conference on Computational Linguistics (CLiC-it 2020)*, Bologna.
- Christian Chiarcos, Katerina Gkirtzou, Fahad Khan, Penny Labropoulou, Marco Passarotti, and Matteo Pellegrini. 2022. [Computational morphology with ontalex-morph](#). *Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference, 20-25 June 2022, Marseille, France*.
- Christian Chiarcos and Maria Sukhareva. 2015. [Olia – ontologies of linguistic annotation](#). *Semantic Web*, 6:379–386.
- Michiel de Vaan. 2008. *Etymological Dictionary of Latin and the other Italic Languages*. Brill, Leiden and Boston.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. [Bert: Pre-training of deep bidirectional transformers for language understanding](#). *CoRR*.

- Margherita Fantoli, Marco Passarotti, Francesco Mambrini, Giovanni Moretti, and Paolo Ruffolo. 2022. [Linking the LASLA Corpus in the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin](#). In *Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference*, pages 26–34. European Language Resources Association.
- 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.
- Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. [How can we know when language models know? on the calibration of language models for question answering](#). *Transactions of the Association for Computational Linguistics*, 9:962–977.
- Anas Fahad Khan. 2018. [Towards the representation of etymological data on the semantic web](#). *Information*, 9(304).
- Ora Lassila and Ralph R. Swick. 1998. [Resource Description Framework \(RDF\) Model and Syntax Specification](#).
- Charlton T. Lewis and Charles Short. 1879. *A Latin Dictionary. Founded on Andrews' edition of Freund's Latin dictionary*. Clarendon Press, Oxford.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2020. [Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, page 7871–7880.
- Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2022. [Pretrained language models for text generation: A survey](#). *arXiv preprint arXiv:2201.05273*.
- Eleonora Litta, Marco Passarotti, and Francesco Mambrini. 2019. [The Treatment of Word Formation in the LiLa Knowledge Base of Linguistic Resources for Latin](#). In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2019)*. 19-20 September 2019, Prague, Czechia, pages 35–43, Prague, Czech Republic. Institute of Formal and Applied Linguistics, Charles University in Prague.
- Ye Liu, Yao Wan, Lifang He, Hao Peng, and S Yu Philip. 2021. [Kg-bart: Knowledge graph-augmented bart for generative common-sense reasoning](#). In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 6418–6425.
- Francesco Mambrini, Marco Passarotti, Eleonora Litta, and Giovanni Moretti. 2021. [Interlinking Valency Frames and WordNet Synsets in the LiLa Knowledge Base of Linguistic Resources for Latin](#). In *Further with Knowledge Graphs. Studies on the Semantic Web 53*, Amsterdam. IOS Press.
- Michael Meier-Brügger. 2003. *Indo-European linguistics*. De Gruyter, Berlin; New York.
- Alan Nussbaum. 1999. [*Jocidus: an account of the latin adjectives in -idus](#). *Compositiones indogermanicae: in memoriam Jochem Schindler*, pages 377–419.
- OED. *Oxford English dictionary online*. Oxford University Press, Oxford.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. [Training language models to follow instructions with human feedback](#). *Advances in Neural Information Processing Systems*, 35:27730–27744.
- 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*, pages 24–31.
- 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(1):177–212.
- Matteo Pellegrini. 2023. [Flexemes in theory and in practice](#). *Morphology*, pages 1–35.
- Matteo Pellegrini, Marco Passarotti, Eleonora Litta, Francesco Mambrini, Giovanni Moretti, Claudia Corbetta, and Martina Verdelli. 2022. [Enhancing derivational information on latin lemmas in the lila knowledge base. a structural and diachronic extension](#). *The Prague Bulletin of Mathematical Linguistics*, (119):67–92.
- 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*.
- Jeremy Rau. 2009. *Indo-European nominal morphology: The decads and the Caland system*. Innsbrucker Beiträge zur Sprachwiss, Innsbruck.
- Helmut Rix, ed. 1998. *LIV. Lexikon der indogermanischen Verben. Die Wurzeln und ihre Primärstammbildungen*. Reichert Verlag, Wiesbaden.
- Helmut Rix, ed. 2001. *LIV. Lexikon der indogermanischen Verben. Die Wurzeln und ihre Primärstammbildungen*, 2nd edition. Reichert Verlag, Wiesbaden.
- Rachele Sprugnoli, Francesco Mambrini, Giovanni Moretti, and Marco Passarotti. 2020. [Towards the Modeling of Polarity in a Latin Knowledge Base](#). In *Proceedings of the Third Workshop on Humanities in the Semantic Web (WHiSe 2020) co-located with 15th Extended Semantic Web Conference (ESWC 2020)*. Heraklion, Greece, June 2, 2020, pages 59–70. CEUR-WS.
- Michael Weiss. 2020. *Outline of the historical and comparative grammar of Latin*. Beech Stave Press.
- Mark Wilkinson et al. 2016. [The fair guiding principles for scientific data management and stewardship](#). *Scientific Data*, 3.
- Jing Zhang, Bo Chen, Lingxi Zhang, Xirui Ke, and Haipeng Ding. 2021. [Neural, symbolic and neural-symbolic reasoning on knowledge graphs](#). *AI Open*, 2:14–35.
- Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. [Ernie: Enhanced language representation with informative entities](#). *arXiv preprint arXiv:1905.07129*.
- CIRCSE. 2020a. *Etymological Dictionary of Latin and the Other Italic Languages*. CIRCSE Research Centre. PID <https://zenodo.org/records/4147500>.
- CIRCSE. 2020b. *Latin Affectus*. CIRCSE Research Centre. PID <https://doi.org/10.5281/zenodo.4022689>.
- CIRCSE. 2020c. *Latin Vallex 2.0*. CIRCSE Research Centre. PID <https://doi.org/10.5281/zenodo.4032430>.
- CIRCSE. 2021a. *Charlton T. Lewis and Charles Short. 1879. A Latin Dictionary*. Clarendon Press, Oxford. CIRCSE Research Centre. PID <https://github.com/CIRCSE/LewisShort>.
- CIRCSE. 2021b. *UDante Treebank*. CIRCSE Research Centre. PID <https://github.com/CIRCSE/UDante>.
- CIRCSE. 2022. *LASLA corpus*. CIRCSE Research Centre. PID <https://github.com/CIRCSE/LASLA>.
- CIRCSE. 2023a. *Lexicon der indogermanischen Verben*. CIRCSE Research Centre. PID <https://github.com/CIRCSE/LIV>.
- CIRCSE. 2023b. *PrinParLat*. CIRCSE research centre. PID <https://github.com/CIRCSE/PrinParLat>.

8. Language Resource References

- CIRCSE. 2006-2024. *The Index Thomisticus Treebank*. CIRCSE Research Centre, ISLRN [105-545-284-528-2](https://zenodo.org/records/105-545-284-528-2).
- CIRCSE. 2018. *Word Formation Latin*. CIRCSE Research Centre. PID <https://doi.org/10.5281/zenodo.1492327>.
- CIRCSE. 2019-2024. *The LiLa Lemma Bank*. CIRCSE Research Centre. PID <https://doi.org/10.5281/zenodo.8300851>.

Defining an Ontology for Museum Critical Cataloguing Terminology Guidelines

Erin Canning

University of Oxford

Oxford, UK

erin.canning@eng.ox.ac.uk

Abstract

This paper presents the proposed ontology for the project “Computational Approaches for Addressing Problematic Terminology” (CAAPT). This schema seeks to represent contents and structure of language guideline documents produced by cultural heritage institutions seeking to engage with critical cataloguing or reparative description work, known as terminology guidance documents. It takes the Victoria and Albert Museum Terminology Guidance Document as a source for the initial modelling work. Ultimately, CAAPT seeks to expand the knowledge graph beyond the context of the Victoria and Albert Museum to incorporate additional terminology guidance documents and linked open data vocabularies. The ontology seeks to bring together scholarly communities in areas relevant to this project, most notably those in cultural heritage and linguistics linked open data, by leveraging existing linked data resources in these areas: as such, OntoLex, CIDOC CRM, and SKOS are used as a foundation for this work, along with a proposed schema from a related project, CULCO. As the CAAPT project is in early stages, this paper presents the preliminary results of work undertaken thus far in order to seek feedback from the linguistics linked open data community.

Keywords: cultural heritage, problematic terminology, linked open data, ontology

1. Introduction

Cultural heritage institutions are increasingly aware of the presence of bias and problematic and offensive language in texts of their catalogue records, as evidenced by the growing interest in critical cataloguing (Watson, 2023). There has been effort in the field to define what is meant by “problematic terminology” and therefore what institutions could, or should, examine in catalogue reviews (Chew, 2022; Cress, 2021; Dalal-Clayton & Rutherford, n.d.; Lawther, 2021; Muñoz, 2021; Museums Association, 2021; Ortolja-Baird & Nyhan, 2022; Rutherford, 2021a, 2021b, 2022). For example, the above authors identify “problematic terminology” as encompassing explicit slurs, euphemisms, and derogatory, objectifying, and dehumanizing language, as well as colonial and incorrect names of peoples, places, and types of objects.¹ However, there is little sector-wide guidance on what all this heading could include and the need to share information between institutions in pursuit of the development of best practices is well known (Chew, 2023; Dalal-Clayton & Rutherford, n.d.; Museums Association, 2020, 2021). At the level of individual institutions, museums are developing—and implementing—terminology guidance documents: these are glossary-like documents that list terms that the institution is interested in looking for in their cataloguing, often accompanied by a description of the term and a history of use that may give context to why the term was used when authoring catalogue records, paired with suggestions for actions to take when the term is found in the record. These suggestions are highly dependent on context, and include options such as to replace the term, to format it in a particular way that indicates its historical nature, or to add specific or general explanatory text, to give

three examples. These documents themselves are objects of potential scholarly interest: in addition to being a way for museums to communicate internally about emerging best practices, they show what terms museums are interested in addressing in their catalogue records and how they are thinking about defining such language. As such, terminology guidance documents may hold interest for linguistics as well as cultural heritage scholarly communities.

The project “Computational Approaches for Addressing Problematic Terminology” (CAAPT) seeks to make the contents of these terminology guidance documents available to institutions looking to engage in critical cataloguing as well as to relevant scholarly communities through the use of linked open data (LOD). The first step in this is to define the structure required to represent this information. This paper introduces the proposed ontology for CAAPT, based off of the Victoria and Albert Museum Terminology Guidance Document.

2. The Victoria and Albert Museum Terminology Guidance Document

The Victoria and Albert Museum (V&A) contains close to 1.7 million works of art and design objects acquired over more than 170 years of collecting activity, and which is still ongoing (Victoria and Albert Museum, 2023). The museum’s catalogue records represent objects from vast reaches of time and place, and as such contain a wide variety of problematics. The V&A holds regular cross-department meetings to discuss terminology questions and concerns raised by staff. This working group, in collaboration with the Interpretation Department and additional staff-led internal advisory groups, has produced and maintains

¹ For practical purposes within context of this paper, the author considers “problematic terminology” to be the terms listed in terminology guidance documents. For conceptual framing, the author will propose that

“problematic terminology” be understood as language which enables a catalogue record to perform or play into Haraway’s (1988) concept of “the god trick”; discussion of this falls outside of the present scope.

resources. Therefore, a small number of classes are proposed, with the focus instead on properties that bring together classes from these four ontologies. Furthermore, all classes and almost all properties—labelled here as “Computational Approaches for Addressing Problematic Terminology” (caapt)—are declared as subclasses and subproperties of elements from one or more of these four ontologies.

4.2 CAAPT Proposed Classes

The six classes proposed for CAAPT are:

1. *TermRoot*: root form of a term or phrase.
rdfs:subClassOf: skos:Concept,
crm:E55_Type, ontolex:LexicalEntry,
culco:ContentiousIssue
2. *Guide*: written guidance on language use.
rdfs:subClassOf: skos:ConceptScheme,
crm:E32_Authority_Document,
ontolex:ConceptSet
3. *TerminologyGuide*: written guidance for addressing problematic terminology created with the intention of assisting the work of or education around reparative description.
rdfs:subClassOf: caapt:Guide,
culco:ContentiousIssueScheme
4. *StyleGuide*: written guidance for language style and use.
rdfs:subClassOf: caapt:Guide
5. *UseContext*: context bounding the meaning intended by the use of a term.
rdfs:subClassOf: ontolex:LexicalSense
6. *Suggestion*: action to be considered or taken when a term is encountered.
rdfs:subClassOf: culco:Suggestion,
crm:E29_Design_or_Procedure

These classes represent core concepts for this model: the term being considered (*TermRoot*), the ways it has been used (*UseContext*), the suggestions written in the guidelines (*Suggestion*), and the guidelines documents themselves (*Guide*, *TerminologyGuide*, and *StyleGuide*). The decision to propose these classes as subclasses of multiple ontologies works to build a bridge between these communities. This is reminiscent of the approach taken by Khan & Salgado (2021) in their work to forge connections between OntoLex, FRBRoo, and CIDOC CRM through the creation of two new classes that inherit from each of these ontologies.

The requirement for new properties for three of these classes also justifies the need to create these classes, as opposed to proposing the use of multiple instantiation in the representation of instances of these classes. For example, in the case of the proposed class *UseContext*, which inherits only from OntoLex (*LexicalSense*), the need to connect to different classes in specific ways—discussed below as subproperties for *ontolex:usage*—drove the need to declare a new class.

Inheriting from multiple classes can also introduce new nuances of meaning, such as in the case of the proposed class *Suggestion*: while CULCO’s *Suggestion* class is defined as “a suggestion gives recommendations on how to use a contentious term” (Nesterov et al., 2022), inheritance from CIDOC CRM’s class *E29_Design_or_Procedure* introduces the specification that suggestions as they are understood in this context are, in fact, “documented plans for the execution of actions in order to achieve a result of a specific quality, form or contents” (Bekiari et al., 2022). Therefore, the *Suggestion* class here is a type of documented plan for how to address specific problematic terminology in a museum’s catalogue records.

Finally, the scope of these new classes is narrower in definition than the combination of meanings introduced by the classes from which they inherit. In the case of *Guide* and its subclasses *TerminologyGuide* and *StyleGuide*, the scope is defined more narrowly than for each of the classes from which *Guide* inherits, and terminology guides are differentiated from other forms of language guides, including writing style guides, in the source materials. This distinction is also reflected by *TerminologyGuide* inheriting from CULCO’s *ContentiousIssueScheme* as well as the *Guide* class, as it is only this specific kind of document that meets the additional criteria.

4.3 CAAPT Proposed Properties

The properties that are proposed, listed below in Table 2, are also connected to existing ontologies where possible: four are declared as subproperties of OntoLex’s *usage* predicate, and three as subproperties of CIDOC CRM’s *P69_has_association_with* predicate.

Property	Draft scope note	Domain	Range	rdfs:subPropertyOf
caapt:used_where	Geographic location in which the use context existed or was/is relevant	caapt:UseContext	crm:E53_Place	ontolex:usage
caapt:used_when	Time period in which the use context existed or was/is relevant	caapt:UseContext	rdfs:literal	
caapt:used_why	Intended purpose of use of the term in a use context	caapt:UseContext	crm:E55_Type	
caapt:about_who	Group of persons intended to be described by a term in a use context	caapt:UseContext	crm:E74_Group	
caapt:preferred	Suggestion that is preferred	caapt:Suggestion	caapt:Suggestion	crm:P69_has_association_with
caapt:if_not_possible_use	Suggestion to be considered if not possible to use preferred suggestion	caapt:Suggestion	caapt:Suggestion	

caapt:use_along_with	Suggestion to be used concurrently	caapt:Suggestion	caapt:Suggestion	
caapt:suggests_replacement	Suggested replacement term used in the suggestion	caapt:Suggestion	caapt:TermRoot	---
caapt:suggests_amendment	Suggested amending term used in the suggestion	caapt:Suggestion	caapt:TermRoot	
caapt:encountered	Suggestion encountered in type of catalogue field	caapt:Suggestion	skos:Collection	

Table 2: CAAPT proposed properties

The four properties proposed as subproperties of usage all refine the notion of the original predicate—to “indicate[] usage conditions or pragmatic implications when using the lexical entry to refer to the given ontological meaning” (Cimiano et al., 2016)—to the specific kinds of use cases discussed in the source materials, where an analysis of the descriptions revealed four main considerations: where the use of the term took place, when the use of the term took place, who the term was intended to describe when it was being used, and the intended purpose or use of the term. These four properties therefore introduce these meanings to the relationships they represent, and narrow the scope of the range from *rdfs:Resource* to specific classes according to the needs of those relationships.

The three properties proposed as subproperties of *P69_has_association_with* similarly refine the generic relationship between different instances of the *E29_Design_or_Procedure* class, of which *Suggestion* is proposed as a subclass, in order to specify three ways in which *Suggestions* are related to each other in the source documentation: two of these are hierarchical, representing a preference order in the listed suggestions, and the third indicates when two suggestions should be used at the same time or as two parts of a larger remediative cataloguing actions. For example, adding contextualising text and adding a content warning are often recommended together, as explaining the use of the problematic term does not negate the need for a warning, and adding a warning to a record does not negate the need to add text explaining what the term means in the context of the record or why it was retained.

Three properties are not proposed as subproperties to existing LOD predicates: *suggests_replacement*, *suggests_amendment*, and *encountered*. The first two connect a *Suggestion* to a *TermRoot* and specify whether a term is suggested to be used to replace a term, or to be included alongside the existing term as an amendment of the text. The final property connects a *Suggestion* with the kind of field in which the term is located in the catalogue record. Initial values for instances of this class are “historical context” (e.g. a Title field) and “contemporary context” (e.g. the current display label text for the object’s online collection page) as this is the language used in the

source materials when suggestions are made according to the location of the term in the record. This is an important element to consider as different suggestions are made depending on what kind of field the term appears in.

5. Conclusions

A knowledge graph structured according to the ontology defined here has been populated with the contents of the V&A Terminology Guidance Document, resulting in an initial graph describing 328 Suggestions for 73 potentially problematic terms. The schema and contents have been reviewed by key stakeholders at the V&A. These validation meetings have been successful and the schema in both theory and practice has been well received. The primary suggestion to come out of the knowledge graph review meeting was to include additional *use_along_with* properties between a greater number of *Suggestions*: only relationships that had been made explicit in the source document had been included in the knowledge graph population, and this review meeting revealed that this kind of relationship was often implicit in the museum’s documentation.

Next steps will be to consider two additional terminology guidance documents for inclusion: the Cultural Heritage Terminology Network Glossary and the glossary section of the *Words Matter* publication (Chew, n.d.; Tropenmuseum, 2018). Integrating these sources will validate the schema as being generalizable beyond the sole context of the V&A, as well as produce a knowledge graph that will begin to allow for inter-institutional comparisons of terms and suggestions. Following this, reconciliation with LOD vocabularies—namely the Getty Art & Architecture Thesaurus and the Homosaurus vocabulary—will take place, as connecting to these two resources will demonstrate integration with a vocabulary that is commonly used in the cultural heritage domain (Getty Art & Architecture Thesaurus) and a community-developed vocabulary that is already working in the space of critical cataloguing (Homosaurus).²

The steps taken thus far have built a solid foundation for this work to proceed. Initial validation of the schema and knowledge graph have been successful, and further feedback alongside the integration and reconciliation work will inform future developments.

² Getty Art & Architecture Thesaurus: <https://vocab.getty.edu/aat/>; Homosaurus: <https://homosaurus.org>

6. Acknowledgements

This research is taking place as part of the AHRC-funded Collaborative Doctoral Partnership scheme grant AH/X004775/1.

7. Bibliographic References

- Bekiari, C., Bruseker, G., Canning, E., Doerr, M., Ore, C.-E., Stead, S., & Velios, A. (2022). *CIDOC Conceptual Reference Model, Version 7.1.2*. <https://cidoc-crm.org/Version/version-7.1.2>
- Chew, C. (n.d.). Welcome to CHTNUK. *Cultural Heritage Terminology Network*. Retrieved April 28, 2023, from <https://culturalheritageterminology.co.uk/>
- Chew, C. (2022, May 27). Inclusive Terminology for the Heritage Sector. *ARLIS UK & Ireland*. https://www.youtube.com/watch?v=bb2hMcKz_aY
- Chew, C. (2023). Decolonising Description: Addressing Discriminatory Language in Scottish Public Heritage and Beyond. *Journal of Irish and Scottish Studies*, 11(1), Article 1. <https://doi.org/10.57132/jiss.213>
- Cimiano, P., McCrae, J. P., & Buitelaar, P. (Eds.). (2016). *Lexicon Model for Ontologies: Community Report*. <https://www.w3.org/2016/05/ontolex/>
- Cress, L. (2021, March 25). An Intern's Investigation on Decolonizing Archival Descriptions and Legacy Metadata. *Bitstreams: The Duke University Libraries Digital Collections Blog*. <https://blogs.library.duke.edu/bitstreams/2021/03/25/an-interns-investigation-on-decolonizing-archival-descriptions-and-legacy-metadata/>
- Dalal-Clayton, A., & Rutherford, A. (n.d.). Against a New Orthodoxy: Decolonised "Objectivity" in the Cataloguing and Description of Artworks. *Paul Mellon Centre Photographic Archive*. <https://photoarchive.paul-mellon-centre.ac.uk/groups/against-a-new-orthodoxy>
- Doerr, M., Light, R., & Hiebel, G. (2020). *Implementing the CIDOC Conceptual Reference Model in RDF*. <https://www.cidoc-crm.org/sites/default/files/issue%20443%20-%20Implementing%20CIDOC%20CRM%20in%20RDF%20v1.1.pdf>
- Haraway, D. (1988). Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies*, 14(3), 575–599.
- Khan, F., & et al. (2021). When Linguistics Meets Web Technologies. Recent advances in Modelling Linguistic Linked Open Data. *Semantic Web*, 0(0), 1–62.
- Khan, F., & Salgado, A. (2021). Modelling Lexicographic Resources using CIDOC-CRM, FRBRoo and Ontolex-Lemon. *SWODCH: Semantic Web and Ontology Design for Cultural Heritage*.
- Lawther, K. (2021, July 7). Decolonising the Database (Part 1). *Acid Free Blog*. <http://acidfreeblog.com/documentation/decolonising-the-database/>
- Miles, A., & Bechhofer, S. (2009). *SKOS Simple Knowledge Organization System*. <https://www.w3.org/2009/08/skos-reference/skos.html>
- Muñoz, G. (2021, April 22). Reframing Reparative Description Initiatives through Critical Race Theory and Black Feminism. *Descriptive Notes*. <https://saadescription.wordpress.com/2021/04/22/reframing-reparative-description-initiatives-through-critical-race-theory-and-black-feminism/>
- Museums Association. (2020). Our statement on Decolonisation. *Museums Association*. <https://www.museumsassociation.org/campaigns/decolonising-museums/our-statement-on-decolonisation/>
- Museums Association. (2021). Supporting Decolonisation in Museums. *Museums Association*. <https://www.museumsassociation.org/campaigns/decolonising-museums/supporting-decolonisation-in-museums/>
- Nesterov, A., Hollink, L., Van Erp, M., & Van Ossenbruggen, J. (2022). *CULCO: Concept Scheme for contentious terminology*. <https://cultural-ai.github.io/wordsmatter/>
- Nesterov, A., Hollink, L., van Erp, M., & van Ossenbruggen, J. (2023). A Knowledge Graph of Contentious Terminology for Inclusive Representation of Cultural Heritage. In C. Pesquita, E. Jimenez-Ruiz, J. McCusker, D. Faria, M. Dragoni, A. Dimou, R. Troncy, & S. Hertling (Eds.), *The Semantic Web* (pp. 502–519). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-33455-9_30
- Ortolja-Baird, A., & Nyhan, J. (2022). Encoding the haunting of an object catalogue: On the potential of digital technologies to perpetuate or subvert the silence and bias of the early-modern archive. *Digital Scholarship in the Humanities*, 37(3), 844–867. <https://doi.org/10.1093/lc/fqab065>
- Rutherford, A. (2021a, March 24). Documentation and Decolonisation. *Museum - Data - Laundry*. <https://museumdatalaundry.com/2021/03/24/documentation-and-decolonisation/>
- Rutherford, A. (2021b, July 7). Decolonising the Database event, Centre for Design History, University of Brighton, July 5th 2021. *Museum - Data - Laundry*. <https://museumdatalaundry.com/2021/07/07/decolonising-the-database-event-centre-for-design-history-university-of-brighton-july-5th-2021/>
- Rutherford, A. (2022, June 10). Provisional semantics, projects and positionality. *ARLIS Cataloguing and Classification Ethics Series 2022*. https://www.youtube.com/watch?v=tjTyi_WViSI
- Tropenmuseum. (2018). *Words Matter: An Unfinished Guide to Word Choices in the Cultural Sector*. <http://www.tropenmuseum.nl/en/about-tropenmuseum/words-matter-publication>
- Victoria and Albert Museum. (2023). Size of the V&A Collections on 31 March 2023. *Victoria and Albert Museum*. <https://vanda-production-assets.s3.amazonaws.com/2023/05/18/12/15/18/adca7f35-1299-48cb-be74-4aa07151525e/V&A%20Size%20of%20Collection%20Statement%20March%202023.pdf>
- Watson, B. M. (2023). Critiquing the Machine: The Critical Cataloging Database. *TCB: Technical Services in Religion & Theology*, 31(1), Article 1. <https://doi.org/10.31046/tcb.v31i1.3216>

The MOLOR Lemma Bank: A New LLOD Resource for Old Irish

Theodorus Fransen[†], Cormac Anderson[‡], Sacha Beniamine[‡], Marco Passarotti[†]

[†]Università Cattolica del Sacro Cuore, Milan, Italy, [‡]University of Surrey, Guildford, United Kingdom
{theodorus.fransen, marco.passarotti}@unicatt.it, {cormac.anderson, s.beniamine}@surrey.ac.uk

Abstract

This paper describes the first steps in creating a Lemma Bank for Old Irish (600-900CE) within the Linked Data paradigm, taking inspiration from a similar resource for Latin built as part of the LiLa project (2018–2023). The focus is on the extraction and RDF conversion of nouns from Goidelex, a novel and highly structured morphological resource for Old Irish. The aim is to strike a good balance between retaining a representative level of morphological granularity and at the same time keeping the number of lemma variants within workable limits, to facilitate straightforward resource interlinking for Old Irish, planned as future work.

Keywords: Old Irish, Lemma Bank, morphology, Linked Data

1. Introduction

While text-and-lexicon interlinking for Irish of the early medieval (and modern) period has been the subject of earlier studies and projects (Nyhan, 2008), these efforts have so far not resulted in a published resource. This is due to the highly variable nature of the language of this period in combination with the absence of sufficiently structured digital resources. The electronic Dictionary of the Irish Language, or eDIL (eDIL 2019), is the standard dictionary for Irish of the medieval period; however, it does not exhaustively list all possible inflections and spelling variants and does not use a consistent orthography for the spelling of headwords.

The lexical database Corpus PalaeoHibernicum (CorPH) “Corpus of Old Irish” (Stifter et al., 2021), covering the period ca. 6th–mid 10th century CE, is a more comprehensive and better-structured resource, but, like eDIL, does not provide word-level links to the source texts. Additionally, due to manual annotation practice, the spelling of headwords in the CorPH database is not entirely consistent, and the way it segments and stores complex morphological forms inhibits easy resource interoperability and interlinking.

The most linguistically principled and best structured lexical resource is Goidelex (Anderson et al., 2024), currently based on the 8th-century Old Irish Würzburg glosses (Kavanagh and Wodtke, 2001), which are not included in CorPH. Goidelex contains normalised lexemes with fine-grained morphophonological information. Like the aforementioned resources, however, it is not available in the Resource Description Framework (RDF) and published as Linked Data yet.¹

¹For RDF see <https://www.w3.org/RDF/>. See also Tim Berners-Lee’s Web Design Issues (Berners-Lee, 1996–present), particularly his Design Principles and his four rules about Linked Data

To alleviate the issues around resource interlinking for medieval Irish, this paper puts forward a Lemma Bank: a collection of canonical forms for interlinking lexical and textual resources. The resource is developed as part of the MOLOR — Morphologically Linked Old Irish Resource — project, which aims to create a new lexicographic model and standard for Old Irish for linking inflected forms in a text with a full-form lexicon, with a focus on the Würzburg glosses. The project takes inspiration from the Linking Latin (LiLa) project,² as part of which a Lemma Bank has been developed (Moretti et al., 2023), which was conceived — and has proven to successfully act — as a hub for linking lexical resources and texts for Latin (Passarotti et al., 2020).

Admittedly, while early medieval Irish sources represent the largest corpus of pre-twelfth-century European vernacular material, other than Latin and Greek (Stacey, 1991; Eska, 2019), a mature Natural Language Processing (NLP) pipeline comparable to Latin does not yet exist for automatic processing of medieval Irish texts, which is an urgent desideratum in the field. Dereza et al. (2023) attribute poor performance of NLP models to the lack of a linguistic and editorial standard for historical Irish and prompt Celticists and historical linguists to engage in further discussion. Promising advancements have nonetheless been made over the last 10–15 years, including on NLP core tasks such as tokenisation, lemmatisation, part-of-speech (POS) tagging, and morphological analysis and generation (Lamb and Fransen, forthcoming).

The creation of digital text archives for medieval Irish, however, goes much further back with the establishment of CELT — Corpus of Electronic Texts — in 1997, Ireland’s longest-running Humanities Computing Project (Ó Corráin et al., 1997); this resource contains 688 source texts in Irish (or Scottish Gaelic), albeit without linguistic annotation.

²<https://lila-erc.eu/>

More recently created digital corpora include the text archive as part of CorPH (Stifter et al., 2021) as well as two Old Irish corpora with a combined total of 98 syntactically annotated glosses following the Universal Dependencies (UD) framework, including 42 glosses from the Würzburg corpus; 3,469 POS-tagged glosses containing 21,749 tokens from the St. Gall UD corpus have been the basis for machine-learning-based POS-tagging experiments on diplomatically edited Old Irish text (Doyle and McCrae, 2024).

Interconnecting medieval Irish corpora with lexical resources using the Linked Data paradigm would be a major boon to medieval Irish studies. Despite the current lack of a mature NLP pipeline, a Lemma Bank, functioning as a central hub and interface, is considered to be a vital component in the envisaged MOLOR Knowledge Base of interlinked textual and lexical resources.

The focus of this short paper is on the first steps in creating a Lemma Bank for Old Irish (600–900CE), with a focus on the Würzburg glosses: the extraction and conversion into RDF of nouns contained in Goidelex (Anderson et al., 2024). We report on the design choices in selecting canonical forms, striking a balance between, on the one hand, linguistic granularity and, on the other hand, a workable amount of canonical forms (i.e. lemmas), while adhering to standards and best practices that have emerged in the area of Linguistic Linked Open Data (LLOD), notably the LiLa project and OntoLex (McCrae et al., 2017).

This paper is structured as follows. Section 2 introduces the resources instrumental in creating an Old Irish Lemma Bank. The LiLa Lemma Bank proved useful as an example for design choices, while a subset of the Goidelex data was used for the Lemma Bank’s content. The conversion of this content into RDF using existing ontologies for linguistic annotation is the topic of section 3. Some preliminary conclusions, as well as planned future research directions, are discussed in section 4.

2. Resource context

2.1. The LiLa Lemma Bank

The goal of the ERC-funded LiLa project (2018–2023) was to interconnect distributed (lexical and textual) resources and NLP tools for Latin by using the Linked Data paradigm, which is the basis of the so-called Semantic Web (Berners-Lee et al., 2001). As Passarotti et al. (2020, 187) have pointed out, “The core of the LiLa Knowledge Base consists of a large collection of Latin lemmas: interoperability is achieved by linking all those entries in lexical resources and tokens in corpora that point to the same lemma”. The resulting Lemma Bank currently

contains 215,102 Latin dictionary forms (Mambrini and Passarotti, 2023).

The design principles of the LiLa Lemma Bank are according to the specification of the lexicon model for ontologies (OntoLex-Lemon) as resulting from the work of the W3C Ontology Lexicon Community Group.³ This specification has emerged as the *de facto* standard for describing the content of lexical resources in the Linked Data framework.

It should be stressed that the LiLa Lemma Bank is not a lexical resource, that is, consisting of individuals belonging to the OntoLex-Lemon Lexical Entry class (`ontolex:LexicalEntry`). Instead, it is merely a collection of entities subsumed under the OntoLex Form class (`ontolex:Form`), for which an in-house class was devised — `lila:Lemma`. Being an OntoLex Form, a LiLa lemma can be linked to a Lexical Entry in any lexical resource via the property `ontolex:canonicalForm`, a subproperty of `ontolex:lexicalForm` (see Figure 1), connecting all other lexical resources compiled using the OntoLex-Lemon formalism (Passarotti et al., 2020).

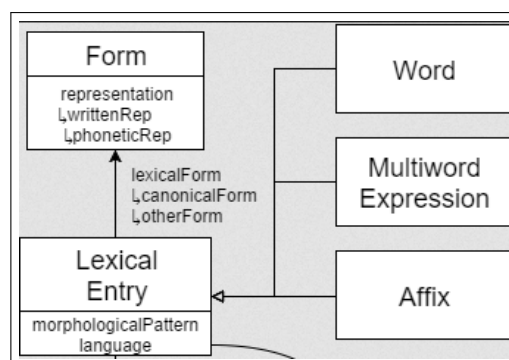


Figure 1: Part of the OntoLex-Lemon core model (Cimiano et al., 2016): the relationship between the classes `ontolex:LexicalEntry` and `ontolex:Form`

According to the OntoLex specification (Cimiano et al., 2016), “A Lexical Entry [...] needs to be associated with at least one form, and has at most one canonical form”.⁴ In order to allow for the use of different canonical forms used in lexical resources and lemmatised corpora, and not impose a lemmatisation criterion, the LiLa project created the symmetric property `lila:lemmaVariant`, whose range is the LiLa lemma class, “making it possible to retrieve from the textual [and lexical] resources connected to LiLa all the tokens that belong to the same lex-

³<https://www.w3.org/community/ontolex/>

⁴For example, the Lexical Entry for Old Irish ‘man’ would be linked to the canonical `nom.sg` form *fer*, while inflected forms such as *fir*, *feraiþ* etc. would have the subproperty `ontolex:otherForm` (not part of the MOLOR Lemma Bank)

ical item, regardless of the lemmatization criteria followed in individual corpora” (Passarotti et al., 2020, 190).

Lemma variants were devised for cases where morphological properties differ as part of the same lexical item, ignoring orthographic and phonetic variation which has no inflectional implications. For example, citation forms such as *claudio*, *claudor*, *claudio* and *claudor* ‘to limp’ constitute four different (LiLa) lemmas (i.e. OntoLex Forms) since each belongs to a different conjugation pattern and may be used as a lemma in lemmatised corpora or lexical resources. The lemma (Form) *claudio*, however, subsumes the graphical variant *cludo* (alongside *claudio*), encoded as written representation belonging to the same lemma (Passarotti et al., 2020). It is important to note that lemma variants, each represented by an OntoLex Form, receive a separate Uniform Resource Identifier (URI),⁵ while written representations (`ontolex:writtenRep`) are encoded as strings of the data property type with range `rdf:langString`, as such being merged as part of the same URI (see Figure 2).

2.2. Goidelex

Goidelex (Anderson et al., 2024) consists of a relational database (currently) containing 574 Old Irish nouns. It provides a normalised representation of lexemes (Fransen et al., 2023) and structured groupings into lexemes and flexemes (Fradin and Kerleroux, 2003; Thornton, 2018; Pellegrini, 2023). Lexemes are defined on the basis of shared meaning and POS type, while inflectional variants that belong to the same lexeme are analysed as separate flexemes. Each lexeme is associated with the corresponding identifiers in eDIL and CorPH (section 1), making Goidelex interoperable with existing resources. Flexemes are accompanied by phonemic transcriptions following Anderson (2016) as well as morphological and phonological properties. This resource also contains information about etymology and derivational morphology. Furthermore, it is designed to be compatible with the Paralex (Beniamine et al., 2023) and the Cross-Linguistic Data Format (Forkel et al., 2018) standards.

3. Conversion Principles

3.1. Motivations

Since Goidelex contains normalised forms and uses a principled approach to dealing with inflectional variation, it was found to be the most suitable resource among the ones mentioned in section 1

⁵A string that uniquely and persistently identifies a resource or concept, most commonly on the web

for starting to populate the Lemma Bank. However, Goidelex is too linguistically granular for the purposes of the Lemma Bank, necessitating a conversion process that often involves more-to-one mappings based on flexemes (see section 3.3).

While Goidelex is intended as a fine-grained morphological resource, the Lemma Bank’s function is rather to offer standardised entities identified by URIs to which other resources can link. Passing through the Lemma Bank, resources referring to these URIs will be made interoperable with each other. Linking to Goidelex is built-in, as its lexemes correspond to `ontolex:LexicalEntry` linked to lemmas in the Lemma Bank. As such, in conjunction with the Lemma Bank, the contents of Goidelex will provide rich morphological and phonological information in a Linked Data-based infrastructure of texts, lexical resources, and tools.

3.2. Lemma Properties and Ontologies

The code snippet in Figure 2 illustrates part of the triples, serialised using Turtle,⁶ that describe the resource `:94459`, which is a `lila:Lemma`, a subclass of `ontolex:Form`.

```
<data/id/lemma/94459>
  a      lila:Lemma ;
  rdfs:label      "claudio" ;
  lila:hasInflectionType
    lila:v3r ;
  lila:hasPOS      lila:verb ;
  lila:lemmaVariant
    <data/id/lemma/94457> ,
    <data/id/lemma/94458> ,
    <data/id/lemma/94456> ;
  dct:isPartOf
    <data/id/lemma/LemmaBank> ;
  ontolex:writtenRep      "claudio"
    , "cludo" .
```

Figure 2: Part of the triples as part of the LiLa lemma *claudio*, serialised using Turtle. Strictly speaking, a language tag for Latin (e.g. ISO 639-3 code `lat`) is required with `rdfs:label "claudio"`

Apart from `ontolex:writtenRep`, LiLa uses its own ontology and namespace (<http://lila-erc.eu/>) for linguistic annotations, which are aligned with OLiA (Chiaros and Sukhareva, 2015). This modelling decision has been emulated in MOLOR (see Figure 3).

3.3. Mappings

Mapping flexemes in Goidelex (section 2.2) indiscriminately onto lemmas would lead to a multitude

⁶<https://www.w3.org/TR/turtle/>

cluster cardinality	LiLa		Goidelex		MOLOR	
	# lemmas	percentage	# flexemes	percentage	# lemmas	percentage
1	156,323	94.81%	467	67.19%	549	91.65%
2	7,344	4.45%	188	27.05%	44	7.35%
3	999	0.61%	36	5.18%	6	1%
4	164	0.10%	4	0.58%		
5	30	0.02%				
6	18	0.01%				
	164,878	100%	695	100%	599	100%

Table 1: Statistics relating to LiLa lemma variant clusters, Goidelex (nominal) flexeme clusters, and MOLOR (nominal) lemma variant clusters

of lemma variants. Performing lemmatisation of corpora at this very fine-grained level would be challenging, in turn impeding straightforward linking between resources. By adopting a coarser granularity, we intend to facilitate the creation of accurate lemmatisers. Following the approach taken in LiLa, it was decided to create separate lemmas (a subclass of `ontolex:Form`) only where flexemes differ in inflectional properties (i.e. inflectional class, gender).

Some statistics may lend support for this decision.⁷ Table 1 shows how many LiLa lemmas are in each lemma variant cluster with a cardinality of 1 (no lemma variants) to 6 (six lemmas for a lexical item).⁸ The same statistics are calculated for Goidelex, but with flexemes rather than lemmas (there are no lexemes with more than 4 flexemes). Contrasting the given percentages for both resources, it becomes clear that a flexeme-to-lemma mapping would translate into about one-third of the lemmas having more than one lemma variant, as opposed to only about 5% in the LiLa Lemma Bank, negatively impacting lemmatisation and straightforward resource interlinking.

The variation seen in the Old Irish data can be categorised according to a four-way typology:

- i phonologically same, morphosyntactically same; e.g. *fer* ‘man’, masculine o-stem — realised by one `ontolex:Form`
- ii phonologically different, morphosyntactically same; e.g. *muinter*, *muntar* ‘community’, feminine ā-stem (see also flexeme 74.1 and 74.2 in Figure 3) — realised by one `ontolex:Form` and two spellings (`ontolex:writtenRep`)
- iii phonologically same, morphosyntactically different; e.g. *fius* ‘knowledge’, neuter or masculine

⁷This comparison should not be understood as solely reflecting the difference in the range of variation found in these languages; different design decisions in the resources concerned undoubtedly play a role as well

⁸The lemma as represented here does not include multiple written representations (e.g. *claudo*, *cludo*), which would result in a higher number

u-stem, alternatively neuter o-stem — realised by three `ontolex:Forms`

- iv phonologically different, morphosyntactically different; e.g. *brith* ‘carrying’, feminine i-stem, alternatively *breith*, feminine ā-stem — realised by two `ontolex:Forms`

The upshot of this decision is that flexemes with the same inflectional properties but a different phonological representation (and hence a different range of possible spellings)⁹ are merged as part of a single `ontolex:Form`. Figure 3 illustrates a case with three flexemes mapped to two Forms, i.e. MOLOR lemmas, which are lemma variants of each other.

Consideration has also been given to lexemes that have different or additional stem classes for singular and plural, leading to inflectional micro-classes, e.g. *duine* ‘person’ (`gen_masc;stem_io;num_sg`) with suppletive plural *doíne* (`gen_masc;stem_i;num_pl`), and *demun/demon* ‘demon, devil’ (`gen_masc;stem_o;num_all`), additionally *demon* with both a different gender and inflectional pattern in the plural (`gen_neut;stem_i;num_pl`). In the case of defective nouns, Goidelex uses a combined class as part of one flexeme, e.g. *aipgitir* ‘alphabet’ (`indecl/i;num_all`; indeclinable in the singular, i-stem inflection in the plural).

We are currently looking for satisfying ways to model this micro-variation, keeping in mind that a lemma is modelled as a form (`ontolex:Form`) rather than a lexeme, and ideally should not predicate features that only apply to parts of the paradigm, perhaps not even to the (lemmatic) form itself — the `nom.sg` in the case of nouns. However, it was decided to provisionally take over the inflectional micro-classes as they are encoded in Goidelex, with the exception of flexemes with different plural inflection, i.e. those encoded with `num_pl`, which were ignored (only 4 cases). This

⁹One spelling in Goidelex since this resource uses normalised spellings, i.e. a one-to-one mapping between phonological form and written representation

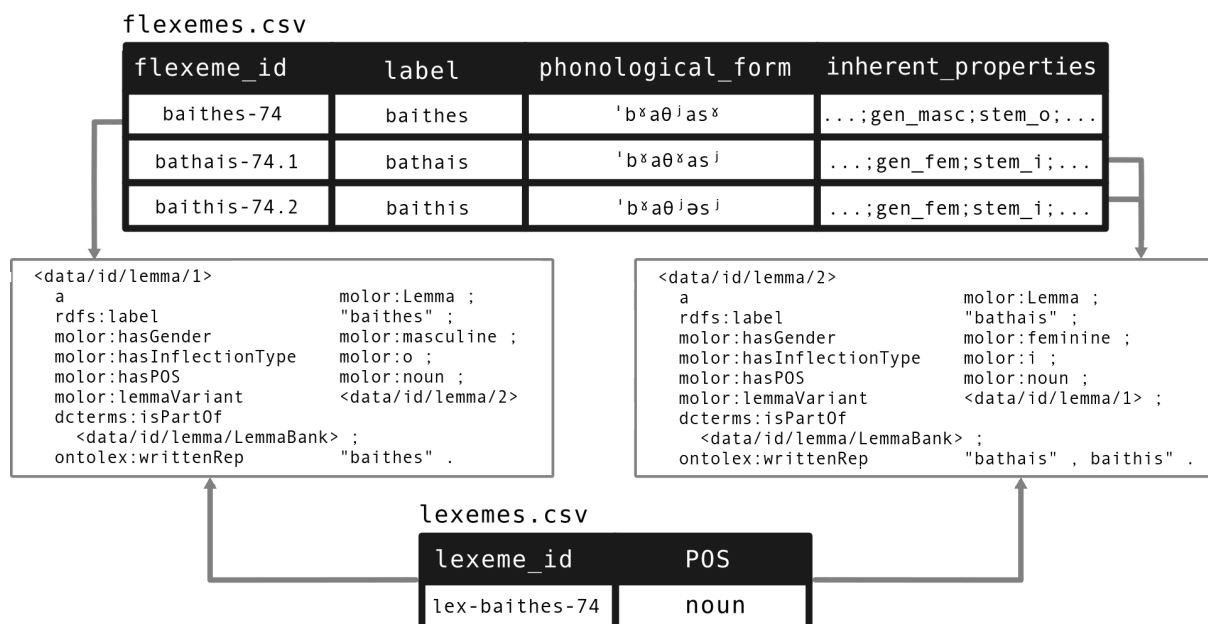


Figure 3: Example entry in Goidelex mapped to RDF-encoded MOLOR lemmas (`ontolex:Forms`). The required Old Irish language tag (e.g. ISO 639-3 code *sga*) with `rdfs:label` is not shown in this example

leads to the statistics given for MOLOR in the two rightmost columns in Table 1, which are now more similar to those of the LiLa Lemma Bank.

Lastly, we currently merge the Goidelex POS types `compound_noun`, `numeral_noun`, `prefixed_noun`, `proper_noun`, and `verbal_noun` into just `noun`; this information could be used at a future stage to establish derivational relationships (also encoded as part of Goidelex), likely to be modelled as an external resource similar to Word Formation Latin (Litta et al., 2020).

4. Conclusion and Future Work

This paper has described the first steps in converting the content of Goidelex (Anderson et al., 2024), a novel and highly structured lexical resource for Old Irish, into a Linked Data Lemma Bank, currently focussing on nouns. For design choices we have relied on the Lemma Bank developed as part of the LiLa Knowledge Base (Passarotti et al., 2020).

The next steps involve adding more lemmas with different POS categories from lexical resources, with the verb being the first in line. There are undoubtedly new challenges to overcome; the Old Irish verbal system (McCone, 1997) is much more complicated than the noun, whose inflectional patterns, as shown in this paper, already show an intricate interplay between morphology and phonology. Since verbs have not yet been systematically incorporated into Goidelex, and in the absence of a resource similar to Goidelex, we will have to resort mostly to other resources following a somewhat less granular approach. Thanks to its comprehen-

siveness and tabular format, CorPH (Stifter et al., 2021), in conjunction with the Würzburg dictionary (Kavanagh and Wodtke, 2001), is considered to be the most suitable starting point.

It is hoped and indeed expected that an Old Irish Lemma Bank will be an important hub in an inter-linked resource framework, making medieval Irish texts and the language’s grammar more accessible to scholars with various backgrounds. In the meantime, the authors gladly receive feedback from the Linked Data community on best practices for modelling under-resourced historical languages, especially in relation to variability and uncertainty.

5. Acknowledgements

Theodorus Fransen has received funding from the European Union’s Horizon Europe scientific research initiative under the Marie Skłodowska-Curie Actions (MSCA), grant agreement No 101106220 (MOLOR — Morphologically Linked Old Irish Resource). Cormac Anderson is funded by a British Academy Grant (GP GP300169) while Sacha Beniamine is funded by a Leverhulme Early Career Fellowship (ECF-2022-286). We would also like to thank the reviewers, especially reviewers 2 and 3, for their feedback and helpful comments.

6. Ethical Considerations and Limitations

To the best of our knowledge, there are no ethical concerns pertaining to this resource.

7. Bibliographical References

- Cormac Anderson. 2016. *Consonant colour and vocalism in the history of Irish*. Ph.D. thesis, Adam Mickiewicz University, Poznań.
- Cormac Anderson, Sacha Beniamine, and Theodorus Fransen. 2024. Goidelex: A lexical resource for Old Irish. Paper accepted to *Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)* at LREC-Coling 2024.
- Sacha Beniamine, Cormac Anderson, Mae Carroll, Matías Guzmán Naranjo, Borja Herce, Matteo Pellegrini, Erich Round, Helen Sims-Williams, and Tiago Tresoldi. 2023. [Paralex: a dear standard for rich lexicons of inflected forms](https://www.paralex-standard.org). In *Presentation at International Symposium of Morphology*. <https://www.paralex-standard.org>.
- Tim Berners-Lee. 1996–present. Design issues: Architectural and philosophical points. <https://www.w3.org/DesignIssues/>. Accessed: March 31, 2024.
- Tim Berners-Lee, James Hendler, and Ora Lassila. 2001. The Semantic Web. *Scientific American*, 284(5):34–43.
- Christian Chiarcos and Maria Sukhareva. 2015. [OLiA - ontologies of linguistic annotation](https://www.semantic-web.org). *Semantic Web*, 6:379–386.
- Philipp Cimiano, John P. McCrae, and Paul Buitelaar. 2016. Lexicon Model for Ontologies: Community report. W3C community group final report, World Wide Web Consortium. <https://www.w3.org/2016/05/ontolex/>. Accessed: March 31, 2024.
- Oksana Dereza, Theodorus Fransen, and John P. McCrae. 2023. [Do not trust the experts - how the lack of standard complicates NLP for historical Irish](https://www.aclweb.org/anthology/W19-01). In *Proceedings of the Fourth Workshop on Insights from Negative Results in NLP*, pages 82–87, Dubrovnik, Croatia. Association for Computational Linguistics.
- Adrian Doyle and John P. McCrae. 2024. Developing a part-of-speech tagger for diplomatically edited Old Irish text. Paper accepted to *Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)* at LREC-Coling 2024.
- Charlene M. Eska. 2019. *A Raven's Battle-cry: The limits of Judgment in the medieval Irish Legal Tract Anfuigell*, volume 27 of *Medieval Law and Its Practice*. Brill, Leiden and Boston.
- Robert Forkel, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymiski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, and Russell D. Gray. 2018. Cross-linguistic data formats, advancing data sharing and reuse in comparative linguistics. *Scientific Data*, 5(180205).
- Bernard Fradin and Françoise Kerleroux. 2003. Troubles with lexemes. In Geert Booij, Janet DeCesaris, Angela Ralli, and Sergio Scalise, editors, *Selected papers from the third Mediterranean Morphology Meeting*, pages 177–196. IULA – Universitat Pompeu Fabra.
- Theodorus Fransen, Cormac Anderson, and Sacha Beniamine. 2023. Towards a normalised orthography for Old Irish. Paper at *36th Irish Congress of Medievalists*, Dublin, 22–23 June 2023.
- Séamus Kavanagh and Dagmar S. Wodtke. 2001. *A lexicon of the Old Irish glosses in the Würzburg manuscript of the epistles of St. Paul*. Verlag der Österreichischen Akademie der Wissenschaften, Vienna.
- William Lamb and Theodorus Fransen. Forthcoming. Irish and Scottish Gaelic language technology. In Joe Eska, Paul Russell, Peadar Ó Muircheartaigh, and Silva Nurmio, editors, *Palgrave Handbook of Celtic Languages and Linguistics*. Palgrave Macmillan, London.
- Eleonora Litta, Marco Passarotti, and Francesco Mambrini. 2020. [Derivations and connections: Word formation in the LiLa Knowledge Base of linguistic resources for Latin](https://www.mathnet.uniroma2.it/). *The Prague Bulletin of Mathematical Linguistics*, 115:163–186.
- Francesco Mambrini and Marco Carlo Passarotti. 2023. [The LiLa Lemma Bank: A Knowledge Base of Latin canonical forms](https://www.ijer.uniroma2.it/). *Journal of Open Humanities Data*, 9(1):28.
- Kim McCone. 1997. *The Early Irish verb*, 2nd edition. An Sagart, Maynooth. Revised edition with *index verborum*.
- John P. McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. [The OntoLex-Lemon model: Development and applications](https://www.aclweb.org/anthology/W17-01). In *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference*, pages 587–597, Brno, Czech Republic. Lexical Computing CZ s.r.o.
- Julianne Nyhan. 2008. [Developing integrated editions of minority language dictionaries: the Irish example](https://www.aclweb.org/anthology/W08-01). *Literary and Linguistic Computing*, 23(1):3–12.

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(1):177–212.

Matteo Pellegrini. 2023. [Flexemes in theory and in practice](#). *Morphology*, 33:361–395.

Robin Chapman Stacey. 1991. Law and order in the very old West: England and Ireland in the early Middle Ages. In Benjamin T. Hudson and Vicki Ziegler, editors, *Crossed paths: methodological approaches to the Celtic aspect of the European Middle Ages*, pages 39–60. University Press of America, Lanham and London.

Anna M. Thornton. 2018. [Troubles with flexemes](#). In Oliver Bonami, Gilles Boyé, H el ene Firaudo, and Fiammetta Namer, editors, *The lexeme in descriptive and theoretical morphology*, pages 202–321. Language Science Press, Berlin.

8. Language Resource References

Cormac Anderson, Sacha Beniamine, and Theodorus Fransen. 2024. *Goidelex: A Lexical Resource for Old Irish*. Zenodo. PID <https://doi.org/10.5281/zenodo.10898228>.

David Stifter, Bernhard Bauer, Elliott Lash, Fangzhe Qiu, Nora White, Siobh an Barrett, Aaron Griffith, Romanas Bulatovas, Francesco Felici, Ellen Ganly, Truc Ha Nguyen, and Lars Nooij. 2021. *Corpus PalaeoHibernicum (CorPH) v1.0*. Maynooth University. PID <http://chronhib.maynoothuniversity.ie>.

Donnchadh  O Corra n, Hiram Morgan, Beatrix F arber, Benjamin Hazard, Emer Purcell, Caoimh n  O D onail, Hilary Lavelle, Julianne Nyhan, and Emma McCarthy. 1997. *CELT: Corpus of Electronic Texts*. University College Cork. PID <http://www.ucc.ie/celt>.

eDIL 2019. *An Electronic Dictionary of the Irish Language*, based on the Contributions to a Dictionary of the Irish Language (Dublin: Royal Irish Academy, 1913–1976) (www.dil.ie 2019). PID <https://www.dil.ie>.

Giovanni Moretti, Marco Passarotti, Rachele Sprugnoli, Paolo Ruffolo, and Francesco Mambrini. 2023. *CIRCSE LiLa Lemma Bank (V1.2)*. Zenodo. PID <https://doi.org/10.5281/zenodo.8300851>.

CHAMUÇA: Towards a Linked Data Language Resource of Portuguese Borrowings in Asian Languages

Anas Fahad Khan¹, Ana Salgado², Isuri Anuradha³, Rute Costa²,
Chamila Liyange⁴, John P. McCrae⁵, Atul Kr. Ojha⁵, Priya Rani⁵,
Francesca Frontini¹

¹CNR-ILC, Italy, ²CLUNL, NOVA University Lisbon, Portugal, ³University of Wolverhampton, UK,

⁴University of Colombo, Sri Lanka, ⁵University of Galway, Ireland

¹{fahad.khan, francesca.frontini}@ilc.cnr.it, ²{anasalgado, rute.costa}@fcs.unl.pt,

³Isuri.Anuradha@wlv.ac.uk, ⁴cml@ucsc.cmb.ac.lk,

⁵{atul.kumar.ojha, priya.rani}@insight-centre.org, john.mccrae@universityofgalway.ie

Abstract

This paper introduces CHAMUÇA, a novel lexical resource designed to document the influence of the Portuguese language on various Asian languages, initially focusing on South Asian languages. Through the utilisation of linked open data and the OntoLex vocabulary, CHAMUÇA provides structured insights into the linguistic characteristics and cultural ramifications of Portuguese borrowings across multiple languages. The article outlines CHAMUÇA's potential contributions to the linguistic linked data community, emphasising its role in addressing the scarcity of resources for lesser-resourced languages and serving as a test case for organising etymological data in a queryable format. CHAMUÇA emerges as an initiative towards the comprehensive catalogisation and analysis of Portuguese borrowings, offering valuable insights into language contact dynamics, historical evolution, and cultural exchange in Asia, one that is based on linked data technology.

Keywords: portuguese, ontolex, language contact, lexicon

1. Introduction

In the current article, we introduce a novel lexical resource titled **Cultural Heritage and Multilingual Understanding through lexiCal Archives (CHAMUÇA)** that is currently under preparation. The intention behind the resource is to describe the impact that the Portuguese language has had on the lexicons of the languages of Asia, with an initial focus on those of South Asia. CHAMUÇA, when complete, will consist of lexicons of Portuguese borrowings in each of the target languages covered by the resource along with a Portuguese language lexicon containing detailed information on each single etymon mentioned in the other lexicons. CHAMUÇA will be published on both in TEI-XML and as linked open data; in the current submission, we will focus on the latter. As we detail below, CHAMUÇA is informed by a number of relevant lexical and scholarly sources including pre-existing dictionaries, research articles and monographs, however, it will be based directly on open-source lexical resources such as *Wiktionary* and *Wikidata*. In turn, it will be published with a Creative Commons Attribution licence. The intention is for CHAMUÇA to be an open-source lexical resource that will be expanded through crowd-sourcing.

We begin this article by presenting the background to the project and motivating the need for such a resource in the first place. Then we will go

into some more details on the planned resource itself, including the languages in which we will begin by covering and the kinds of information which we plan to include. We also highlight those aspects of CHAMUÇA which are potentially of most interest to the linguistic linked data community. In addition, an example is presented from the Portuguese and Hindi lexicon to illustrate the content of CHAMUÇA.

2. Historical and Linguistic Background

Portuguese has a lengthy history of influence in Asia, stemming from the presence of Portuguese traders and colonists on the continent, traceable back to the 15th century and figures such as Pêro da Covilhã and Vasco de Gama. It is arguable that, with the very obvious exception of English, no other modern European language has had as much impact as Portuguese on the lexicons of the languages of, at least, South Asia. This influence can often manifest itself culturally in interesting and perhaps unexpected ways. One such example is the lexical unit *balti* which refers to a variety of Punjabi cuisine which is popular in the United Kingdom¹.

This borrowing, which entered British English

¹<https://visitbirmingham.com/inspire-me/areas/balti>

from Hindi/Urdu a few decades ago, ultimately derives from the Portuguese lexical unit *balde* 'bucket'. A detailed history of language contact between Portuguese and the languages of Asia and the formation of Portuguese language creoles, as well as a survey of previous work in this area, can be found in Cardoso's seminal article (Cardoso, 2016).

In the current work, we focus on borrowings into pre-existing Asian languages resulting, directly or indirectly, from this historical contact rather than on Portuguese creoles. These borrowings range from a handful of lexical units in languages such as Tibetan to languages with hundreds of Portuguese borrowings. It is interesting to note that although Hindi and Urdu, two of the most widely spoken languages in South Asia, only feature a few dozen borrowings from Portuguese (and these are generally shared by both languages), a good number of these are common everyday words: e.g., those for key (*chabi*), room (*kamra*), and even the word for English (*ingrez*). Other languages, such as Sinhala and Malayalam exhibit a much more substantial Portuguese lexical influence, reflecting a greater level of contact with Portuguese traders and colonists. Cardoso's article (Cardoso, 2016), and indeed research in this area in general, is heavily in debt to the work of the turn of the century scholar Sebastião Rodolfo Dalgado, and in particular his lexicon of Portuguese borrowings in Asian languages (Dalgado, 1913), a work which has had a significant influence on CHAMUÇA.

3. CHAMUÇA as Lexical Resource

3.1. The Why and How

Many interesting questions arise from the borrowings discussed in the previous section, considering various linguistic, historical, and cultural factors. While it is true that some of the information that could be used to respond to such questions is currently only available in print (non-digitized) resources or behind paywalls, a lot of it is currently available online and, in many cases, under an open license via sites as Wiktionary and Wikipedia². In this latter case, however, the information can either be incomplete, or unavailable in a structured form that can be easily queried using formal languages such as SPARQL. This is where CHAMUÇA enters the scene. The idea is precisely to create a structured lexical resource of Portuguese borrowings into Asian languages: one that is initially bootstrapped using open publicly available sources. In particular, we will make

²see for instance https://en.wikipedia.org/wiki/List_of_Sinhala_words_of_Portuguese_origin

use of Wiktionary, and its RDF version DBnary (Sérasset, 2015), for basic linguistic and grammatical information. This will be augmented by further relevant lexical information, such as corpus frequency data for the borrowed words using contemporary corpora for the languages in question, example sentences, more detailed domain label information, and alternative etymologies. It is important to emphasise that the authors of this submission – who are also the core contributors to this work – include not only speakers of the languages covered by the first version of CHAMUÇA but linguists and lexicographers who have worked with the languages in question as experts (including Portuguese) and will be able to curate the information that is included in CHAMUÇA, thereby adding scholarly value to the resource.

We have initiated our work on CHAMUÇA by focusing on the South Asian languages Urdu/Hindi, Sinhala, Tamil, Gujarati and Bengali. The plan is to open CHAMUÇA up to crowd-sourcing (initially via Github) to allow the addition of more words, more lexical information and more languages (again, this information will be checked and curated by the experts working on CHAMUÇA). The plan would be eventually to create an updated version of Dalgado's lexicon of Portuguese borrowings in Asian languages. One could ask whether such a resource is really necessary in the age of LLMs. However, after having carried out several experiments with ChatGPT, we found that it was very often unreliable with the kind of lexical information we were interested in; in short, then, the answer is yes.

From a high-level, architectural, perspective, CHAMUÇA is a *lexical resource*, where we understand this term as it is defined in the 2008 version of the Lexical Markup Framework standard (Francopoulo, 2013), that is, as a container for one or more lexicons. In our case, each separate lexicon belongs to a different language and consists of lexical units borrowed from Portuguese, or at least units which can plausibly be said to have been borrowed from Portuguese (since some words have conflicting etymologies)³. We decided to publish our resource in linked data because aside from the more general benefits of publishing data in a structured format and using a recognised standard⁴, the graph-based RDF model seems to be ideal for a resource structured in the way that CHAMUÇA is –

³That is, aside from the obvious case of the CHAMUÇA lexicon for Portuguese which contains lexical information on the Portuguese etymons which are featured in the other CHAMUÇA lexicons.

⁴Benefits which we would also have from publishing the resource solely in TEI-XML, a format which humanists and especially lexicographers tend to be more comfortable with, or at least less suspicious of.

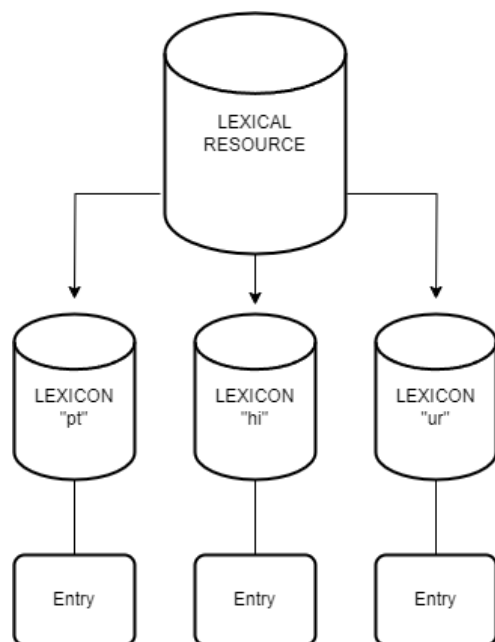


Figure 1: CHAMUÇA as a lexical resource

this becomes especially clear when one considers the power of the graph traversal-based SPARQL query language. In addition, RDF makes it easy to link to other kinds of resources (we intend to link CHAMUÇA to other, non-linguistic linked data resources including historical/geographical ones). In modelling CHAMUÇA as linked data, and more generally, as a structured dataset in the first place, we began by thinking about the kinds of questions (competency questions) that a user might ask of such a resource, e.g., those relating to which domains the borrowed words tend to belong to in a given language and how this changes across languages (and what this can tell us about the particular historical conditions of cultural contact for that given language) or those relating to the extent to which phonological, grammatical and semantic features are preserved (such as gender) or altered in different languages. This determined both what we intended to include as well as how it would be structured. At the time of writing, we have generated the first version of our Portuguese, Hindi and Urdu lexicons in RDF. Before we describe the dataset itself, we note the following points of interest for the linguistic linked data community:

- CHAMUÇA will cover (non-European and in some cases non-European and non-Indo-European) languages that currently don't have many resources dedicated to them in the LLOD cloud (as well as being lesser-resourced languages more generally). Building OntoLex lexicons for these languages will help us, among other things, in understanding the extent to which different kinds of linguistic

phenomena associated with these languages can be described by this model.

- CHAMUÇA is a kind of specialised lexical resource (a lexical resource consisting of lexical borrowings from a single language that has a strong cultural and historical interest) that so far has not been represented in the LLOD cloud, and which hasn't yet been covered in any existing OntoLex reports or sets of guidelines and best practices.
- CHAMUÇA will serve as a test case for the structuring of etymological information in a way that can be easily queryable.
- CHAMUÇA will allow us to further develop previous work on domain labelling⁵ carried out by some of the authors of the current submission as part of a Short Term Scientific Mission for the Nexus Linguarum COST action⁶ – since we plan to add domain labels explicitly to our data, informed by the approach set out in (Salgado, 2022).

In particular, we intend to contribute to current efforts in the BPMLOD W3C group⁷ on the creation of guidelines and best practices for LLD for tasks related to each single point listed above (Khan et al., 2022). In particular, we intend to create a series of metadata patterns for specifying the relationship of single resources with others both within the resource (in our case a single lexical resource and component lexicons) and external resources from which a given LLD lexical resource has been derived.

3.2. Generating a First Version of Chamuça

As a first experiment, we converted our initial dataset, composed of lexicons for three languages, Portuguese, Hindi and Urdu into linked data using the OntoLex vocabulary; for now the information in these lexicons derives principally from Wiktionary, although as mentioned above we plan to augment this with additional information in future. Our data was originally stored as a TSV file which was used to generate the RDF sources (and which will be used to generate the TEI-XML too) via a Python script⁸. The result is a first

⁵<https://github.com/anasfkhan81/EncodingDomainLabelsRDF/blob/main/Guidelines.md>

⁶<https://nexuslinguarum.eu/>

⁷<https://www.w3.org/community/bpmlod/>

⁸We intend to make the RDF files available by the time of the workshop, for various logistical reasons we weren't able to make them available by the time of submission.

version of Chamuca-RDF which consists of four separate files `chamuca_lexical_resource`, `chamuca_pt_lexicon`, a lexicon of Portuguese etymons, `chamuca_hi_lexicon`, a lexicon of Portuguese borrowings into Hindi, and `chamuca_ur_lexicon`, a lexicon of Portuguese borrowings into Urdu. As mentioned above `chamuca_lexical_resource` is a container for the three OntoLex lexicons, and will contain lexicons for other languages when they are ready. Since there is no specific class for lexical resources in OntoLex we have made `chamuca_lexical_resource` a subclass of `DCAT:dataset` from the Data Category Vocabulary⁹. We link `chamuca_lexical_resource` to its component lexicons using the Dublin Core `hasPart`.

```
:chamuca_lexical_resource a dcat:dataset ;
  dct:hasPart
    chamuca_hi_lex:,
    chamuca_ur_lex: ;
  dct:language
    "hi", "pt", "ur" ;
  dct:license
    <https://creativecommons.org/licenses/by/4.0/>;
  dct:title
    "chamuça"@eng .
```

In order to show the relationships between separate lexicons and the kinds of information which this first iteration of the language resource contains, we look at a single entry in Portuguese and its corresponding entry in the Hindi lexicon. The entry for *câmara* meaning 'chamber' (at least in its primary sense `câmara_sense_1`) in `chamuca_pt_lex` is as follows:

```
:câmara_entry a ontolex:LexicalEntry,
  ontolex:Word ;
  lexinfo:gender lexinfo:feminine ;
  lexinfo:partOfSpeech
    lexinfo:commonNoun ;
  ontolex:canonicalForm :câmara_lemma ;
  ontolex:lexicalForm :câmara_plural ;
  ontolex:sense :câmara_sense_1,
    :câmara_sense_2,
    :câmara_sense_3,
    :câmara_sense_4,
    :câmara_sense_5,
    :câmara_sense_6,
    :câmara_sense_7 .
```

The entry for कमरा (*kamra*) 'room' the Hindi word corresponding to *câmara* is as follows:

```
:कमरा_entry a ontolex:LexicalEntry,
  ontolex:Word ;
  lexinfo:etymologicalRoot
    chamuca_pt_lex:câmara ;
  lexinfo:gender lexinfo:male ;
  lexinfo:partOfSpeech
    lexinfo:commonNoun ;
  rdfs:seeAlso
    chamuca_ur_lexicon:kamra ;
  ontolex:canonicalForm
    :कमरा_lemma ;
  ontolex:lexicalForm
    :कमरे_dp_form_कमरा,
    :कमरे_os_form_कमरा,
    :कमरे_vs_form_कमरा ;
    :कमरो_vp_form_कमरा,
    :कमरो_op_form_कमरा ;
  ontolex:sense
    :कमरा_sense .
```

From the preceding, one can see that the word switched its grammatical gender in entering Hindi, this is not unusual since the '-a' ending in Hindi and Urdu is usually associated with masculine nouns (with the opposite being true in Portuguese). Our immediate plans are to add a fuller etymology for each Portuguese etymon, as well as having an example sentence for each word in the target languages along with corpus frequency and attestation data, using the Frequency Attestation and Corpus module of OntoLex, currently under development.

4. Future Work and Conclusion

In this article, we have introduced our ongoing development of CHAMUÇA, a novel lexical resource documenting the Portuguese influence on various Asian languages, with an initial focus on South Asian languages. By leveraging linked data principles and the OntoLex vocabulary, we have structured CHAMUÇA to facilitate accessibility, interoperability, and queryability. Through our efforts, we have transformed initial datasets into Chamuca-RDF, comprising lexicons for Portuguese, Hindi, and Urdu. This structured representation will potentially enable us to explore relationships between lexicons and delve into borrowed word domains across languages. Moving forward, CHAMUÇA holds the promise of being a valuable resource for linguistic research, historical inquiry, and cultural understanding. Ultimately, CHAMUÇA is intended to stand as a testament to the collaborative efforts of linguists, lexicographers, and language enthusiasts in preserving and exploring the rich tapestry of linguistic interactions between Portuguese and Asian languages.

⁹<https://www.w3.org/TR/vocab-dcat-3/>

5. Acknowledgements

The research of Ana Salgado and Rute Costa is supported by the Portuguese national funding through the FCT – Portuguese Foundation for Science and Technology, I.P. as part of the project UIDB/LIN/03213/2020; 10.54499/UIDB/03213/2020 and UIDP/LIN/03213/2020; 10.54499/UIDP/03213/2020 – Linguistics Research Centre of NOVA University Lisbon (CLUNL).

John P. McCrae, Atul Kr. Ojha and Priya Rani would like to acknowledge the support of the Science Foundation Ireland (SFI) as part of Grant Number SFI/12/RC/2289_P2 Insight_2, Insight SFI Centre for Data Analytics.

References

- Hugo C Cardoso. 2016. O português em contacto na ásia e no pacífico. *Manual de linguística portuguesa*, pages 68–97.
- Philipp Cimiano, John P McCrae, and Paul Buitelaar. 2016. Lexicon model for ontologies: Community report. *Final community group report, 10 may 2016, W3C*.
- Sebastião Rodolfo Dalgado. 1913. *Influência do vocabulário português em línguas asiáticas:(abrangendo cêrca de cinquenta idiomas)*. Impr. da Universidade.
- Gil Francopoulo. 2013. *LMF lexical markup framework*. John Wiley & Sons.
- 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.
- Ana Salgado. 2022. *Terminological Methods in Lexicography: Conceptualising, Organising, and Encoding Terms in General Language Dictionaries*. Ph.D. thesis, Universidade NOVA de Lisboa.
- Gilles Sérasset. 2015. Dbnary: Wiktionary as a lemon-based multilingual lexical resource in rdf. *Semantic Web*, 6(4):355–361.

LODinG: Linked Open Data in the Humanities

Jacek Kudera, Claudia Bamberg, Thomas Burch, Folke Gernert, Maria Hinzmann, Susanne Kabatnik, Claudine Moulin, Benjamin Raue, Achim Rettinger, Jörg Röpke, Ralf Schenkel, Kristin Shi-Kupfer, Doris Schirra, Christof Schöch, Joëlle Weis

Trier University, Am Universitätsring 15, 54286 Trier, Germany
{kudera, bamberg, burch, gernert, hinzmannm, kabatnik, moulin, raue, rettinger, roepke, schenkel, shikupfer, schirra, schoech, weis}@uni-trier.de

Abstract

We are presenting *LODinG – Linked Open Data in the Humanities* (abbreviated from *Linked Open Data in den Geisteswissenschaften*), a recently launched research initiative exploring the intersection of Linked Open Data (LOD) and a range of areas of work within the Humanities. We focus on effective methods of collecting, modeling, linking, releasing and analyzing machine-readable information relevant to (digital) humanities research in the form of LOD. LODinG combines the sources and methods of digital humanities, general and computational linguistics, digital lexicography, German and Romance philology, Sinology, translatology, cultural and literary studies, media studies, information science and law to explore and expand the potential of the LOD paradigm for such a diverse and multidisciplinary field. The project's primary objectives are to improve the methods of extracting, modeling and analyzing multilingual data in the LOD paradigm; to demonstrate the application of the linguistic LOD to various methods and domains within and beyond the humanities; and to develop a modular, cross-domain data model for the humanities.

Keywords: Linked Open Data, Humanities, Digital Humanities, Knowledge Graphs

1. Background

The potential of implementing the Linked Open Data (LOD) paradigm in the field of (Digital) Humanities is immense and has already been discovered by many scholars (Zhao, 2023). The interconnectedness of data from different, even seemingly unrelated disciplines has already allowed for a more comprehensive description of linguistic, cultural, sociological, and/or historical phenomena, often providing previously unnoticed contexts. A key finding of (Zhao, 2023) is that fields such as linguistics or actors like libraries have been producing new LOD resources for some time; however, projects emerging from other areas of the humanities primarily use existing LOD resources to uniquely identify and disambiguate entities relevant to their domain, but only rarely produce substantial new LOD resources themselves.

Despite the availability of conceptual reference models (such as CIDOC-CRM; (Faraj and Micsik, 2021)), ontology representation frameworks (e.g., OWL) and Semantic Web technologies (e.g., RDF; (Hitzler, 2021)), different disciplines in the humanities develop independent ways of categorizing entities. In the humanities, the domain-specific terminology often circulates within a particular area of research and rarely takes advantage of conceptual interlinking of uniquely-identified items. Descriptive studies that refrain from placing their entities within the structures of a formalized ontology significantly reduce their interdisciplinary potential and leave the opportunities that lie in knowledge networks undiscovered. Furthermore, the lack of presence of

linguistic LOD (LLOD) across the disciplines canonically associated with the humanities limits the existing data to digitally-structured sources only and renounces the methods that can benefit from the exploration of robust knowledge graphs (KGs).

Relevant work specifically at the intersection of literary studies and LOD is emerging recently, such as the *GOLEM* project (Graphs and Ontologies for Literary Evolution Models) at Groningen University (Pianzola et al., 2023) or the *MEDIATE* project at Radboud University (Montoya, 2021). Similarly, initiatives such as the one aiming to 'lodyfy' the *European Literary Text Collection* (ELTeC) Odebrecht et al. (2021) are building bridges between linguistics and literary studies (Ikonić Nešić et al., 2022; Schöch et al., 2021). LODinG, however, takes up and expands on the questions opened by its preceding research initiative, i.e., *MiMoText – Mining and Modeling Text* (2019-2023) hosted by Trier University and coordinated by the Trier Center for Digital Humanities (TCDH). LODinG's predecessor has developed a KG for the domain of the French novel of the Enlightenment, the *MiMoTextBase*. The project team used computational methods to extract information from a wide range of sources – from bibliographic resources and primary texts from the 18th century to current research literature (Schöch et al., 2022). The information ranges from bibliographic data (such as places and dates of publication) to book formats, themes, narrative locations, protagonists and sentiment trajectories or stylistic similarities between texts. The LOD paradigm allows this heterogeneous information to be linked to form a common body of knowledge. Its contents are

formally modeled and linked to each other. In addition, the extracted information is also linked to external knowledge resources such as Wikidata. The numerous query options that this allows – including federated queries originating from both *MiMoTextBase* and *Wikidata* – open up entirely new perspectives on both well-known and lesser-known literary-historical knowledge. With LODinG, we can now extend this paradigm to a much wider range of domains within and beyond the humanities.

Against this background, the presented project *LODinG* – initiated by a group of researchers at Trier University, Germany, and coordinated by the Trier Center for Digital Humanities – aims to explore the potential of the LOD paradigm at the intersection of qualitative and quantitative studies in the humanities. The project seeks to enrich the methods of annotation and information extraction applied to domain data relevant to a range of fields in the humanities. The initiative is currently exploring the potential of bridging multiple semi-structured datasets using formally-modeled, domain-adapted and modular ontologies and taxonomies pertinent to literary studies, linguistics, digital lexicography, scholarly editing, media studies, scientometrics and law. The LOD paradigm is a cornerstone of innovation in (digital) humanities. It enables the linking of multidisciplinary data using a coherent ontological classification and interoperable formats. The project aims to promote the interdisciplinary and transparent research supported by state-of-the-art data management infrastructure driven by knowledge networks. Overall it aims to bring a new quality of interdisciplinary reasoning to the area of data science and the humanities.

2. Objectives

The presented project emphasizes an interdisciplinary approach to building a modular ontology using LOD. It combines the methods commonly used to build, explore and query knowledge networks with the apparatus traditionally employed in Natural Language Processing and Information Retrieval. Furthermore, LODinG aims not only to explore the potential of existing digital resources in the context of LOD but also to generate, publish and integrate new resources. Finally, the project aims to demonstrate the potential of linguistic LOD for innovative research endeavors in the humanities. To further present the potential of LOD in interdisciplinary research, LODinG bridges multimodal and multilingual areas of study, including Romance studies, German studies, Sinology and law, and demonstrates the possibilities arising from computing multimodal KGs embedded in linguistic, literary, cultural and legal contexts.

The LODinG project conceptualizes and deploys

the knowledge networks that enable querying, statistical analysis, data visualization, and linking to open datasets via formal modeling of entities and properties and the application of a modular, human-generated ontology. At the conceptual level, we use named entity recognition and other information retrieval tasks to provide entities (such as people, places, organizations, motifs, methods, works or themes, etc.) with unique identifiers and labels that allow us to build robust linked knowledge networks. Furthermore, with the application of LOD, the already existing criteria for classifying the entities and creating typologies can be easily inspected to reconsider their analytical value. Such an approach helps to strike a balance between typologies driven by too many categories, resulting in overly specific information, on the one hand, and too few categories, carrying the risk of deriving inaccurate generalizations, on the other.

3. Areas of activity

LODinG is organized in closely interlinked areas of activity, each focusing on different aspects of linguistic LOD (LLOD). In the scope of LODinG, we will explore the enormous potential of LOD for innovative research in the humanities with a focus on linguistics, law, as well as literary, cultural and media studies.

The first research area focuses on the lexical level of the language system and examines the neologisms coined as a result of the Covid pandemic crisis. This work bridges the fields of German lexicology and digital lexicography. In addition to studying recent lexical phenomena, this subproject introduces a diachronic perspective by taking into account historical lexicons rich in pandemic-related vocabulary (Zacherl, 2022). This sub-project applies LOD and LLOD methods based on the *semantics by reference* (McCrae et al., 2012; Cimiano et al., 2020) framework. The research agenda of this subproject also includes the exploration of the semantic domain related to infections and diseases through the prism of standard Semantic Web data (Wandl-Vogt and Declerck, 2014). The diachronic perspective on established synsets would enable tracing of lexical changes of the analyzed set of lexemes as well as multiword expressions semantically related to infectious diseases. Furthermore, this subproject of LODinG emphasizes the importance of the representation of dictionary entries in the linguistic LOD framework and provides support for the integration of lexicons into the Semantic Web (Passarotti et al., 2020; Khan et al., 2022; Lindemann et al., 2022).

The second area of research in LODinG focuses on the terminology frequently used in historical medical and botanical sources from the early modern

period of Romance and Germanic texts. In addition to linguistic LOD, this subproject is based on the methodological apparatus of translatology and scholarly editing in a LOD context (Spadini et al., 2021). This work package aims to compare historical sources containing standardized botanical *nomina propria* from different languages. The comparison will be done through direct source-target translation and interlanguage interference. The use of the latter approach may reveal the influence of an intermediate language that is typologically and phylogenetically distant from source and target languages on equivalent matching in translation. Several historical sources dating back to the 16th century have already been digitized in the preparatory stage for this analysis (Moulin, 2018). The use of supervised OCR and LLOD methods will enable the triangulation of quantitative and qualitative analyses planned for this project.

The third area of work combines the methods of digital humanities and computer science. This subproject focuses on extracting statements related e.g. to datasets, methods and results, utilizing, in part, the OpenAlex resource. Currently, the production of scientific works far exceeds the reading capacity of researchers and research teams. Traditional indexing solutions, such as keywords, abridged abstracts and reviews often fail to address central questions of publications. Algorithmic, scalable methods of information extraction and synthesis are becoming increasingly important, supporting *semantic publishing* (Shotton, 2009; Schöch, 2021; Verma et al., 2023). To address the question of the lack of so-called *semantic statements*, this subproject aims to employ the LLOD methods in combination with manual tagging and machine content retrieval to semantically annotate a collection of works spanning across the field of the humanities. The subproject focuses on transforming abstracts and keywords into a limited number of machine-readable LOD statements (Metzger et al., 2011). This subproject partly departs from the OpenAlex platform, that contains metadata and LOD statements, but so far is lacking a systematic domain-dependent content modelling and a solid architecture of a formal ontology. The identified area for improvement and LOD-inG's involvement is to supplement the available data with content analyses and provide additional domain-specific context.

The fourth area of work builds upon the previous subproject and combines sinology with computer science. It focuses on scientific literature published in Modern Standard Chinese and converts its findings into machine-readable synthetic LOD statements. This work will compare the performance of entity extraction methods excerpted from the source language on available non-Chinese sources. The analyses of information extraction will

be conducted from a cross-linguistic perspective. The project aims to develop language-specific tools for information extraction and synthesis. The goal is to provide non-Chinese readers with a toolkit to discover Chinese-language scholarly literature based on linguistic LOD. This approach provides an alternative to common machine translation solutions, which often lack high-quality training data that involves matching specialized terminology across multiple languages.

The fifth subproject introduces the synergy between cultural studies and Natural Language Processing. The material of focus, namely wine labels, consists of items that are constrained to a specific domain and often combine text and image. The coherence between the text and image on wine labels varies, sometimes resulting in the juxtaposition of opaque concepts that do not clearly correspond with one another. This is a good starting point for investigating the potential of combining text and image analyses to build more robust KGs. The subproject employs textual content and images from wine labels to develop robust multimodal knowledge representation networks. Today, some Large Language Models (LLMs) are trained using multimodal data that combine lexical input and images (Wu et al., 2023; Zhang et al., 2024). This means they support multimodal indexing processes, where text and image recognition can benefit from each other. The challenge is to harness the power of generative LLMs to make them sufficiently robust and predictable to create standards-based LOD. Thus, the objective of this subproject is to investigate the potential of this method for indexing collections of wine labels scraped from the web. The subproject aims to create a more generalizable process by starting with in-domain source material that could then be extended to other types of sources such as postcards, geographical maps, advertisements, or book illustrations.

The sixth area of work focuses on the conceptual indexing of multilingual European texts. The primary goal of this subproject is to develop a multilingual parallel corpus of European legal texts thematically related to digitization processes, datasets and digital data processing (such as Digital Services Act, Regulation EU 2022/2065). This work package aims to identify differences in legally-binding terminology among all official EU languages. Additionally, using LOD, the subproject will develop methods for transparent equivalent matching, supported by conceptual indexing. To accomplish these objectives, this work package includes the following consecutive steps: automatic sentence-level alignment of legal texts; identification of key jurisprudential concepts and their integration into the LOD framework; multilingual annotation (both manual and automatic) of the identified concepts; and concept-

driven search to identify mismatches across selected languages. This workflow is based on a set of similar tasks involving a multilingual corpus that comprises translations of an 18th-century legal text (Bretschneider et al., 2020). Such an approach enables the identification of contextual or conceptual discrepancies across multiple languages, which is particularly important in legal texts.

The seventh area of work in LODinG exhibits a cross-sectional character. Its aim is to integrate the standardized entries of all above-mentioned work areas into a modular ontology that supports federated queries (Shimizu et al., 2023, 2022; Cimini et al., 2020). This will be accomplished by using established techniques for constructing ontology and metadata structures, glossing, cataloging, employing semantic networks, and comparing taxonomies (Borek et al., 2021). We plan to contribute, wherever possible, to recent and emerging initiatives that strengthen the alignment of vocabularies and KGs and the potential of federation (Steller et al., 2024). The subproject aims to integrate both domain-specific and cross-domain general elements. Currently, only a limited set of predicates are being used across domains, such as person, place, publication, discipline, century, country, or continent. To bridge domain- and discipline-specific entities, a larger set of predicates will be implemented including methods, procedures, epochs, subdomains, phenomena etc. (Bodard, 2021; Burrows and Nurmikko-Fuller, 2020). We aim to employ Wikidata identifiers along with other available authority file data allowing for the enhancement of KGs. Striving for a balance between a project-specific micro-perspective and an overarching macro-perspective, the objective is twofold: (1) to utilize and develop domain-specific resources to generate new research perspectives, and (2) to support and promote overarching integration in the LOD paradigm enabling cross-disciplinary and cross-domain linking and reuse of information (Brown, 2022; Santini et al., 2024).

The final work area focuses on developing technical solutions necessary to achieve the project's goals. This cross-cutting area aims to create an environment that promotes interoperability of data curated in the above-mentioned subprojects. This work area also aims to ensure the quality of research data management strategies used by LODinG. Furthermore, this infrastructural project aims to create interfaces for unstructured data, enabling non-standard formats to adapt to the LOD framework. The tools developed within the scope of this work package will allow for data annotation using standardized classifiers and facilitating interlinking between disciplines traditionally associated with the research area of the humanities.

A further goal of this work area is to examine

the intersections between information extraction, ontology design, KG engineering and artificial intelligence. We plan to host several local Wikibase instances and to anchor LODinG in the *Wikibase ecosystem* (Diefenbach et al., 2021; Faulhaber, 2022; Simons, 2023; Rossenova, Lozana et al., 2023). Through our membership in the Wikibase Stakeholder Group, we aim to further contribute to this environment and develop its workflows and tool chains, for example, to semi-automatically convert exports from the semantic annotation tool INCEPTION or manual annotations from the virtual research environment FuD into LOD statements (Klie et al., 2018; Bamberg et al., 2023).

4. Anticipated Results

The LODinG initiative is still in its early stages, having been launched only at the beginning of 2024. However, we can outline several areas in which we anticipate results and outcomes. These areas pertain to domain-specific results from various work areas, capacity building and networking. Additionally, a modular cross-domain ontology for the humanities is proposed. LODinG aims to research, design, and publish interconnections between various standards, practices, knowledge domains, tools, and models with different focal points. The research will introduce and exemplify new applications of LOD within and across various disciplines in the humanities, while also rethinking LOD as a research paradigm. The LOD framework is expected to be suitable for both qualitative (hermeneutic and contextualizing) and quantitative (algorithmic and statistical) research, making it attractive to scholars from various areas of the humanities.

LODinG focuses on using LOD for multilingual and multimodal resources. The project aims to enrich a set of indexed entities involving several languages, addressing the dominance of English in models and tools for information extraction. Furthermore, LODinG has potential in multimodal applications of LOD, such as combining lexical information with images.

The experiences from the undertaken tasks will be documented extensively to provide guidance for other projects or published as best practices guidelines. The presented project offers an interdisciplinary platform to explore the effectiveness and necessity of domain-specific solutions, as well as the potential for holistic linguistic LOD infrastructures.

The novelty of LODinG lies in combining quantitative and qualitative research methods from various fields in the humanities with the LOD paradigm. We believe that such a multidisciplinary orientation of our research project has a potential to open a new chapter in digital humanities.

5. Bibliographical References

- Claudia Bamberg, Cornelia Bögel, Thomas Bürger, Thomas Burch, Ruth Golyschkin, Bianca Müller, Radoslav Petkov, Thomas Stern, Jochen Strobel, and Olivia Varwig. 2023. [Digitale Edition der Korrespondenz August Wilhelm Schlegels – Ergebnisse, Erfahrungen, Entwicklungen](#). *Zeitschrift für digitale Geisteswissenschaften (ZfdG)*.
- Gabriel Bodard. 2021. [Linked Open Data for Ancient Names and People](#). *ISAW Papers*, 20(4).
- Luise Borek, Canan Hastik, Vera Khramova, Klaus Illmayer, and Jonathan D. Geiger. 2021. [Information Organization and Access in Digital Humanities: TaDiRAH Revised, Formalized and FAIR](#). In *Information between Data and Knowledge. Information science and its neighbors from data science to digital humanities (ISI 2021)*. Universität Regensburg.
- Falk Bretschneider, Rainer Maria Kiesow, Claudine Moulin, and Christof Schöch. 2020. Les mots de Beccaria. Métalexicographie des langues du droit à partir de Dei delitti e delle pene (1764) et ses traductions en Europe (Metalex). *Beccaria. Revue d'histoire du droit de punir*, 5:117–125.
- Susan Brown. 2022. [Same Difference: Identity and Diversity in Linked Open Cultural Data](#). *International Journal of Humanities and Arts Computing*, 16(1):1–16.
- Simon Burrows and Terhi Nurmikko-Fuller. 2020. [Charting Cultural History through Historical Bibliometric Research. Methods, concepts, challenges, results](#). In *Routledge International Handbook of Research Methods in Digital Humanities*. Routledge.
- Philipp Cimiano, Christian Chiarcos, John P. McCrae, and Jorge Gracia. 2020. [Linguistic Linked Open Data Cloud](#). In Philipp Cimiano, Christian Chiarcos, John P. McCrae, and Jorge Gracia, editors, *Linguistic Linked Data: Representation, Generation and Applications*, pages 29–41. Springer, Cham.
- Dennis Diefenbach, Max De Wilde, and Samantha Alipio. 2021. [Wikibase as an Infrastructure for Knowledge Graphs: The EU Knowledge Graph](#). In *The Semantic Web – ISWC 2021*, Lecture Notes in Computer Science, pages 631–647, Cham. Springer.
- Ghazal Faraj and András Micsik. 2021. [Representing and Validating Cultural Heritage Knowledge Graphs in CIDOC-CRM Ontology](#). *Future Internet*, 13(11):277.
- Charles B. Faulhaber. 2022. [PhiloBiblon y el mundo wiki](#). *Magnificat Cultura i Literatura Medievals*, 9:203.
- Pascal Hitzler. 2021. [A review of the semantic web field](#). *Communications of the ACM*, 64(2):76–83.
- Milica Ikonić 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 Language Resources and Evaluation Conference*, pages 7–16, Marseille, France. ELRA.
- Anas Fahad Khan, Christian Chiarcos, Thierry Declerck, Daniela Gifu, Elena González-Blanco García, Jorge Gracia, Maxim Ionov, Penny Labropoulou, Francesco Mambri, John P. McCrae, Émilie Pagé-Perron, Marco Passarotti, Salvador Ros Muñoz, and Ciprian-Octavian Truică. 2022. [When linguistics meets web technologies. Recent advances in modelling linguistic linked data](#). *Semantic Web*, 13(6):987–1050.
- 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 *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. ACL.
- David Lindemann, Penny Labropoulou, and Christiane Klaes. 2022. [Introducing LexMeta: a meta-data model for lexical resources](#). In *Dictionaries and Society. (EURALEX 2022)*. IDS-Verlag.
- John McCrae, Guadalupe Aguado-de Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez-Pérez, Jorge Gracia, Laura Hollink, Elena Montiel-Ponsoda, Dennis Spohr, et al. 2012. [Interchanging lexical resources on the semantic web](#). *Language Resources and Evaluation*, 46:701–719.
- Steffen Metzger, Shady Elbassuoni, Katja Hose, and Ralf Schenkel. 2011. [S3K: seeking statement-supporting top-k witnesses](#). In *Proceedings of CIKM 2011*, pages 37–46. ACM.
- Alicia C. Montoya. 2021. The Shark in the Library: Books and Non-book Artifacts in Private Library Auction Catalogues, 1665–1830. In *Figurations animalières à travers les textes et l'image en Europe*, pages 249–265. Brill.
- Claudine Moulin. 2018. [Textwandlungen – Eucharis Rösslin, Der Swangern Frauwen und hebammen Rosegarten als sprachhistorische Quelle](#). In Christina Waldvogel Luise Czajkowski,

- Sabrina Ulbrich-Bösch, editor, *Sprachwandel im Deutschen*, pages 319–336. De Gruyter, Berlin/Boston.
- Marco Carlo 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(1):177–212.
- Federico Pianzola, Xiaoyan Yang, Noa Visser, Michiel van der Ree, and Andreas van Cranenburgh. 2023. [Constructing the GOLEM: Graphs and Ontologies for Literary Evolution Models](#). In *Digital Humanities Conference (DH2023)*.
- Rossenova, Lozana, Duchesne, Paul, and Blümel, Ina. 2023. [Wikidata and Wikibase as complementary research data management services for cultural heritage data](#). In *Proceedings of the 3rd Wikidata Workshop 2022 co-located with ISWC2022*.
- Cristian Santini, Nele Garay, Etienne Posthumus, and Harald Sack. 2024. [The Art of Relations](#). In *DHd 2024*.
- Christof Schöch. 2021. [Open Access für die Maschinen](#). In Maria Effinger and Hubertus Kohle, editors, *Die Zukunft des kunsthistorischen Publizierens*. arthistoricum.net-ART-Books.
- Christof Schöch, Maria Hinzmann, Julia Röttgermann, Katharina Dietz, and Anne Klee. 2022. [Smart Modelling for Literary History](#). *International Journal of Humanities and Arts Computing*, 16(1):78–93.
- Christof Schöch, Roxana Patras, Tomaž Erjavec, and Diana Santos. 2021. [Creating the european literary text collection \(eltec\): Challenges and perspectives](#). *Modern Languages Open*, 1.
- Cogan Shimizu, Andrew Eells, Seila Gonzalez, Lu Zhou, Pascal Hitzler, Alicia Sheill, Catherine Foley, and Dean Rehberger. 2022. [Ontology Design Facilitating Wikibase Integration – and a Worked Example for Historical Data](#).
- Cogan Shimizu, Karl Hammar, and Pascal Hitzler. 2023. [Modular ontology modeling](#). *Semantic Web*, 14(3):459–489.
- David Shotton. 2009. [Semantic publishing: the coming revolution in scientific journal publishing](#). *Learned Publishing*, 22(2):85–94.
- Olaf Simons. 2023. [Keine Selbstverständlichkeit: Citizen Science auf der FactGrid Wikibase-Plattform](#). In René Smolarski, Hendrikje Carius, and Martin Prell, editors, *Citizen Science in den Geschichtswissenschaften*. V&R unipress, Göttingen.
- Elena Spadini, Francesca Tomasi, and Georg Vogeler, editors. 2021. [Graph data-models and semantic web technologies in scholarly digital editing](#). Number Band 15 in *Schriften des Instituts für Dokumentologie und Editorik*. BoD – Books on Demand, Norderstedt.
- Jonatan Jalle Steller, Linnaea Charlotte Söhn, Julia Tolksdorf, Oleksandra Bruns, Tabea Tietz, Etienne Posthumus, Heike Fliegl, Sarah Pittroff, Harald Sack, and Torsten Schrade. 2024. [Communities, Harvesting, and CGIF: Building the Research Data Graph at NFDI4Culture](#). In *DHd 2024*.
- Shilpa Verma, Rajesh Bhatia, Sandeep Harit, and Sanjay Batish. 2023. [Scholarly knowledge graphs through structuring scholarly communication: a review](#). *Complex & Intelligent Systems*, 9(1):1059–1095.
- Eveline Wandl-Vogt and Thierry Declerck. 2014. [Cross-linking Austrian dialectal dictionaries through formalized meanings](#). In *Proceedings of EURALEX XVI*.
- Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S. Yu. 2023. [Multimodal large language models: A survey](#). In *2023 IEEE International Conference on Big Data (BigData)*.
- Florian Zacherl. 2022. [Digitale Tiefenerschließung traditioneller Lexikographie – am Beispiel des Romanischen Etymologischen Wörterbuchs](#).
- Duzhen Zhang, Yahan Yu, Chenxing Li, Jiahua Dong, Dan Su, Chenhui Chu, and Dong Yu. 2024. [Mm-llms: Recent advances in multimodal large language models](#).
- Fudie Zhao. 2023. [A systematic review of Wikidata in Digital Humanities projects](#). *Digital Scholarship in the Humanities*, 38(2):852–874.

6. Language Resource References

- Odebrecht, Carolin and Burnard, Lou and Schöch, Christof. 2021. [European Literary Text Collection \(ELTeC\)](#). COST Action 16204, 1.1.0.

DigitAnt: a platform for creating, linking and exploiting LOD lexica with heterogeneous resources

Michele Mallia*, **Michela Bandini**, **Andrea Bellandi***, **Francesca Murano†**,
Silvia Piccini*, **Luca Rigobianco◇**, **Alessandro Tommasi***, **Cesare Zavattari***,
Mariarosaria Zinzi†, **Valeria Quochi***

*Istituto di Linguistica Computazionale "A. Zampolli", Consiglio Nazionale delle Ricerche, Pisa, Italy
Area della Ricerca, Pisa, Italy
name.surname@ilc.cnr.it

†Dipartimento di Lettere e Filosofia, Università di Firenze
Firenze, Italy
name.surname@unifi.it

◇ Dipartimento di Studi Umanistici, Università Ca' Foscari
Venezia, Italy
luca.rigobianco@unive.it

Abstract

Over the past few years, the deployment of Linked Open Data (LOD) technologies has witnessed significant advancements across a myriad of sectors, linguistics included. This progression is characterized by an exponential increase in the conversion of resources to adhere to contemporary encoding standards. Such transformations are driven by the objectives outlined in "ecological" methodologies, notably the FAIR data principles, which advocate for the reuse and interoperability of resources. This paper introduces the solutions devised within a nationwide collaborative research project aimed at integrating techniques and methodologies from the conventional study of epigraphic materials, computational lexicography, semantic web, and other digital humanities subfields. It details its services, utilities, and data types and shows how it manages to produce, exploit, and interlink LLOD and non-LLOD datasets in ways that are meaningful to its intended target disciplinary context, i.e. historical linguistics over epigraphic data. The paper also introduces how DigitAnt services and functionalities will contribute to the empowerment of a recently started Italian infrastructure cluster project devoted to the construction of a nationwide federation of research infrastructures for the humanities and cultural heritage, and in particular to its pilot project towards establishing an authoritative LLOD platform.

Keywords: Historical linguistics, Services for linguistics technologies, LLOD, Ontolex-lemon, Digital epigraphy

1. Introduction

The recent years have witnessed a significant technological evolution, accompanied by a parallel methodological development in data processing and utilization. Linguistic technologies, in particular, have seen substantial growth, exemplified by the advancement of expansive linguistic models and tools like ChatGPT. This growth extends to various technological domains within linguistics, including the creation and enhancement of linguistic resources. The increasing adherence to FAIR principles (Wilkinson et al., 2016) and the utilization of Linked Open Data (LOD) (Yu, 2011) have facilitated the emergence of numerous projects, generating valuable resources that have enriched the current data landscape.

Guided by the strategic roadmaps of the European Union and directives from higher institutions, the prevailing policy direction emphasizes data sustainability (European Commission and Directorate-

General for Research and Innovation, 2016). The principle here is not to generate data from scratch but to reuse and encode data in a standard format that ensures interoperability for specific applications.

Within the ItAnt project (Marinetti et al., 2021), the DigitAnt platform positions itself within this scientific framework. It aims to establish methodologies and services for creating linguistic resources in LLOD compliant formats for a specific and multidisciplinary area such as digital epigraphy, with a particular focus on historical linguistic aspects.

This initiative, which will be discussed in detail in subsequent paragraphs, is also becoming part of a large infrastructural project named H2IOSC (Humanities and Heritage Italian Open Science Cloud)¹, the ambition of which is to federate all national research nodes into a single entity. DigitAnt's role within H2IOSC is to contribute to piloting the CLARIN-IT LLOD platform by providing a set of web

¹<https://www.h2iosc.cnr.it/home/>

tools that would allow users to create/update/revise LOD compliant lexical resources (for digital epigraphy) and interlink them with other materials such as digital editions of testimonies, other available LOD lexical and/or conceptual datasets, bibliographic information and common shared vocabularies.

2. Context

This work has been carried out within a 3-year collaborative research project dedicated to expand and advance existing scientific knowledge about the archaic languages of ancient Italy. The *Languages and Cultures of ancient Italy. Historical Linguistics and Digital Models* project (ItAnt henceforth) is thus situated at the crossroad between digital epigraphy and historical linguistics, fields that have experienced significant advancements through numerous interesting projects. In many of these projects, the utilization or publication of linked data is described as presenting opportunities for further growth. However, tools like EFES (Bodard and Yordanova, 2020)² and INCEPTION (Klie et al., 2018)³ facilitate the publication and creation of resources - mostly annotated text corpora - using encoding standards such as TEI-Epidoc (Bodard et al., 2014)⁴ or CoNLL, but currently lack the capability to directly produce Linked Open Data (LOD) outputs. Similarly, resource access tools like Institutional Cretan Inscriptions (Vagionakis, 2021) rely on XML technologies like EpiDoc without intending to generate LODified outputs. Some initiatives such as the Epigraphic Database Heidelberg⁵ and iSicily⁶ (Prag and Chartrand, 2019) recently have leveraged the ability to link data from inscriptions to other data sources (e.g., DbPedia⁷) and have used controlled vocabularies (Pleiades⁸, Geonames⁹, Trismegistos¹⁰) for semantically precise and updated metadata annotation (Grieshaber, 2019), but still deliberately do not produce or publish LOD datasets. Within these contexts, spanning epigraphy and other linguistic fields, a need has emerged to tackle one of the most compelling challenges from both a technological and methodological standpoint: to provide (web/virtual) environments enabling scholars to more easily create and access resources available to the humanities public following Open Science paradigms and methodologies promoting interoperability and re-usability. A

²<https://github.com/EpiDoc/EFES>

³<https://inception-project.github.io/>

⁴<https://epidoc.stoa.org/>

⁵<https://edh.ub.uni-heidelberg.de/>

⁶www.isicily.org

⁷<https://www.dbpedia.org/>

⁸<https://pleiades.stoa.org/>

⁹<https://www.geonames.org/>

¹⁰<https://www.trismegistos.org/>

project related to historical linguistics (specifically to Latin) that fully adheres to LOD standards is the *LiLa: Linking Latin* project (LiLa, for short)¹¹, which led to the development of various Latin lexical and textual resources, alongside with a suite of tools for analysis, resource linking, and utilization (Pasarotti and Mambrini, 2021). Regarding tools, a successful editor for RDF terminological resources is VocBench (Stellato et al., 2015)¹², which has become one of the most comprehensive tools for editing linked data resources in various formats (primarily SKOS, but also Ontolex-lemon (McCrae et al., 2017)), offering collaborative and infrastructural functionalities. The DigItAnt platform positions itself between traditional databases and portals in use in digital epigraphy environments and advanced tools like VocBench.

It represents the first endeavor to integrate functionalities typical of epigraphic databases and web annotation tools into a unified web environment alongside lexicographic tools, facilitating the creation and editing of lexica, vocabularies, and thesauri, as well as to facilitate the interlinking of heterogeneous datasets and publish them as LLOD.

3. DigItAnt Architecture

The DigItAnt platform is developed within the ItAnt project, in collaboration with the Ca' Foscari University of Venice and the University of Florence. Its main goal is to provide scholars with an online environment for creating LOD-ready lexica for the languages of ancient Italy starting from corpora of inscriptions, either already published or autopsically investigated by the project, encoded in TEI-EpiDoc format, and further enrich lexical information by means of linking it to other existing relevant datasets, such as bibliographies and possibly related external lexical resources.

Its service-oriented architecture showcases a dual nature: on the one side, a web application, EpiLexo (Mallia et al., 2023) has been developed to facilitate the editing and accessing of lexical-conceptual data from a triple store (access and manipulation of this data is mediated by a back-end module called LexO-server (Bellandi, 2019), which manages the database triples) and data ingested from XML editions of inscriptions encoded according to the TEI-EpiDoc standard (access and manipulation of this data is handled by the back-end module CASH-server (Zavattari and Tommasi, 2021)). On the other side, a second web application retrieves the data edited and produced by the previous editing interface, making it accessible to users without any need for authentication. To support

¹¹<https://lila-erc.eu/>

¹²<https://vocbench.uniroma2.it/doc/dev/>

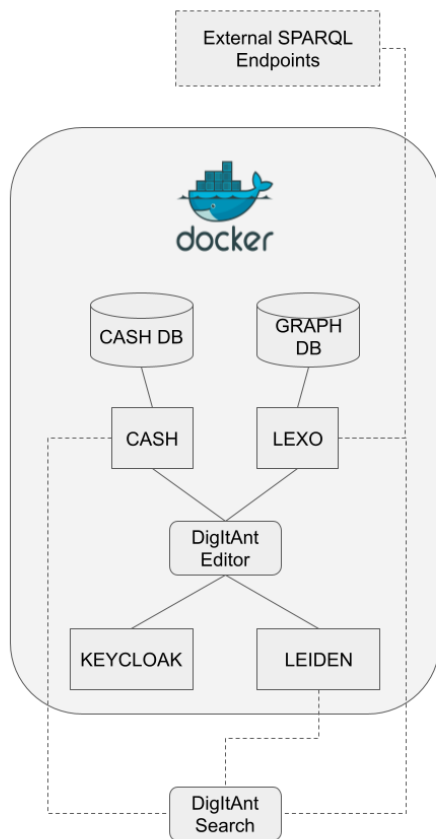


Figure 1: The DigItAnt software architecture

these back-end services, two types of APIs have been prepared: public and private. Only the APIs that allow data retrieval have been made openly available, while editing APIs require an authentication token obtained through user registration¹³. The modular architecture proposed for this project, moreover, as opposed to existing monolithic solutions, potentially allows for various customization, esp. on the front-end side, and improves the application's usability should any of the various back-end services cease to function or become superseded. Furthermore, container technology (specifically, Docker) was chosen to make all applications and services "atomic" and independent from each other. This approach enabled the step-by-step construction of essential components, ranging from the graphical interface to the services for managing LOD data and inscriptions, ultimately leading to the underlying schema (see Figure 1).

In addition, the implementation of authentica-

¹³The exploration interface will be publicly launched and opened upon finalization of the corpus and lexical data at the end of the project (July 2024).

tion via KeyCloak¹⁴ facilitates role mapping among users and makes the platform easily integratable into federated infrastructural environments. Another important functionality, currently embedded in the LexO-server, is the ability to query external SPARQL endpoints to facilitate linking internal items to external salient resources. The current system offers as a proof-of-concept direct querying to the LiLa endpoint¹⁵ for linking Latin cognate words, etymons and etymologies; specifically, Proto-Indoeuropean and Protp-Italic etyma can be represented by linking directly to the corresponding roots encoded in the *The Etymological Dictionary of Latin and the other Italic Languages in LiLa (EDLIL)* (Mambrini and Passarotti, 2020), while Latin cognates can be linked to the corresponding lemmas in the *LiLa Lemma Bank* (Passarotti et al., 2020). However, the potential to connect with other SPARQL endpoints exists¹⁶.

Beyond this stack lies the exploration and search interface, which makes the data produced with the editing tools accessible in a user-friendly way, and offers a different user experience in comparison to the default SPARQL endpoint, and thus potentially serves different user profiles. In addition to retrieving, filtering and visualizing data from single back-ends of data sources, this interface acts as a kind of middle layer, combining data from different data sources/providers for conducting advanced searches (which, in the current DigItAnt implementation include lexical data in LOD, inscriptions encoded in TEI-EpiDoc, and bibliographic references from Zotero¹⁷).

Currently, the platform adopts a relatively simple solution for authentication, lacking a genuine federated recognition system. User accounts are custom-created, with rules and authorizations assigned at various levels for resource usage and management. An interface panel facilitates the utilization of these functions, closely integrated with the Keycloak environment. Keycloak possesses the technological capabilities to handle various federated access types through support for multiple secure and legally compliant authentication protocols. Such capabilities should ensure smooth future integration into existing research infrastructures' AAI systems.

Additionally, certain aspects of both front-end interfaces could be improved, particularly regarding the mesh-up and integration of data coming from different heterogeneous sources. For DigItAnt Search, in particular, exploring different data

¹⁴<https://www.keycloak.org/>

¹⁵<https://lila-erc.eu/sparql/>

¹⁶For more details on the architecture, interfaces and functionalities see also Quochi et al. (2022a) and Quochi et al. (2022b)

¹⁷https://www.zotero.org/groups/2552746/itant_project/library

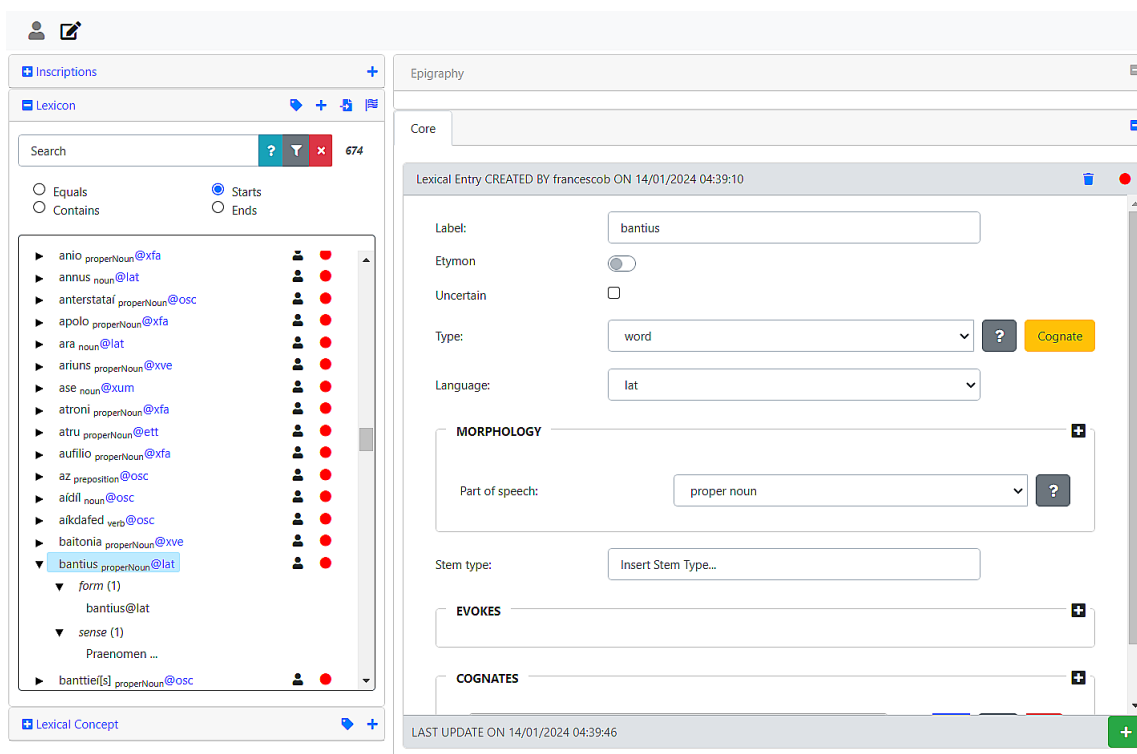


Figure 2: The editing environment

visualization and arrangement possibilities is necessary, depending on the linguistic context in which this service is applied, such as developing tools for information representation based on language or linguistic material type. For the editing platform, an instrumental tool would facilitate managing parameters for LOD material management (e.g., namespace management, workflows, projects, repositories, etc.), making the service accessible to a broader user base.

Finally, it would be advantageous to explore the capability to process a broader typologies of data types and perform multiple computations using certain parameters, such as the selection of metadata types ingested by the server handling texts, and the ability to ingest significantly larger textual corpora compared to the typically small inscriptions.

4. DigItAnt Data

Concerning data models, the platform mainly deals with three heterogeneous data types:

1. lexical data modeled according to the Ontolex-lemon and persisted in a GraphDB instance via the LexO-server;
2. digital scholarly editions of inscriptions encoded according to the TEI-EpiDoc model specifications and ingested from their XML serializations;

3. a bibliographic dataset created and managed via Zotero.

While inscriptions are encoded independently of the editing platform and subsequently ingested as ancillary resources to facilitate the representation of lexica with appropriate attestations, lexica are generated through the platform itself and natively linked to the inscriptions. Additionally, they are linked to relevant bibliographic references via Zotero, and to external lexical-conceptual resources (e.g., through direct queries to the LiLa knowledge-base SPARQL endpoint). Outputs primarily adhere to LLOD formats, with the exception being the ability to export the original XML editions of inscriptions annotated with links to the related lexical forms at the token level. Further details on the EpiDoc customization adopted to address the specificity of the target epigraphical documentation can be found in [Murano et al. \(2023\)](#). Notably, linking with the lexicon is facilitated by the tokenization of each word form in the original XML and the assignment of an @xmlid to each token.

Lexica are at the core of the editing platform. They are designed to be inherently LLOD-ready by adhering to Ontolex-lemon for the model, and LexInfo ([Cimiano et al., 2011](#)) for linguistic descriptors, with minimal adjustments to accommodate the special requirements for handling archaic, highly fragmented languages, defined in a project specific ontology.



Figure 3: The exploration and search environment

Although digital humanities projects more often adopt TEI XML formats for encoding dictionary data, with TEI Lex-0 becoming a widespread choice, we deliberately chose to model our lexica in Ontolex mainly because: 1. our goal in ItAnt is not to retrodigitize any traditional dictionary, rather to encode the linguistic knowledge that expert scholars formulate on the basis of their interpretation and analysis of the epigraphic texts; 2. for the sake of economy and FAIRness, we wanted to be able to reuse (by linking) available existing (LOD) knowledge; and 3. we wanted to make our outcome actionably available to others. However, because in this project we are dealing with *Restsprachen*, i.e. highly fragmentary attested languages, from ancient Italy –such as Oscan, Faliscan, Venetic, and Cisalpine Celtic– we had to face and find solutions to a number of lexicographic challenges.

First and foremost, because a full paradigm is lacking, it is difficult to retrieve a ‘traditional’ lemma. Therefore, lexical entries are associated with non-normalized linguistic realization, and no canonical form is formalized. Lexical Entries however still have a label, which is used by the interface for visualization purposes. Due to our limited knowledge of these languages, it is also impossible to provide a thorough description of the syntactic and semantic features typically found in (computational) lexica, such as lexical/syntactic relations or syntactic/semantic roles and frames. From the historical linguistic perspective of ItAnt, etymological information and its level of certainty are instead fundamental. For these reasons, the DigItAnt lexical model

uses a subset of the Ontolex Core: i.e. Lexical Entry, Form, Lexical Sense and Lexical Concept; and represents etymological data by exploiting the *lemonEty* extension proposed by Khan (2018) and already used in some important projects, among which the LiLa.

Morphosyntactic representation *Lexical Entry* is the container grouping all the attested forms of a lexical unit. Figure 4 below shows an example¹⁸. Apart from language and part-of-speech, two additional non-standard data properties are introduced for this class: *stemType*, which roughly indicates noun and adjective classes¹⁹ and *uncertain* for indicating whether the Entry is uncertain.

Form, exemplified in Figure 5, is the key pivotal element of our lexica and encodes standard formal features such as written representation and morphological properties. Word forms in DigItAnt, in fact, correspond to the attested forms, coming from the editor’s reading and including the editorial interventions (such as, for example, the restoration of damaged or missing letters). Linking lexical information with the corpus becomes, therefore, fundamental also to ensure reliability. To this end, attestations need to be recorded and encoded for every form, as is usually done in traditional (histori-

¹⁸The code has been simplified and the URIs have been removed to meet space and template requirements.

¹⁹For instance, *ā-stems*, i.e. stems ending in *-ā* < PIE *-eh2*, belonging to a specific declension type. LiLa makes use of a similar custom property *inflectionType*.


```

<!-- Lexical Entry-->
ItAntlex:upsed_entry
  a ontolex:Word;
  rdfs:label "upsed"@osc ;
  lime:language "osc" ;
  lexinfo:partOfSpeech lexinfo:verb ;
  :uncertainty "certain" ;
  ontolex:sense ItAntlex:upsed_sense ;
  ontolex:evokes ItAntlex:toWorkToil_
    semfield_concept .
  ontolex:lexicalForm
    ItAntlex:upsed_opsens_form ;
    ItAntlex:upsed_osins_form ;
    ItAntlex:upsed_upsed_form ;
  lemonEty:etymology ItAntlex:etym_upsed.
...

```

Figure 4: Simplified code snippet of the Lexical Entry for the Oscan verb *upsed*

```

<!-- Lexical Forms -->
...
ItAntlex:upsed_upsed_form
  a ontolex:Form ;
  ontolex:writtenRep "upsed"@osc .
  lexinfo:mood lexinfo:indicative;
  lexinfo:person lexinfo:thirdPerson;
  lexinfo:number lexinfo:singular;
  lexinfo:tense lexinfo:past;
  lexinfo:voice lexinfo:active voice ;
  :cites lexbib:upsed_verb_osc_upsed_
    form_bib583715
...

```

Figure 5: A sample of lexical forms encoded in the Entry for the Oscan verb *upsed*

cal) dictionaries. To represent and describe attestations, we plan to adopt and adapt the FrAC extension to Ontolex (Chiarcos et al., 2022). Currently, each form of a lexical entry is associated to its exact occurrence(s) in the ItAnt transcribed inscription(s), based on the ingested EpiDoc documents. Attestations are persisted in the CASH-server as text annotations and are enriched with optional information about certainty, authorship, relevant bibliographic citations, and free text notes.

Semantics representation Because for *Restsprachen* it is often not possible to retrieve the accurate semantic content of the words, the provided meanings are mostly generic, and entries generally have one sense. *Lexical Sense* encoding is therefore minimal; it is specified via a definition, can be indicated as uncertain, and can be associated with a *Lexical Concept*, used in DigItAnt to represent semantic fields. For this purpose, we created a SKOS taxonomy of semantic fields based on Buck's list of semantic fields (Buck, 1949). Among the works concerning the Indo-European semantics, Buck's list is one of the few to have organized

the Indo-European lexicon by categories, following a taxonomy²⁰.

Etymology. As anticipated above, etymology is represented via a subset of classes and properties from *lemonEty*, as exemplified in Figure 6. Etymological information, via *Etymology*, is attached to a *Lexical Entry* and applies to all of its forms. For each lexical entry either or both the Proto-Italic and Proto-Indo-European reconstructed roots are represented and encoded as instances of the class *Etymon*, i.e. Lexical Entries with a special status. Similarly, loanwords may also be reported as such, specifying the relationship with related forms such as *borrowing* rather than *inheritance*. Cognate words attested in sister languages are encoded as instances of another subtype of Lexical Entry established by *lemonEty*, the class *Cognate*. In accordance with the Linked Data principles and so as to avoid to produce data islands, Latin cognates as well as etymons and when deemed relevant Etymologies are linked to the LiLa knowledge base (respectively to the LiLa Lemma Bank (Passarotti et al., 2020) and the EDLIL (Mambrini et al., 2020)

Cognates can be encoded in two ways: 1. by linking externally to another linked data compliant lexicon²¹ or 2. by linking internally to a Lexical Entry of a different language, see Figure 7²².

Finally, bibliographic references and citations of relevant literature can be added/linked to any of the above elements to provide literature regarding the particular lexical information expressed. Currently, *Bibliography* is a system-internal data structure which links directly to the target in the ItAnt Zotero library specifying author, title and date. Furthermore, it makes it possible to specify additional citational information such as page spans and to add free text notes. Ontologies such as CITO (Peroni and Shotton, 2012) are under consideration for exporting citations related to both lexical classes and attestations in the lexicon. Work is also in progress for the mapping of the whole Zotero bib-

²⁰The taxonomy, created within the platform, also includes references to the Semantic Index of the Indo-European Lexicon, accessible at <https://lrc.la.utexas.edu/lex/semantic>, which served as inspiration for our adaptation. It will be disseminated at the end of the project along with the other project outcomes.

²¹This option is viable as regards Latin cognates, for which direct links to a canonical form in the LiLa Lemma Bank can be established directly by means of the *cognate* property. For instance, the Latin cognate of osc. *upsed* 'to erect, to set up, to produce' is represented by the URI of the corresponding lemma in the LiLa knowledge base, namely lat. *opus*.

²²This option is necessarily used for cognates in languages other than Latin for which LLOD lexica are not available, or when there is no satisfactory match in LiLa.

RIHS²⁶, CLARIN-IT²⁷, and OPERAS²⁸ research infrastructures. Its goal is to provide researchers with wide access to virtual laboratories, data centers and advanced tools for storing, processing, and visualizing digital resources, transcending disciplinary barriers to foster interdisciplinary innovative research.

The collaboration between ItAnt and H2IOSC exemplifies efforts to federate and optimize national research infrastructural resources, incorporating projects that overcome disciplinary boundaries and promote data-driven research in the humanities. This collaboration shall bring mutual benefits to both parties. For ItAnt this partnership ensures the sustainability. Interested scholars will be able not only to explore the project outcomes in the long term, but also to enrich the knowledge (graph) about ancient languages by contributing new data. On the other side, the project serves as a testing ground for H2IOSC's federation solutions and workflows, particularly toward its Linked Open Data (LOD) platform, one of H2IOSC's pilot projects.

DigitAnt may act as a test case for the planned workflows that assist scholars from depositing a (LOD compliant) resource to publishing it in the national endpoint.

6. Conclusion

In this paper we have presented the technological results of a research project that is concluding its activities in July 2024: the current implementation of a platform for creating and exploring linked data about ancient languages and cultures. This platform aims to assist historical linguists in representing their knowledge about these languages and cultures digitally, masking the complexities of dealing with digital models and formats. Centered around lexical data, the unique characteristic of this platform lies in the attempt to mesh-up and interlink heterogeneous datasets. In particular, the platform aims to integrate digital scholarly editions of epigraphic inscriptions, lexical data, citations, bibliographic references, and other relevant external resources. These resources vary not only in type, but also in their representational models and serialization formats (e.g., XML TEI, RDF Ontolex, and Zotero exports). Section 4 briefly described and exemplified their characteristics. The integration and meshing-up of heterogeneous and independent resources are made possible by the underlying Service-Oriented Architecture (SOA), which allows different back-ends to implement suitable technologies for handling various data types and models individually. The orchestration of integrated editing,

visualizations, and exports is then delegated to the front-ends and/or middle layers.

The DigitAnt platform will soon be finalized and released as an ItAnt project outcome. It will include export functionalities for the lexicon, attestations, and bibliography, as discussed in Sections 3 and 4, so that the resulting linked datasets may be versioned and deposited in an H2IOSC repository in compliance with the FAIR principles. Within the LOD-platform pilot project, this last event might trigger a procedure that automatically publishes the dataset on the CLARIN-H2IOSC SPARQL endpoint.

In the evolving landscape of digital humanities and cultural heritage research, the integration and optimization of research infrastructures (RI) have emerged as pivotal elements in enhancing interdisciplinary studies and overcoming traditional barriers. Web environments like DigitAnt, which offer sets of web tools for the creation or revision, enrichment, linking, LLOD publication, exploration and search of interconnected digital materials and knowledge about ancient cultures and languages, are good candidates for integration into RIs with mutual benefits. Indeed, an integral component of the H2IOSC vision is the development and refinement of services catering to the diverse needs of the research community. This includes the introduction of novel services. The collaborative paradigm exemplified by the H2IOSC initiative and the integration of projects such as DigitAnt can serve as a model for future developments and integration of data and services into infrastructure clouds. By advocating for a federated approach to research infrastructure, H2IOSC underlines the importance of accessibility, interoperability, and the collective utilization of digital resources, an aspect which will be strengthened by the (L)LOD platform pilot.

Finally, from our list of desired improvements that can further enhance the robustness of the system, we plan to prioritize those that may facilitate the integration into the CLARIN-IT/H2IOSC infrastructure and the LOD pilot. These may include allowing DigitAnt to ingest and manipulate other annotated text formats than TEI EpiDoc, exporting a new version of the original scholarly critical edition of the inscriptions enriched/annotated with the URIs of the lexical items attested, and allowing federated AAI to access the editing functionalities.

7. Acknowledgements

This work is carried out in the context of the PRIN 2017 "Languages and Cultures of Ancient Italy. Historical Linguistics and Digital Models" (no. 2017XJLE8J) funded by the Italian Ministry of University and Research. The DigitAnt platform is also supported by CLARIN-IT. Work on the DigitAnt

²⁶<https://www.e-rihs.it/>

²⁷<https://www.clarin-it.it/it>

²⁸<https://operas-eu.org/>

platform will continue within the H2IOSC Project - Humanities and cultural Heritage Italian Open Science Cloud funded by the European Union NextGenerationEU - National Recovery and Resilience Plan (NRRP) - Mission 4 “Education and Research” Component 2 “From research to business” Investment 3.1 “Fund for the realization of an integrated system of research and innovation infrastructures” Action 3.1.1 “Creation of new research infrastructures strengthening of existing ones and their networking for Scientific Excellence under Horizon Europe” - Project code IR0000029 - CUP B63C22000730005. Implementing Entity CNR.

8. Bibliographical References

- Gabriel Bodard, Greta Franzini, Simona Stoyanova, and Charlotte Tupman. 2014. [Introducing the epidoc collaborative: TEI XML and tools for encoding classical source texts](#). In *9th Annual International Conference of the Alliance of Digital Humanities Organizations, DH 2014, Lausanne, Switzerland, 8-12 July 2014, Conference Abstracts*. Alliance of Digital Humanities Organizations (ADHO).
- Gabriel Bodard and Polina Yordanova. 2020. [Publication, Testing and Visualization with EFES: A tool for all stages of the EpiDoc XML editing process](#). *Studia Universitatis Babeş-Bolyai Digitalia*, 65(1):17–35.
- C.D. Buck. 1949. *A Dictionary of Selected Synonyms in the Principal Indo-European Languages: A Contribution to the History of Ideas*. Linguistics/Reference. University of Chicago Press.
- 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, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Philipp Cimiano, Paul Buitelaar, John McCrae, and Michael Sintek. 2011. [LexInfo: A Declarative Model for the Lexicon-Ontology interface](#). *SSRN Electronic Journal*.
- European Commission and Directorate-General for Research and Innovation. 2016. [Report on the consultation on long term sustainability of research infrastructures](#). Publications Office.
- Frank Grieshaber. 2019. *Epigraphic Database Heidelberg—Data Reuse Options*. Universitätsbibliothek Heidelberg.
- Anas Fahad Khan. 2018. [Towards the Representation of Etymological Data on the Semantic Web](#). *Information*, 9(12).
- 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 *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018): System Demonstrations*, pages 5–9. Association for Computational Linguistics.
- Francesco Mambrini, Flavio Massimiliano Cecchini, Greta Franzini, Eleonora Litta, Marco Carlo Passarotti, and Paolo Ruffolo. 2020. [LiLa: Linking Latin Risorse linguistiche per il latino nel Semantic Web \(AIUCD 2019\)](#). *Umanistica Digitale*, 8.
- Anna Marinetti, Francesca Murano, Valeria Quochi, Monica Ballerini, Federico Boschetti, Angelo M. Del Grosso, Silvia Piccini, Luca Rigobianco, and Patrizia Solinas. 2021. [Languages and Cultures of Ancient Italy. Historical Linguistics and Digital Models](#). In *Decimo convegno annuale dell’Associazione per l’Informatica Umanistica e la Cultura Digitale (Pisa, 19 - 22 gennaio 2021)*, pages 528–532, Pisa. Associazione per l’Informatica Umanistica e la Cultura Digitale.
- John P McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. [The Ontolex-Lemon model: development and applications](#). In *Proceedings of the eLex 2017 conference*, pages 19–21.
- Francesca Murano, Valeria Quochi, Angelo Mario Del Grosso, Luca Rigobianco, and Mariarosaria Zinzi. 2023. [Describing Inscriptions of Ancient Italy. The ItAnt Project and Its Information Encoding Process](#). *Journal on Computing and Cultural Heritage*, 16(3):1–14.
- 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 Carlo Passarotti and Francesco Mambrini. 2021. [Linking Latin: Interoperable Lexical Resources in the LiLa Project](#). In Erica Biagetti, Chiara Zanchi, and Silvia Luraghi, editors, *Building new resources for historical linguistics*, pages 103–124. Pavia University Press.

Silvio Peroni and David Shotton. 2012. FaBiO and CiTO: ontologies for describing bibliographic resources and citations. *Journal of Web Semantics*, 17:33–43.

Jonathan R. W. Prag and James Chartrand. 2019. I. Sicily: Building a Digital Corpus of the Inscriptions of Ancient Sicily. In *Crossing Experiences in Digital Epigraphy: From Practice to Discipline*, pages 240–252. De Gruyter Open Poland.

Valeria Quochi, Andrea Bellandi, Fahad Khan, Michele Mallia, Francesca Murano, Silvia Piccini, Luca Rigobianco, Alessandro Tommasi, and Cesare Zavattari. 2022a. From Inscriptions to Lexicon and Back: A Platform for Editing and Linking the Languages of Ancient Italy. In *Proceedings of Second Workshop on Language Technologies for Historical and Ancient Languages LT4HALA 2022*, pages 59–67. European Language Resources Association (ELRA).

Valeria Quochi, Andrea Bellandi, Michele Mallia, Alessandro Tommasi, and Cesare Zavattari. 2022b. Supporting Ancient Historical Linguistics and Cultural Studies with EpiLexO. In *CLARIN Annual Conference Proceedings*, page 39.

Pat Riva and Maja Žumer. 2018. FRBRoo, the IFLA Library Reference Model, and Now LRMoo: A Circle of Development. In *Transform Libraries, Transform Societies*, Kuala Lumpur, Malaysia.

Armando Stellato, Sachit Rajbhandari, Andrea Turbati, Manuel Fiorelli, Caterina Caracciolo, Tiziano Lorenzetti, Johannes Keizer, and Maria Teresa Pazienza. 2015. Vocbench: A web application for collaborative development of multilingual thesauri. In *The Semantic Web. Latest Advances and New Domains - 12th European Semantic Web Conference, ESWC 2015, Portoroz, Slovenia, May 31 - June 4, 2015. Proceedings*, volume 9088 of *Lecture Notes in Computer Science*, pages 38–53. Springer.

Irene Vagionakis. 2021. Cretan Institutional Inscriptions: A New EpiDoc Database. *Journal of the Text Encoding Initiative [Online]*.

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, et al. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3.

Liyang Yu. 2011. *A Developer's Guide to the Semantic Web*. Springer.

9. Language Resource References

Andrea Bellandi. 2019. [LexO - Lexicographic Editor for Ontolex-lemon resources](#). ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa.

Michele Mallia, Andrea Bellandi, Alessandro Tommasi, Cesare Zavattari, Michela Bandini, and Valeria Quochi. 2023. [EpiLexO](#). ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa.

Francesco Mambrini and Marco Passarotti. 2020. [The Etymological Dictionary of Latin and the other Italic Languages in LiLa \(EDLIL\)](#). ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa.

Marco Passarotti, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo, and Rachele Sprugnoli. 2020. [LiLa Lemma Bank - Turtle format](#). ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa.

Cesare Zavattari and Alessandro Tommasi. 2021. [CASH - Corpus, Annotation and Search](#). A corpus and annotations management server.

Appendix: a sample entry in turtle format

```
@prefix itant:
</itantproject/ontologies/itant.owl> .
@prefix rdf:
<http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs:
<http://www.w3.org/2000/01/rdf-schema#> .
@prefix ns:
<http://www.w3.org/2003/06/sw-vocab-status/ns#> .
@prefix ontolex:
<http://www.w3.org/ns/lemon/ontolex#> .
@prefix lime:
<http://www.w3.org/ns/lemon/lime> .
@prefix lilaLemma:
<http://lila-erc.eu/data/id/lemma/> .
@prefix edlil:
<http://lila-erc.eu/data/lexical
Resources/BrilledL> .
@prefix lemonEty:
<http://lari-datasets.ilc.cnr.it/
```



```

lemonEty> .
@prefix crm:
<http://www.cidoc-crm.org/cidoc-crm/> .
@prefix lexinfo:
<http://www.lexinfo.net/ontology/3.0/lexinfo> .
@prefix skos:
<http://www.w3.org/2004/02/skos/core> .
@prefix ItAntlex:
</itantproject/data/lexicon#> .
@prefix lexbib: <http://itantproject/data/lexicon/bibliography#> .
@prefix semfield: <http://lrc.la.utexas.edu/lex/semantic/field/> .

<!-- Lexical Entry-->
ItAntlex:upsed_entry
  a ontolex:Word;
  dct:creator "Edoardo Middei" ;
  dct:contributor "Mariarosaria Zinzi" ;
  ns:term_status "editing";
  rdfs:label "upsed"@osc ;
  lime:language "osc" ;
  lexinfo:partOfSpeech lexinfo:verb ;
  ontolex:sense ItAntlex:upsed_sense ;
  ontolex:evokes ItAntlex:toWorkToil_
    semfield_concept .

<!-- Etymological info about cognates-->
lemonEty:cognate lilaLemma:115170 ;
lemonEty:cognate ItAntlex:upsaseter_pgn ;

<!-- Forms list -->
ontolex:lexicalForm
  ItAntlex:upsed_opsens_form ;
  ItAntlex:upsed_osins_form ;
  ItAntlex:upsed_upsed_form ;
  ... .

<!-- Lexical Sense -->
ItAntlex:upsed_sense1
  a ontolex:LexicalSense ;
  dct:creator "Edoardo Middei" ;
  skos:definition "to erect, to set up,
  to produce" ;
  ontolex:lexicalConcept
    ItAntlex:toWorkToil_semfield_concept .

<!-- Lexical Concept -->
ItAntlex:toWorkToil_semfield_concept
  a ontolex:LexicalConcept ;
  owl:sameAs semfield:PA_WV
  (https://lrc.la.utexas.edu/lex/semantic/field/PA\_WV) .

<!-- Lexical Forms -->
ItAntlex:upsed_opsens_form
  a ontolex:Form ;
  dct:creator "Edoardo Middei" ;
  dct:contributor "Mariarosaria Zinzi" ;
  ontolex:writtenRep "opsens"@osc .
  lexinfo:mood lexinfo:indicative;
  lexinfo:person lexinfo:thirdPerson;
  :cites lexbib:upsed_verb_osc_opsens_
    form_bib682785 .

ItAntlex:upsed_osins_form
  a ontolex:Form ;
  dct:creator "Edoardo Middei" ;
  dct:contributor "Mariarosaria Zinzi" ;
  ontolex:writtenRep "osins"@osc .
  lexinfo:mood lexinfo:subjunctive ;
  lexinfo:person lexinfo:thirdPerson ;
  :cites lexbib:upsed_verb_osc_osins_
    form_bib345190 .

ItAntlex:upsed_upsed_form
  a ontolex:Form ;
  dct:creator "Edoardo Middei" .
  dct:contributor "Mariarosaria Zinzi" .
  ontolex:writtenRep "upsed"@osc .
  lexinfo:mood lexinfo:indicative;
  lexinfo:person lexinfo:thirdPerson;
  lexinfo:number lexinfo:singular;
  lexinfo:tense lexinfo:past;
  lexinfo:voice lexinfo:active voice ;
  :cites lexbib:upsed_verb_osc_upsed_
    form_bib583715 .
  ... .

<!-- Etymology -->
ItAntlex:upsed_entry
  lemonEty:etymology ItAntlex:etym_upsed .
ItAntlex:etym_upsed
  a lemonEty:Etymology ;
  a crm:E89 ;
  rdfs:label "Etymology of: upsed@osc" ;
  lemonEty:etymon ItAntlex:he3p@PIE_entry ;
  lemonEty:hasEtyLink ItAntlex:etyLupsed-PIE ;
  lexbib:cites lexbib:etymology_412923bib412923 .

ItAntlex:he3p@PIE_entry
  a lemonEty:Etymon ;
  seeAlso edlil:etymon_pie0847 ;
  (https://lila-erc.eu/data/lexicalResources/BrilledL/id/etymon/pie0847)

ItAntlex:etyLupsed-PIE
  a lemonEty:EtyLink ;
  lemonEty:etyLinkType "inheritance" ;
  lemonEty:etySource ItAntlex:he3p@PIE_entry ;
  lemonEty:etyTarget ItAntlex:upsed_entry .

```


Teanga Data Model for Linked Corpora

John P. McCrae, Priya Rani, Adrian Doyle, Bernardo Stearns

SFI Insight Centre for Data Analytics, Data Science Institute, University of Galway, Ireland

john@mccr.ae, priya.rani@insight-centre.org,

adrian.doyle@insight-centre.org, bernardo.stearns@insight-centre.org

Abstract

Corpus data is the main source of data for natural language processing applications, however no standard or model for corpus data has become predominant in the field. Linguistic linked data aims to provide methods by which data can be made findable, accessible, interoperable and reusable (FAIR). However, current attempts to create a linked data format for corpora have been unsuccessful due to the verbose and specialised formats that they use. In this work, we present the Teanga data model, which uses a layered annotation model to capture all NLP-relevant annotations. We present the YAML serializations of the model, which is concise and uses a widely deployed format, and we describe how this can be interpreted as RDF. Finally, we demonstrate three examples of the use of the Teanga data model for syntactic annotation, literary analysis and multilingual corpora.

Keywords: corpora, natural language processing, linked data, formats

1. Introduction

Corpus data is vital to modern natural language processing and is often annotated with many layers of extra information from part of speech to complex structural and semantic categories. There are several standards for publishing corpora including the Text Encoding Initiative (Ide, 1994, TEI) and the Linguistic Annotation Framework (Eckart, 2012, LAF), however, none of these have become widely accepted in natural language processing. In contrast, for lexico-semantic data, linked data models based on RDF have had great success through models such as OntoLex-lemmon (McCrae et al., 2017; Cimiano et al., 2016). However, attempts to produce RDF models for representing corpus information such as the NLP Interchange Framework (Hellmann et al., 2013, NIF) and POWLA (Chiarcos, 2012) have had less success. A major reason for the failure of these models to have sufficient traction in NLP communities is that the RDF models adopted for linked data and the XML models used for TEI and LAF are very verbose and do not fit in with modern natural language processing pipelines. As such, most natural language processing data does not satisfy the FAIR principles, particularly in relation to reusability as the use of custom parsers, which may be difficult for others to reuse. Similarly, the adoption of a linked data paradigm will increase the findability and accessibility of the resource by providing methods where corpora can be connected with lexicographic, terminological and encyclopaedic resources.

In this paper, we introduce a new model called

the Teanga¹² data model, which aims to provide a simple, low-overhead method for sharing text corpora and interacting with linked data. The Teanga data model can be simply serialized as JSON or YAML, allowing it to be easily loaded and worked with in modern programming languages. The Teanga data model also develops a new method of annotation called *layered annotation*, which combines the best of stand-off annotation and in-line (XML-style) annotation to enable data to be quickly handled. Finally, the Teanga data model defines a method of annotation that provides a conversion to RDF and can be converted into standard RDF. We note the Teanga JSON serialization is inspired by JSON-LD (Sporny et al., 2020), but is not directly a JSON-LD model. The Teanga data model is being developed as part of Teanga 2, a new platform for NLP based on the previous Teanga platform (Ziad et al., 2018).

The rest of this paper is as follows: firstly, we will introduce the Teanga data model and layered annotation and then we will describe the technical implementation of the model, including serialization as YAML, an implementation in Python and the conversion to RDF. We will then provide three examples of conversions of data from Universal Dependencies (de Marneffe et al., 2021), conversion of TEI data, such as from the ELTeC (Schöch et al., 2021) corpus, and an example of parallel corpora with word-level alignment data. We will then conclude with a discussion of the Teanga data model in comparison to other corpus models.

¹Teanga is Irish for tongue/language and is pronounced t'anga

²<https://teanga.io/>

2. Design of the model

2.1. Layered Annotation Model

A corpus in the Teanga data model, as depicted in Figure 1, is composed of a metadata section and a list of documents. Each of these documents has layers that are defined in the metadata layers and may have some or all of these layers. All documents must have at least one **character** layer, which consists of a single Unicode string containing the text of the document. This means that Teanga preserves the plain text version of the document, in contrast to XML annotation where annotations must be inserted into the document. Currently, Teanga only supports text corpora, but the introduction of new base layer types would allow the model to extend to multimodal corpora. The remaining layers of annotation consist of a reference mechanism and (optionally) a data value. The referencing mechanism refers to an annotation in another layer (the *base layer*), which is defined in the metadata. For character layers, the elements are the Unicode characters in the layers. All indexes in layers start from zero. The referencing mechanisms are as follows:

- **Span Layer:** A span layer gives two indexes corresponding to the start and end of the annotation.
- **Division Layer:** The division layer divides the base layer into non-overlapping segments
- **Element Layer:** An element layer refers to a single element in the base layer
- **Sequence Layer:** A sequence layer corresponds to the annotation layer in a one-to-one manner so that there is one annotation for each element of the base layer.

In most cases, a span layer is used to divide the lower layer into words and other annotations are based on this word layer. Division layers are used to divide the text by sentences, paragraphs or chapters.

Each annotation in a layer must have the same data value, the values are defined as follows:

- **None:** No data is associated with an annotation, for example, in tokenization.
- **String:** A single string is associated with each annotation
- **Enumeration:** The annotation may have one value from a list of values given in the metadata section
- **Link:** A reference is defined to another annotation in the same layer, or in a secondary layer called the *target layer*.

- **Typed Link:** Combines the data of the enumeration and the link layer.

In addition to layers, each document or layer may have any number of *meta-properties*. The most important of which is the `_uri` property which gives the URI to interpret the document as linked data.

Each document in a Teanga corpus is associated with an identifier that ensures that the document content is valid. This check means that if the text content is changed, we can detect this and not proceed with annotations that have become invalid. The methodology for deducing the identifier is as follows: each document is indexed by initial characters the Base64 encoding of the SHA-256 of the UTF-8 representation of the text. The text representation consists of all character layers ordered alphabetically by their key with the key appended before the text. Keys and text should be separated by a zero byte (Unicode 0000). In most cases, the key should be at least four characters long and should be the shortest representation that is unique in the corpus. As such, it is not possible to have documents with duplicate text in a Teanga corpus. This can be avoided if necessary by adding an extra field with an identifier.

Finally, each corpus is associated with an *order* that gives the order of the documents in the corpus. This order is simply the list of identifiers in the document. This may be omitted in some serializations if the order is implicit in the serialization, however, if given it overrides the order in the serialized document.

3. Technical Implementation

The preferred method for representing Teanga corpora is as YAML documents. Teanga documents can also be easily represented as JSON documents. We provide a Python implementation of the Teanga model that ensures that models are easily serialized and provides useful features. In addition, a database model is provided for serializing and sharing the models as compact binary models. Further, we support the export of the model into RDF in both a generic method and methods compatible with NIF (Hellmann et al., 2013) and Web Annotation (Sanderson et al., 2017).

3.1. YAML form

The preferred serialization of Teanga is as YAML. Teanga YAML documents consist of a dictionary with one special key `_meta`, an optional second special key `_order` and the remaining keys consist of the document identifier and a dictionary of the layers in the document. The `_order` key is normally omitted in YAML as the order of documents

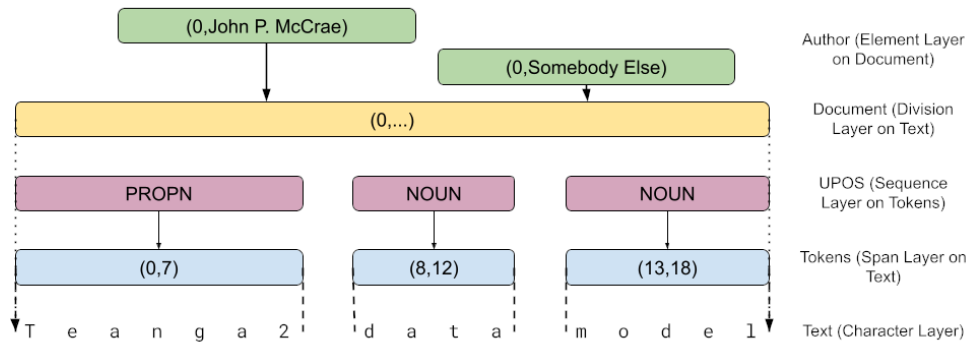


Figure 1: An example of the annotation layers of a Teanga document

in the text can be inferred by the order in the document. For example, a simple Teanga YAML document is as follows:

```
_meta:
  text:
    type: characters
Lz1r:
  text: This is an example document.
```

Each field of the metadata may have the following values

- `type`: The type of the layer (referencing mechanism)
- `base`: The name of the base layer (omitted for character layers)
- `data`: The data type. A list of values indicates an enumeration
- `target`: The target layer
- `link_types`: The enumeration of values used for links
- `default`: A default value for this layer if omitted
- `_uri`: The URI of the RDF property that documents this layer

The `_order` element of a corpus is a simple list of document IDs. It is not required in YAML and may instead be inferred from the order of the words in the document

Each non-character layer consists of a list of lists³, where each list consists of zero, one or two indexes and the data consisting of an optional integer for the target of a link and a string for the data or enumeration type. As such, each element may

³This efficient representation cannot be supported by JSON-LD

consist of one to five elements⁴. It is important to note that the indexes refer to the order of annotations in the base layer, not the absolute character index, unless the base layer is a character layer. For division layers, the index indicates the start index of each section. So for example to provide tokens and sentences we may have the following document.

```
_meta:
  text:
    type: characters
  tokens:
    type: span
    base: text
  sentences:
    type: div
    base: tokens
hDRz:
  text: Hello there! Goodbye!
  tokens: [[0, 5], [6, 11], [11, 12],
    ↪ [13, 20], [20, 21]]
  sentences: [0, 3]
```

It is important to note here that the indexes in the sentence layer are in terms of the tokens, so the second sentence starts from the 4th token (index=3) and this can be mapped into characters by reference to the token layer.

3.2. Python Implementation

Teanga is provided as a Python library on GitHub⁵ this library supports the basic operation of the library including adding and removing documents and updating metadata layers. In addition, it provides support for mapping indexes from a base layer to lower layers, which is a specific challenge as complex multi-layer annotations may make it difficult to reach the actual characters the annotations are referring to.

⁴Zero elements are allowed but meaningless

⁵<https://github.com/teangaNLP/teanga2>

In addition to providing a strong in-memory version of the Teanga data model, a secondary implementation⁶ in Rust using the Sled⁷ library provides a simple method for working with large corpora in Teanga. This persists large corpora to disk, allowing them to be searched and queried efficiently. A full interface is available for this in Python and as such there is minimal change to use this version of the interface. Due to this technology, it will be possible for Teanga to handle very large corpora, of the order of billions of tokens, and load and parse such corpora rapidly. We will further investigate this alongside tools for improving query time of the corpora based on use cases of the systems in future work focussed on these aspects.

3.3. Conversion to RDF

Support for conversion to linked data is a key goal of Teanga and it is expected that Teanga corpora could be used as targets for linking of other resources. If `_uri` properties are given for layers these can be used to map the resource to RDF. Each document in the model is given a URI based on its identifiers and these are included in the fragment identifier. So, for example, we can indicate an identifier as follows for a document available at <http://www.example.com/corpus.yaml>.

```
_meta:
  text:
    type: characters
    _uri: https://teanga.github.io/\
teangaNLP/teanga.rdf#text
hDRz:
  text: Hello there! Goodbye!
```

Is converted to Turtle as follows:

```
<http://www.example.com/corpus.yaml#hRDz>
  teanga:text "Hello there! Goodbye!"
```

Annotations in Teanga are modelled with the use of two special properties `teanga:idx`, `teanga:ref` and `teanga:data`⁸, which give the order of the annotation and a reference to the base layer and the data, respectively.

Following RFC 5147, references to text layers can be made with `char=` elements in the fragment, for example:

```
<#hRDz>
  ex:tokens [
    teanga:idx 0 ;
    teanga:ref <#hRDz&char=0,5>
  ] .
```

⁶<https://github.com/teangaNLP/teanga.rs>

⁷<https://docs.rs/sled/latest/sled/>

⁸The namespace `teanga` is defined as <https://teanganlp.github.io/teanga2/teanga.rdf>

References to any other layer can be made with `n=` fragment.

The default URI for a document is given by adding the Teanga document identifier to the URI, but can alternatively be specified by giving a `_uri` property on the individual document.

In addition to this direct export to RDF using the Teanga RDF vocabulary, it is also possible to export to NIF and WebAnnotation style vocabularies. The RDF generated in these exports is generally more verbose than the Teanga RDF model. For example, the NIF export looks like this:

```
<#hRDz&char=0,5> a
  nif:OffsetbasedString ;
  nif:anchorOf "Hello" ;
  nif:beginIndex 0 ;
  nif:endIndex 5 ;
  rdf:value ex:tokens .
```

Similarly, the annotation in the WebAnnotation model is as follows in JSON-LD:

```
{
  "@context":
    "http://www.w3.org/ns/anno.jsonld",
  "id": "#anno_1",
  "type": "Annotation",
  "body": {
    "value": {
      "@id":
        "http://www.example.com/tokens"
    }
  },
  "target": {
    "source": "#hRDz",
    "selector": {
      "type":
        "TextPositionSelector",
      "start": 0,
      "end": 5
    }
  }
}
```

Note that we use the fully expanded URI for `ex:tokens` in this example.

4. Examples

We present three examples of NLP data and how they can be represented in Teanga by means of examples. Conversion tools for these formats are already published or under development.

4.1. CoNLLU Data

CoNLLU data format is the representation of the linguistic data developed to train the dependency parser once at a time for many different languages.

The annotation of the CoNLLU is encoded in plain text format where a break line or LF character is used for representing a new line. The data has three different types of lines: - comment line: This line represents any sentence-level comments. It is represented with '#' and is usually at the beginning of the sentence. - token/words: This line contains the annotation of a word/token/node in 10 fields separated by single tab characters - newline: This is a blank line at the end of each sentence, which indicates the sentence boundary.

In the Teanga model, we include the UD data with the same annotation features; however, the annotation representation starts with the character level. As described in Section 3.1, we convert the CoNLLU data in Teanga model format, which is represented in YAML format as shown in the following example.

This is converted to Teanga as follows. We note that the header information is fixed and as such the document has a similar size without the header⁹

```
_meta:
  text:
    type: characters
  tokens:
    base: text
    type: span
  comm:
    base: text
    type: characters
    data: string
  upos:
    data: ["ADJ", "ADP", "ADV",
           ↪ "AUX", "CCONJ", "DET",
           ↪ "INTJ", "NOUN", "NUM",
           ↪ "PART", "PRON", "PROPN",
           ↪ "PUNCT", "SCONJ", "VERB",
           ↪ "X" ]
    base: tokens
    type: seq
sjKY:
  text: _Bhojpuri text_
  tokens: [[0, 7], [8, 9], [10, 20],
           ↪ [21, 24], [25, 26]]
  comm: lokaramjana ā sāṃskrtika
           ↪ gīta -
  upos: ["NOUN", "CCONJ", "ADJ",
           ↪ "NOUN", "PUNCT" ]
u40k:
  text: _Bhojpuri text_
  tokens: [[0, 3], [4, 5], [6, 13],
           ↪ [14, 17], [18, 19]]
  comm: āīm ā saporivāra āīm .
```

⁹Note the original example uses Devanagari script but these could not be reproduced in the PDF and have been replaced with 'Bhojpuri text'. This is not a limitation of Teanga.

```
upos: ["VERB", "CCONJ", "NOUN",
       ↪ "VERB", "PUNCT"]
```

In the above example ¹⁰, we can see that each sentence of the text file is tokenised at the character level and consists of only text. The rest of the features, including the ten fields, are categorised in the span layers as shown in `upos`¹¹; similarly, the other morphological features will be included in the span layer. As we tokenize the text into character, there is no need to include the newline to show the sentence boundary, as in CoNLLU data. The tanga format also makes it easier to extract each feature of the text through the span layers. For example, if a task needs to use only parts of speech information of the given text, then the user can easily extract only the upon layer of the text rather than the whole document.

4.2. TEI Conversion

TEI tags can be used to annotate a variety of text features, as well as information about a text:

```
<div type="prose">
  <p>
    <supplied>B</supplied>ui
    ↪ oeng<expan>us</expan>
    ↪ hindaidqi naile inachotlud
    ↪ confacca ni hinningin chuici
    ↪ <expan>ar</expan> cranssiuil
    ↪ do.
  </p>
</div>
```

When a TEI-encoded text is converted to Teanga, the character layer consists of the text with all XML tags removed leaving only the text of the document as a single string of characters. The information which had been encoded using these TEI tags is instead preserved in a span layer according to the Teanga data model. Thus, the information represented in the TEI encoded text above may be represented in Teanga as follows:

```
_meta:
  text:
    type: characters
  div:
    type: span
    base: text
  div_type:
    type: element
```

¹⁰The example is taken from the Bhojpuri UD data https://github.com/UniversalDependencies/UD_Bhojpuri-BHTB/tree/master.

¹¹Note that it is possible to assign URIs to each of these values and as such, they can be mapped to other schemes such as OLiA or LexInfo


```

    base: div
    data: string
  p:
    type: span
    base: text
  supplied:
    type: span
    base: text
  expan:
    type: span
    base: text
7nkN:
  text: Bui oengus hindaidqi naile
    → inachotlud confacca ni
    → hinningin chuici ar crannsiuil
    → do.
  p: [[0, 84]]
  div: [[0, 84]]
  div_type: [[0, "prose"]]
  supplied: [[0, 1]]
  expan: [[8, 10], [67, 69]]

```

Aside from formatting tags like `<div>` or `<p>`, TEI annotation can also be used to preserve specific information about small portions of a text, often at word-level or smaller granularity. Repositories containing historical texts, for example, may use TEI tags to identify snippets of digital text which were added or changed by modern editors, but which were not present in an earlier manuscript. The example above, which was taken from *Thesaurus Linguae Hibernicae* (Kelly et al., 2006), uses `<supplied>` tags to identify text which has been supplied by the editors, and `<expan>` tags to show where manuscript abbreviations have been expanded. As with word-level annotations, this kind of information is captured by the Teanga data model in the span layer, though such annotations often apply to portions of text at a sub-word level. We also see how attributes of XML elements are mapped to layers that are dependent on the tag annotation, for example, the `div_type` layer represents the `type` attribute of the `div` tag.

```

<title>The Sign of Four</title>
<author>Doyle, Arthur Conan
  → (1859-1903).</author>
<publisher>London: Spencer
  → Blakett</publisher>
<date>1890</date>
<ref target=
"http://archive.org/detail..."/>

```

Where TEI tags are used to annotate metadata which is unrelated to any specific span of text, for example, information pertaining to authorship or publishing (as shown above), this can be preserved in the Teanga data model at the document level. This is done by creating a `document` layer which refers to all information in the text.

```

_meta:
  text:
    type: characters
  document:
    type: div
    base: characters
    default: [[0]]
  title:
    type: seq
    base: document
abcd:
  text: "... "
  title: ["The Sign of Four"]
  author: ["Doyle, Arthur Conan
    → (1859-1903)."]

```

Note that by specifying the default of the `document` layer as `[[0]]`, we give a default value of a division that starts at character 0 and ends at the end of the document. More complex linguistic annotations, such as those found in Level 2 of EL-TeC (Schöch et al., 2021), can be modelled in a similar manner to what is done in the UD example in section 4.1.

4.3. Parallel Texts

Word alignment is the task of assigning words from one sentence (the source sentence) to words in a target sentence when given two parallel sentences that are translations of each other. Typically, the datasets for this task are distributed as bitext corpus, for example, a few sentences extracted from the Spanish-English Europarl v7 corpus:

```

¿Hay alguna objeción ? ||| Are there
  → any comments ?
Muchas gracias ||| Thank you very
  → much .
Apruebo esta petición. ||| I agree
  → with this request.

```

When converting parallel sentences into Teanga, each sentence is represented through a character layer, with one layer for the source language and another for the target language. Following this, since tokenization is required for the alignment task, each character layer is annotated with a token span layer. Ultimately, the word alignments are annotated as an element-linking layer, connecting an element in the source tokens layer to a corresponding element in the target tokens layer.

```

_meta:
  align:
    type: element
    base: en_tokens
    data: link
    target: de_tokens

```

```

en:
  type: characters
en_tokens:
  type: span
  base: en
es:
  type: characters
es_tokens:
  type: span
  base: es

```

8I9N:

```

es: ¿Hay alguna objeción?
en: Are there any comments?
es_tokens: [[0, 4], [5, 11],
  → [12, 20], [20, 21]]
en_tokens: [[0, 3], [4, 9], [10,
  → 13], [14, 22], [22, 23]]
align: [[0, 0], [1, 2], [2, 3],
  → [3, 4]]

```

JmZn:

```

es: Muchas gracias.
en: Thank you very much.
es_tokens: [[0, 6], [7, 14],
  → [14, 15]]
en_tokens: [[0, 5], [6, 9], [10,
  → 14], [15, 19], [19, 20]]
align: [[0, 2], [0, 3], [1, 0],
  → [1, 1], [2, 4]]

```

SQ/9:

```

es: Apruebo esta petición.
en: I agree with this request.
es_tokens: [[0, 7], [8, 12],
  → [13, 21], [21, 22]]
en_tokens: [[0, 1], [2, 7], [8,
  → 12], [13, 17], [18, 25],
  → [25, 26]]
align: [[0, 0], [0, 1], [0, 2],
  → [1, 3], [2, 4], [3, 5]]

```

In this example, we highlight the capability of Teanga to link elements in different layers of annotation, facilitating the representation of linked data in a coherent and interconnected manner, demonstrating its broader potential in handling complex annotation relationships within parallel texts. Further, we note that sentence alignment can also be modelled the same way if sentence annotations are available as in Section 3.1

5. Related Work and Discussion

Teanga is a new model for annotated corpora that aims to be able to represent all kinds of natural language processing data in a single, consistent manner. The most widely used formats for annotated corpora are either limited models, such as CoNLL, which can only represent token-level annotations

and links between tokens (dependency parses) and as demonstrated Teanga can represent these kinds of data in a manner that is not substantially more verbose than these specific formats. As such, Teanga is a flexible data model that can be parsed without the need for external libraries except for a YAML parser which is widely available (although the Teanga library provides some additional features). Thus, this avoids the development of custom extensions of formats such as CoNLL (Chiarcos and Glaser, 2020; Graën et al., 2019), which requires the development of new parsers and avoids the risk of using proprietary formats that may be hard to access in the future.

The most widely used model that allows for general annotation of a corpus is TEI (Ide, 1994), however, this is a model based on XML and as such is destructive of the original text content. Further, extensions on TEI are not easy to write and the interface with RDF and linked data is not clear (Burrrows et al., 2021). Also, as demonstrated Teanga is able to efficiently and correctly represent complex annotations found in TEI.

The Teanga data model is more closely related to attempts to create linked data corpus models. Two of these models have risen to particular prominences. Firstly, the NLP Interchange Format introduced by Hellmann et al. (2013) has seen adoption for tasks such as named entity recognition (Röder et al., 2014), question answering (Latifi and Sánchez-Marré, 2013) and frame semantics (Alexiev and Casamayor, 2016). However, this model proves very verbose in practical applications and the project is not actively maintained anymore, with version 2.0 of the model being released in 2013 and very few updates in any of the provided tooling since 2016.

The Web Annotation data model (Sanderson et al., 2017), was introduced as a model for annotating documents on the web using RDF. Web annotations consist of annotations that link bodies with targets. The body can be either a literal value or structured content and is used to give the value of the annotation. The targets can be selected by various methods including character offsets, as well as through mechanisms such as XPointer (for XML documents). This annotation is used by the INCEpTION (Klie et al., 2018) platform for annotating documents. Teanga annotations are exportable to Web Annotation, however, the format is generally much more verbose than the Teanga model.

6. Conclusion

In this paper, we have presented a data model for Teanga, a new framework for NLP based on the previous Teanga model (Ziad et al., 2018). The

layer annotation model proposed by this model allows the representation of all NLP-relevant corpus data and does so in a manner that is efficient and readable. Further, this framework integrates with linked data, both as a linked data format in its own right and also by exporting to other RDF serializations such as Turtle and JSON-LD. This data model will simplify the publishing corpora as linked data, by providing tooling and a self-documenting format that satisfies FAIR principles.

7. Acknowledgements

This work has been supported by Science Foundation Ireland under Grant Number SFI/12/RC/2289_P2 Insight_2, Insight SFI Centre for Data Analytics.

8. Bibliographical 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.
- Toby Burrows, Matthew Holford, David Lewis, Andrew Morrison, Kevin Page, and Athanasios Veliou. 2021. Transforming TEI manuscript descriptions into RDF graphs. *Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing*, 15:143.
- 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, May 27-31, 2012. Proceedings*, volume 7295 of *Lecture Notes in Computer Science*, pages 225–239. Springer.
- Christian Chiarcos and Luis Glaser. 2020. [A tree extension for CoNLL-RDF](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7161–7169, Marseille, France. European Language Resources Association.
- Philipp Cimiano, John P. McCrae, and Paul Buitelaar. 2016. [Lexicon model for ontologies: Community report](#). Technical report.
- Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. [Universal dependencies](#). *Comput. Linguistics*, 47(2):255–308.
- Kerstin Eckart. 2012. [A standardized general framework for encoding and exchange of corpus annotations: The linguistic annotation framework, LAF](#). In *11th Conference on Natural Language Processing, KONVENS 2012, Empirical Methods in Natural Language Processing, Vienna, Austria, September 19-21, 2012*, volume 5 of *Scientific series of the ÖGAI*, pages 506–515. ÖGAI, Wien, Österreich.
- Johannes Graën, Tannon Kew, Anastassia Shaitarova, and Martin Volk. 2019. [Modelling large parallel corpora: The zurich parallel corpus collection](#). In *Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache.
- 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, October 21-25, 2013, Proceedings, Part II*, volume 8219 of *Lecture Notes in Computer Science*, pages 98–113. Springer.
- Nancy Ide. 1994. [Encoding standards for large text resources: The text encoding initiative](#). In *15th International Conference on Computational Linguistics, COLING 1994, Kyoto, Japan, August 5-9, 1994*, pages 574–578.
- Patricia Kelly, Niall Brady, and Hugh Fogarty. 2006. [TLH: Thesaurus Linguae Hibernicae](#). Online Resource.
- 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 *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico. Association for Computational Linguistics.
- Majid Latifi and Miquel Sànchez-Marrè. 2013. [The use of NLP interchange format for question answering in organizations](#). In *Artificial Intelligence Research and Development - Proceedings of the 16th International Conference of the Catalan Association for Artificial Intelligence, Vic, Catalonia, Spain, October 23-25, 2013*, volume 256 of *Frontiers in Artificial Intelligence and Applications*, pages 235–244. IOS Press.
- 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*, pages 587–597.

- 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).
- Robert Sanderson, Paolo Ciccarese, and Benjamin Young. 2017. Web Annotation Data Model. W3C Recommendation. W3C Recommendation.
- Christof Schöch, Roxana Patras, Tomaž Erjavec, and Diana Santos. 2021. [Creating the european literary text collection \(eltec\): Challenges and perspectives](#). *Modern Languages Open*.
- Manu Sporny, Dave Longley, Gregg Kellogg, Markus Lanthaler, Pierre-Antoine Champin, and Niklas Lindström. 2020. JSON-LD 1.1: A JSON-based Serialization for Linked Data. W3C Recommendation. W3C Recommendation.
- Housam Ziad, John P. McCrae, and Paul Buitelaar. 2018. [Teanga: A linked data based platform for natural language processing](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

The Services of the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin

Marco Passarotti, Francesco Mambrini, Giovanni Moretti

Università Cattolica del Sacro Cuore

Milan, Italy

{marco.passarotti,francesco.mambrini,giovanni.moretti}@unicatt.it

Abstract

This paper describes three online services designed to ease the tasks of querying and populating the linguistic resources for Latin made interoperable through their publication as Linked Open Data in the LiLa Knowledge Base. As for querying the KB, we present an interface to search the collection of lemmas that represents the core of the Knowledge Base, and an interactive, graphical platform to run queries on the resources currently interlinked. As for populating the KB with new textual resources, we describe a tool that performs automatic tokenization, lemmatization and Part-of-Speech tagging of a raw text in Latin and links its tokens to LiLa.

Keywords: Latin, Linked Open Data, SPARQL

1. Introduction

Over the past two decades, the scientific community that focuses on Linguistic Linked Open Data (LLOD) has worked in two main closely connected directions. First, it has developed numerous vocabularies and ontologies for representing various types of linguistic (meta)data as Linked Open Data (LOD) (Khan et al., 2022). Secondly, these vocabularies and ontologies have been applied to (meta)data extracted from various linguistic resources for publishing them as LOD: the LLOD Cloud (Cimiano et al., 2020, 29-41)¹ provides a synoptic view of the resources published so far.

One challenge that the LLOD community must now address is to make the interoperable (meta)data of the resources easily accessible and fully exploitable. Such task is challenging as it must fit the needs and expertise of diverse user communities besides computer scientists and computational linguists. However, this challenge is unavoidable, especially because many semantic web technologies (like RDF, OWL or SPARQL) have a (not entirely undeserved) reputation of being too abstruse or hard to learn for the general public.

The current availability of projects like the LiLa Knowledge Base (KB)², which has published several lexical and textual resources for the Latin language as LOD, or, more in general, the increasing success of the LOD paradigm in the Digital Humanities communities (Khan et al., 2022, 991-2) has highlighted the need to enable also specialists from areas like Classics to access and query the resources, as well as to encourage the production of new LOD-compliant resources.

While developing LiLa, we built a number of ser-

vices to address such needs. After introducing the LiLa KB (Section 2), this paper describes those services, all developed as web applications with the backend managed via servlets and the interface developed using the React javascript framework. The source code for all applications is published in Github under an open-source license. As for querying the KB, we present an interface to search the collection of lemmas that represents the core of the KB (Section 3.1), and an interactive, graphical platform to run queries on the resources interlinked therein (Section 3.2). As for populating the KB with new textual resources, we describe a tool that performs automatic tokenization, lemmatization and Part-of-Speech tagging of a raw text in Latin and links its tokens to LiLa (Section 4). Finally, we draw some conclusion and sketch future works (Section 5).

2. The LiLa Knowledge Base

The LiLa Knowledge Base (Passarotti et al., 2020) achieves interoperability between linguistic resources for Latin by adopting a set of ontologies widely used to model linguistic information, as well as Semantic Web and Linked Data standards. Among the former, OLiA is used to model linguistic annotation (Chiarcos and Sukhareva, 2015), Ontolex-Lemon for lexical data (McCrae et al., 2017) and POWLA for corpus data (Chiarcos, 2012). As for the latter, the Resource Description Framework (RDF) is the data model used to describe information in terms of triples (McBride, 2004).

The architecture of the LiLa Knowledge Base is highly lexically-based, as it exploits the lemma as the most productive interface between resources and tools. Indeed, its core is the so-called Lemma

¹<https://linguistic-lod.org/llod-cloud>

²<https://lila-erc.eu>

Bank (Mambrini et al., 2023) (CIRCSE, 2019-2024), a collection of around 200,000 lemmas taken from the database of the morphological analyzer LEMLAT (Passarotti et al., 2017) and constantly extended. A `lila:Lemma`³ is a sub-class of `ontolex:Form`⁴, whose individuals are the inflected forms of a lexical item. In particular, the lemma is a form that can be linked to a `ontolex:LexicalEntry`⁵ via the property `ontolex:canonicalForm`⁶, which identifies the form that is canonically used to represent a lexical entry. To overcome divergent lemmatization criteria that may possibly be adopted in resources, LiLa exploits three key properties. The symmetric property `lila:lemmaVariant`⁷ connects different forms of the same lexical item that can be used as lemmas for that item, like for verbs with an active and a deponent inflection (e.g., *sequo* and *sequor* ‘to follow’). The property `ontolex:writtenRep`⁸ registers different spellings or graphical variants (called “written representations”) of one lemma, like for instance *conditio* and *condicio* ‘condition’. For forms that can be reduced to multiple lemmas like participles – that can be considered either part of the verbal inflectional paradigm or as independent lemmas – a special sub-class of `lila:Lemma` called `lila:Hypolemma`⁹ is defined.

The LiLa Knowledge Base has already a wide coverage in terms of interlinked resources, including corpora, and dictionaries. Among the former are the *Opera Latina* corpus by LASLA, which features 130 Classical Latin texts (Fantoli et al., 2022), and two dependency treebanks, namely the *Index Thomisticus* Treebank, which comprises texts by Thomas Aquinas (1225–1274) (Mambrini et al., 2022) (CIRCSE, 2006-2024), and the *UDante* treebank, which encompasses Medieval Latin works written by Dante Alighieri (Passarotti et al., 2021) (CIRCSE, 2021b). Among the latter are the bilingual Latin-English dictionary by Lewis and Short, whose primary focus is on Classical Latin (Mambrini et al., 2021a) (CIRCSE, 2021a), and the *Dictionary of Medieval Latin in the Czech Lands*, a lexical resource that collects the Latin vocabulary (pro-

vided with translations into Czech) as it emerged in Eastern Europe during the Middle Ages (Gamba et al., 2023) (CIRCSE, 2023a). Currently, the LiLa RDF graph includes a total of more than 80 million triples, which can be queried from the SPARQL endpoint of the KB, where a few ready-made queries are provided¹⁰.

3. Querying LiLa

This Section describes two services for querying, respectively: a) the Lemma Bank (3.1), and b) the textual resources and a selection of lexical resources currently linked to the LiLa KB (3.2).

3.1. The Lemma Bank Query Interface

The Lemma Bank query interface¹¹ allows users to interrogate the collection of Latin lemmas utilized in LiLa to interlink the linguistic resources published therein.

Relevant lemmas from the Lemma Bank can be selected based on various filters, including the lemma string, the presence of a specific affix (either prefix or suffix), the connection with a lexical base, the gender (for nouns), the part of speech (PoS), and the inflectional category. The lemma string search is performed by entering the desired string in a free text-box that supports regular expressions. The values for the other filters are provided through a dropdown menu.

The Lemma Bank query interface was designed to keep the search for lemmas as light as possible, by breaking down the query into blocks. Such query decomposition ensures that the minimum number of null results is obtained, by recalculating dynamically the values of all the fixed-value boxes every time the user adds a value in the query. For instance, if the lemmas of the verbs of the second conjugation are selected, the system cascades a series of SPARQL queries that update the values of the fixed-value boxes (i.e., prefix, suffix, gender, PoS, and inflectional category) only with those values that are compatible with the lemmas of the verbs of the second conjugation. The results of the query are then obtained by concatenating the selected values into a single SPARQL query, which can be downloaded.

Results are presented in the form of an alphabetically ordered list of lemmas, which can be downloaded along with the SPARQL query that produced it. For each lemma in the list, its written representation(s) and its PoS are shown, followed by two kinds of icons:

³<https://lila-erc.eu/lodview/ontologies/lila/Lemma>

⁴<http://www.w3.org/ns/lemon/ontolex#Form>

⁵<http://www.w3.org/ns/lemon/ontolex#LexicalEntry>

⁶<http://www.w3.org/ns/lemon/ontolex#canonicalForm>

⁷<http://lila-erc.eu/ontologies/lila/lemmaVariant>

⁸<http://www.w3.org/ns/lemon/ontolex#writtenRep>

⁹<https://lila-erc.eu/lodview/ontologies/lila/Hypolemma>

¹⁰<https://lila-erc.eu/sparql/>

¹¹<https://lila-erc.eu/query/>;
https://github.com/CIRCSE/LiLa_LB_QueryInterface.

- if the lemma is linked to a lexical entry of (a) the derivational lexicon *Word Formation Latin* (Pellegrini et al., 2021) (CIRCSE, 2018), (b) a manually checked subset of the *Latin Word-Net* enhanced with valency information taken from the *Latin Vallex* lexicon (Mambrini et al., 2021b) (CIRCSE, 2020b) (CIRCSE, 2023b), (c) the *LatinAffectus* polarity lexicon (Sprugnoli et al., 2020) (CIRCSE, 2020a), or (d) the Lewis and Short dictionary, an icon for each of these resources opens a window that provides an overview of the information reported by the lexical entry for the lemma in the resource selected (e.g., the derivational cluster of the lemma from *Word Formation Latin*);
- two icons show the triples connected to the selected lemma in the LiLa KB, respectively presenting the triples in a datasheet and in a network-like graphical representation, where nodes are individuals (e.g., the lemma) and edges are properties connecting individuals¹².

Figure 1 shows the datasheet for the verb *admiror* ‘to admire’, presenting the triples where *admiror* is in the domain (i.e., it is the subject of the property). Among the information shown in the datasheet is that the lemma: (a) has 2 written representations (*admiror* and *ammirror*), (b) pertains to the lexical base of *mirus*, which connects the lemmas in the Lemma Bank that share this base (property `lila:hasBase`¹³), (c) is a first conjugation verb (property `lila:hasInflectionType`¹⁴), (d) is formed with the prefix *ad-* (property `lila:hasPrefix`¹⁵), and (e) has as lemma variant the first conjugation not deponent form *admiro* (property `lila:lemmaVariant`¹⁶).

In the bottom of the datasheet, the inverse relations for the lemma are shown, namely those where the lemma is in the range (i.e., it is the object of the property). These are the cases where the lemma is linked to: (a) a lexical entry in a lexical resource (property `ontolex:canonicalForm`), (b) an hypolemma (property `lila:isHypoLemma`¹⁷), (c) a lemma variant (property `lila:lemmaVariant`), or (d) a token in a textual resource (property

¹²Graphical representations are shown using the LodiLive navigator (Camarda et al., 2012).

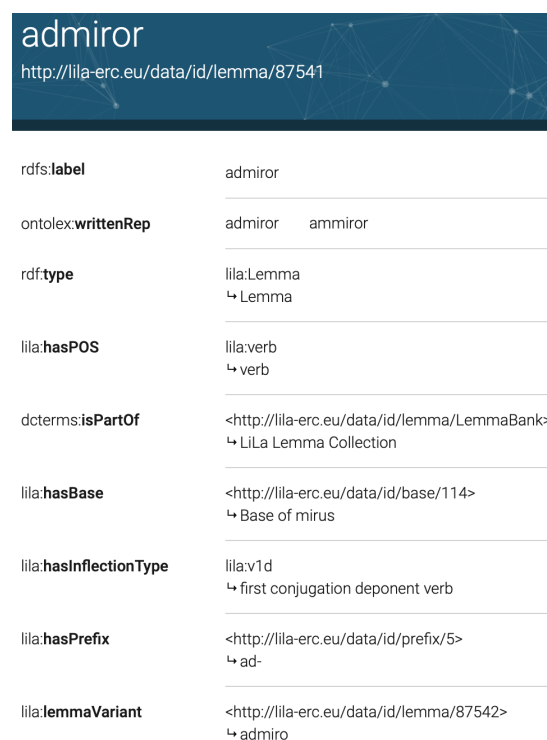
¹³<http://lila-erc.eu/ontologies/lila/hasBase>

¹⁴<http://lila-erc.eu/ontologies/lila/hasInflectionType>

¹⁵<http://lila-erc.eu/ontologies/lila/hasPrefix>

¹⁶<http://lila-erc.eu/ontologies/lila/lemmaVariant>

¹⁷<http://lila-erc.eu/ontologies/lila/isHypoLemma>



Property	Value
<code>rdfs:label</code>	admiror
<code>ontolex:writtenRep</code>	admiror ammirror
<code>rdf:type</code>	lila:Lemma ↳ Lemma
<code>lila:hasPOS</code>	lila:verb ↳ verb
<code>dcterms:isPartOf</code>	< http://lila-erc.eu/data/id/lemma/LemmaBank > ↳ LiLa Lemma Collection
<code>lila:hasBase</code>	< http://lila-erc.eu/data/id/base/114 > ↳ Base of mirus
<code>lila:hasInflectionType</code>	lila:v1d ↳ first conjugation deponent verb
<code>lila:hasPrefix</code>	< http://lila-erc.eu/data/id/prefix/5 > ↳ ad-
<code>lila:lemmaVariant</code>	< http://lila-erc.eu/data/id/lemma/87542 > ↳ admiro

Figure 1: The datasheet for *admiror*.

`lila:hasLemma`¹⁸). By clicking on the URI of a token linked to the lemma, its datasheet is shown, where also the sentence-based context of the token and its citation reference is provided.

3.2. The LiLa Interactive Search Platform

The LiLa Interactive Search Platform (LISP)¹⁹ is an interactive graphical interface to perform SPARQL queries on the textual resources and a subset of the lexical resources interlinked in the LiLa RDF triple store.

Like the Lemma Bank query interface, LISP relies on a SPARQL endpoint, although it works on a larger scale, performing searches on all the graphs present in the LiLa triple store. The interface of LISP was developed in react-js and it replicates the macro structure of the graphs of the resources interlinked in LiLa, representing graphically the connections between them via nodes (for the Lemma Bank and the resources) and directed edges (for their relations). Such network-like representation helps the user to select the nodes that make up the search and to visualize the various levels on which to act to refine the results of the query.

For example, to retrieve in a selection of the corpora interlinked in LiLa all the tokens of those lem-

¹⁸<http://lila-erc.eu/ontologies/lila/hasLemma>

¹⁹<https://lila-erc.eu/LiLaLisp/>;
https://github.com/CIRCSE/LiLa_LISP.

mas that feature certain properties reported by a corresponding entry in a specific lexical resource, LISP completes the path from the node for the tokens to that for the lexical resource in question. In particular, by applying a Depth-first search algorithm on the descriptive tree of the LiLa graphs, LISP adds the nodes for the Documents²⁰ and for the Lemma Bank along the path. Like for the Lemma Bank query interface, the values of each node restrict the configurable values of the others in the query. To reduce the amount of data obtainable by querying the entire LiLa triple store, each node contains only the instances of the class it represents. Then, each node executes a SPARQL query that recovers the data by concatenating backwards the descriptive SPARQL queries of all the nodes present in the generated tree.

On the left part of the screen, the platform features a few buttons organized in three areas. From top to bottom, they are the following:

- area for textual resources, which can be queried by Authors, Corpora, Documents, and Tokens;
- area for the Lemma Bank;
- area for lexical resources. Currently, it includes *Word Formation Latin*, *LatinAffectus*, *Latin WordNet*, the Lewis and Short dictionary and *Latin Vallex*.

LISP helps to combine information taken from different resources, by filtering their (meta)data, using the buttons from the three areas described above. For instance, by using the Documents, one can make a selection of the works (or sections of works) to query. Once works are selected, one can add information taken from a lexical resource, thus narrowing the query further. Typically, the last button to use is that of tokens, as it shows the list of tokens in the works selected that present the lexical properties taken from the lexical resources interlinked. As mentioned, the query is represented graphically in network-like fashion, showing the complete query path leading to tokens, according to the Lemma Bank based architecture of the LiLa KB.

Figure 2 shows the graphical representation of a query that searches in the documents whose authors are Catullus (taken from the LASLA corpus), Thomas Aquinas (from the *Index Thomisticus* Treebank), or Dante (from *UDante*). The node for the authors is linked to that for the tokens by the node for the Documents, which is connected to the lexical resources by passing through the Lemma Bank. The lexical resources provide lexical information to restrict furthermore the tokens to search. In

²⁰Documents are single works, or sections of works (e.g., books). Corpora are collections of Documents.

the example, two resources are used: from *Word Formation Latin* the deverbal verbs formed with the prefix *de-* are selected; from *Latin Vallex* those words that have an *Adresse* in at least one of their valency frames (passing through the node for the *Latin WordNet*, as the two resources share the lexical entries). This query results in 1,225 tokens, which LISP presents as an alphabetically ordered list, where each token is followed by the title of the work in which it occurs (see Figure 3). By clicking on a token, its datasheet is shown, where its full reference and a KWIC-like visualization is provided.

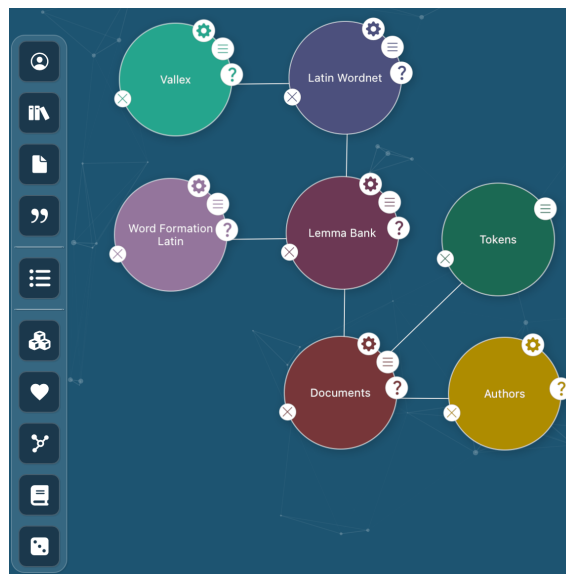


Figure 2: A graphical query in LISP.

Token	Document
Allegat	De Vulgari Eloquentia
Asserit	De Monarchia
Assumite	Epistole
Assummunt	De Monarchia
Assummunt	De Monarchia
Assumpta	De Monarchia

Figure 3: Results of a query in LISP.

4. Populating LiLa. The Text Linker

The LiLa Text Linker²¹ is a web application designed to assist users in the every step of the workflow to produce RDF editions of Latin texts fully

²¹<https://lila-erc.eu/LiLaTextLinker/>;
https://github.com/CIRCSE/LiLa_TextLinker.

integrated with the LiLa KB. The Text Linker integrates components to perform the text-processing stage, the manual editing and the creation of the RDF output.

The workflow starts from the raw text of a Latin work²². In the text-processing stage, after a minimal normalization step that takes care of spelling conventions such as the use of characters *u* or *j* for *v* and *i*, the input is lemmatized and PoS-tagged with the help of a custom model for the UDPipe (v.1.3) annotation pipeline (Straka and Straková, 2017). The ad-hoc model was trained on approximately 3,400,000 tokens, including data from 4 of the Latin treebanks distributed in Universal Dependencies (*Index Thomisticus* Treebank, *PROIEL*, *Perseus*, and *UDante*)²³, the *Opera Latina* published by LASLA, the Latin text database *Computational Historical Semantics*, which is part of the Latin Text Archive²⁴, and a series of lemmatized works curated by the CIRCSE, either published²⁵, or in publication²⁶. Data were harmonized as for both lemmatization criteria and PoS tagging, using the Universal PoS tagset (Petrov et al., 2011).

	Prec.	Recall	F1	AligndAcc
UPOS	94.02	94.02	94.02	94.02
Lemmas	93.70	93.70	93.70	93.70

Table 1: Performance of the ad-hoc model used by the LiLa Text Linker.

This corpus was randomly partitioned into a training (70%), development (20%) and test (10%) set. We evaluated the performances of the model on the test set. The results are reported in Table 1.

In a second step, the lemmatized tokens are matched against the lemmas in the LiLa KB. The Text Linker’s matching algorithm is set to be strict, returning only candidates whose lemma string and PoS-tag fully match the output of the annotation via the UDPipe model.

The result of the lemmatization and linking phase is returned to the users, who have the opportunity to perform any manual edits or correction that they desire. A screenshot of the interface is shown in Figure 4. The tokens in the text are coloured accord-

ing to the results from the previous stage: tokens that were matched with one single entry of the LiLa KB are visualized in green. Grey is used for tokens that were matched to more than one candidate; tokens in orange could not be matched.

By clicking on any linked token in the text, it is always possible to modify the automatic match by removing the suggested link and searching for candidates in the KB manually. In case of ambiguous matches (tokens in gray), it is also possible to select the appropriate candidate (or search for the right lemma by unlinking any of the proposed options), thus manually turning a 1:many match into a 1:1. Figure 4 shows an example of this process: the right pane of the interface shown in the screenshot is triggered by clicking on the ambiguous word *litora*, which is automatically assigned lemma *litus* and PoS NOUN²⁷. For all the matching lemmas in the LiLa KB, the interface displays a series of information (including the senses for the *Latin WordNet* and the Lewis and Short dictionary, if available). These data are retrieved via a chain of SPARQL queries to the LiLa triple store executed in the background. By selecting one of the lemma candidate, users have the opportunity to save the link. A pie chart on the top-right corner visualizes the statistics of the matching phase, showing the number of unique, ambiguous or missing matches; the counts are updated after any manual intervention of the editor.

The lemmatization and linking process can also be performed using a REST API for the service. The API returns a JSON output with the tokenized and sentence-split text. For each token, the output includes the PoS-tag, the lemma string produced by UDPipe and the list of URIs of all candidates for matching in the LiLa KB. It is also possible to use the API via the Language Resource Switchboard of the CLARIN consortium (Zinn, 2018), where the tool can be selected from the menu of the lemmatizers for Latin²⁸.

Once that the users are satisfied of the results, they can use the Text Linker to export the text as RDF. In order to generate a RDF serialization, the interface requires a series of metadata, which the users can enter by filling the short form shown in Figure 5.

²²At the moment, the application only accepts simple text (txt) as input. A future development could be to support also other formats and standards that are commonly used for digital editions, including in particular TEI-compliant XML. On TEI see <https://tei-c.org/>.

²³<https://universaldependencies.org/>

²⁴<https://lta.bbaw.de>

²⁵Augustine’s *Confessions* (<https://github.com/CIRCSE/AugustiniConfessiones>), Sabellicus’ *De Latinae Linguae Reparatione* (<https://github.com/CIRCSE/Sabellicus>).

²⁶Avianus’ *Fabulae*, Cicero’s *De Divinatione*.

²⁷There are three lemmas *litus* (NOUN) in the Lemma Bank: <http://lila-erc.eu/data/id/lemma/110686> (meaning: ‘a landing place’), <http://lila-erc.eu/data/id/lemma/62506> (meaning: ‘a servant’), and <http://lila-erc.eu/data/id/lemma/111141> (meaning: ‘a smearing’).

²⁸At the moment, the integration is still in progress, and the tool is only available in the testing interface of the Switchboard: <https://beta-switchboard.clarin.eu/>.



Figure 4: Correcting the lemmatization/linking output with the LiLa Text Linker.

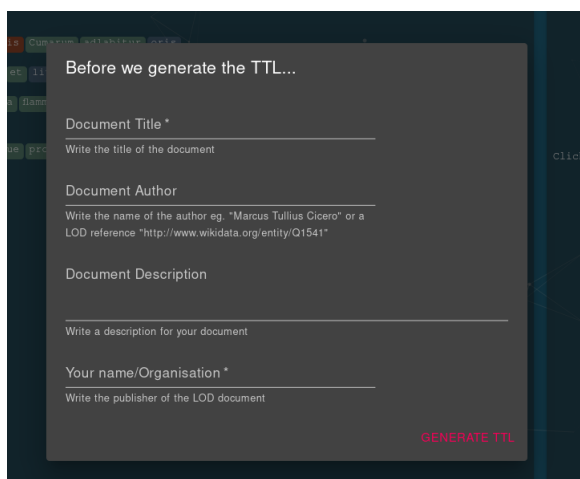


Figure 5: Metadata for the RDF output of the LiLa Text Linker.

5. Conclusion and Future Work

One of the main challenges for the LLOD world is to make fully exploitable the wealth of data and metadata from linguistic resources that, over the

last decade, has been made interoperable through the application of the principles of the Linked Data paradigm.

In this paper, we have presented some services of the LiLa KB, developed with the aim of enabling scholars to make the most out of the interactions between the Latin resources made available by the KB. Indeed, more specifically, the challenge concerns the impact that the computational treatment of linguistic data can and should have on Classical language studies. For this impact to occur, it is necessary for digital resources and computational analysis tools to be made more easily accessible, and for computational skills to be provided to humanists, especially classicists. The LiLa services described in this paper allow classicists to collect empirical results that could not be obtained previously. They represent a good showcase demonstrating the utility of interoperability between different linguistic resources. The hope is that classicists not only use the services but also strive to go beyond, becoming autonomous in both querying and publishing linguistic data.

To this goal, testing and improving usability is

a key factor. Of the three tools, the LiLa Text Linker has been demoed and showcased to a series of events for professionals in the Digital Humanities, including the *LinkedPast 6* workshop (2020),²⁹ and a dedicated tutorial at the 2nd Conference of the European Association for Digital Humanities (EADH21).³⁰ The other two, on the other hand, are still to be presented to the wider public. In the future, we intend to monitor the users more closely and to run usability tests for the interfaces involving representatives from the different communities of our target users.

Another important aspect that we want to explore is that of the adaptability of the software. The suite of tools that we presented here was designed specifically for the LiLa knowledge base; therefore, it is not ready to be used “out of the box” with data modeled according to other ontologies or structured differently from the LiLa paradigm. However, due to the way our tools were developed, we expect that only limited effort would be required to adapt the software to other projects, especially those that adopt the community standard Ontolex-Lemon. The fact that the tools work with linked data and are (mostly) based on interactions with a SPARQL endpoint is crucial in ensuring adaptability. More specifically, the LiLa query interface and the LiLa Lisp interface retrieve their data via SPARQL and can be re-modulated to query different triple stores. The LiLa Text Linker is the only tool that, at the moment, relies on an SQL database for reasons of efficiency; that application too, however, can be modified to interface with a triple store in order to increase its portability. Such aspects of portability must still be tested concretely, and any requirement for adapting the tools to different data must still be documented properly.

Beside linking new lexical and textual resources and keeping on expanding the coverage of the Lemma Bank, we also plan to update the trained model of the Text Linker, using a larger training set and version 2 of UDPipe. Furthermore, in LISP we will add access to further lexical resources, such as the *Lexikon der indogermanischen Verben* (Zimmer, 2002) (Boano et al., 2023) (CIRCSE, 2023c), by generalizing the query process that we already developed for querying similar resources. Indeed, so far the facets that describe the nodes used in LISP have been developed ad-hoc for each single resource included in the platform. However, in the near future, we expect to reuse the nodes as modeling templates for adding more resources.

²⁹<https://lila-erc.eu/linked-pasts-6-activity/>.

³⁰<https://lila-erc.eu/eadh-2021/>.

6. Acknowledgements

The “LiLa - Linking Latin” 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.

7. Bibliographical References

- Valeria Irene Boano, Francesco Mambrini, Marco Carlo Passarotti, and Riccardo Ginevra. 2023. Modelling and publishing the “lexicon der indogermanischen verben” as linked open data. In *Proceedings of CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30—Dec 02, 2023, Venice, Italy*, pages 1–7. CEUR-WS.
- Diego Valerio Camarda, Silvia Mazzini, and Alessandro Antonuccio. 2012. Lodlive, exploring the web of data. In *Proceedings of the 8th International Conference on Semantic Systems*, pages 197–200.
- Christian Chiarcos. 2012. Powla: Modeling linguistic corpora in owl/dl. In *Extended Semantic Web Conference*, pages 225–239. Springer.
- Christian Chiarcos and Maria Sukhareva. 2015. Olla—ontologies of linguistic annotation. *Semantic Web*, 6(4):379–386.
- Philipp Cimiano, Christian Chiarcos, John P. McCrae, and Jorge Gracia. 2020. *Linguistic Linked Data: Representation, Generation and Applications*. Springer, Cham.
- Margherita Fantoli, Marco Passarotti, Francesco Mambrini, Giovanni Moretti, and Paolo Ruffolo. 2022. Linking the lasla corpus in the lila knowledge base of interoperable linguistic resources for latin. In *Proceedings of the Linked Data in Linguistics Workshop@ LREC2022*, pages 26–34.
- Federica Gamba, Marco C Passarotti, and Paolo Ruffolo. 2023. Linking the dictionary of medieval latin in the czech lands to the lila knowledge base. In *Proceedings of CLiC-it 2023: 9th Italian Conference on Computational Linguistics*, pages 1–8. CEUR Workshop Proceedings.
- Anas Fahad Khan, Christian Chiarcos, Thierry Declerck, Daniela Gifu, Elena González-Blanco García, 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 data. *Semantic Web*, 13(6):987–1050.
- Francesco Mambrini, Eleonora Litta, Marco Passarotti, and Paolo Ruffolo. 2021a. 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)*, pages 214–220. Accademia University Press.
- Francesco Mambrini, Marco Passarotti, Eleonora Litta, and Giovanni Moretti. 2021b. [Interlinking Valency Frames and WordNet Synsets in the LiLa Knowledge Base of Linguistic Resources for Latin](#). In *Further with Knowledge Graphs*, volume 53 of *Studies on the Semantic Web*, pages 16–28.
- 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*, pages 4022–4029.
- Francesco Mambrini, Marco Carlo Passarotti, et al. 2023. The lila lemma bank: A knowledge base of latin canonical forms. *JOURNAL OF OPEN HUMANITIES DATA*, 9(28):1–5.
- Brian McBride. 2004. The resource description framework (rdf) and its vocabulary description language rdflits. In *Handbook on ontologies*, pages 51–65. Springer.
- 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.
- 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*, pages 24–31.
- Marco Passarotti, Flavio Massimiliano Cecchini, Rachele Sprugnoli, Giovanni Moretti, et al. 2021. Udante. l’annotazione sintattica dei testi latini di dante. *Studi Danteschi*, 86:309–338.
- 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(1):177–212.
- Matteo Pellegrini, Eleonora Litta, Marco Passarotti, Francesco Mambrini, and Giovanni Moretti. 2021. [The Two Approaches to Word Formation in the LiLa Knowledge Base of Latin Resources](#). In *Proceedings of the Third International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2021)*, pages 101–109, Nancy, France. ATILF.
- Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*.
- Rachele Sprugnoli, Francesco Mambrini, Giovanni Moretti, and Marco Passarotti. 2020. [Towards the Modeling of Polarity in a Latin Knowledge Base](#). In *Proceedings of the Third Workshop on Humanities in the Semantic Web (WHiSe 2020)*, volume 2695, pages 59–70, Heraklion, Greece. eur-ws.org.
- Milan Straka and Jana Straková. 2017. [Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipes](#). In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.
- Stefan Zimmer. 2002. *Lexikon der indogermanischen Verben (LIV)*. Walter de Gruyter GmbH & Co. KG.
- Claus Zinn. 2018. [Squib: The language resource switchboard](#). *Computational Linguistics*, 44(4):631–639.

8. Language Resource References

- CIRCSE. 2006-2024. [The Index Thomisticus Treebank](#). CIRCSE Research Centre, ISLRN [105-545-284-528-2](#).
- CIRCSE. 2018. [Word Formation Latin](#). CIRCSE Research Centre. PID <https://doi.org/10.5281/zenodo.1492327>.
- CIRCSE. 2019-2024. [The LiLa Lemma Bank](#). CIRCSE Research Centre. PID <https://doi.org/10.5281/zenodo.8300851>.
- CIRCSE. 2020a. [Latin Affectus](#). CIRCSE Research Centre. PID <https://doi.org/10.5281/zenodo.4022689>.
- CIRCSE. 2020b. [Latin Vallex 2.0](#). CIRCSE Research Centre. PID <https://doi.org/10.5281/zenodo.4032430>.

CIRCSE. 2021a. *Charlton T. Lewis and Charles Short. 1879. A Latin Dictionary. Clarendon Press, Oxford.* CIRCSE Research Centre. PID <https://github.com/CIRCSE/LewisShort>.

CIRCSE. 2021b. *UDante Treebank.* CIRCSE Research Centre. PID <https://github.com/CIRCSE/UDante>.

CIRCSE. 2023a. *Dictionary of Medieval Latin in the Czech Lands.* CIRCSE Research Centre. PID <https://github.com/CIRCSE/LexiconBohemorum>.

CIRCSE. 2023b. *Latin WordNet (revised version).* CIRCSE Research Centre. PID <https://doi.org/10.5281/zenodo.7561689>.

CIRCSE. 2023c. *Lexicon der indogermanischen Verben.* CIRCSE Research Centre. PID <https://github.com/CIRCSE/LIV>.

An Annotated Dataset for Transformer-based Scholarly Information Extraction and Linguistic Linked Data Generation

Vayianos Pertsas[◊], Marialena Kasapaki[◊], Panos Constantopoulos[◊]

[◊]Athens University of Economics and Business, Department of Informatics

[◊]Athena R.C., Institute for the Management of Information Systems

vpertsas@aueb.gr, kasapakimariael@gmail.com, panosc@aueb.gr

Abstract

We present a manually curated and annotated, multidisciplinary dataset of 15,262 sentences from research articles (abstract and main text) that can be used for transformer-based extraction from scholarly publications of three types of entities: 1) research methods, named entities of variable length, 2) research goals, entities that appear as textual spans of variable length with mostly fixed lexico-syntactic-structure, and 3) research activities, entities that appear as textual spans of variable length with complex lexico-syntactic structure. We explore the capabilities of our dataset by using it for training/fine-tuning various ML and transformer-based models. We compare our finetuned models as well as LLM responses (chat-GPT 3.5) based on 10-shot learning, by measuring F1 scores in token-based, entity-based strict and entity-based partial evaluations across interdisciplinary and discipline-specific datasets in order to capture any possible differences in discipline-oriented writing styles. Results show that fine tuning of transformer-based models significantly outperforms the performance of few-shot learning of LLMs such as chat-GPT, highlighting the significance of annotation datasets in such tasks. Our dataset can also be used as a source for linguistic linked data by itself. We demonstrate this by presenting indicative queries in SPARQL, executed over such an RDF knowledge graph.

Keywords: Information Extraction from Text, Transformer-based Information Extraction, Scholarly Annotation Corpus, Linguistic Linked Data, RDF Knowledge Graph

1. Introduction

The steep increase of research publications in every major discipline (Bommann et al., 2021) makes it increasingly difficult for experts to maintain an overview of their domain, increases the risk of missing new work or reinventing solutions, and makes it harder to relate ideas from different domains. To address this problem new “strategic reading” methodologies can be applied in order to transform the essence of knowledge encoded in textual form into structured format comprising concepts and relations that address the information needs of researchers, thus changing the ways in which they engage with literature (Renear & Palmer, 2009). This type of encoded information can alleviate the task of keeping up to date in a specific domain, while maintaining a bird’s-eye-view over a discipline or across disciplines, something particularly useful in interdisciplinary fields. To this end, entities representing the encoded information need to be appropriately identified and extracted from text through the use of various NLP and ML methods. This task has been significantly alleviated by the recent advancements in Deep Learning, where the application of transformer-based models in various NLP tasks (Vaswani et al., 2017) enabled the extraction of semantically complex information from text, while at the same time increased the demand for large annotated datasets for fine-tuning the millions of parameters of those models.

Indeed, information extraction (IE) from scientific papers has attracted a lot of interest over the past

years, as testified by the recent creation of various challenges on Scientific Information Extraction (ScienceIE). This constant challenge for new ML methods for ScienceIE calls for additional new datasets, capable of demonstrating and benchmarking the new capabilities of those methods.

In addition, despite the recent advancements in Large Language Models (LLMs) such as chat-GPT¹ and its remarkable ability to generate text that resembles human-like language, as demonstrated by numerous studies (Gao et al., 2023; Jimenez Gutierrez et al., 2022; X. Li et al., 2023; Ma et al., 2023; Qin et al., 2023; Qiu & Jin, 2024), when it comes to NLP tasks like IE and NER, these models underperform significantly compared to DL models that are finetuned in task specific annotated datasets, thus showcasing even more the significance of the latter in IE tasks.

In this paper we present such a manually curated dataset comprising of 15,262 sentences sampled from 3,500 research publications and 172 research subfields, that is specifically designed for extracting various types of entities of varied semantic complexities and lexico-syntactic characteristics. Specifically, we offer annotations for three different types of entities: 1) research methods, named entities of variable length, 2) research goals, entities that appear as textual spans of variable length with mostly fixed lexico-syntactic-structure, and 3) research activities, entities that appear as textual spans of variable length with complex lexico-syntactic structure.

¹ <https://chat.openai.com/chat>

The concepts in this dataset are designed to be general enough so that they can be applied across disciplines and, at the same time, be capable of representing essential knowledge of “who has done what, why and how” in a research paper. Extracting such information can lead to creating RDF Knowledge Graphs capable of answering complex semantic queries like: “find all papers that address a given problem”; “how was the problem solved”; “which methods are employed by whom in an activity addressing particular research goals”, etc. (Pertsas & Constantopoulos, 2023). This goes beyond the retrieval features of search engines widely used by researchers, such as Google Scholar², Scopus³ or Semantic Scholar⁴ that mostly leverage bibliographic metadata, while knowledge expressed in the actual text is exploited mostly by matching query terms to documents.

We explore the capabilities of our dataset along four dimensions: 1) *Classification Method*: we experiment with training/fine-tuning various ML and DL models as well as LLMs (chat-GPT 3.5) through prompting; 2) *Linguistic Characteristics*: we explore the performance of our methods across interdisciplinary and discipline-specific subsets in order to capture any possible differences in discipline-oriented writing styles as demonstrated in (Alluqmani & Shamir, 2018; Leong, 2024); 3) *Processing Granularity*: we test the effectiveness of classification at three levels of granularity: token-based, entity-based strict and entity-based partial. In addition, the included entities represent three levels of lexico-syntactic complexity: named entities of variable length, “non-named” entities (i.e. non real world objects that can’t be denoted with proper names) that are of variable length with mostly fixed lexico-syntactic-structure and variable length with complex lexico-syntactic structure; 4) *Linguistic Linked Data Generation*: we demonstrate the capabilities of our dataset as a source for linguistic linked data, through semantically complex queries in SPARQL that can be executed over such an RDF Knowledge Graph.

The rest of the paper proceeds as follows: in Section 2 we present related work regarding the creation of datasets for Science IE; in Section 3 we present the characteristics of our dataset and describe the methodology for its creation; in Section 4 we demonstrate the capabilities of the dataset through various experiments with ML, DL transformer-based and LLM prompting methods; in Section 5 we discuss the performance of the dataset based on the evaluation experiments and demonstrate its capabilities as a source for linguistic linked data and in Section 6 we conclude the paper with insights for future work.

2. Related Work

Information extraction from scientific text constitutes an active research field where ML and DL models are trained/fine-tuned on annotated corpora designed for

capturing specific knowledge according to the task at hand. Entity extraction is usually treated as a token classification or sequence labeling task where a classifier predicts whether each token belongs to the entity in question or not, based on the corresponding token-based annotations. In addition, recent advancements in LLMs have given rise to new methodologies regarding prompting techniques for interacting with these models based on few or even zero demonstrating examples in few / zero-shot learning (Brown et al., 2020; Das et al., 2022; Lu et al., 2022; Perez et al., 2021; X. Wei et al., 2023), while others implement chain-of-thought (CoT) reasoning (Ashok & Lipton, 2023; J. Wei et al., 2023) that can help in reasoning tasks such as solving mathematical problems, or works like (P. Li et al., 2023; Wang et al., 2023) that experiment with code generation. In our work, for comparison purposes, we include in our dataset experiments, a prompt template for LLMs (chat-GPT 3.5) that leverages both few-shot and code structure transformation.

Concerning the creation of datasets that can be used for IE, in domain specific fields like Biology and Bioinformatics, works like the BioText project (Rosario & Hearst, 2004) offer semantically annotated corpora, consisting of 3500 sentences drawn from MEDLINE abstracts labelled for *Disease* and *Treatment* and seven types of relation holding between them. In (Franzén et al., 2002; Kim et al., 2003) the Yapex and GENIA corpora offer annotated sentences with named entities of proteins and specific biological entities and events respectively. Regarding Medicine and Health Sciences, in (Roberts et al., 2009) the authors present a dataset from clinical texts, annotated with domain specific entities like *Condition*, *Investigation*, *Drug*, *Locus* etc. interrelated with relations: *has_target*, *has_type*, *location*, *modifies*. In (Borchert et al., 2022) the authors present a dataset of annotated named entities regarding Oncology (e.g. *Finding*, *Substance*, *Procedure*), which then evaluate using transformer-based models. In (Cheng et al., 2022) the authors present a manually annotated dataset from Japanese clinical reports with entities representing medical terms like *Diseases and Symptoms* and *Medicine*, as well as medical and temporal relations among them, which they evaluate using ML models. In Material Science, the authors of (Mullick et al., 2022) annotate a corpus with entities of type: *Code*, *Material*, *Method*, *Parameter* and *Structure* in order to train and evaluate their ML pipeline architecture.

In interdisciplinary ScienceIE projects, works like (Jain et al., 2020; Luan et al., 2018) present SciREC and SciREX, datasets from paper abstracts containing annotations of scientific entities (*Task*, *Method*, *Metric*, *Material*, *Other-ScientificTerm* and *Generic*). In (Qasemi, Zadeh & Schumann, 2016) a corpus of paper abstracts is manually annotated with terms classified into categories like *Method*, *Tool*,

² <https://scholar.google.com/>

³ <https://www.scopus.com/home.uri>

⁴ <https://www.semanticscholar.org/>

A classification analysis employing the unweighted paired group method using arithmetic
ACTIVITY METHOD
 average (UPGMA) was conducted in order to reveal the main zoogeographical zones.
GOAL

Figure 1: Example of Activity in passive voice, Method and Goal

For performing stylistic analysis we used the PCA method on 1000 samples of song lyrics.
GOAL ACTIVITY METHOD

Figure 2: Example of Activity in active voice, Method and Goal

Language Resource, Product, etc. In (Osenova et al., 2022) the authors present the Bulgarian Event corpus with annotations of named entities like *Locations, Events, Products*, etc. derived from the CIDOC-CRM Ontology and oriented mainly to Social Sciences and Humanities. In (Augenstein et al., 2017) the authors present a dataset with annotations of named entities like *Process, Task, Material* and relations like *hyponym-of* and *synonym-of*.

Compared to these works, we use a multidisciplinary dataset deriving from more than 170 research subfields in order to capture potential differences in writing styles among disciplines (Alluqmani & Shamir, 2018), since we use concepts that are general enough to be applied in any scientific field. In addition, to the best of our knowledge, our dataset is the first to contain entities of such lexico-syntactic complexity and variation in form and length. In this sense, it can be used for showcasing the capabilities of ML models in capturing various attributes of English language in a scholarly publication and not only those contained in a form of a named entity or an entity of relatively small length and fixed lexico-syntactic structure. The use of such semantically complex and -of highly variable length- entities, makes the problem of IE more challenging when it comes to employing prompting techniques for LLMs (as demonstrated in Section 4), thus showcasing the value of creating large, annotated datasets that can instead fine-tune DL transformer-based models with higher performance in such tasks.

3. Dataset Creation Methodology

For the creation of our dataset, we initially gathered a set of 25,681 papers spanning years 2000-2021 from JSTOR repository using the Constellate⁵ portal. This initial material after various NLP processes for OCR Noise removal, text cleaning, tokenization and sentence segmentation, yielded in total 3,700,000 cleaned sentences. From those, we randomly sampled a total of 15,262 sentences deriving from 3,500 papers which, according to articles' metadata (fields: "publisher" and "tdmCategory") were published under 352 different publishers and derived from 172 different disciplines and subfields. The dataset is in English language since this is most commonly used in academia. The aim was to create

a multidisciplinary corpus capturing as many different writing styles as possible.

The conceptual model behind the annotation schema is Scholarly Ontology (SO) (Pertsas & Constantopoulos, 2017), a domain-independent ontology of scholarly/scientific work. A specialization, in fact precursor, of SO already applied to the domain of Digital Humanities (that being an interdisciplinary field itself) is the NeDiMAH Methods Ontology (NeMO) (Constantopoulos et al., 2016). A brief overview of the definitions of SO concepts that were used in the annotation schema and guidelines is given below. For a full account see (Pertsas & Constantopoulos, 2017).

3.1 Annotation Schema

The Annotation schema used for the creation of this dataset was based on the following SO concepts and relations:

Activity: Instances of the Activity class represent research processes or steps thereof such as an experiment, a medical or social study, an archaeological excavation, etc. They usually manifest in text as spans of phrases in passive or active voice in first person singular or plural, according to the number of authors who are their actual participants.

Method: In contrast to activities, which are actual events carried out by actors, instances of the *Method* class denote procedures, such as an algorithm, a technique or a scheme that can be employed during an activity and describe how this was carried out. They are usually designated by single or multiple word terms, e.g. "ANOVA", "radio-carbon dating", etc., so their manifestations in text are mostly identified as named entities of variable length.

Goal: Goals represent the objectives of the activities and describe the intentional framework in which they were carried out. In addition, instances of the Goal class can represent general research goals of the paper that summarize the research objectives of all the activities described in it. In either case, they manifest in text as spans that declare purpose and are mostly introduced with purpose clauses like "for", "to" or "in order to".

⁵ <https://constellate.org/>

Indicative examples of all the above textual manifestations of SO classes and relations can be seen in Figures 1 and 2.

3.2 Annotation Process

The annotation process was based on protocols described in (Roberts et al., 2009) and involved a trial phase during which three annotators, after appropriate training in the SO concepts, participated in 5 consecutive annotation trials covering in total 500 sentences from 300 papers. Each trial was followed by review of the entire batch by the group, discussion on the results and differences among annotations, re-adjustment of the annotation guidelines and evaluation of the inter-annotator agreement (IAA) using the Cohen’s Kappa metric for IAA between annotator couples and Fleiss’ Kappa for the group of three. We used the Prodigy⁶ annotation tool for all the annotations and developed a Prodigy recipe for calculating the IAA scores.

After the trials, the best IAA scores reached 0.89 for *Activity*, 0.91 for *Method* and 0.92 for *Goal*, yielding sufficient agreement levels so that annotators could subsequently work on separate datasets. The entire annotator training process lasted approximately 25 hours.

As a general comment regarding the annotation of different types of entities, the most difficult type to agree upon was the *Activity* class. This can be attributed to the complexity of the lexico-syntactic structure of that particular entity type that produced differences among annotators, especially in the identification of boundaries in cases of very large lengths (compound phrases). On the other hand, *Methods* and *Goals* with clearer lexico-syntactic structures were easier to agree upon as can be seen from the higher agreement levels starting even from the first trial.

In addition to the annotation labels for the entities, the annotators used three “meta” labels for all the annotation sentences / spans: 1) *Accept*, where the annotator was confident for the annotation and the sentence/span is OK to be included in the dataset; 2) *Reject*, for the cases where the sentence/span was incomprehensible due to high noise from non-Unicode artifacts or non-English language and thus were to be excluded from the dataset; 3) *Ignore*, for the cases where the sentence/span was comprehensible but it wasn’t clear if the annotation fulfils the specifications of the task at hand. The latter were agreed to be included in the dataset, since they can provide valuable material for other experiments, but not to be counted for the experiments mentioned in this paper since they were considered as prone to create outliers due to their ambiguity. Nevertheless, these cases were very few, counting less than 3% of the entire dataset.

When the annotation task was completed, the entire dataset was adjudicated by one annotator in order to maintain a constant annotation style throughout the entire dataset. Analytical results (group IAA) for each annotation trial and entity/relation type that show the progress in the agreement of the annotation tasks are presented in Table 1.

	Trial1	Trial2	Trial3	Trial4	Trial5
Activity	0.69	0.73	0.78	0.81	0.89
Method	0.71	0.78	0.84	0.89	0.91
Goal	0.81	0.86	0.92	0.90	0.92

Table 1: IAA scores per entity type for each annotation trial

3.3 Dataset Statistics

The annotation statistics of the final dataset, after adjudication, are shown in Table 2. In total, the dataset comprises 15,262 sentences and 517,499 tokens. At sentence level, the dataset contains 10,754 labeled sentences (i.e. sentences that contain at least one label). At span level (as a span we consider each individual textual chunk that is annotated as an entity) there are in total 19,173 entity labels (i.e., labels assigned to spans to denote them as activities, methods or goals). At token level (as tokens we consider individual lexical units like words, punctuation marks, etc.) the dataset contains in total 192,087 labeled tokens (i.e. annotation labels assigned to tokens, to denote them as part of a textual span representing an activity, goal and/or a method). Compared to other published benchmarks in ScienceIE tasks (Augenstein et al., 2017; Jain et al., 2020; Luan et al., 2018; Qasemi, Zadeh & Schumann, 2016) our dataset shows similar or higher numbers of annotations, which renders it a good source for ground truth in such experiments. The annotated dataset in jsonl format can be accessed from GitHub⁷.

	Activity	Method	Goal	Total
Sent-level	6,610	6,028	4,029	10,754
Span-level	7,211	7,415	4,547	19,173
Token-level	126,702	14,036	51,349	192,087

Table 2: Dataset statistics for entity extraction

4. Experimental Setup

In order to evaluate the capabilities of the dataset in terms of how well it can fine-tune / train different types of ML models for performing the task at hand, we designed a total of 36 experiments measuring performance in entity extraction task.

4.1 Models and Methods

From the annotated dataset after random shuffling, we held out 20% for the evaluation set and the rest we split into training and development sets with the latter being 10% of the training set. We balanced our training sets but left unbalanced the evaluation sets so that we could measure performance in real case scenarios.

⁶ <https://prodi.gy/>

⁷ <https://github.com/athenarc/ScholarlyIE-Datasets/>

	Training/development			Total Evaluation set			H&B Subset			Humanities Subset		
	Act	Meth	Goal	Act	Meth	Goal	Act	Meth	Goal	Act	Meth	Goal
Sent	4,329	4,259	2,492	2,281	1,769	1,537	1,242	1072	699	1,008	658	889
Span	4,727	5,250	2,840	2,484	2,165	1,707	1,357	1,338	781	1,095	755	984
Token	84,469	9,716	32,237	42,233	4,320	19,112	24,577	2,706	9,665	15,944	1,489	10,237

Table 3: Number of annotated spans of the train/dev and eval subsets at sentence, span and token level.

From each of the following inputs included in "text" field, identify the textual spans representing entities that are defined as follows:

ACTIVITY: a research process like an experiment or a survey that is carried out by the author of the text.

GOAL: an objective of a research activity or a general research objective of the author of the text.

METHOD: the name of a research method denoting a procedure or a technique that was employed during an activity.

Your output should be in jsonl format containing the fields: "text" for the text in the input and "spans", a list of all the annotated spans each in a dictionary with the following fields: "start": character-based pointer to the start of the span, "end": character-based pointer to the end of the span, "token_start": token-based pointer to the start of the span, "token_end": character-based pointer to the end of the span, "label": the label of the annotated span, "span": the textual span of the annotation.

Below are some indicative examples that can be used as a guide.

Examples:

```
{
  "text": "A classification analysis employing the unweighted paired group method using arithmetic average (UPGMA) was conducted in order to reveal the main zoogeographical zones.",
  "spans": [
    {
      "start": 0,
      "end": 117,
      "token_start": 0,
      "token_end": 16,
      "label": "ACTIVITY",
      "span": "A classification analysis employing the unweighted paired group method using arithmetic average (UPGMA) was conducted"
    },
    {
      "start": 40,
      "end": 103,
      "token_start": 5,
      "token_end": 14,
      "label": "METHOD",
      "span": "unweighted paired group method using arithmetic average (UPGMA)"
    },
    {
      "start": 130,
      "end": 167,
      "token_start": 20,
      "token_end": 24,
      "label": "GOAL",
      "span": "reveal the main zoogeographical zones"
    }
  ]
}
```

....

....

Input:

```
{
  "text": "All bryophyte species were collected and identified in the laboratory with the aid of a stereo microscope and a light microscope (Leica DMLB, Leica Microsystems SAS, Rueil Malmaison, France)."
}
```

....

....

Figure 3: Indicative example of the prompt template. Each section is highlighted in different color.

In addition, in order to explore further possible differences in the writing styles of various disciplines, we created two subsets of the evaluation set: one with the sentences that were derived from papers in Humanities disciplines and another with sentences from papers in Health Sciences and Biology (H&B). Detailed statistics of the annotated entities contained in each subset are given in Table 3.

Regarding the entity extraction task, we used the above datasets to train / evaluate two different DL models for each entity: 1) a DL entity recognizer employing a Bert-base-NER transformer model that uses self-attention to process input sequences and generate contextualized representations of words in a sentence and 2) a DL entity recognizer employing a Roberta-base transformer, a variant of BERT, trained on a much larger dataset (10 times larger) and using a dynamic masking technique during training that helps the model learn more robust and generalizable representations of words. Both models came from the Hugging-Face library⁸ and were used for vector representation in combination with a transition-based parser for the sequence labeling part. For the latter we used the development set for hyperparameter optimization (dropout=0.1, Adam optimizer -L2=0.01). All of the transformer models and the transition-based parsers were fine-tuned / trained on the same datasets. These are the models **A-BERT-base-NER**, **A-RoBERTa-base**, for the extraction of Activities, **M-BERT-base-NER** and **M-RoBERTa-base** for the extraction of Methods and **G-BERT-base-NER**, **G-RoBERTa-base** for the extraction of Goals.

⁸ <https://huggingface.co/models>

In addition, for comparison reasons, we used the same dataset for training/evaluation of the spaCy default Named Entity Recognizer⁹ consisting (at the time of writing this paper) of a CNN with Bloom Embeddings that utilize a stochastic approximation of traditional embeddings in order to provide unique vectors for a large number of words without explicitly storing a separate vector for each of them (Miranda et al., 2022). These are the models **A-CNN**, **M-CNN**, **G-CNN**.

Furthermore, we designed a prompt template that leverages k-shot learning and text-to-structure capabilities of chat-GPT (GPT 3.5), in order to recast the structured output in the form of code instead of natural language. More specifically, we used the development set for experimenting with various combinations in prompt, such as different number of included examples (k=3,5,10,20), inclusion or not of the actual entity spans and inclusion or not of the reasoning for each entity extraction. Responses of the LLM into various prompt types during development stage showed that: i) describing the type of output in combination with specific examples helps the LLM to understand how to perform the output transformation and the classification task; ii) Although the increase in the number of examples helps performance, the added computational (and budget) costs from the larger prompts need to be taken into account when setting the threshold for the number of included examples (in our case k=10 proved to be a fair threshold); iii) using only the reasoning field without any demonstrating examples didn't contribute

⁹ <https://spacy.io/api/entityrecognizer>

	Humanities			Health & Biology			Total		
	Token	Partial	Strict	Token	Partial	Strict	Token	Partial	Strict
A-10-shot-GPT	48.53	42.66	12.37	69.64	44.91	15.17	56.17	45.32	13.25
A-CNN	64.99	61.12	47.21	71.46	65.71	50.51	68.15	63.06	48.36
A-Bert-base-NER	86.93	81.58	78.26	86.19	86.78	80.26	86.43	84.07	79.08
A-Roberta-base	88.10	84.36	79.67	89.26	88.00	81.06	89.01	86.62	80.06
G-10-shot-GPT	44.28	44.99	11.26	49.76	47.05	12.34	47.54	45.11	12.27
G-CNN	82.65	70.63	54.87	80.36	67.15	47.72	81.94	69.49	52.38
G-Bert-base-NER	86.99	80.68	71.63	87.12	78.61	66.29	86.98	79.97	69.51
G-Roberta-base	87.03	81.45	73.01	88.84	82.20	70.11	88.59	80.62	72.79
M-10-shot-GPT	40.03	33.96	18.87	43.89	34.03	19.19	43.11	34.31	19.74
M-CNN	74.41	72.86	64.85	76.33	74.49	66.95	75.54	73.75	65.84
M-Bert-base-NER	82.83	79.29	73.18	82.61	79.63	73.70	83.03	79.80	74.01
M-Roberta-base	83.59	80.47	75.10	83.61	80.60	74.43	83.79	80.81	74.97

Table 4: Evaluation results (F1 Scores). Prefixes A, G & M denote Activities, Goals & Methods respectively.

significantly to the overall performance increase (in comparison to adding more examples), as could be the case with other tasks like solving mathematical problems. Also it is to be noted that, similarly to (Fatemi & Hu, 2023), we experienced inconsistent performance across all experiments with variations in the output when the same input was repeated, even from a single account. Based on these observations, our proposed template consists of five sections: 1) description of the task at hand; 2) definitions of the entities, requested for extraction; 3) description of the requested output; 4) inclusion of 10 indicative examples for guidance; 5) input of the text to be annotated in the desired format. Using this template, the input is inserted as json lines (jsonl), each consisting of a dictionary containing the keys: "text" - with the actual text of the sentence and "spans" - a list of dictionaries, each containing the "label" denoting the type of the extracted entity, the entity span and pointers for the token-based and/or character-based entity boundaries, respectively. The LLM is enforced to recast the output in the same format, thus enabling easy integration with other workflows (through the Open AI API) and annotation tools such as Prodigy. The template is displayed in Figure 3. We used the same evaluation set in order to measure the performance of GPT 3.5 in the tasks at hand. These are the models **A-10-shot-GPT** for the extraction of Activities, **G-10-shot-GPT** for the extraction of Goals and **M-10-shot-GPT** for the extraction of Methods respectively.

4.2 Evaluation

The evaluation of Information Extraction methods involves comparing classifier results against a "gold standard" produced by human annotators. To this end, a confusion matrix is calculated based on the true positives (TP) -correctly classified predictions-, false positives (FP) -incorrectly classified predictions-, true negatives (TN) -correctly non-classified predictions and false negatives (FN) -incorrectly non-classified predictions. Performance scores are then

measured based on Precision (P), Recall (R) and F1 as usual.

For the entity extraction task, we conducted three types of evaluation experiments following the guidelines in (Segura-Bedmar et al., 2013) and using the `nerevaluate 0.1.8`¹⁰ and the `scikit-learn`¹¹ python libraries: 1) *token-based*, where a true positive (TP) is a token correctly classified as part of a chunk representing the entity, etc.; 2) *entity based -partial matching*, where some overlap between the tagged entity and the "golden" entity is required, but counts as half compared to the exact matches and 3) *entity-based -strict matching*, where only exact boundaries of the entities are counted for the match. Detailed results for all the evaluation experiments (reported here as F1 scores per entity type, classification method, evaluation method and dataset) are shown in Table 4.

5. Discussion

As a general remark regarding all the evaluation experiments, overall performance suggests that the dataset can be used adequately for finetuning DL models like transformers.

5.1 Classification Method

Regarding the performance of each methodology, fine-tuned transformer-based models showed superior performance in comparison to the rest of the models.

Specifically, compared to the CNN, higher performance was expected since transformer-based models can capture far more language attributes from the textual context and thus "understand" better the individual characteristics even for syntactically complex entity types.

Performance of the LLM was also inferior, something expected since, as demonstrated in (Gao et al., 2023; Jimenez Gutierrez et al., 2022; X. Li et al., 2023; Ma et al., 2023; Qin et al., 2023; Qiu & Jin, 2024), when it

¹⁰ <https://pypi.org/project/nerevaluate/>

¹¹ <https://scikit-learn.org/stable/>

comes to NLP tasks like IE and NER, these models underperform significantly compared to DL models like BERT that are finetuned in task specific annotated datasets. This situation is expected to become worse when it comes to the extraction of entities with more complex lexico-syntactic structures than standard named entities and of variable length, as is the case in our dataset. This is demonstrated in particular by the low performance in the entity-strict evaluations, where probably due to the aforementioned reasons and the lack of massive training data that is available in fine-tuning methods, the LLM failed to capture the exact boundaries of the spans. Nevertheless, LLM's performance in partial- and token-based evaluations suggests their potential use in distance learning techniques, since they can easily yield massive (but noisy) annotations that could further be manually corrected, or filter candidate sentences for annotation, thus easing the total annotation cost in time and effort.

Regarding the fine-tuned transformer-based models, the difference in performance among the RoBERTa and the BERT models can be attributed to the fact that the former is pretrained on much larger datasets and in a more efficient way than the latter. The high performance of transformer-based models, with F1 reaching up to 89.26 in "lenient" token-based evaluation and up to 81.06 in strict entity-based evaluation, is also evidence of the adequacy and quality of the annotations in our dataset for fine-tuning/training.

5.2 Linguistic Characteristics

Regarding the variations in performance with respect to the different discipline-focused evaluation subsets, the biggest differences appear in the extraction of activities (F1=88.00 in H&B compared to F1=84.36 in Humanities subset). Apart from the difference in the number of labeled tokens between the two subsets, which could lead to lower performance, visual inspection of the errors showed that in Humanities disciplines (e.g. in Archeology, History, Paleontology, etc.) there are a lot of mentions of historical events which, being events themselves, have textual descriptions that bear similar lexico-syntactic structures with those of research activities. Such cases, especially in passive voice with missing agent, are more difficult to discern. A similar situation arises in certain cases of research goals extraction, especially when these are goals of those "misclassified activities".

Based on visual inspection of more than 1000 sentences from the evaluation set and their comparison the rest of the dataset, the aforementioned cases could be considered as "extreme scenarios" of the dataset, since in these situations, the semantics for discerning a textual span representing a general activity or a goal (that are irrelevant of the research described in the paper) are not enough for the classifier to be able to make the correct prediction. Nevertheless, these errors could

probably be resolved with heuristics that analyze only specific sections of the paper (e.g. excluding related work, background, historical references sections, etc.).

5.3 Processing Granularity

Analyzing the results of each entity type, showed that the highest performance was achieved in Activity extraction. This can be attributed to the differences in the number of labeled tokens for each entity that follows the overall differences in performance. So, the extraction of Methods -having the fewest labeled tokens per sentence on average-, despite being the simplest of all, in terms of lexico-syntactic structure, yielded lower performance compared to Goals which, in turn, fared slightly lower than Activities.

Regarding the extraction of instances of the Goal class, analysis showed that, despite the fewer labeled tokens compared to activities spans, the overall good performance could be attributed to the fact that textual manifestations of goals have a concrete and consistent lexico-syntactic representation that allows for easier generalization of the corresponding DL models. Errors mainly occurred in cases of textual spans representing purpose that was not attributed to the author of the paper and thus should not be classified as a research goal according to SO definitions (e.g.: "The consortium's survey of East Los Angeles was one of the first holistic efforts to document historic and cultural resources in the community.").

Similar performance was also observed in the recognition of Methods. Analysis showed that the errors mainly occurred in cases of named entities other than methods, which, however, appear in similar textual contexts. For example, consider the sentence: "In May 2005 two of us traveled to the Angolan provinces of Namibe and Bengo, where we employed a geographic information system (GIS) to model the potential distribution of new species.". Here the tool: "geographic information system (GIS)" is erroneously annotated as a method by the classifier, probably due to the similar lexical form or the textual context of the sentence.

Regarding the extraction of textual spans referring to the Activity class, errors were observed in some instances of the Activity class in passive voice, not recognized as such by the classifier. For instance, in the sentence: "In this study carbon isotope discrimination was performed to assess the growing conditions of fossil cereal grains", the classifier failed to recognize the activity span. These errors could be attributed to the inclusion of negative training samples (i.e., cases of sentences in passive voice referring to historical events or activities not performed by the authors and thus not being annotated as activities) in the training set.

Regarding the variation in performance across different evaluation experiments (token-based, entity-based-partial, and entity-based-strict evaluations) it can be seen that the exact boundaries of the entity are difficult to capture even for the highest performing

models. Analysis indicates that such errors mostly occur in cases where one type of entity overlaps with another. E.g., “As a consequence of different growth behavior of trees in the juvenile phase, two different methods to **estimate the juvenile rings were used**.” Here, the boundaries of the enclosed entity were incorrectly detected, just like the tokens “were used” (in bold) were erroneously recognized as part of the goal span, although they are part of the overlapping activity. Other cases of erroneous boundary detection involved the inclusion of punctuation marks immediately following the entity inside the textual span. E.g., “To fulfill this purpose, we analyzed cranial discrete traits from this population.”. Especially concerning the Method class, such cases also involved the inclusion of information inside parentheses or brackets adjacent to the entities, probably due to the similarity in form with cases where the acronym of a method inside parentheses follows the method name (e.g., “Evaluation of Logistic Regression (P, R, F1) yielded good performance results...”, see also Fig. 1 for another example).

5.4 Linguistic Linked Data Generation

Apart from fine tuning transformer-based models for information extraction, this dataset can be used directly as a source for linguistic linked data by itself. Specifically, using the methodologies described in (Pertsas & Constantopoulos, 2023) the dataset can be transformed into an RDF Knowledge Graph (KG) adhering to Linked Data Standards. Such a KG can offer structured semantic views of the content of publications, which enhance our capability for comprehensive exploration of research work. This can be demonstrated through semantically complex queries executed over the KG. Indicative such queries, expressed in SPARQL are presented below:

Query 1: Retrieve all researchers that participate in activities or have research objectives that deal with linguistic analysis.

```
SELECT DISTINCT ?p_label
WHERE {
  ?p rdfs:label ?p_label
  ?p so:hasGoal / rdfs:label ?g_label
  ?p so:participatesIn / rdfs:label ?a_label
  filter contains(ucase(?g_label),?a_label),
  "linguistic analysis".}
```

Here, through the use of property chains in SPARQL and the *filter contains* SPARQL expression, all the methods employed in activities that have objectives with labels (i.e. textual spans) that contain the words “linguistic analysis” can be retrieved.

Query 2: For a specific paper (e.g. “Paper1”) retrieve all the research activities, conducted by the authors, along with their objectives and the methods they employed.

```
SELECT ?m_label ?a_label ?g_label
WHERE {
  ?a so:isDocumentedIn so:Paper1.
  ?a rdfs:label ?a_label.
  ?g so:isDocumentedIn so:Paper1.
  ?g rdfs:label ?g_label.
```

```
?m so:isDocumentedIn so:Paper1.
?m rdfs:label ?m_label. }
```

Here, the overall activity reported in a paper is decomposed into a series of activities denoting “what” the authors have done, associated with the methods they employed, and the goals they were trying to accomplish. Through this way, basic questions of “what”, “how” and “why” regarding information described in a research publication can be answered. Using such queries, the reader has access to an enhanced “bird’s-eye” view of what is described in a paper before actually reading it. Additional information regarding the authors, their research interests or the abstract can also be retrieved using the appropriate SO classes and relations.

6. Conclusion

In this paper we presented a manually curated dataset of 15,262 sentences in English, derived from 3,500 research articles (abstract and main text) and 172 different disciplines and subfields. The dataset contains in total 23,562 labels for three types of entities: 1) research methods, named entities of variable length, 2) research goals, entities that appear in text as textual spans of variable length with mostly fixed lexico-syntactic-structure, and 3) research activities, entities that appear as textual spans of variable length with complex lexico-syntactic structure.

We explored the capabilities of our datasets along four dimensions: 1) *Classification Method*: we experimented with training/fine-tuning various ML and DL models as well as LLMs (chat-GPT 3.5) through prompting; 2) *Linguistic Characteristics*: we explored the performance of our methods across interdisciplinary and discipline-specific subsets in order to capture any possible differences in discipline-oriented writing styles; 3) *Processing Granularity*: we tested the effectiveness of classification at three levels of granularity: token-based, entity-based strict and entity-based partial. In addition, the included entities represent three levels of lexico-syntactic complexity: named entities of variable length, “non-named” entities of variable length with mostly fixed lexico-syntactic-structure and variable length with complex lexico-syntactic structure; 4) *Linguistic Linked Data Generation*: we explored the capabilities of our dataset as a potential source for linguistic linked data through the use of SPARQL queries that can be executed over an RDF KG that can be created from it.

Evaluation scores showed high performance in all the experiments, especially with transformer-based models, showcasing the capabilities of our dataset in fine-tuning / training transformer models that can achieve very high results in entity extraction reaching up to F1=89.26 in “lenient” token-based evaluation and up to F1=81.06 in strict entity-based evaluation, even for entities of complex lexico-syntactic structure and variable length like the ones of research activities.

Future work includes expansion of our dataset with annotation of other types of entities and relations of the Scholarly Ontology concerning research publications. Specifically, we intend to provide annotations as well as trained DL models for the relations among SO entities, such as *employs(Activity,Method),hasObjective(Activity,Goal)* for interrelating the extracted activities with their corresponding methods and goals respectively, thus enhancing the produced linguistic linked data.

In addition, we intend to produce annotations for the research findings, arguments that describe various experiment results and interrelate them with their associated research activities that provide the supporting evidence or premise for those findings.

7. Bibliographical References

- Alluqmani, A., & Shamir, L. (2018). Writing styles in different scientific disciplines: A data science approach. *Scientometrics*, 115. <https://doi.org/10.1007/s11192-018-2688-8>
- Ashok, D., & Lipton, Z. C. (2023). *PromptNER: Prompting For Named Entity Recognition* (arXiv:2305.15444). arXiv. <http://arxiv.org/abs/2305.15444>
- Bornmann, L., Haunschild, R., & Mutz, R. (2021). Growth rates of modern science: A latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, 8(1), 224. <https://doi.org/10.1057/s41599-021-00903-w>
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., & Henighan, T. (2020). Language Models are Few-Shot Learners. *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, 1877–1901.
- Constantopoulos, P., Hughes, L. M., Dallas, C., Pertsas, V., & Christodoulou, T. (2016). Contextualized Integration of Digital Humanities Research: Using the NeMO Ontology of Digital Humanities Methods. *Digital Humanities 2016*, 161–163.
- Das, S. S. S., Katiyar, A., Passonneau, R., & Zhang, R. (2022). CONTaiNER: Few-Shot Named Entity Recognition via Contrastive Learning. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6338–6353. <https://doi.org/10.18653/v1/2022.acl-long.439>
- Fatemi, S., & Hu, Y. (2023). *A Comparative Analysis of Fine-Tuned LLMs and Few-Shot Learning of LLMs for Financial Sentiment Analysis* (arXiv:2312.08725). arXiv. <http://arxiv.org/abs/2312.08725>
- Gao, J., Zhao, H., Yu, C., & Xu, R. (2023). *Exploring the Feasibility of ChatGPT for Event Extraction* (arXiv:2303.03836). arXiv. <http://arxiv.org/abs/2303.03836>
- Jimenez Gutierrez, B., McNeal, N., Washington, C., Chen, Y., Li, L., Sun, H., & Su, Y. (2022). Thinking about GPT-3 In-Context Learning for Biomedical IE? Think Again. *Findings of the Association for Computational Linguistics: EMNLP 2022*, 4497–4512. <https://doi.org/10.18653/v1/2022.findings-emnlp.329>
- Leong, A. P. (2024). Marked Themes in academic writing: A comparative look at the sciences and humanities. *Text & Talk*, 0(0). <https://doi.org/10.1515/text-2022-0188>
- Li, P., Sun, T., Tang, Q., Yan, H., Wu, Y., Huang, X., & Qiu, X. (2023). CodeIE: Large Code Generation Models are Better Few-Shot Information Extractors. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15339–15353. <https://doi.org/10.18653/v1/2023.acl-long.855>
- Li, X., Chan, S., Zhu, X., Pei, Y., Ma, Z., Liu, X., & Shah, S. (2023). *Are ChatGPT and GPT-4 General-Purpose Solvers for Financial Text Analytics? A Study on Several Typical Tasks* (arXiv:2305.05862). arXiv. <http://arxiv.org/abs/2305.05862>
- Lu, Y., Bartolo, M., Moore, A., Riedel, S., & Stenetorp, P. (2022). Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8086–8098. <https://doi.org/10.18653/v1/2022.acl-long.556>
- Ma, Y., Cao, Y., Hong, Y., & Sun, A. (2023). Large Language Model Is Not a Good Few-shot Information Extractor, but a Good Reranker for Hard Samples! *Findings of the Association for Computational Linguistics: EMNLP 2023*, 10572–10601. <https://doi.org/10.18653/v1/2023.findings-emnlp.710>
- Miranda, L. J., Kádár, Á., Boyd, A., Van Landeghem, S., Søgaard, A., & Honnibal, M. (2022). *Multi hash embeddings in spaCy* (arXiv:2212.09255). arXiv. <http://arxiv.org/abs/2212.09255>
- Perez, E., Kiela, D., & Cho, K. (2021, May). *True Few-Shot Learning with Language Models*. <https://doi.org/10.48550/arXiv.2105.11447>
- Pertsas, V., & Constantopoulos, P. (2017). Scholarly Ontology: Modelling scholarly practices. *International Journal on Digital Libraries*, 18(3), 173–190. <https://doi.org/10.1007/s00799-016-0169-3>
- Pertsas, V., & Constantopoulos, P. (2023). Ontology-Driven Extraction of Contextualized Information from Research Publications: *Proceedings of the 15th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, 108–118. <https://doi.org/10.5220/0012254100003598>

- QasemiZadeh, B., & Schumann, A.-K. (2016). The ACL RD-TEC 2.0: A Language Resource for Evaluating Term Extraction and Entity Recognition Methods. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 1862–1868.
- Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., & Yang, D. (2023). *Is ChatGPT a General-Purpose Natural Language Processing Task Solver?* (arXiv:2302.06476). arXiv. <http://arxiv.org/abs/2302.06476>
- Qiu, Y., & Jin, Y. (2024). ChatGPT and finetuned BERT: A comparative study for developing intelligent design support systems. *Intelligent Systems with Applications*, 21, 200308. <https://doi.org/10.1016/j.iswa.2023.200308>
- Renear, A. H., & Palmer, C. L. (2009). Strategic Reading, Ontologies, and the Future of Scientific Publishing. *Science*, 325(5942), 828–832. <https://doi.org/10.1126/science.1157784>
- Segura-Bedmar, I., Martinez, P., & Zazo, M. H. (2013). SemEval-2013 Task 9: Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013). *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, 2, 341–350.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6000–6010.
- Wang, X., Li, S., & Ji, H. (2023). Code4Struct: Code Generation for Few-Shot Event Structure Prediction. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3640–3663. <https://doi.org/10.18653/v1/2023.acl-long.202>
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Findings of the Association for Computational Linguistics: ACL 2023*, 6519–6534. <https://doi.org/10.18653/v1/2023.findings-acl.408>
- Wei, X., Cui, X., Cheng, N., Wang, X., Zhang, X., Huang, S., Xie, P., Xu, J., Chen, Y., Zhang, M., Jiang, Y., & Han, W. (2023). *Zero-Shot Information Extraction via Chatting with ChatGPT* (arXiv:2302.10205). arXiv. <http://arxiv.org/abs/2302.10205>
- 8. Language Resource References**
- Augenstein, I., Das, M., Riedel, S., Vikraman, L., & McCallum, A. (2017). SemEval 2017 Task 10: ScienceIE - Extracting Keyphrases and Relations from Scientific Publications. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 546–555. <https://doi.org/10.18653/v1/S17-2091>
- Borchert, F., Lohr, C., Modersohn, L., Witt, J., Langer, T., Follmann, M., Gietzelt, M., Arnrich, B., Hahn, U., & Schapranow, M.-P. (2022). GGPONC 2.0—The German Clinical Guideline Corpus for Oncology: Curation Workflow, Annotation Policy, Baseline NER Taggers. *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, 3650–3660.
- Cheng, F., Yada, S., Tanaka, R., Aramaki, E., & Kurohashi, S. (2022). JaMIE: A Pipeline Japanese Medical Information Extraction System with Novel Relation Annotation. *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, 3724–3731.
- Franzén, K., Eriksson, G., Olsson, F., Asker, L., Lidén, P., & Cöster, J. (2002). Protein names and how to find them. *International Journal of Medical Informatics*, 67(1), 49–61. [https://doi.org/10.1016/S1386-5056\(02\)00052-7](https://doi.org/10.1016/S1386-5056(02)00052-7)
- Jain, S., Van Zuylen, M., Hajishirzi, H., & Beltagy, I. (2020). SciREX: A Challenge Dataset for Document-Level Information Extraction. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7506–7516. <https://doi.org/10.18653/v1/2020.acl-main.670>
- Kim, J.-D., Ohta, T., Tateisi, Y., & Tsujii, J. (2003). GENIA corpus—A semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl_1), i180–i182. <https://doi.org/10.1093/bioinformatics/btg1023>
- Luan, Y., He, L., Ostendorf, M., & Hajishirzi, H. (2018). Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3219–3232. <https://doi.org/10.18653/v1/D18-1360>
- Mullick, A., Pal, S., Nayak, T., Lee, S.-C., Bhattacharjee, S., & Goyal, P. (2022). Using Sentence-level Classification Helps Entity Extraction from Material Science Literature. *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, 4540–4545.
- Osenova, P., Simov, K., Marinova, I., & Berbatova, M. (2022). The Bulgarian Event Corpus: Overview and Initial NER Experiments. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 3491–3499. <https://aclanthology.org/2022.lrec-1.374>
- Roberts, A., Gaizauskas, R., Hepple, M., Demetriou, G., Guo, Y., Roberts, I., & Setzer, A. (2009). Building a semantically annotated corpus of clinical texts. *Journal of Biomedical Informatics*, 42(5), 950–966. <https://doi.org/10.1016/j.jbi.2008.12.013>
- Rosario, B., & Hearst, M. A. (2004). Classifying semantic relations in bioscience texts. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics - ACL '04*, 430-es. <https://doi.org/10.3115/1218955.1219010>

Linguistic LOD for Interoperable Morphological Description

Michael Rosner, Maxim Ionov

University of Malta, University of Cologne
mike.rosner@um.edu.mt, mionov@uni-koeln.de

Abstract

Interoperability is frequently cited as one important rationale underlying the use of LLOD representations and is generally regarded as highly desirable. However, the concept is generally taken for granted, and rarely analysed or exemplified. In this paper we attempt to remedy these shortcomings by concentrating on morphology, distinguishing three different kinds of interoperability which are relevant to that field. We provide practical implementations making extensive use of the vocabulary offered by Ontolex Morph.

Keywords: Morphological Resources, Interoperability, Linked Linguistic Open Data

1. Introduction

In general, interoperability is a characteristic of a product or system that seamlessly works with another product or system. For example, when you plug in your toaster to a wall socket, the two systems are (i) the toaster plug, and (ii) the power network that provides the toaster with electricity. The plug is interoperable in the sense that it works with any compatible socket. The advantage of this interoperability is that you can use the toaster anywhere. Conversely, we experience the problem of non-interoperable plugs when we go abroad. Typically, the solution is to use an appropriate *adaptor*, but for the well-travelled this means the additional burden of carrying an array of adaptors in order to guarantee world-wide functionality of your device. The moral here is clear: interoperability of the device implies a maximum degree of independence from the context of use.

Turning to the use of Linked (Open) Data representations, interoperability is frequently cited as the main underlying rationale. It is claimed that due to the use of open standards, such representations facilitate the use of resources in a variety of different contexts with zero or minimal adaptation, either in the resource itself or in the context that uses it. This is in contrast to hand-crafted representations which may require arbitrary adaptation to the resource to be used successfully.

Although these general considerations apply widely and are often mentioned, they are rarely analysed, exemplified, or specifically applied to the narrower scope of Linguistic LOD (LLOD). In this paper we propose to remedy this lacuna.

Some preliminary considerations reveal that when it comes to LRs, there exist other kinds of interoperability. Below we consider three ways of dividing up the landscape: *task interoperability*, *language interoperability*, and *domain interoperability*. In the following sections we take a closer look at each of these types, assess how well they can be ad-

dressed using LLOD and where it falls short.

1.1. Task Interoperability

The basic idea underlying task interoperability is that the same machinery is applied without modification to different tasks. There are essentially two kinds of LRs: data-oriented and process-oriented. Data-oriented resources express static facts e.g. an annotated text corpus, or a lexicon containing facts about the words of a language, whilst process-oriented ones (often referred to as tools), such as parsers, translators and chatbots, the focus is on the achievement of tasks or on explicit behaviours which are of interest to us. Importantly, the two are connected: process-oriented resources make use of data-oriented ones. Thus, a chatbot might make use of a lexicon to identify the current topic and determine its important characteristics.

In this paper we focus on the interoperability of data-oriented LRs. As our working example, we choose the lexicon because a lexicon is not only a clear example of a data resource but also one which is crucial for practically every NLP task. Examples are parsing; sentiment analysis; translation. Each of these tasks use a lexicon to associate information with words - but for each task the information is different. Thus for parsing, it concerns part-of-speech information; for sentiment analysis, sentiment values; and for translation, a translational equivalent in another language.

Given this diversity of information types associated with words, there is a tendency to create diverse representations, that for the sake of efficiency, require specialised access and processing procedures. This is a perfect scenario for developing a series of specialised lexicons, one for each task. SentiWordnet (Baccianella et al., 2010) for example, contains opinion information on terms extracted from WordNet, providing a database of term/sentiment information for English. Bilingual lexicons are crucial to the operation of translation

systems such as Apertium (Forcada et al., 2011), and DBnary (Sérasset, 2015), a multilingual lexicon based on Wiktionary. The variety of formalisms to represent such data can lead to a lack of compatibility. We explore ways to address this problem using LLOD.

1.2. Linguistic Interoperability

Linguistic interoperability refers to a language processing system that is language agnostic in the sense that it operates correctly with any natural language.

A clear example is the Unicode¹ text encoding standard designed to support the representation of text written in all of the world's major writing systems. Prior to its introduction, different, sometimes proprietary encodings for individual language and language groups were used. This kind of non-interoperability yields, for example, text documents in language L1 which work fine on text processing systems W1 and W2, whilst for language L2 they only work for W2. In contrast, any system that is Unicode compliant can operate with text in any language.

Another example is furnished by Universal Dependencies (Nivre et al., 2017) (UD), a framework for consistent annotation of grammar (parts of speech, morphological features, and syntactic dependencies) across different human languages. UD trees are an interoperable representation for which language-independent tools can be developed that operate on the syntactic structures of different languages.

1.3. Domain Interoperability

By extension to linguistic interoperability, domain interoperability means that a system or platform is “domain agnostic”: it operates correctly and without adaptation in different domains. For this to be possible, the system must have some advance knowledge of the abstract structure of the domain, so that it can unpick the parts that need to be processed. The notion of domain is extremely general, but within language processing communities it refers to a subject area, theme or topic, typically associated with a characteristic vocabulary of words. Examples of such domains are finance, biomedicine, justice. More formally, we can regard this as something close to an *ontology*, i.e. a set of related concepts together with an associated set of terms that are used for naming them. Indeed, for all the above examples, and many others, we find such ontologies: e.g. FIBO (Bennett, 2013) (finance); Hu (2006) (biomedicine); Engers et al. (2008) (justice).

¹<https://home.unicode.org/>

1.4. Related Work

There are many approaches and initiatives that aim to increase the interoperability of LRs. Probably one of the most famous ones is Universal Dependencies, already mentioned in Section 1.2. Its success led to many other related projects. One of them is Unimorph (Batsuren et al., 2022), an initiative aimed at creating a unified framework for morphological data across languages. It seeks to provide a standardised format for encoding morphological information, such as inflections, derivations, and other morphological processes, across different languages. Unlike LLOD, the project relies on lists of wordforms combined with the list of grammatical categories². Such a simple representation is very appealing and, at first sight, provides interoperability, since the format is very easy to read and write. On the other hand, any operation requires the creation of a custom solution, or the use of ad-hoc tools created specifically for this. At the same time, simplifying the annotation standard to a flat list does not work for all languages, which is evident from the fact that the format becomes more complex over time when inconsistencies arise. And the more complex the format becomes in order to increase language interoperability, the less straightforward it becomes to parse the dumps, and the lower is the overall interoperability.

On the other spectrum of interoperability and human readability there is another related technology: *xfst* (Beesley and Karttunen, 2003), *foma* (Hulden, 2009) and other Finite-State Transducer (FST) frameworks, powerful computational tools used for modelling and analysing natural language phenomena. These frameworks provide scripting languages that allow users to create transducers that can encode a wide range of linguistic phenomena, including morphological analysis, phonological rules, and syntax. FSTs provide functionalities for composing, intersecting and manipulating these transducers, allowing researchers to model complex linguistic processes in a formal and computationally tractable manner.

Transducers provide task interoperability (they are bidirectional) and, given that there are rules for all the languages of interest, language interoperability. They can generally be adapted to any domain. However, they operate on strings, so there is almost no way to enrich the dataset with additional information while staying within the formalism. For example, when dealing with homonyms of the same syntactic category (e.g. *bank*), there is no principled way to encode the particular sense of the word as would be the case when using LLOD or other semantic representations.

²The latest version uses a more complex format than a flat list in order to be able to deal with complex cases where a flat list leads to ambiguous parses.

In addition, different FST frameworks have slightly different formats and languages, which makes them non-interoperable with each other. To combat this, Chiarcos et al. (2022) shows that OntoLex-Morph can be used to encode FSTs and that it can function as an interchange format to convert between them.

1.5. Structure of the paper

We have introduced and described three kinds of interoperability. The remaining sections illustrate how all three can be achieved for LRs using an approach based on LLOD. Our examples centre around morphology resources and associated tools for morphological processing and for this reason we rely heavily on the extension of the OntoLex vocabulary for representations of morphology, OntoLex-Morph. Section 2 provides some background on morphology. Section 3 gives an overview of the OntoLex vocabulary and its extension for morphological descriptions. Sections 4–7 explore types of interoperability created by this vocabulary using two examples, and the final sections give an overview of related work and some conclusions.

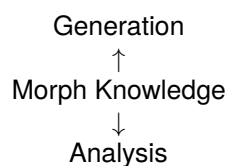
2. Morphology: facts versus processes

Morphology is the study of forms. Within linguistics, it denotes the study of linguistic forms, i.e. words, word parts, and their relationship to other words. A morphological description of a particular language assigns a structure to all the valid words and provides the rationale for grouping words e.g. into paradigms like verb conjugations or lexical entries that share the same sense. Alongside this purely structural information is the association of word forms with grammatically relevant information concerning e.g. part-of-speech or agreement features like number, gender and person.

From the perspective of NLP (in contrast to that of theoretical linguistics), the morphological description of a language is a data-oriented resource, as identified above. We should be careful to notice that although such a description *assigns* a morphological structure to a valid wordform, it does not tell us how to actually *discover* that structure. The same principle holds in the reverse direction (i.e. for generation of wordforms) which is to say that the morphological description contains enough information to assign one or more wordforms to a valid (but possibly underspecified) morphological structure, but it does not tell you how to go about computing it.

Here then, are two concrete examples of task interoperability: the language description is the static, data-oriented resource, whilst morphological analysis and generation are each oriented toward distinct processes. The basic idea is that the morphological

knowledge should be able to interoperate between the two computational tasks, i.e.



Here the up and down arrows denote distinct computational processes that respectively produce (i) a morphological generator and (ii) a morphological analyser for a given language. These are specified by two programs (one for analysis; another for generation) that each need to make certain assumptions about the format of the underlying morphological knowledge. Our claim is that OntoLex-Morph, a vocabulary designed for representing morphological information as LLOD, is a good choice of representation because we can specify both processes using an appropriately configured SPARQL query that embodies the structural assumptions that are made explicit in OntoLex-Morph.

3. 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 model revolves around the concept of a `LexicalEntry` — a lexeme or a dictionary entry. It must have at least one (word)form (`canonicalForm`) and can have a number of other forms, as well as lexical senses, which can be linked to either lexical concepts or entities in an ontology (Fig. 1). Basic morphological information such as part of speech and grammatical categories can be provided for lexical entries and forms using elements of any suitable vocabulary, such as LexInfo.³

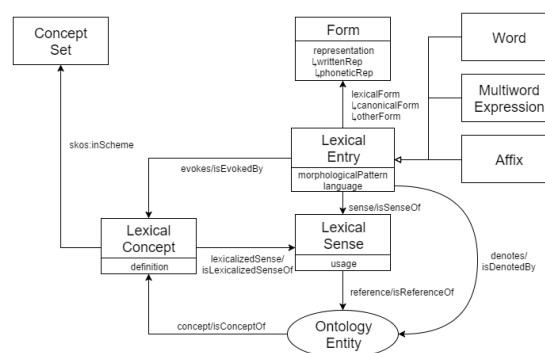


Figure 1: OntoLex-Lemon core model

Although there is a place for including basic morphological information in the core model, the standard does not give a clear way to represent paradigmatic relationships between lexical entries and forms (inflectional morphology) or derivational relationships

³<https://lexinfo.net/>.

between lexical entries. To close this gap and establish a standard way to represent this, an extension to the core module, OntoLex-Morph, is being developed.⁴

The model (Fig. 2) consists of three parts: derivation (left), inflection (right), and information on how to generate new forms, both for inflection and derivation (top). The central part of the module is the class `Morph`, which corresponds to a specific realisation of a morpheme. It is a subclass of `LexicalEntry`, which might be a bit counterintuitive at first, but this allows for modelling resources which have morphs as entries of their own.

The representation of rules for generating new forms (inflection) in the model works as follows: (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 not strictly set, but the one described in the guidelines is a (POSIX-compatible) regular expression.

For generating new lexical entries (derivation), we need to specify (i) word formation relations specifying what pairs of lexical entries should form a particular relation and potentially having additional information related to it, and (ii) derivation rules that specify how the parts of the words should be attached to each other.

4. Morphological Knowledge: An Illustrative Example

To illustrate how OntoLex-Morph facilitates different types of interoperability, we will use a toy dataset⁵ that models a part of a regular verb paradigm of an Italian verb *parlare*. The morphological knowledge we need to represent is summarised in the table below:

person/number	present
1SG	<i>parlo</i>
2SG	<i>parli</i>
3SG	<i>parla</i>
1PL	<i>parliamo</i>
2PL	<i>parlate</i>
3PL	<i>parlano</i>

Table 1: A fragment of conjugation of *parlare*

We model the lexical entry and its canonical form in the following way:

```
:parlare a ontolex:LexicalEntry ;
    lexinfo:partOfSpeech :lexinfo:verb ;
```

⁴<https://www.w3.org/community/ontolex/wiki/Morphology>.

⁵Throughout the paper we have, for the sake of clarity and brevity, adopted the simplest possible examples.

```
morph:morphologicalPattern :v-are_paradigm;
ontolex:canonicalForm :parlare_form ;
morph:baseForm :parlare_form .
```

```
:parlare_form a ontolex:Form ;
    ontolex:writtenRep "parlare"@ita .
```

Properties `baseForm` and `morphologicalPattern` specify a base form that is used to create inflected forms and the type of conjugation (i.e. a set of rules that can be applied to the form), respectively.

For each cell in the paradigm, we need to provide an affix and a rule that describes how to create a corresponding form:

```
:suff_o_lsg a ontolex:Affix ;
    rdfs:label "-o"@ita ;
    morph:grammaticalMeaning [
        lexinfo:number lexinfo:singular ;
        lexinfo:person lexinfo:firstPerson ;
    ] .

:v-are_ind_lsg a morph:InflectionRule ;
    morph:paradigm :v-are_paradigm ;
    morph:involves :suff_o_lsg ;
    morph:replacement [
        a morph:Replacement ;
        morph:source "are$" ;
        morph:target "o" ;
    ] .
```

5. Task Interoperability

Using the example described above,⁶ we can now examine the state of task interoperability in the OntoLex-Morph vocabulary.

5.1. Generation

As described in Section 3, part of OntoLex-Morph was designed to allow the representation of generation rules for both inflection and derivation. In this way, it is possible to store lexical entries with their dictionary forms in the dataset, along with instructions on how to generate the rest rather than having a complete set of pre-generated forms. The example above provides the data necessary to generate the forms for the case of inflection.

The generation process can be built on top of native RDF technologies (i.e. SPARQL). This provides a general level of interoperability, as these technologies follow open standards, so the implementation does not depend on proprietary tools or particular products that might be discontinued in the future. In real-life applications, of course, the implementation should contain at least a wrapper around RDF technologies, but for the purposes of this paper, the raw output of SPARQL queries is enough.

To show the generation capabilities, we created a SPARQL SELECT query that, when applied to the data described in the previous section, outputs

⁶The data is available at <https://github.com/max-ionov/ldl-2024-morph-interoperability/blob/main/italian.ttl>.

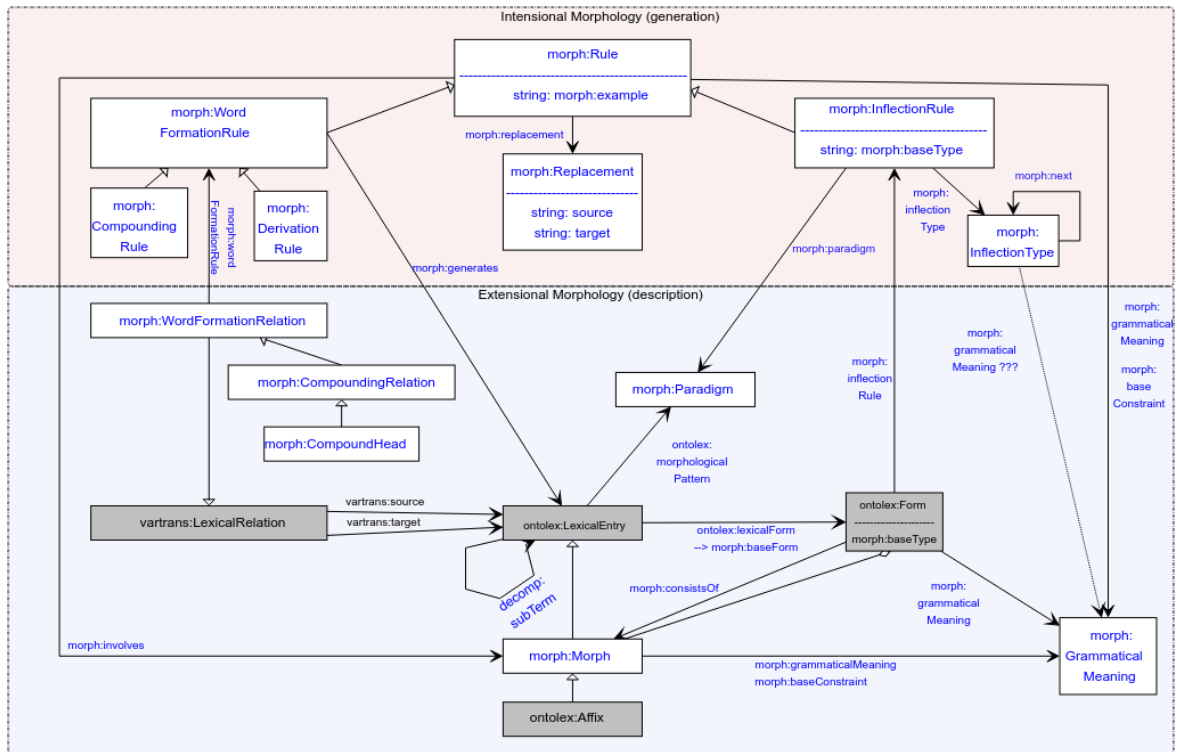


Figure 2: OntoLex-Morph draft model

generated inflected forms with their assigned grammatical categories (see Table 2).⁷

entry	form	gram. cat.	value
parlare	parlo	person number	firstPerson singular
parlare	parli	person number	secondPerson singular
parlare	parla	person number	thirdPerson singular
...

Table 2: Generation of inflected forms of *parlare*

Alternatively, it is possible to modify it to a SPARQL CONSTRUCT query which outputs RDF with the generated forms that can be used saved in a local graph and used in subsequent queries.

5.2. Analysis

An additional level of interoperability, task interoperability, is provided both by the way the Morph module was designed and the nature of RDF technologies: with a slight modification of the SPARQL query used for generation, we can revert the pro-

⁷All queries and the full version of the outputs are available at <https://github.com/max-ionov/ldl-2024-morph-interoperability/blob/main/sparql/>.

cess and get possible morphological analyses for a word, without having these forms pre-generated in advance:

entry	form	gram. cat.	value
parlare	parlo	person number	firstPerson singular

Table 3: Analyses for a wordform *parlo*

Note that this procedure differs from a simple table lookup: the final list of forms is never created, but each is dynamically computed and tested against search criteria. In most cases, it might be impractical (especially since SPARQL endpoints are generally slow and unreliable), but this might be useful, for example, when the rules are not stable or come from multiple external sources.

This also differs from generational approaches, whether statistical or rule-based (e.g., finite-state transducers). While most generational approaches operate with strings, this procedure finds URIs of lexical entries,⁸ which, in turn, may contain more information, both paradigmatic and syntagmatic.

⁸URIs are omitted in the tables above for formatting reasons.

6. Linguistic Interoperability

Having established task interoperability, we turn to argue similarly for *linguistic* interoperability. In our previous research, we have shown that OntoLex-Morph can encode inflectional rules for languages with different types of word formation, i.e., non-concatenative morphology of the Maltese language (Ionov and Rosner, 2023). Further examples exist showing the applicability of the model to agglutinative and polysynthetic languages.⁹ This alone is an example of linguistic interoperability. Furthermore, we can apply the queries that we applied to the Italian dataset in the previous section to the aforementioned Maltese dataset¹⁰ almost without any adaptation:

entry	form	gram. cat.	value
kiteb	ktibt	person number aspect	firstPerson singular perfective
kiteb	ktibt	person number aspect	secondPerson singular perfective

Table 4: Analyses of a Maltese wordform *ktibt*

In this way, we can use the same machinery for both generation and analysis for both languages, and generally this should be extensible for any other language.

However, there are some caveats. Most importantly, the queries we present only account for cases where a form is created by adding only one affix: we do not account for agglutination, where multiple rules can be applied to a single entry to create a wordform (cf. Finnish noun inflection, with separate suffixes for number and grammatical case). Another problem with the queries we used with regard to linguistic interoperability is that there are character classes for vowels and consonants hard-coded into them, and they only account for the Maltese alphabets since the classes are used only in the Maltese set of rules.

Finally, there are some minor inconsistencies in the way different SPARQL engines implement the standards, which leads to slightly different outputs. But all these are limitations of the specific (quite rudimentary) implementation and can be solved by a more complex way of applying the rules, with pre- and post-processing.

7. Domain Interoperability

To show an example of domain interoperability, we need another example. We will focus on chemi-

cal nomenclature, developed to facilitate communication by providing a methodology for assigning descriptors to chemical substances so that they can be identified without ambiguity. This domain exhibits a three-level structure: (i) actual *chemical compounds* with a definite molecular structure which is the underlying semantic interpretation, (ii) a *formula* which notates that structure and (iii) terms which are composite strings with their own systematic morphological structure. A few simple examples from *Nomenclature of Inorganic Chemistry* (Connelly et al., 2005), the so-called “Red Book” illustrate this.

Chemical Term	Formula
trioxygen	O_3
sodium chloride	$NaCl$
iron dichloride	$FeCl_2$
trisodium pentabismuthide	Na_3Bi_5
magnesium chloride hydroxide	$MgCl(OH)$

The wording describing the principles of nomenclature is highly reminiscent of that used in linguistic morphology:

Generally, nomenclature systems require a root [...] Names are constructed by joining other units to these roots. Among the most important units are affixes. These are syllables added to words or roots and can be suffixes, prefixes or infixes according to whether they are placed after, before or within a word or root.
(Connelly et al., 2005, p. 5)

For example, the name iron dichloride for the substance $FeCl_2$ involves the juxtaposition of element names (iron, chlorine), their ordering in a specific way (electropositive before electronegative), the modification of an element name to indicate charge (the ‘ide’ ending designates an elementary anion and, more generally, an element being treated formally as an anion), and the use of the multiplicative prefix ‘di’.

In practice, we might utilise this knowledge in one of the two (non-exclusive) ways: we might extend a general-purpose lexicon that does not have some of these terms, or we can add information about word relations between the terms. In both cases, it is possible to utilise OntoLex-Morph. As an example, we are going to look at derivatives of the word *chlorine*: *chloride* and *dichloride* and multiword expressions (MWE) that contain them: *sodium chloride* and *iron dichloride*.

We start with definitions of the lexical entry *chlorine* and its canonical form:

⁹<https://github.com/ontolex/morph/tree/master/data>.

¹⁰<https://raw.githubusercontent.com/max-ionov/maltese-morph/main/lexical-entries-small.ttl>.

```
:chlorine a ontolex:LexicalEntry ;
          ontolex:canonicalForm :chlorine_form .

:chlorine_form a ontolex:Form ;
              ontolex:writtenRep "chlorine"@en-GB .
```


To add information on how to generate new derivate entries, we add the following instances of `WordFormationRelation` and `DerivationRule`:

```
:rel_chlorine_ide a morph:WordFormationRelation ;
    vartrans:source :chlorine ;
    vartrans:target :chloride ;
    morph:WordFormationRule :ine_ide_rule .

:ine_ide_rule a morph:DerivationRule ;
    morph:replacement [
        morph:source "ine$" ;
        morph:target "ide"
    ] .

:di_rule a morph:DerivationRule ;
    morph:replacement [
        morph:source "^" ;
        morph:target "di"
    ] .

:rel_di_chloride a morph:WordFormationRelation ;
    vartrans:source :chloride ;
    vartrans:target :dichloride ;
    morph:WordFormationRule :di_rule .
```

This additional information can be created and stored independently from the main lexicon and can be used to extend the original data with the new domain-specific words and constructions. Using SPARQL federated queries, it can be queried together with the main lexicon, providing the desired results without changing the original dataset. The data can be further expanded by adding string representations in chemical notation:

```
:chlorine_form a ontolex:Form ;
    ontolex:writtenRep "chlorine"@en-GB,
                      "Cl"@en-x-chem .

:di_form a ontolex:Form ;
    ontolex:writtenRep "di"@en-GB,
                      "_2"@en-x-chem .
```

With this modelling, it is possible to use SPARQL to generate (a small subset of) chemical formulas from its components, as long as we use written representations with the corresponding language tag. In addition, it is also possible to translate chemical formulas from their notation to natural language and back, regardless of the natural language in question, which is a combination of all types of interoperability discussed in this and the previous sections.

8. Discussion and future work

8.1. Discussion

A key issue is the practical feasibility of using OntoLex-Morph in the way presented in this paper. The performance of SPARQL and the relatively low adoption of LLOD technologies make it a bad contender for an interchange format used on the level of UD or UniMorph. On the other hand, more and more people are becoming familiar with the field, and small and medium-sized datasets work relatively well, even under pressure.

For large datasets, or for services where availability is paramount, pre-generated tables are potentially

a better alternative. There are also hybrid solutions such as aggressive caching.

Another problem that we have not touched on in this paper is the heterogeneity of OntoLex datasets. According to [Bosque-Gil et al. \(2018\)](#), this can be because “authors have developed their own ad-hoc extensions due to the actual lack of existing models that account for the specific features of the resource they aim to convert, due to the lack of awareness of a partially similar resource, or even due to the difficulty of finding the appropriate documentation”. Inconsistencies between the datasets require querying the datasets separately, creating more complex queries that account for all the possible configurations, or inserting additional unifying statements into the datasets.

Of course, sometimes this might be due to under-specification by design. One example is representation of grammatical categories in OntoLex: Although the *LexInfo* vocabulary is recommended, it is not required to use it, since there might be categories that are not represented there.

8.2. Future work

At the outset of this paper we suggested that interoperability is widely cited but rarely exemplified quality of LOD. In the preceding sections we have tried to demonstrate that the three types of interoperability described actually make sense and can be illustrated concretely at least for LLOD in the domain of morphology. The main lesson is that it is difficult to demonstrate interoperability without narrowing the focus of application precisely because the inherent characteristics of LOD – linkedness and openness – are very general, giving rise to a very general and therefore weak notion of interoperability. We feel that we have progressed by narrowing the application to linked language data in the particular area of description systems for the morphological structure of terms.

We foresee two main directions of interest for future research. The first is to deepen the coverage by advancing from a few choice examples to a deeper and wider coverage of tasks, languages and domains. This would bring some much needed detail concerning the adequacy of the underlying descriptive framework. There is clearly a lot of work involved in any of these.

The second direction concerns the inherently useful idea of measuring the degree of interoperability displayed by a given set of resources and associated tools. Such measurement would enable us to investigate whether one system of description is more interoperable than another or whether there is a measurable tradeoff between interoperability and efficiency. At present, these questions are too complex to answer in any exact sense.

One possible direction would be to look at the

amount or the complexity of queries required to extract all the relevant concepts from two or more datasets (e.g. lexical entries, forms, written representations, etc.), or the intersection between these queries for each: the more there is in common, the more interoperability there is.

However, we need to be careful to distinguish between interoperability and lack of flexibility. Having different datasets fully interoperable is not very useful if this comes at a cost of them not representing the data properly.

9. Bibliographical References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. [SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining](#). In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Abbott Lane, Mohit Raj, Matt Coler, Jaime Rafael Montoya Samame, Delio Siticonatzí Camaiteri, Esaú Zumaeta Rojas, Didier López Francis, Arturo Oncevay, Juan López Bautista, Gema Celeste Silva Villegas, Lucas Torroba Hennigen, Adam Ek, David Guriel, Peter Dirix, Jean-Philippe Bernardy, Andrey Scherbakov, Aziyana Bayyr-ool, Antonios Anastasopoulos, Roberto Zariquiey, Karina Sheifer, Sofya Ganieva, Hilaria Cruz, Ritván Karahóga, Stella Markantonatou, George Pavlidis, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Candy Angulo, Jatayu Baxi, Andrew Krizhanovsky, Natalia Krizhanovskaya, Elizabeth Salesky, Clara Vania, Sardana Ivanova, Jennifer White, Rowan Hall Maudslay, Josef Valvoda, Ran Zmigrod, Paula Czarnowska, Irene Nikkarinen, Aelita Salchak, Brijesh Bhatt, Christopher Straughn, Zoey Liu, Jonathan North Washington, Yuval Pinter, Duygu Ataman, Marcin Wolinski, Totok Suhardijanto, Anna Yablonskaya, Niklas Stoehr, Hossep Dolatian, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Aryaman Arora, Richard J. Hatcher, Ritesh Kumar, Jeremiah Young, Daria Rodionova, Anastasia Yemelina, Taras Andrushko, Igor Marchenko, Polina Mashkovtseva, Alexandra Serova, Emily Prud'hommeaux, Maria Nepomniashchaya, Fausto Giunchiglia, Eleanor Chodroff, Mans Hulden, Miikka Silfverberg, Arya D. McCarthy, David Yarowsky, Ryan Cotterell, Reut

Tsarfaty, and Ekaterina Vylomova. 2022. [UniMorph 4.0: Universal Morphology](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 840–855, Marseille, France. European Language Resources Association.

Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI Studies in Computational Linguistics. CSLI Publications.

Mike Bennett. 2013. [The financial industry business ontology: Best practice for big data](#). *Journal of Banking Regulation*, 14.

Julia Bosque-Gil, Asuncion Gómez-Pérez, Elena Montiel-Ponsoda, and Jorge Gracia. 2018. Models to represent linguistic linked data. *Natural Language Engineering*, 24(6):811–859.

Christian Chiarcos, Christian Fäth, and Maxim Ionov. 2022. [Unifying morphology resources with OntoLex-morph. a case study in German](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4842–4850, Marseille, France. European Language Resources Association.

Neil G. Connelly, Ture Damhus, Richard M. Hartshorn, and Alan T. Hutton. 2005. [Nomenclature of inorganic chemistry – iupac recommendations 2005](#). *Chemistry International – Newsmagazine for IUPAC*, 27(6).

Tom Engers, Alexander Boer, Joost Breuker, Andre Valente, and Radboud Winkels. 2008. [Ontologies in the Legal Domain](#), pages 233–261. Springer.

Mikel L. Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144.

Xiaohua Hu. 2006. [Natural language processing and ontology-enhanced biomedical literature mining for systems biology](#). *Computational Systems Biology*, pages 39–56.

Mans Hulden. 2009. [Foma: a finite-state compiler and library](#). In *Proceedings of the Demonstrations Session at EACL 2009*, pages 29–32, Athens, Greece. Association for Computational Linguistics.

Maxim Ionov and Mike Rosner. 2023. [Beyond concatenative morphology: Applying OntoLex-morph to Maltese](#). In *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 385–391, Vienna, Austria. NOVA CLUNL, Portugal.

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*, pages 19–21.

Joakim Nivre, Daniel Zeman, Filip Ginter, and Francis Tyers. 2017. [Universal Dependencies](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain. Association for Computational Linguistics.

Gilles Sérasset. 2015. DBnary: Wiktionary as a Lemon-based multilingual lexical resource in RDF. *Semantic Web*, 6(4):355–361.

Modeling linking between text and lexicon with OntoLex-Lemon: a case study of computational terminology for the Babylonian Talmud

Flavia Sciolette

Institute for Computational Linguistics "A.Zampolli"
Via Giuseppe Moruzzi, 1, 56124 Pisa PI, Italia
flavia.sciolette@ilc.cnr.it

Abstract

This paper illustrates the first steps in the creation of a computational terminology for the Babylonian Talmud. After introducing the motivation for this work and the state of the art, the paper exposes the choice of using OntoLex-Lemon and the new FrAC module for encoding the attestations and quantitative data of the terminology extraction. The Talmudic terminological base is introduced, with an example of an entry populated with the above-mentioned data. The choices for modeling are motivated by the rich representation the model allows and also for future needs for the management of the link between text and lexical entries.

Keywords: Computational Terminology, Linked Open Data, Talmud, OntoLex-Lemon, FrAC

1. Introduction

Over time, the gap between lexicographic and terminological practices has narrowed (Salgado et al., 2022) in terms of models and methodologies, thanks to the 'linguistic turn' of the 2000s (Bellandi et al., 2020), which is now well established in several studies. A term is the linguistic realisation of a domain concept (Buitelaar et al., 2005); in many contexts, however, the exact correspondence between term and concept - between actual use and norm - is not taken for granted (Soffritti, 2010). Therefore, for many resources, the text becomes a necessary starting point for the observation of the term, a manifestation of the word (Chiocchetti and Ralli, 2022), and consequently a basis for its extraction, in order also to subsequently build a knowledge base (Buitelaar et al., 2005). Although the debate on the creation of these resources is also inevitably experiencing the influence of Large Language Models (LLM)¹, we point out that the link between the text as source and term analysis is still fundamental when dealing with historical or highly specialised languages, less rich in resources suitable for training specific models. The preservation of this link allows to represent useful information, all the more so when considering quotations from corpora as examples of authentic linguistic usages (Klosa, 2015), which can convey many linguistic,

historical, and cultural information.

1.1. Motivation

We choose to present the case study offered by the Babylonian Talmud - a fundamental text for Jewish religion and culture - and the creation of a terminological resource, currently under development, for the project of Italian translation for this text. A complex task like translation allows for an in-depth discussion on the need for resources based on state-of-the-art models and formats, as well as on shared standards that guarantee broad use and interoperability of data. The resource is indeed built according to the Linked Open Data (LOD) principles (Bizer et al., 2023) and recent good practices for modeling quotations and attestations. In the following sections, we consider the related work about the chosen case study, and consequently, a section is dedicated to the choice of adopting the OntoLex-Lemon model and the recent FrAC module for modeling attestations and frequency values; an example of an entry is provided, linked to contexts extracted from the treatises of which the Talmud is composed; finally, future developments are outlined in the conclusions. The resulting terminological resource is intended to be a useful tool to deepen the study of the languages used in the Talmud and to help translators in their choices.

2. Related Work

The Talmud represents a fundamental text for Judaism and constitutes a veritable mine of historical, cultural, social, legal, and scientific information. Among religious texts, it appears to be one of the

¹For considerations on ontologies and ontology learning, see (Neuhaus, 2023) and (Babaei Giglou et al., 2023). For experiments on term and entity extraction, see (Meoni et al., 2023), (Liu et al., 2023), with general considerations on the use of these models in low-resourced contexts. For translation, similarly see (Robinson et al., 2023).

richest for cultural and linguistic information, as well as one of the most complex, considering also that it is multilingual, with a formulaic structure, and it is further enriched with several commentaries. Before the Italian translation project, there were no specific resources available for the Talmud in Italian, neither printed nor digital. About the latter in other languages, for an overview, see (Giovannetti et al., 2020) and (Saponaro et al., 2022), in particular for resources available as LOD. For the Italian language, the terminological extraction conducted on the translated treatises of the Talmud is also described in (Giovannetti et al., 2020). This extraction was also used in a terminology graph visualisation application (Marchi et al., 2022). Other experiments on the extraction of named terms and entities were mainly concerned with the creation of an ontology on master rabbis (Giovannetti et al., 2021). Also noteworthy is the study of the networks of relationships between master rabbis by (Satlow and Sperling, 2022).

3. Ontolex-Lemon

Linguistic Linked Open Data (LLOD) (Cimiano et al., 2020) constitute a subset of LODs and represent a set of best practices that facilitate the sharing and reuse of linguistic data in various applications and research domains, according to FAIR principles (Wilkinson et al., 2016). OntoLex-Lemon (McCrae et al., 2017) is the most well-known and widely used vocabulary for the creation, publication, and sharing of lexical and terminological resources such as LLOD. The model includes several extensions that have already been published or are currently under development; these include the module for representing frequency, attestation, and corpus information (FrAC).

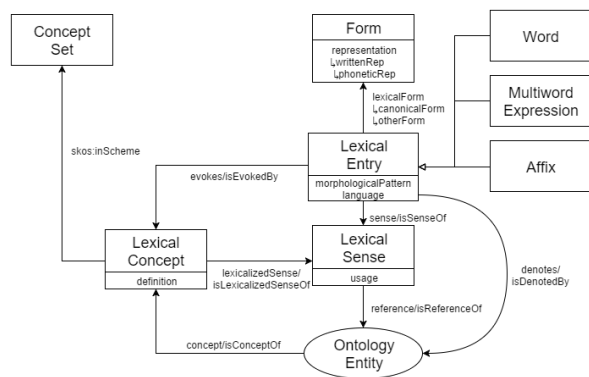


Figure 1: Module Core of OntoLex-Lemon.

3.1. FrAC

FrAC is currently at an advanced phase (Chiarcos et al., 2022) and undergoing final revision². Here we mention only the part of module used for the examples in this paper: the class `Attestation`, which constitutes "a special form of citation that provides evidence for the existence of a certain lexical phenomenon"; the property `attestation`, which associates `Attestation` with an `Observable` of FrAC. To this is added the relation `quotation`, to insert the text of the quotation in natural language; the `Frequency` for the absolute number of times the term appears in an attestation; the property `measure` to indicate the term frequency-inverse document frequency (tf-idf³).

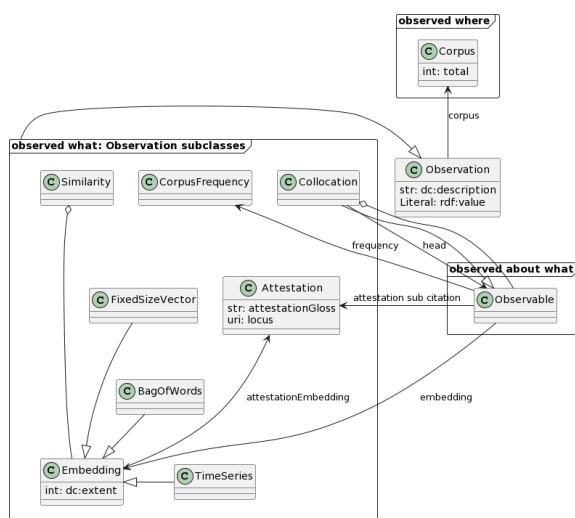


Figure 2: Diagram of FrAC.

4. Talmudic Terms

4.1. First steps of the terminological base

The starting point was the extraction of terms from the translated treatises of the Talmud. In this context, the considered definition was "a (candidate) term was defined as a simple (single-word) or complex (multi-words) nominal structure with modifiers" (Giovannetti et al., 2020). It is therefore a 'procedural' definition, functional for querying the text using regular expressions. The extraction was conducted with the Term To Knowledge tool (T2K) (Dell'Orletta et al., 2014), on the Italian translation⁴. For each

²<https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/index.md>

³<https://it.wikipedia.org/wiki/Tf-idf>

⁴There are no tools for the automatic analysis of Biblical and Mishnaic Hebrew. For the use of tools for Modern Hebrew, see (Pecchioli et al., 2018).

term, the absolute frequency and tf-idf were provided. A high tf-idf value implies that the term frequently appears in a few documents and is therefore specific, whereas a low tf-idf value means that the term is distributed in many different documents (as is the case, for example, with 'rabbi', which appears extensively throughout the Talmud treatises). Consequently, it was decided to model all terms with high tf-idf values, which were then manually checked by domain experts (approximately 4000 terms). These data were exported in a .csv format. From the .csv format, through a specially created Python script, the data structures for lexical entry were created, including language, canonical form, sense, absolute frequency, tf-idf, and the treatise to which the term belongs. The natural language definition and example quotation content were entered manually.

4.2. An example of entry

It follows an example of an entry modeled according to the OntoLex-Lemon model for Talmud terminology. The term is the Hebrew 'shemà' which indicates one of the obligatory readings to be performed during the day. The main entry is the lexical entry :shema, which is associated with various types of morpho-syntactic information (part of speech, gender, number, etc.).

```
:shema_entry a ontolx:lexicalEntry;
dct:language <http://www.lexvo.org/page/iso639-3/ita>;

lexinfo:partOfSpeech lexinfo:commonNoun ;
lexinfo:gender lexinfo:masculine ;
lexinfo:number lexinfo:singular.
```

The lexical entry is associated with the sense, a canonical form corresponding to the lemma contained in the glossaries in use by the translators of the Talmud project, and the absolute frequency value. According to the description of the frequency class in FrAC, it is possible to associate the value with both the entry and the form; in this case, it was chosen to associate the value with the entry and to specify it further in the description of the individual forms if necessary.

The observedIn relation clarifies the source of the data (in this case, the frequency) in the form of a URI. A second frequency information is modeled with the frac:measure relation to providing the value of tf-idf as an rdf:value. Finally, a certain number of individuals are associated to the entry with the attestation relation (three examples of segments are given in the modeled entry). Figure 3 provides a visual representation of the scheme for the entry.

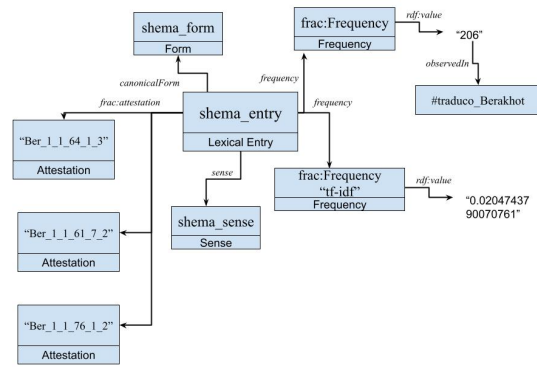


Figure 3: Encoding of the entry "shemà".

```
ontolx:sense :shema_sense;
ontolx:canonicalForm :shema_form;
frac:frequency [
a frac:Frequency;
rdf:value "206"^^xsd:int;
frac: observedIn <#traduco_berakhot >
];
frac:frequency [
a frac:Frequency; frac:measure "tf-idf" ;
rdf:value "0.0204743790070761"
].
frac:attestation :Ber_1_1_64_1_3,
Ber_1_1_61_7_2, Ber_1_1_76_1_2.
```

The definition is provided with the relation defined in SKOS⁵ to provide a natural language description of sense, taken from the Talmudic glossaries compiled in the project. In this way, we can preserve the attestations and the definitions written by domain experts, as in the case of the shemà: "Three passages from the Pentateuch: 'Hear, O Israel' (Deut. 6:4-9), 'And if you will listen' (Deut. 11:13-21), 'And the Lord spoke to Moses' (Num. 15:37-41), the reading of which is obligatory twice a day, in the morning and the evening." The selected example is contained in the first treatise of the Talmud, Berakhot: "One may stand and recite the Shemà."

```
:shema_sense a ontolx:LexicalSense;

skos:definition "Treubrani del Pentateuco
: Ascolta Israele (Deut.6:4-9)
, E use ascolterai (Deut.
11:13-21), E il Signore disse a
Moshè (Num.15:37-41), la cui
lettura è obbligatoria due volte al
giorno, alla mattina e alla sera."@it

:Ber_1_1_64_1_3 frac:Attestation ;
frac:quotation "Si può stare in piedi e
recitare lo Shemà".@it.
```

In this way, it is possible to link different sources of

⁵<https://www.w3.org/TR/skos-reference/>

data, even if not originally LLOD. Individuals can be described as an attestation; through the *quotation*, it is possible to provide the content of the attestation in natural language (a single segment as an example). The adoption of FrAC thus makes it possible to enrich the knowledge graph of lexical entries with quantitative information, potentially useful for various tasks such as topic modeling. In this way, other knowledge graphs outside the terminology or text annotations can also be linked to the entries (see next section).

4.3. Text management and lexical linking

The rich amount of information related to Jewish culture in the Talmudic text is systematised in the glossary entries prepared within the project. These glossaries have been produced in a translation-oriented manner and are therefore based on the annotation of the term in both the original text and its Italian counterpart. Maintaining this link is therefore fundamental in the elaboration of a terminological resource; the adoption of FrAC enables the linking of attestations to elements outside the graph, including annotations to the text itself. We call this task 'lexical linking', which consists, similarly to entity linking, in linking words in texts with linguistic entities (lemma, meanings, etc.) encoded in another resource. These annotations may include terms, but also information of a different nature, distributed on different annotation layers (e.g. to encode specific formulae in which terms are inserted). Currently, this phase is handled manually through an editor, named "Maia" prepared for this purpose, under development⁶.

5. Conclusion and future works

In this paper, we presented a case study, offered by the creation of Talmudic terminology, for the encoding of dictionary attestations and quantitative data. The starting point was offered by an extraction of terminology for the Italian translation of the Talmud. We, therefore, presented an example of an entry using the new FrAC module, currently undergoing final revision, to show its productivity, also with a view to future linking with the annotated text, thanks also to a specific editor currently being developed. Future work includes linking terminology entries and specific senses to an Italian reference lexicon in LLOD, Compl-It⁷ currently available on CLARIN; linking to ontological references of individual terms and occurrences of Hebrew terms in the untranslated text.

⁶<https://github.com/klab-ilc-cnr/Maia>

⁷<https://dSPACE-clarin-it.ilc.cnr.it/repository/xmlui/handle/20.500.11752/ILC-1007>

6. Acknowledgments

This work was conducted in the context of the TALMUD project, the scientific cooperation between S.c.a r.l. PTTB and CNR-ILC, and the contribution of a short-term mobility (STSM) grant within the COST Action "Nexus Linguarum", at the University of Augsburg.

7. Bibliographical References

- Hamed Babaei Giglou, Jennifer D'Souza, and Sören Auer. 2023. *LLMs4OL: Large language models for ontology learning*. page 408–427, Berlin, Heidelberg. Springer-Verlag.
- Andrea Bellandi, Emiliano Giovannetti, Simone Marchi, Silvia Piccini, and Flavia Sciolette. 2020. Come dare senso a un termine? caratteristiche, potenzialità e opportunità dello strumento Lexo. In *Comunicazione al XXX Convegno Ass.I.Term.*, Eurac Research, Bolzano.
- Christian Bizer, Tom Heath, and Tim Berners-Lee. 2023. Linked data – the story so far. In *Linking the World's Information: Essays on Tim Berners-Lee's Invention of the World Wide Web*, pages 115–143. Association for Computing Machinery, New York, NY, United States.
- Paul Buitelaar, Philippe Cimiano, and Bernardo Magnini. 2005. Ontology learning from text: An overview. In *Ontology Learning from Text: Methods, Evaluation and Applications*. IOS Press, Amsterdam.
- 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, Gyeongju, Republic of Korea.
- Elena Chiochetti and Natascia Ralli. 2022. Introduzione. In *Risorse e strumenti per l'elaborazione e la diffusione della terminologia in Italia*. Eurac Research, Bolzano.
- Philipp Cimiano, Christian Chiarcos, John P. McCrae, and Jorge Gracia. 2020. Linguistic linked data in digital humanities. In *Linguistic Linked Data in Digital Humanities*, pages 229–262. Springer International Publishing, Cham.
- Felice Dell'Orletta, Giulia Venturi, Andrea Cimino, and Simonetta Montemagni. 2014. *T2K²: a system for automatically extracting and organizing*

- knowledge from texts. In *Proceedings of 9th Edition of International Conference on Language Resources and Evaluation*, pages 178–190, Reykjavik. ACL.
- Emiliano Giovannetti, Davide Albanesi, Andrea Bellandi, David Dattilo, Angelo Maria Del Grosso, and Simone Marchi. 2021. [An ontology of masters of the Babylonian Talmud](#). *Digital Scholarship in the Humanities*, fqab043.
- Emiliano Giovannetti, Andrea Bellandi, David Dattilo, Angelo Maria Del Grosso, Simone Marchi, Alessandra Pecchioli, and Silvia Piccini. 2020. The terminology of the Babylonian Talmud: Extraction, representation and use in the context of computational linguistics. *Materia Giudaica*, XXV.
- Annette Klosa. 2015. On corpus citations in monolingual general dictionaries. *Dictionaries: Journal of the Dictionary Society of North America*, 36(1):72–87.
- Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Lin Zhao, Dajiang Zhu, Xiang Li, Ning Qiang, Dinggang Shen, Tianming Liu, and Bao Ge. 2023. [Summary of ChatGPT-related research and perspective towards the future of large language models](#). *Meta-Radiology*, 1(2):100017.
- Simone Marchi, Marianna Colombo, David Dattilo, and Emiliano Giovannetti. 2022. Un esperimento di visualizzazione grafica della terminologia del Talmud Babilonese. In *AIUCD 2022 - Proceedings*, Lecce, Italy.
- 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.
- Simon Meoni, Eric De la Clergerie, and Theo Ryffel. 2023. [Large language models as instructors: A study on multilingual clinical entity extraction](#). In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 178–190, Toronto, Canada. Association for Computational Linguistics.
- Fabian Neuhaus. 2023. [Ontologies in the era of large language models: a perspective](#). *Applied Ontology*, 18(4):399–407.
- Alessandra Pecchioli, Davide Albanesi, Andrea Bellandi, Emiliano Giovannetti, and Simone Marchi. 2018. Annotazione linguistica automatica dell'ebraico mishnaico: Esperimenti sul Talmud Babilonese. *Materia Giudaica*, XXIII.
- Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. [ChatGPT MT: Competitive for high- \(but not low-\) resource languages](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.
- Ana Salgado, Rute Costa, and Toma Tasovac. 2022. Applying Terminological Methods to Lexicographic Work: Terms and Their Domains. In *Dictionaries and Society. Proceedings of the XX EURALEX International Congress*, pages 181–195, Mannheim. IDS-Verlag.
- Davide Saponaro, Emiliano Giovannetti, and Flavia Sciolette. 2022. From religious sources to computational resources: Approach and case study on hebrew terms and concepts. *Materia Giudaica*, 27.
- Michael L. Satlow and Michael Sperling. 2022. [The rabbinic citation network](#). *AJS Review: The Journal of the Association for Jewish Studies*, 46(2):291–319.
- Marcello Soffritti. 2010. Termontografia e innovazione della terminologia plurilingue. In Franco Bertaccini, Sara Castagnoli, and Francesco La Forgia, editors, *Terminologia a colori*, pages 31–5. Bononia University Press, Forlì.
- Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabriel Appleton, Myles Axton, Arie Baak, and Niklas Blomberg. 2016. [The FAIR guiding principles for scientific data management and stewardship](#). *Scientific data*, 3.

OntoLex Publication Made Easy: A Dataset of Verbal Aspectual Pairs for Bosnian, Croatian and Serbian

Ranka Stanković¹, Maxim Ionov², Medina Bajtarević³, Lorena Ninčević⁴

¹University of Belgrade, Faculty of Mining and Geology, Serbia, ranka.stankovic@rgf.bg.ac.rs

²University of Cologne, Cologne Center for eHumanities, Germany, mionov@uni-koeln.de,

³Bosnia and Herzegovina, medina.bajtarevic@gmail.com,

⁴University of Zagreb, Faculty of Humanities and Social Sciences, Croatia, lnincevi@ffzg.unizg.hr

Abstract

This paper introduces a novel language resource for retrieving and researching verbal aspectual pairs in BCS (Bosnian, Croatian, and Serbian) created using Linguistic Linked Open Data (LLOD) principles. As there is no resource to help learners of Bosnian, Croatian, and Serbian as foreign languages to recognize the aspect of a verb or its pairs, we have created a new resource that will provide users with information about the aspect, as well as the link to a verb's aspectual counterparts. This resource also contains external links to monolingual dictionaries, Wordnet, and BabelNet. We believe it will be useful for research in the field of aspectology, as well as machine translation and other NLP tasks. Using this resource as an example, we also propose a sustainable approach to publishing small to moderate LLOD resources on the Web, both in a user-friendly way and according to the Linked Data principles.

Keywords: verbal aspect, Linguistic Linked Open Data, BCS

1. Introduction

One of the most difficult properties for the learners of Slavic languages is the verbal aspect. When speaking or writing, the L2 learners/speakers of BCS have to choose whether the appropriate aspect in the given context is perfective or imperfective. After that, they have to recall the form for the appropriate aspect and then conjugate it accordingly. When reading or listening, the learners have to recognize the aspect used. The lack of learning resources makes this task even more difficult. Therefore, we have decided to create a resource to help BCS learners learn the verbal aspect.

1.1. Motivation

Generally, there is a lack of language learning resources for Bosnian, Croatian, and Serbian. The available digital dictionaries do provide information about a verb's aspect, however, their aspectual counterparts are not included in the entry. The users simply need to know the counterpart. Moreover, if learners encounter a new verb, there are no ways to distinguish its aspect. Indeed, perfectivization happens mostly by prefixation, but there can be a secondary imperfectivization in which we would have a verb with a prefix but with a suffix added subsequently, which then makes it imperfective. Sometimes perfective verbs are longer, and other times imperfective verbs are longer. There can be cases where both verbs are of the same length, e.g. 'odmarati se' (to be resting, imperfective) and 'odmoriti se' (to rest, perfective).

There are also cases where the perfective pair is formed with the prefix but it is also made reflexive. There are no ways for learners to know all this, and this is the reason why the aspect can be one of the most disliked features of Slavic languages for learners.

In addition to being a language learning resource, our dataset can also be used in research. For instance, retrieving all the verbs derived with a certain prefix can be used for much more systematic research of the nuanced meanings given by it.

With the aspectual information provided, our resource can also be used to aid in NLP applications, such as machine translation. Sometimes, the aspectual information is not translated well, and it is not possible to infer additional temporal information. Aspect also conveys information about the duration of an event, completion, or frequency, and if it is not translated well, much of the meaning is lost. Therefore, having aspectual information included in the translation process could greatly enrich it. It would also greatly improve temporal reasoning in NLP.

For example, Google Translate gives the following translations of the sentences below. In example (1) the aspect is perfective and the translation is correct. However, example (2) is imperfective and the translation should be '*Were you reading the book?*' or '*Have you been reading the book?*' Although there are indeed contexts in which the translation in the example (2) would be appropriate. In example (3), even with the aspectual marker '*how long*', which makes the aspect imperfective,

we get a perfective translation.

(1) *'Jesi li pročitala knjigu?'* – *'Have you read the book?'*

(2) *'Jesi li čitala knjigu?'* – *'Have you read the book?'*

(3) *'Koliko dugo si čitala knjigu?'* – *'How long did you read the book?'*

1.2. About Aspect

Aspect is a grammatical category that provides information on whether an action is completed, repeated, or in progress. Slavic languages can have perfective and imperfective verbal aspects. To simplify, the perfective aspect is used for single, completed actions, while the imperfective aspect is used to express actions that are ongoing, habitual or repeated. Perfective verbs are formed from their imperfective counterparts mostly by adding a prefix. Imperfective verbs are formed from perfective ones mainly by adding suffixes. There are also biaspectual verbs that can be used as both perfective and imperfective.

Generally, when prefixes are added, the meaning of the root is changed. Semantic changes caused by prefixes can be neutral, sublexical, and genuine lexical modification (Sussex and Cubberley, 2006). Richardson (2007) calls them purely perfectivizing, superlexical, and lexical prefixes, respectively. Most verbs usually have only one neutral prefix (Sussex and Cubberley, 2006). For example, in the case of *'pisati'* (write, imperfective) prefix *'na-'* would be the neutral prefix which produces *'napisati'* (to complete writing, perfective). If we add the prefix *'pre-'* - we would get *'prepisati'* (to copy by writing, perfective) which changes the meaning of the verb slightly and that would be a sublexical change.

For our resource, we decided to collect only the perfectives formed by purely perfectivizing and sublexical prefixes. Since this resource is made for learners of these languages, we decided not to match the verbs that have undergone a considerable semantic modification. For example, we decided not to match *'ispraviti'* (to correct, to align, perfective) with *'praviti'* (to make, imperfective). In cases like this, we decided that there is no sense match. These decisions were made using native speaker intuition with a subsequent check for archaic verbs.

2. Related Work and Resources

A similar resource has been created for the Russian — Database of Russian Verbal Aspect (Borik and Janssen, 2012). It is a part of the Open Source Lexical Information Network (OSLIN) for

Russian (Janssen, 2005).¹ Verbs with the same derivational base are linked to each other and classified as perfective, imperfective or biaspectual. However, there are no definitions of the verbs and no links to other resources. The database, however, provides clusters of verbs, that is, a list of all verbs that are morphologically or aspectually related to a base verb.

Samardžić and Miličević-Petrović (2013) have proposed a learner-friendly dictionary of verbal aspects in Serbian, but so far the resource has not been created. In later work, Samardžić and Miličević (2016) have also proposed a framework for the automatic classification of verbal aspect. A dataset of 2000 verbs based on this framework has been created for testing automatic classification. We are planning on using this resource in the future for our database, as it progresses.

3. Data Modelling with OntoLex-Morph

Since we originally decided to produce and publish our dataset as LLOD, we chose OntoLex-Lemon (McCrae et al., 2017) to model it, since it is the *de facto* standard for publishing lexical resources in RDF in accordance to the Semantic Web standards. The central element of the model is a `LexicalEntry`, which corresponds to lexemes or dictionary entries (Fig. 1).

Each verb in a pair should be modelled as such and there should be a relation between them, showing that they form an aspectual pair.

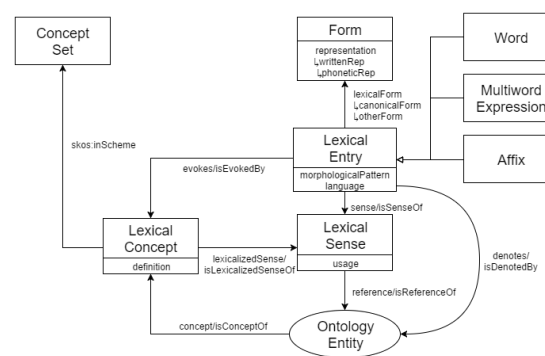


Figure 1: OntoLex-Lemon core model

Generally, to express lexico-semantic relations between two lexical entries, one would use a subclass of a class `LexicalRelation` from the *vartrans* module. But since this relation describes word formation, we decided to use the new OntoLex-Morph module (Chiarcos et al., 2022), which has a way of expressing the derivational relations between words.

¹<http://ru.oslin.org/>.

OntoLex-Morph (Fig. 2) consists of three parts: derivation (left), inflection (right), and information on how to generate new forms, both for inflection and derivation (top).

In order to represent this relation, we need to create a `WordFormationRelation` between the two entries and additionally specify a `DerivationRule` that provides information on how to create a written representation of a canonical form of the target entry. Below is an example of a relation between an imperfective verb *'parati'*, to tear apart and its perfective counterpart *'proparati'*, to tear apart successfully:

```
:parati a ontolex:LexicalEntry ;
  ontolex:canonicalForm [
    ontolex:writtenRep "parati"@sr
  ] .
:rel_pro_parati a morph:WordFormationRelation ;
  vartrans:source :parati ;
  vartrans:target :proparati ;
  morph:WordFormationRule
    :pro_prefix_rule .
:pro_prefix_rule a morph:DerivationRule ;
  morph:replacement [
    morph:source "^" ;
    morph:target "pro"
  ] .
```

Additionally, we create `Morph` objects for each prefix and add information about them to the corresponding rules:

```
:pro_prefix a ontolex:Morph, lexinfo:Prefix ;
  rdfs:label "pro-@sr ;
  morph:grammaticalMeaning [
    lexinfo:aspect lexinfo:Perfective
  ] .
:pro_prefix_rule morph:involves :pro_prefix_morph .
```

Having this along with triples describing the original lexical entries, we can either generate lexical entries of the pairs (pre-generate or create them on the fly every time they are requested) using the provided rules, or create them any other way (e.g., if we extract both entries from a database).

Here is an example of a SPARQL CONSTRUCT query that adds the rest based on the data described above:²

```
# PREFIXes are removed for brevity
CONSTRUCT {
  ?new_entry a ontolex:LexicalEntry ;
    ontolex:canonicalForm ?new_form ;
    decomp:subTerm ?prefix, ?source_entry .
  ?new_form a ontolex:Form ;
    ontolex:writtenRep ?new_string ;
}
WHERE {
  ?source_entry ontolex:canonicalForm ?source_form ;
  ?source_form ontolex:writtenRep ?base_string .

  ?wfRel a morph:WordFormationRelation ;
    vartrans:source ?source_entry ;
    vartrans:target ?new_entry ;
    morph:WordFormationRule ?rule .

  ?rule a morph:DerivationRule ;
    morph:replacement/morph:source ?srcPattern;
```

²The fragment described in this section, SPARQL query for generating derived forms and the full dataset are available at <https://github.com/max-ionov/aspect-db/tree/main/public/docs/ld12024/>.

```
morph:replacement/morph:target ?dstPattern;
morph:involves ?prefix .

?prefix morph:grammaticalMeaning [
  ?pred ?obj ;
] .

BIND (URI (CONCAT (STR (?new_entry), "_form"))
  AS ?new_form)
BIND (REPLACE (?base_string, ?srcPattern, ?dstPattern)
  AS ?new_string)
}
```

The output of the query for this example is the following:

```
:proparati a ontolex:LexicalEntry;
  ontolex:canonicalForm :proparati_form;
  decomp:subTerm :pro_prefix, :parati.
:proparati_form a ontolex:Form;
  ontolex:writtenRep "proparati"@sr.
```

4. Dataset Description and Conversion

For this paper, we decided to limit the database to aspectual pairs only formed by perfectivization. We extracted prefixes and the verbs they modify from Leximirka (Stanković et al., 2021), a lexicographic database with a web application for developing, managing and exploring lexicographic data in Serbian. It enables lexical entry control, automatic vocabulary enrichment, multiuser work, and establishment of relations among lexical entries.

Figure 3 shows the source data in the interface: on the left there is an imperfective verb *'raditi'* and it can be seen that it has several perfective pairs linked to it: *'proraditi'*, *'izraditi'*, *'zaraditi'*, *'naraditi'*, *'odraditi'*, *'doraditi'*. In the lower part of the panel, there are available markers: '+Imperf+Tr+It+Iref' meaning (i) imperfective, (ii) can be both transitive and intransitive, and (iii) non-reflexive. For the verb on the right, *'naraditi'*, markers are '+Perf+It+Ref' meaning (i) perfective, (ii) intransitive, and (iii) reflexive.

The rule-based system enables automatic linking between lexical entries in several different ways. One of them related to the linking perfective-imperfective pairs was used for this research.

Leximirka is interlinked with corpora, enabling developers and users to consult concordances and frequency lists for each lexical entry, being single- or multi- word unit, and its collocations. Currently, Leximirka supports Serbian Morphological Dictionaries (Krstev and Vitas, 2006) (Krstev, 2008), but it can support any other language as long as lexical data conform to the lexical database model (Stanković et al., 2018).

The Leximirka data category thesaurus controls the morphological, domain, syntactic, and semantic features that describe the lexical entries. The pronunciation markers are 'ljk' for ijekavian words and 'Ek' for ekavian-specific words. For example, child *'dijete'* has a marker 'ljk' for ijekavian, *'dete'* has 'Ek' for the ekavian pronunciation. In addition

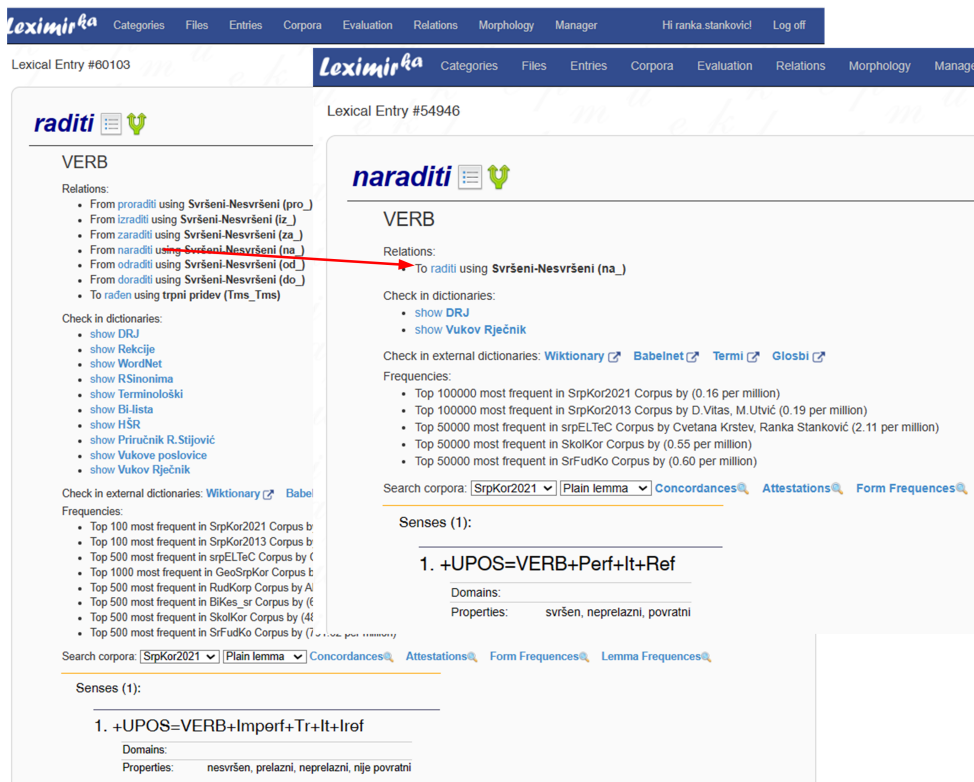


Figure 3: Leximirka: lexical entry for imperfective lexical entry left and one of the perfective pairs on the right side.

itself a part of the LLOD cloud (Orešković, 2019).

5. Deployment as a Static LLOD Resource

One of the problems that often arises in discussions about the adoption of LLOD technology is the lack of infrastructure and high technical requirements for publishing datasets in a way that they can be easily queried (Chiarcos, 2021; Gromann et al., in press, p. 27). The problem can be summarised as follows:³

- Data consumers want to be able to access the data in a convenient way, without downloading data dumps and setting up LD infrastructure on their side;
- Data providers do not want to have the burden of supporting SPARQL endpoints, or do not have the resources for sustainable long-term solution.

Most proposals to remedy this argue towards large infrastructures that could take the technological burden, and some already do (e.g., Databus,⁴

³These points were confirmed by a poll conducted at a plenary meeting of the COST Action NexusLinguarum which hosted both data consumers and data providers.

⁴<https://databus.dbpedia.org/>.

TriplyDB,⁵ and Semantic Media Wiki⁶). On the opposite side, Linked Data Fragments⁷ is an effort to put computational load on the side of the end-user, without them needing to pay the price of setting up the infrastructure (Heling and Acosta, 2020).

We argue that this direction — moving the computation to the side of the end-user — has more promise than relying on big infrastructure projects, both from the theoretical and practical sides: On the theoretical side, this is much more aligned with the decentralisation spirit of the World Wide Web and the LOD cloud; on the practical side, creating small independent services that do not have much technical requirements would make publishing Web Data more accessible.

More specifically, our approach is to use statically generated web pages with serialised RDF datasets and client-side SPARQL engine that queries these local datasets (with federated queries to remote ones, if needed).

In this way, just by serving static web sites, data providers can simultaneously distribute their datasets in three different ways with (i) different levels of availability, (ii) usability, and (iii) oriented towards different groups:

⁵<https://triple.cc/>

⁶<https://www.semantic-mediawiki.org/>

⁷<https://linkeddatafragments.org/>

- RDF dumps that can be downloaded and used independently most availability, least usable for end-users, oriented towards people who need unrestricted access to data for parsing or converting;
- A remote endpoint that can be loaded or queried with a SPARQL engine less availability, more usable for end-users, oriented towards people who want to query the data;
- Web page with predefined functionality defined by the data provider least available, most usable for the end-users. For people who want to use the service provided on top of the data.

This approach significantly lowers the requirements to host a website showcasing a dataset in RDF. Due to the fact that many organisations provide free hosting solutions for static web sites, it is not necessary to have access to a server with specialised software installed. And unlike relying on specialised solutions like Triply, this does not create vendor lock since there are many options for hosting a static site.

To showcase our approach, we deployed the database of aspectual pairs as a static website hosted on GitHub Pages.⁸

The page presents the project and allows interaction with the data: searching for an aspectual pair for a verb and for getting all the verbs that use a certain prefix for perfectivisation. Both functions correspond to a SPARQL query that is being run on live data on the side of the client. The dataset does not need to be downloaded and the user does not need to set anything up, since JavaScript-based SPARQL engine Comunica⁹ queries the remote file.

The website is being regenerated with every commit to the repository, and using the dynamic routing system, VitePress,¹⁰ the static site generator that we use generates static pages that dereference local URIs of the dataset.

6. Conclusion and Future Work

In this paper, we have presented a new openly available language learning resource for learning verbal aspect in Bosnian, Croatian, and Serbian. The resource is available as an RDF dump, as an endpoint and as an interface, built on top of the endpoint.

Currently, the dataset consists of aspectual pairs in which perfective forms are formed by prefixation. In the future, we will expand this resource

to include other types of word formation, for example, imperfectivization done by suffixation.

The second result of this paper is a proposed approach to democratise publishing LLOD datasets by using client-based RDF technology. We believe that this is the way to increase the number of accessible usable LLOD resources and help their adoption for end-user applications.

The possible limits of this approach — how much data can be handled in this way and how performant it can be — is still an open question.

Regardless of that, this approach is a working solution for small- and medium-scale datasets, so it could be adopted by student and small research projects as a way to present and preserve the results.

7. Acknowledgements

This research was supported by the COST Action NexusLinguarum (CA18209) – “European network for Webcentered linguistic data science”. The authors also want to thank the organizers and the lecturers of the 5th Summer Datathon on Linguistic Linked Open Data held in Zaprešić, Croatia in June 2023, where this project started.

8. Bibliographical References

- Olga Borik and Maarten Janssen. 2012. A database of russian verbal aspect. In *Proceedings of the conference Russian Verb, St. Petersburg, Russia*.
- Christian Chiarcos. 2021. [Get! Mimetype! Right!](#) In *3rd Conference on Language, Data and Knowledge (LDK 2021)*, volume 93 of *Open Access Series in Informatics (OASIs)*, pages 5:1–5:4, Dagstuhl, Germany. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.
- Christian Chiarcos, Katerina Gkirtzou, Anas Fahad Khan, Penny Labropoulou, Marco Passarotti, and Matteo Pellegrini. 2022. Computational morphology with ontalex-morph. In *LREC 2022 Workshop Language Resources and Evaluation Conference 20-25 June 2022*, page 78.
- Dagmar Gromann, Elena-Simona Apostol, Christian Chiarcos, Marco Cremaschi, Jorge Gracia, Katerina Gkirtzou, Chaya Liebeskind, Verginica Mititelu, Liudmila Mockiene, Michael Rosner, et al. in press. [Multilinguality and LLOD: A Survey Across Linguistic Description Levels](#). *Semantic Web Journal*.

⁸<https://ionov.me/aspect-db/>.

⁹<https://comunica.dev/>.

¹⁰<https://vitepress.dev/>.

- Lars Heling and Maribel Acosta. 2020. Cost- and robustness-based query optimization for linked data fragments. In *The Semantic Web–ISWC 2020: 19th International Semantic Web Conference, Athens, Greece, November 2–6, 2020, Proceedings, Part I* 19, pages 238–257. Springer.
- Maarten Janssen. 2005. Open source lexical information network. In *Third international workshop on generative approaches to the lexicon*, pages 400–401. Citeseer.
- Cvetana Krstev. 2008. *Processing of Serbian. Automata, texts and electronic dictionaries*. Faculty of Philology of the University of Belgrade.
- 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.
- Marko Orešković. 2019. *An Online Syntactic and Semantic Framework for Lexical Relations Extraction Using Natural Language Deterministic Model*. Ph.D. thesis, University of Zagreb. Faculty of Organization and Informatics.
- Kylie R Richardson. 2007. *Case and aspect in Slavic*. Oxford University Press.
- Tanja Samardžić and Maja Miličević. 2016. A framework for automatic acquisition of croatian and serbian verb aspect from corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).
- Tanja Samardžić and Maja Miličević-Petrović. 2013. Constructing a learner-friendly corpus-based dictionary of serbian verbal aspect. *Primenjena lingvistika*, 14:77–89.
- Ranka Stanković, Cvetana Krstev, Biljana Lazić, and Mihailo Škorić. 2018. Electronic dictionaries—from file system to lemon based lexical database. In *6th Workshop on Linked Data in Linguistic (LDL-2018), Towards Linguistic Data Science*.
- Roland Sussex and Paul Cubberley. 2006. *The slavic languages*. Cambridge University Press.
- Ranka Stanković and Mihailo Škorić and Biljana Lazić and Cvetana Krstev. 2021. *Leximirka lexical database*. ELG, <https://live.european-language-grid.eu/catalogue/tool-service/17356>.
- [european-language-grid.eu/catalogue/lcr/17355](https://live.european-language-grid.eu/catalogue/lcr/17355), 1.0.

9. Language Resource References

- Cvetana Krstev and Duško Vitas. 2006. *SrpMD - Serbian morphological dictionaries*. ELG, <https://live.european-language-grid.eu/catalogue/lcr/17355>.

Towards Semantic Interoperability: Parallel Corpora as Linked Data Incorporating Named Entity Linking

Ranka Stanković*^{id}, Milica Ikončić Nešić[†]^{id}, Olja Perišić[‡]^{id},
Mihailo Škorić*^{id}, Olivera Kitanović*^{id}

*University of Belgrade, Faculty of Mining and Geology, Serbia
{ranka.stankovic,mihailo.skoric,olivera.kitanovic}@rgf.bg.ac.rs

[†]University of Belgrade, Faculty of Philology, milica.ikoncic.nesic@fil.bg.ac.rs,

[‡]University of Turin, Italy, olja.perisic@unito.it

Abstract

The paper presents the results of the research related to the preparation of parallel corpora, focusing on transformation into RDF graphs using NLP Interchange Format (NIF) for linguistic annotation. We give an overview of the parallel corpus that was used in this case study, as well as the process of POS tagging, lemmatization, and named entity recognition (NER). Next, we describe the named entity linking (NEL), data conversion to RDF, and incorporation of NIF annotations. Produced NIF files were evaluated through the exploration of triplestore using SPARQL queries. Finally, the bridging of Linked Data and Digital Humanities research is discussed, as well as some drawbacks related to the verbosity of transformation. Semantic interoperability concept in the context of linked data and parallel corpora ensures that data exchanged between systems carries shared and well-defined meanings, enabling effective communication and understanding.

Keywords: parallel corpora, named entity linking, named entity recognition, NER, NEL, linked data, NIF, Wikidata

1. Introduction

The motivation for publishing parallel corpora as linked data lies in the benefits of increased accessibility, interoperability, semantic enrichment, community collaboration, and the promotion of open science. These motivations collectively contribute to advancing linguistic research, language technology, and cross-disciplinary insights.

Parallel corpora are essential for multilingual studies, and publishing them as linked data simplifies cross-lingual research. Researchers can efficiently compare and analyze texts in multiple languages, enabling more comprehensive linguistic and cultural studies. Linked data enables semantic enrichment through the integration of annotations, linguistic metadata, and cross-lingual alignments. This enrichment provides deeper context and insights for linguistic research, machine translation, and language technology development.

Previous successful use cases of representation of the linguistic annotations of textual data in RDF (Stanković et al., 2023; Stanković et al., 2024) using NLP Interchange Format (NIF) (Hellmann et al., 2013) inspired this research. NIF facilitates the annotations of various types of linguistic data, e.g. part-of-speech, lemmas, and named entities. By using string-based URIs (Uniform Resource Identifier), NIF additionally accommodates multilingual text materials, allowing the annotations of translation equivalents across different languages via RDF properties. This is directly aligned with the activities of Nexus Linguarum COST Action (Declerck

et al., 2020), devoted to the creation, interlinking, enrichment, and evolution of linguistic resources, especially in the context of under-resourced languages and domains. In this paper, the showcase of the Italian-Serbian parallel corpus will be used to illustrate previously mentioned possibilities for annotation and linking.

The Serbian language boasts a rich and intricate morphology, allowing for the declension of toponyms and other proper nouns, which foreign students may not always find easy to identify and derive to their basic form (lemma) searchable in dictionaries and encyclopedias. Some of the difficulties are the transcription of proper names e.g. *Džon* (John), *Đovani* (Giovanni), their declension (*Đovaniju*, loc./dat., *Džona*, acc.), the formation of possessive adjectives from personal names such as *Đovanijev* (m.sg., Đovani's) and *Džonove* (f.pl., John's) all subject to declension. Conversely, Italian, lacking grammatical cases, conveys numerous syntactic relationships through the use of prepositions. For instance, "di Giovanni" can be rendered in Serbian as a possessive adjective, such as "*Đovanijev(a/o/i/e/a)*", or as a genitival phrase, "*od Đovanija*" with its precise semantic interpretation heavily contingent on the context (of Giovanni, by Giovanni...).

To overcome these and similar problems the project "It-Sr-NER: Web services for named entity recognition, linking, and mapping" was implemented as part of CLARIN's "Bridging Gaps" call in 2022 (Perisic et al., 2023). Within this project, web services were developed for annotating named

entities in text, namely personal names, places, organizations, ethnicities, events, and works of art.

The project participants were experts from Serbian and Italian academic institutions: University of Turin and the Society for Language Resources and Technologies JeRTeh. The result was the creation and publication of web applications and services for annotating named entities (NE) in monolingual and bilingual parallel texts for 24 languages, with a case study focused on Italian and Serbian parallel texts. Furthermore, an Italian-Serbian parallel corpus comprising 10,000 segments of extracted and aligned sentences, chosen from classic works of Italian and Serbian literature, was also created and made publicly available (Perišić et al., 2022b).¹

The main research objective and contribution of this paper was to provide an existing parallel corpus as linked data that adheres to standardized formats and structures, ensuring interoperability with other datasets and systems. This interoperability will allow researchers to integrate parallel corpora into larger linguistic databases or use them in conjunction with other linked data resources for more comprehensive analyses. The developed procedure can be used for other monolingual or parallel corpora, and thus serve as a point of orientation for future publication workflows of multilingual corpus data publication on the web.

Several aligned corpora exist in which Serbian is one of the languages. In most cases, the second language is English or French, while corpora including the Serbian-Italian combination are rare. Additionally, we see a special contribution to our work in discussing how to establish bridges between Linked Data technologies developed for NLP and data produced and consumed in digital humanities.

In Section 2 we give a short overview of related work concerning the preparation and annotation of parallel corpora, named entity recognition and their linking, linked data standards for corpora, and NLP Interchange Format (NIF) for linguistic annotation. Section 3 brings an overview of the parallel corpus that was used in this case study, the process of POS-tagging and lemmatization, as well as named entity recognition (NER). Section 4 describes integration results: NEL, data conversion to RDF, incorporation of NIF annotations, while in Section 5 validation of produced NIF through the exploration of triplestore using SPARQL is described and the NER and linking is presented. Section 6 is dedicated to the bridging of Linked Data and Digital Humanities research. The concluding remarks and plans for future research are given in Section 7.

¹It-Sr-NER: CLARIN compatible NER and geoparsing web services for parallel texts: case study Italian and Serbian

2. Related Work

2.1. Parallel corpora

More than ten years ago Zanettin (2012) emphasized the limited availability of readily accessible sources of parallel corpora across various domains and text genres. The availability of parallel corpora remains limited even for languages with a large number of speakers and a wide range of digital resources despite the ever-increasing demand for them. These available parallel corpora often serve as examples for testing new tools and methods for the less spoken languages with limited resources and for which translations of literary works and other texts are primarily in print, going slowly through digital conversion (Jenn and Fraise, 2022).

Although the significance of parallel corpora in literary and translation studies has been confirmed (Moratto and Li, 2022), literary parallel corpora are particularly challenging to create due to the increased resources required for their development and concerns related to copyright issues (Dimiitroulia, 2023). If recent research has shown that the potential of parallel corpora remains invisible and unknown to most literary translators, the introduction of these technologies into the education of future translators could bring about a change in this trend. At the same time, the exploration of parallel corpora can improve reciprocal language learning from a contrastive perspective that enables the observation of different cross-cultural and linguistic asymmetries (Hunston, 2002).

2.2. Corpus Linked Data Standards

NIF and Web Annotation are two well-known RDF standards for linguistic annotation. Both specifications use URIs (or IRIs) to address corpora, which is similar to how URIs are used in other formats. The ‘Best Practices for Multilingual Linked Open Data’ (BPMLOD) W3C community group and the LIDER project’s² results were used in addition to NIF standards, since standards themselves are somewhat technical and not very user-friendly. This document describes NIF as a format for corpus data.³

In (Hellmann et al., 2013) NIF was employed as the corpus format to ensure compatibility with DBpedia through Linked Data and to facilitate interoperability with NLP tools. DBpedia abstracts were one of the first implementations of NIF (Brümmer, 2015; Brümmer et al., 2016) on an open, large-scale corpus of annotated Wikipedia texts in six languages, with over 11 million texts and more than 97 million entity links.

²<https://lider-project.eu>

³BPMLOD-NIF, <http://bpmlod.github.io/report/nif-corpus/index.html>

FrameNet (FN) lexical database for English has been published as RDF Linked Open Data (LOD), including the corpus of text that has been annotated using FN. [Alexiev and Casamayor \(2016\)](#) compared FN-LOD with NIF, and proposed to integrate FN into NIF. The widely used standards for linguistic annotations in RDF are: 1) Annotation ([Sanderson et al., 2013](#)), published as a W3C standard (recommendation) in 2017;⁴ 2) POWLA ([Chiarcos, 2012](#)), a reconstruction of the Linguistic Annotation Framework ([Ide and Suderman, 2014](#)) in OWL2/DL; 3) CoNLL-RDF focusing on the compatibility with tabular ('CoNLL') formats as used in NLP ([Chiarcos and Glaser, 2020](#)).

While describing principles for annotating text data using RDF-compliant formalism to be accessible from the LLOD ecosystem, [Cimiano et al. \(2020a\)](#) recommended including the full text of the annotated document in the RDF data, to preserve interoperability.

After studying the relevant literature and taking into consideration the characteristics of our data, we decided to follow the BPMLOD draft recommendation and apply NIF (version 2.0) to our data, similar to our approach in the previous project ([Stanković et al., 2023](#); [Stanković et al., 2024](#)). We are working with an annotated parallel corpus, which opens up opportunities to explore the potential of RDF technology for cross-lingual linking, as well as for the linking of corpora with annotations and lexical resources.

2.3. NLP Interchange Format (NIF) for Linguistic Annotation

NIF is a community standard developed through a series of research projects at the AKSW Leipzig, Germany, and maintained by the same group. A typical URI/IRI consists of two main components, a base name that serves to locate the document, and an optional fragment identifier. For different media types and 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. \(2022\)](#) report that this is one area where there is a real necessity for documentation that provides clear guidelines (GLs) and BPs. The research we present could be a showcase for the use of NIF and the transformation of parallel corpora to NIF. This paper contributes to this effort by providing a case study on the use of 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

⁴<https://www.w3.org/TR/annotation-model/>

2.0.⁵ The reason not to use the latest version 2.1 of NIF Ontology is the lack of full documentation.

3. Data Preparation and Preprocessing

3.1. Description of the Parallel Corpus

The Italian-Serbian corpus It-Sr-NER ([Perišić et al., 2022b](#)) consists of 10,000 aligned segments (sentences) taken from Italian and Serbian translations of ten different novels. For the presented work, 1000 aligned sentences from various novels were selected. Table 1 presents an overview of the novels in It-Sr-NER, where the last column designates the novels whose sentences belong to the 1000-sentence corpus.

The novels were aligned and converted into the TMX (Translation Memory eXchange) ([Serge, 2020](#)) format using the ACIDE program for creating parallel corpora ([Obradović et al., 2008](#); [Krstev and Vitas, 2011](#)). Each segment in Italian and Serbian is numbered and paired with the corresponding language segment(s) indicated by the "xml:lang" attribute.⁶ The It-Sr-NER corpus is available on the CLARIN Center and can be accessed through the VLO (Virtual Language Observatory) and Language Resource Switchboard. The corpus includes the aligned bilingual version, as well as individual monolingual versions, and named entities that have been automatically annotated ([Perišić et al., 2022a](#)). Additional information can be found in the *GitHub*⁷ and it is searchable in the Bibliša digital library ([Stanković et al., 2018](#)) ([Stanković et al., 2017](#)).⁸

The resources developed in this project are open and accessible to researchers, teachers, and students, but the biggest benefit will be for those interested in the Italian language in Serbia and the Serbian language in Italy. Given the polycentrism of the Serbo-Croatian language, the students and teachers in Croatian, Montenegrin, and Bosnian universities and schools could also benefit from this corpus and web services.

3.2. POS tagging and lemmatization

The complete parallel corpus was annotated with part-of-speech (POS) tags using *Universal POS* tagset, and lemmas.

⁵<https://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core/nif-core.html>

⁶TXM file of the novel "Around the World in Eighty Days"

⁷It-Sr-NER GitHub repository

⁸Bibliša digital library

Name of Novel	Novel Name Translation	Samples in NIF
Il nome della rosa	The Name of the Rose	✓
Le avventure di Pinocchio	The Adventures of Pinocchio	✓
Storia di chi fugge e di chi resta, L'amica geniale	Those Who Leave and Those Who Stay	✓
Uno, nessuno e centomila	One, None and a Hundred Thousand	✓
Anikina vremena	Legends of Anika	✓
Na drini ćuprija	The Bridge on the Drina	✓
Nečista krv	Impure Blood	
Opštinsko dete	Municipal child	
Bašta, pepeo	Garden, Ashes	
Le Tour du monde en quatre-vingts jours	Around the World in Eighty Days	

Table 1: An overview of the novel samples included in the corpus.

The Serbian part of the corpus was annotated using a multi-model tagger for the Serbian language, *BEaST* (Stanković et al., 2022) which uses both *TreeTagger* (Schmid, 2013) and *spaCy*⁹ models trained on part-of-speech tagging task using the manually annotated, publicly available corpus *Srp-Kor4Tagging* (Vitas et al., 2021). The lemmatization is performed after the POS-tagging step, using electronic morphological dictionaries for Serbian (Krstev and Vitas, 2006) (Krstev, 2008; Vitas and Krstev, 2012), incorporated through the aforementioned *TreeTagger* model.

The Italian part of the corpus was annotated using *spaCy* model for Italian (Explosion, 2022), using the UD annotation scheme obtained by conversion from the Italian Stanford Dependency Treebank, released for the dependency parsing shared task of Evalita-2014 (Bosco et al., 2014).

3.3. Named Entity Recognition

For NER in Serbian texts *Jerteh-355-tesla* (Ikonić Nešić et al., 2024), a version of *Jerteh-355* (Škorić, 2024) language model was used. *Jerteh-355*, based on the RoBERTa-large architecture (Liu et al., 2019), was pre-trained for Serbian on a diverse corpus of approx 4 billion tokens. *Jerteh-355-tesla* was fine-tuned specifically for NER task, using *spaCy* framework on the corpus of Serbian novels published between 1840 and 1920, named *SrpELTeC-gold* (Krstev et al., 2021), newspaper articles and sentences generated from the Wikidata (Wikimedia, 2023) and *Leximirka* lexical database (Stanković et al., 2021). It achieves an F_1 score of approx 96% on the test dataset.

For the Italian language texts, *spaCy* model *it_core_news_sm-3.4.0* (Explosion, 2022) was used, which was trained on a synthetic NER corpus *WikiNER*, based on the text and structure of Wikipedia (Nothman et al., 2013). The model achieved F_1 score of 86% on the test set.

After automatic annotation, the *INCEpTION* (Klie et al., 2018) was used for manual correction and linking of named entities. In this paper, the focus was on the three most frequent types of named entities across language-specific models: persons

(<PERS>), locations (<LOC>), and organizations (<ORG>), as explained in Subsection 2.3.

Table 2 presents statistics of several named entities per class in Serbian (sr) and Italian (it) datasets, with explanations of entity types.

4. Integration

4.1. Named Entity Linking

After annotating the parallel corpus as described in the previous section, the next step was to link entities belonging to one of the NE classes with Wikidata (Wikimedia, 2023). Extracted PERS entities refer mostly to the characters of novels, LOC entities designate places where the action of a novel takes place (geopolitical locations) while ORG represents organizations mentioned in novels. Entries in Wikidata didn't exist for characters of some novels; thus, similar to the approach in (Ikonić Nešić et al., 2021), the *OpenRefine* (David Huynh, 2022) and *QuickStatements* (Manske, 2019) were used to create 111 appropriate items for 5 novels (56 characters of the novel "*Storia di chi fugge e di chi resta, L'amica geniale*" (Q55517451) by Elena Ferante). For novel "*Le avventure di Pinocchio*" (Q8065468) all characters were already in Wikidata.

The named entities for both languages were linked with Wikidata in additional layer of annotation a Wikidata identifier is assigned to each entity. For example, *Jakša*, a character from the novel "Legends of Anika" (wd:Q61133860), is recognized as a person, assigned NE tag <PERS> and linked with URL <http://www.wikidata.org/entity/Q122730462>. The annotation and linking with Wikidata using the *INCEpTION* tool is presented in Figure 1. Two more entities are recognized in the text presented in this figure: PERS *Anika* (wd:Q122730455) and LOC *Višegrad* (wd:Q239266).

The full process of linking entities with knowledge bases using the *INCEpTION* annotation platform is described in (Klie et al., 2020).

For annotating named entities (NE), several ontologies were consulted. The following NE type equivalents were used

⁹SpaCy

Entity	Explanation	sr	it
PERS Personal names	First names, surnames, nicknames and their combinations (of real people and fictional characters, gods and saints).	901	1036
LOC Locations	Continents, countries, regions, populated places, oronyms, water surfaces, names of celestial bodies, city locations.	257	310
ORG Occupations and titles	Names of companies, political parties, educational institutions, sports teams, hospitals, museums, libraries, hotels, cafes, and places of worship.	31	30

Table 2: Number of named entities per class

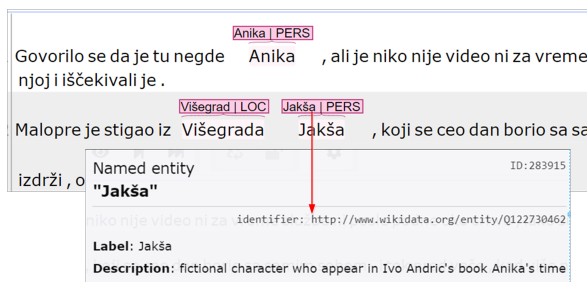


Figure 1: An annotated example

from OLIA:¹⁰ `olia:Person`, `olia:Space`, `olia:Organization`. `dbo`¹¹ namespace was introduced to link NEs with DBpedia, and `wd` namespace for Wikidata. The following classes were used to link types of recognized NEs: `dbo:Person = wd:Q5`, `dbo:Place = wd:Q7884789`, `dbo:Organisation = wd:Q43229`.

4.2. Data Conversion to RDF

A collab notebook was prepared for the transformation of the parallel corpus into NIF. The library `rdflib`¹² was used for RDF management.¹³ Code comprises classes `Corpus_mono`, `Sentence`, `Word`, `NamedEntity`, `Corpus_bili` for necessary transformations and a set of additional functions. `Corpus_mono` takes as input TSV file with annotations and produces a RDF graph (a `ttl` file) for one language, instantiating further for each sentence an object of a `Sentence` class, that produces RDF triples related to the object of `nif:Sentence` type. Further, class `Word` manages tokens from the file and generates RDF triples for objects of the `nif:Word` type, while `NamedEntity` finds the words (and tokens) that belong to one name entity, specify its type, and link it to Wikidata, if exists.

¹⁰http://purl.org/olia/discourse/olia_discourse.owl

¹¹<https://dbpedia.org/ontology/>

¹²<https://rdflib.readthedocs.io/en/stable/>

¹³The code is available in the GitHub repository.

For interlinking sentences that are translation units, class `Corpus_bili` is used.

Two monolingual corpora consist of the same number of segments, that are aligned as translation equivalents. Since NIF does not support translation units and translation unit variants (as TMX standard), the sentence class `nif:Sentence` is used, as the most similar NIF concept.

The main function `write_gcorpus_mono` instantiate RDF Graph with the following namespaces: `itsrdf`, `nif`, `olia`, `dc`, `dct`, `ms`, `wd`, `wdt`, `dbo`, `eltec`. After `Corpus_mono` is created, the first set of triples is introduced to the monolingual corpus.

Figure 2 presents an outline of the model for a parallel corpus in NIF.

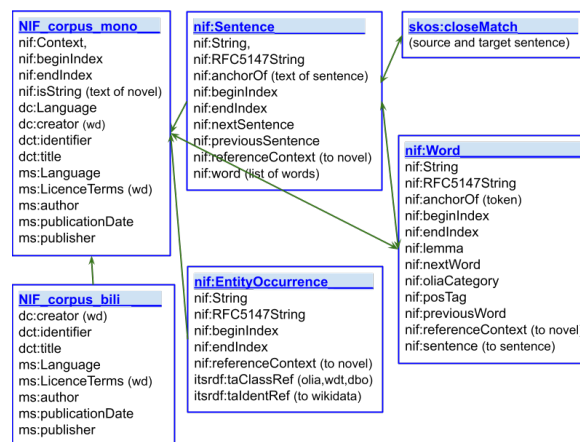


Figure 2: A data model for a parallel corpus in NIF

For establishing links between translation equivalents in different languages we used `skos:closeMatch` from SKOS (Simple Knowledge Organization System).¹⁴ The `skos:closeMatch` property indicates that the two objects are sufficiently similar that they can be used alternately in applications.

4.3. Incorporating NIF Annotations

NIF Terse RDF Triple Language (`ttl`) was used as a serialization for transformation into linked data.

¹⁴<https://www.w3.org/TR/skos-reference/>

Dataset with 1000 aligned sentences within six *ttl* files derived from the corpus described in Subsection 3.1, is published and included in LRE map.¹⁵

The core class `nif:String` is used for the monolingual corpus content itself (a text in Italian and a corresponding text in Serbian), described by `nif:beginIndex` and `nif:endIndex`. Dublin Core vocabulary is used for predicates related to the language, author, identifier, and title. META-SHARE ontology¹⁶ is used to describe language, license terms, author, publisher, and publication year.

For illustration, we will present a part of an Italian sentence “- *Dunque, compar Geppetto, - disse il falegname in segno di pace fatta, - qual è il piacere che volete da me ?*”¹⁷ from the novel “The Adventures of Pinocchio”¹⁸ and discuss some of its parts. The main class `nif:String` represents strings of Unicode characters. The subclass of `nif:String` is `nif:Context`, that represents a text in its entirety and holds the characters of this text in the `nif:isString` property. A substring of the `nif:Context` can be: a single word, a sentence, or a named entity that is linked to the relevant `nif:Context` resource via `nif:referenceContext`. Beginning and end indices refer to the string content (sentence) represented by the context. The previous and the next sentence are references as well as a list of words.

```
<http://url/it1.txt#char=105530,105643>
a nif:RFC5147String, nif:Sentence,
nif:String;
nif:anchorOf "- Dunque , compar
Geppetto , - disse il falegname
in segno di pace fatta , - qual è il
piacere che volete da me ?" ;
nif:beginIndex "105530" ;
nif:endIndex "105643" ;
nif:nextSentence
<http://url/it1.txt#char=105644,105851>;
nif:previousSentence
<http://url/it1.txt#char=105356,105529>;
nif:referenceContext
<http://url/it1.txt>;
nif:word
<http://url/it1.txt#char=105530,105531>,
<http://url/it1.txt#char=105532,105538>,
...
<http://url/it1.txt#char=105642,105643>;
dct:identifier "585" .
```

¹⁵Uncompressed files are accessible at: [URL](#), with a CCA 4.0 International license. Zipped files will be available also at CLARIN, the European Language Grid portal, and other language repositories.

¹⁶<http://w3id.org/meta-share/meta-share/2.0.0>

¹⁷“So, Compare Geppetto, - said the carpenter as a sign of peace made, - what pleasure do you want from me?”

¹⁸[Le-avventure-di-Pinocchio.xml](#)

The following classes: `nif:Word`, `nif:Phrase`, `nif:Sentence` represent a segment of a text, depending on the unit of annotation. The property `nif:referenceContext` points to the respective `nif:Context` instance of the text segment. The segment position inside the context is specified using the `nif:beginIndex` and `nif:endIndex` properties. The actual text segment can be specified using the `nif:anchorOf` property.

The following listing presents triplets for tokens (words). Apart from text segments (indices), additional grammatical information and relations can be included. The information about the part of speech can be linked using the `nif:posTag` property, while for the canonical form the `nif:lemma` property is used. Previous and next words are linked with the following properties: `nif:previousWord` and `nif:nextWord`. To link a word or a named entity with its sentence the `nif:sentence` property is used.

```
<http://url/it1.txt#char=105541,105547> a
nif:RFC5147String, nif:String, nif:Word;
nif:anchorOf "compar";
nif:beginIndex "105541";
nif:endIndex "105547";
nif:lemma "compar";
nif:nextWord
<http://url/it1.txt#char=105548,105556>;
nif:oliaCategory olia:CommonNoun ;
nif:posTag "NOUN";
nif:previousWord
<http://url/it1.txt#char=105539,105540>;
nif:referenceContext
<http://url/it1.txt> ;
nif:sentence
<http://url/it1.txt#char=105530,105643>.
```

In this particular scenario, it is evident that `itsrdf:taClassRef` is employed to connect with the relevant category of named entities, such as individuals, places, or organizations. When dealing with individuals (person), various ontologies are utilized, including `olia:Person` from Olia ontology, `wdt:Q5` from Wikidata, and `dbo:Person` from DBpedia.

```
<http://url/it1.txt#char=105007,105015> a
nif:RFC5147String, nif:String, nif:Word;
nif:anchorOf "Geppetto";
nif:beginIndex "105007";
nif:endIndex "105015";
...
itsrdf:taClassRef olia:Person,
wdt:Q5, dbo:Person ;
itsrdf:taIdentRef wdt:Q1428120 .
```

Figure 3 presents the transformation of novels into aligned TMX-XML, annotation in TSV files (NER+NEL) into RDF (NIF).

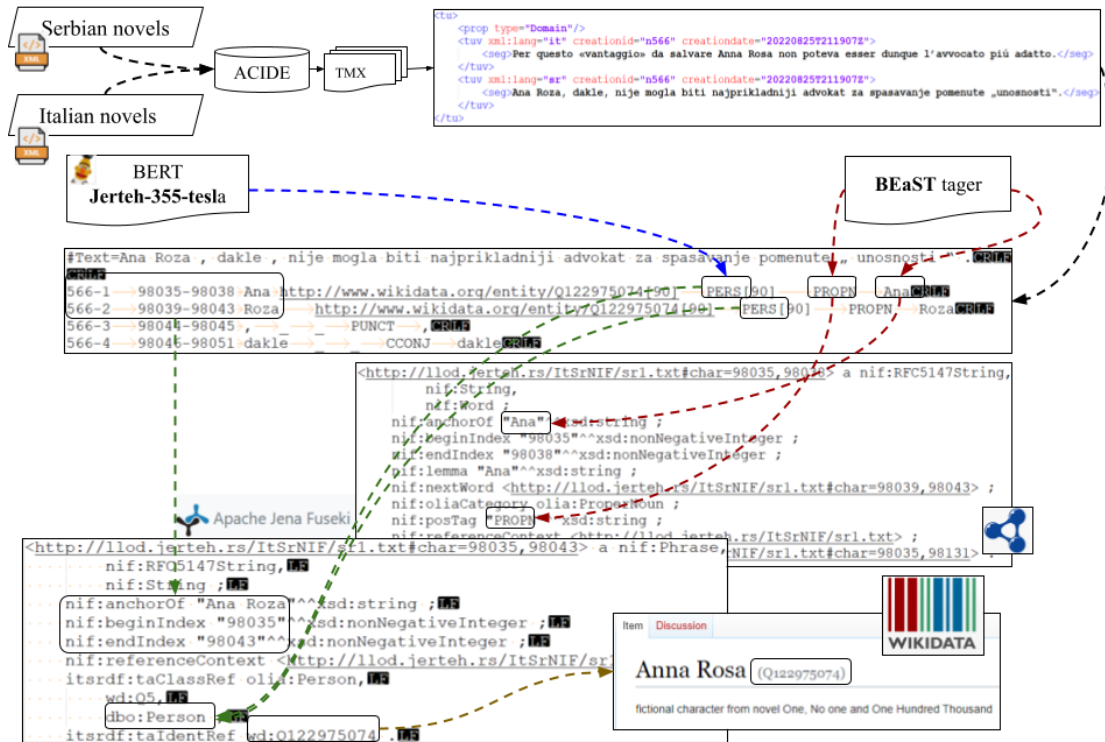


Figure 3: Workflow of the transition from a novel to LOD

5. Querying It-Sr-NER using SPARQL

Apache Jena Fuseki (Apache Software Foundation, 2023) is used for the management of the RDF graphs in the form of *ttl* files. Dataset *ItSrNIF* was created by uploading all files which generated 1,002,834 triples for 1000 sentences in each language. The Italian part of the corpus has 36,457 words, 1036 persons (*wd:Q5*), 310 toponyms (*wd:Q7884789*), 30 organizations (*wd:Q43229*), while Serbian part has 33,514 words, 901 persons, 257 toponyms, 31 organization.

The following query presents SPARQL query in Fuseki presenting retrieved result with aligned sentences.

```
SELECT ?sr ?srt ?it ?itt
WHERE {
  ?sr a nif:Sentence ;
    nif:anchorOf ?srt .
  ?it a nif:Sentence ;
    nif:anchorOf ?itt .
  ?it skos:closeMatch ?sr .
}
```

The query retrieves Serbian sentences represented by variables *?sr* (sentence ID) and *?srt* (sentence itself), Italian sentences represented by variables *?it* and *?itt*, while the query constraint demanding a link of a type *skos:closeMatch* between the sentence identifiers *?sr* and *?it* ensures that sentences are translation equivalents.

Figure 4 presents a Fuseki screenshot with

SPARQL query for counting and presenting aligned named entities in Serbian, Italian, and their Wikidata URI.

Figure 4: SPARQL query with aligned sentences

6. Discussion

The presented research connects the previous results from the fields of Digital Humanities (Ikonić Nešić et al., 2022) and Linked Data (Hell-

Lng	txt	tsv	tfl	Fuseki
it	0.17	1.7	24.4	/
sr	0.19	1.5	26.5	/
All	0.36	3.2	51.0	317

Table 3: Size of files in MB. **txt** - plain text, **tsv** - tab separated INCEpTION format (POS, lemmas, NER, NEL), **tfl** - NIF files, **Fuseki** -whole repository.

mann et al., 2012; Brümmer, 2015; Alexiev and Casamayor, 2016; Cimiano et al., 2020b) which are traditionally considered separate areas of research. Parallel corpora are widely used in translation studies, while Linked Data focuses on interlinking and integrating diverse datasets. The integration of parallel resources with the broader Linked Data ecosystem, described in this paper, contributes to the efforts to bridge the gap between these two areas.

We are aware that NIF has some potential downsides, one of which is a high degree of verbosity. Therefore, the scalability issues for such kinds of data should be carefully planned. Table 3 gives an overview of the differences in size that can be expected for different levels of annotation and formats, taking as an example the data set with 1000 sentences. It can be seen that the size of NIF files is 16 times larger than TSV version with similar information, while the Fuseki repository size for both languages and for the same dataset is more than 6 times larger than the repository with tfl files.

The presented pipeline transforming parallel corpus into NIF-linked data (Figure 3), offers several benefits: multilingual research and translation, cross-lingual information retrieval, multilingual information extraction, cultural and societal insights, and bridging language barriers. In summary, the benefits of parallel corpus NIF linked data, extend to various domains, including machine translation, linguistics, language learning, and cross-lingual information access, making it a valuable resource for researchers, businesses, and individuals seeking to bridge language gaps and expand their global reach. Analyzing parallel corpus data in a distributed environment using federated SPARQL queries can reveal cultural and societal differences in how topics are discussed and portrayed across languages.

The greatest benefits will be in the field of translation, encompassing teaching and lexicography, especially in resolving cases of lexical anisomorphism. This phenomenon results not only from linguistic asymmetry but also from cultural differences, so this insight can be valuable for cross-cultural studies and international business strategies. The varied lexical realization of a concept or its lack of lexicalization creates lexical gaps that can be identified, understood, and translated by applying

targeted translation strategies. These strategies are made possible through data linking with other layered multilingual resources. Through this approach, the semantic essence of every word can be grasped, beginning from individual concepts and extending to their functional manifestation within the context.

7. Conclusion

One way to achieve semantic interoperability is by leveraging parallel corpora and incorporating NEL. By representing parallel corpora as linked data, we can establish links between equivalent concepts or entities in different languages, thereby enhancing cross-lingual information exchange. This paper demonstrated NEL for people, organizations, and locations by linking their references in texts to their corresponding entries in Wikidata. By linking these entities to standardized identifiers or ontologies, the interoperability of data is greatly improved. Incorporating NEL into parallel corpora as linked data not only enhances cross-lingual interoperability but also fosters better integration with the broader semantic web. When parallel corpora are exposed as linked data, they become part of the larger network of linked open data, allowing for a more comprehensive and coherent exchange of information. Further research will include NEL model training (Upadhyay et al., 2018), as well as the publication of all 10,000 aligned segments in NIF.

8. Acknowledgements

The authors extend their gratitude to Prof. Cvetana Krstev for her invaluable contributions. This research was supported by the Science Fund of the Republic of Serbia, #7276, Text Embeddings - Serbian Language Applications - TESLA and COST Action NexusLinguarum (CA18209)

9. Bibliographical References

- Vladimir Alexiev and Gerard Casamayor. 2016. FN goes NIF: integrating FrameNet in the NLP interchange format. In *Proc. of the LDL 5th Workshop on Linked Data in Linguistics: Managing, Building and Using Linked Language Resources*, pages 1–10.
- Cristina Bosco, F Dell’Orletta, S Montemagni, Manuela Sanguinetti, Maria Simi, et al. 2014. *The Evalita 2014 Dependency Parsing task*. In *Proc. of the 4th Int. Workshop EVALITA 2014*, pages 1–8. Pisa University Press.

- Martin Brümmer, Milan Dojchinovski, and Sebastian Hellmann. 2016. DBpedia abstracts: a large-scale, open, multilingual NLP training corpus. In *Proc. of the 10th Int. 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.
- Christian Chiarcos. 2012. POWLA: Modeling linguistic corpora in OWL/DL. In *The Semantic Web: Research and Applications: 9th Extended Semantic Web Conference, ESWC 2012, 2012. Proc. 9*, pages 225–239. Springer.
- Christian Chiarcos and Luis Glaser. 2020. A tree extension for CoNLL-RDF. In *Proc. of the 12th LREC*, pages 7161–7169.
- Philipp Cimiano, Christian Chiarcos, John P. McCrae, and Jorge Gracia. 2020a. *Linguistic Linked Open Data Cloud*, pages 29–41. Springer International Publishing, Cham.
- Philipp Cimiano, Christian Chiarcos, John P. McCrae, and Jorge Gracia. 2020b. Linked Data-Based NLP Workflows. *Linguistic Linked Data: Representation, Generation and Applications*, pages 197–211.
- Thierry Declerck, Jorge Gracia, and John P. McCrae. 2020. COST Action “European network for Web-centred linguistic data science”(NexusLinguarum). *Proc. del Lenguaje Natural*, 65:93–96.
- Titika Dimitroulia. 2023. Corpora and literary translation. In *Advances in Corpus Applications in Literary and Translation Studies*, pages 103–118. Taylor & Francis.
- 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, 2013, Proc., 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 Int. Conference, EKAW 2012, 2012. Proc. 18*, pages 446–449. Springer.
- Susan Hunston. 2002. *Corpora in applied linguistics*. Cambridge University Press.
- Nancy Ide and Keith Suderman. 2014. The linguistic annotation framework: a standard for annotation interchange and merging. *Language Resources and Evaluation*, 48:395–418.
- Milica Ikonić Nešić, Ranka Stanković, Christof Schöch, and Mihailo Skoric. 2022. *From ELTeC Text Collection Metadata and Named Entities to Linked-data (and Back)*. In *Proc. of the 8th Workshop on Linked Data in Linguistics the 13th LREC*, pages 7–16, France. ELRA.
- Milica Ikonić Nešić, Ranka Stanković, and Biljana Rujević. 2021. ELTeC Corpus in Wikidata. *Infotheca - Journal for Digital Humanities*, 21(2).
- Ronald Jenn and Amel Fraisse. 2022. Benefits of a Corpus-based Approach to Translations: The Example of Huckleberry Finn. In *Advances in Corpus Applications in Literary and Translation Studies*, pages 176–190. Routledge.
- Fahad Khan, Christian Chiarcos, Thierry Declerck, et al. 2022. *A Survey of Guidelines and Best Practices for the Generation, Interlinking, Publication, and Validation of Linguistic Linked Data*. In *Proc. of the 8th Workshop on Linked Data in Linguistics the 13th LREC*, pages 69–77. ELRA.
- Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. 2020. *From Zero to Hero: Human-In-The-Loop Entity Linking in Low Resource Domains*. In *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6982–6993. Association for Computational Linguistics.
- Cvetana Krstev. 2008. *Processing of Serbian. Automata, texts and electronic dictionaries*. Faculty of Philology of the University of Belgrade.
- Cvetana Krstev and Duško Vitas. 2011. An Aligned English-Serbian Corpus. In *ELLSIIR*, volume I, pages 495–508, Belgrade. Faculty of Philology, University of Belgrade.
- Yinhan Liu, Myle Ott, Naman Goyal, et al. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Riccardo Moratto and Defeng Li. 2022. *Advances in Corpus Applications in Literary and Translation Studies, Introduction*, pages 1–9. Taylor & Francis.
- Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. 2013. *Learning multilingual named entity recognition from Wikipedia*. *Artificial Intelligence*, 194:151–175. AI, Wikipedia and Semi-Structured Resources.
- Ivan Obradović, Ranka Stanković, and Miloš Utvić. 2008. Integrisano okruženje za pripremu paralelizovanog korpusa. *Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen*, pages 563–578.

Olja Perisic, Stanković Ranka, Ikonić Nešić Milica, Škorić Mihailo, et al. 2023. *It-Sr-NER: CLARIN Compatible NER and Geoparsing Web Services for Italian and Serbian Parallel Text*. In *Selected Papers from the CLARIN Annual Conference 2022, Czechia, 2022*, pages 99–110. Linköping University Electronic Press.

Robert Sanderson, Paolo Ciccarese, and Herbert Van de Sompel. 2013. Designing the W3C open annotation data model. In *Proc. of the 5th Annual ACM Web Science Conference*, pages 366–375.

Helmut Schmid. 2013. Probabilistic part-of-speech tagging using decision trees. In *New methods in language processing*, page 154.

Ranka Stanković, Christian Chiarcos, Miloš Utvić, and Olivera Kitanović. 2023. Towards ELTeC-LLOD: European Literary Text Collection Linguistic Linked Open Data. In *Proc. of the 4th Conference on Language, Data and Knowledge*, pages 180–191.

Ranka Stanković, Cvetana Krstev, Duško Vitas, Nikola Vulović, and Olivera Kitanović. 2017. Keyword-based search on bilingual digital libraries. In *Semantic Keyword-Based Search on Structured Data Sources*, pages 112–123, Cham. Springer International Publishing.

Ranka Stanković, Mihailo Škorić, and Branislava Šandrih Todorović. 2022. *Parallel Bidirectionally Pretrained Taggers as Feature Generators*. *Applied Sciences*, 12(10).

Ranka Stanković, Christian Chiarcos, and Milica Ikonić Nešić. 2024. Leveraging Linked Data, NIF, and CONLL-U for Enhanced Annotation in Sentence Aligned Parallel Corpora. In *Book of Abstracts of the UniDive 2nd general meeting, 8-10 February 2024, Naples*.

Shyam Upadhyay, Nitish Gupta, and Dan Roth. 2018. *Joint Multilingual Supervision for Cross-lingual Entity Linking*. In *Proc. of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2495, Brussels, Belgium. Association for Computational Linguistics.

Duško Vitas and Cvetana Krstev. 2012. Processing of Corpora of Serbian Using Electronic Dictionaries. *Prace Filologiczne*, XVIII:279–292.

Federico Zanettin. 2012. *Translation practices explained: translation-driven corpora*. St Jerome Publishing.

Mihailo Škorić. 2024. Roberta: A robustly optimized bert pretraining approach. *Infotheca - Journal for Digital Humanities*.

10. Language Resource References

Apache Software Foundation, Apache. 2023. *Apache Jena Fuseki*. The Apache Software Foundation, <https://jena.apache.org/documentation/fuseki2/>.

David Huynh. 2022. *OpenRefine*. Metaweb Technologies, Inc, <https://openrefine.org/>.

Explosion. 2022. *it_core_news_sm spaCy pipeline model*. spaCy, https://github.com/explosion/spacy-models/releases/tag/it_core_news_sm-3.4.0.

Milica Ikonić Nešić and Saša Petalinkar and Mihailo Škorić and Ranka Stanković. 2024. *Jerteh-355 Tesla - model for Named Entity Recognition*. Hugging Face, https://huggingface.co/Tanor/sr_pln_tesla_j355.

Klie, Jan-Christoph and Bugert, Michael and Boullosa, Beto and Eckart de Castilho, Richard and Gurevych, Iryna. 2018. *The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation*. Association for Computational Linguistics, <https://inception-project.github.io>.

Cvetana Krstev and Duško Vitas. 2006. *SrpMD - Serbian morphological dictionaries*. ELG, <https://live.european-language-grid.eu/catalogue/lcr/17355>, 1.0.

Cvetana Krstev and Branislava Šandrih Todorović and Ranka Stanković and Milica Ikonić Nešić. 2021. *SrpELTeC-gold - Named Entity Recognition Training Corpus for Serbian*. ELG, <https://live.european-language-grid.eu/catalogue/corpus/9485>, 1.0.

Magnus Manske. 2019. *QuickStatements*. Free Software Foundation, <https://quickstatements.toolforge.org/>, 2.0.

Perišić, Olja and Stanković, Ranka and Ikonić Nešić, Milica and Škorić, Mihailo and Vitas, Duško and Krstev, Cvetana. 2022a. *It-Sr-NER*. ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa.

Perišić, Olja and Stanković, Ranka and Vitas, Duško and Krstev, Cvetana and Moderc, Saša. 2022b. *It-Sr-NER: CLARIN compatible NER and geoparsing web services for parallel texts: case study Italian and Serbian*. CLARIN-IT, <http://hdl.handle.net/>

20.500.11752/OPEN-980. ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa.

Heiden Serge. 2020. *The TXM Platform*. National Center for Scientific Research, <https://txm.gitpages.huma-num.fr/textometrie/>.

Ranka Stanković and Olivera Kitanović and Nikola Vulović and Cvetana Krstev. 2018. *Bibliša: Aligned Collection Search*. JeRTeh, <http://biblisha.jerteh.rs/>.

Ranka Stanković and Mihailo Škorić and Biljana Lazić and Cvetana Krstev. 2021. *Leximirka lexical database*. ELG, <https://live.european-language-grid.eu/catalogue/tool-service/17356>.

Duško Vitas and Cvetana Krstev and Ranka Stanković and Miloš Utvić and Mihailo Škorić. 2021. *SrpKor4Tagging*. ELG, <https://live.european-language-grid.eu/catalogue/corpus/9295>, 1.0.

Wikimedia. 2023. *Wikidata*. Wikimedia, <https://www.wikidata.org/>.

Author Index

- Anderson, Cormac, 37
Anuradha, Isuri, 44
Apostol, Elena-Simona, 1
Armaselu, Florentina, 1
- Bajtarević, Medina, 108
Bamberg, Claudia, 49
Bandini, Michela, 55
Banerjee, Shubhanker, 11
Bellandi, Andrea, 55
Beniamine, Sacha, 37
Boano, Valeria Irene, 22
Burch, Thomas, 49
- Canning, Erin, 32
Chakravarthi, Bharathi Raja, 11
Constantopoulos, Panos, 84
Costa, Rute, 44
- Doyle, Adrian, 66
- Fransen, Theodorus, 37
Frontini, Francesca, 44
- Gernert, Folke, 49
Gifu, Daniela, 1
Ginevra, Riccardo, 22
- Hinzmann, Maria, 49
- Ikonić Nešić, Milica, 115
Ionov, Maxim, 94, 108
- Kabatnik, Susanne, 49
Kasapaki, Marialena, 84
Khan, Fahad, 44
Kitanović, Olivera, 115
Kudera, Jacek, 49
- Liebeskind, Chaya, 1
Liyanage, Chamila, 44
- Mallia, Michele, 55
Mambrini, Francesco, 75
Marongiu, Paola, 1
McCrae, John P., 44, 66
- McCrae, John Philip, 11
McGillivray, Barbara, 1
Moretti, Giovanni, 75
Moulin, Claudine, 49
Murano, Francesca, 55
- Ninčević, Lorena, 108
- Ojha, Atul Kumar, 44
- Passarotti, Marco, 22, 37, 75
Perisic, Olja, 115
Pertsas, Vayianos, 84
Piccini, Silvia, 55
- Quochi, Valeria, 55
- Rani, Priya, 44, 66
Raue, Benjamin, 49
Rettinger, Achim, 49
Rigobianco, Luca, 55
Röpke, Jörg, 49
Rosner, Michael, 94
- Salgado, Ana, 44
Schenkel, Ralf, 49
Schirra, Doris, 49
Schöch, Christof, 49
Sciolette, Flavia, 103
Shi-Kupfer, Kristin, 49
Škorić, Mihailo, 115
Stanković, Ranka, 108, 115
stearns, bernardo, 66
- Tommasi, Alessandro, 55
Truica, Ciprian-Octavian, 1
- Valunaite Oleskeviciene, Giedre, 1
- Weis, Joëlle, 49
- Zavattari, Cesare, 55
Zinzi, Mariarosaria, 65