CLiC-it 2023

The Nineth Italian Conference on Computational Linguistics

Proceedings of the Conference

November 30 - December 2, 2023

These proceedings have been published in the CEUR Workshop Proceedings series.
The original papers are available at: https://ceur-ws.org/Vol-3596.
The papers are mirrored in the ACL Anthology.

ISBN 979-12-550-0084-6

Table of Contents

Preface to the CLiC-it 2023 Proceedings
Federico Boschetti, Gianluca E. Lebani, Bernardo Magnini and Nicole Novielli 1

When the Lab of CompLing Was Started at the University of Venice - Preface to the Proceedings of the First Workshop Held in 1982
Rodolfo Delmonte 4

Legal Summarization: to Each Court its Own Model
Flavia Achena, David Preti, Davide Venditti, Leonardo Ranaldi, Cristina Giannone, Fabio Massimo Zanzotto, Andrea Favalli and Raniero Romagnoli 10

CorpusCompass: A Tool for Data Extraction and Dataset Generation in Corpus Linguistics
Muhadj Adnan and Nicolo' Brandizzi 17

Harnessing 'Il Manifesto' Newspaper Archive for Knowledge Base Creation: Techniques and Findings in the 'MeMa' Project
Robert J. Alexander, Matteo Bartocci, Oriana Persico and Guido Vetere 28

Unmasking the Wordsmith: Revealing Author Identity through Reader Reviews
Chiara Alzetta, Felice Dell'Orletta, Chiara Fazzone, Alessio Miaschi and Giulia Venturi 36

Integrated Gradients as Proxy of Disagreement in Hateful Content
Alessandro Astorino, Giulia Rizzi and Elisabetta Fersini 46

Challenging Specialized Transformers on Zero-shot Classification
Serena Auriemma, Mauro Madeddu, Martina Miliani, Alessandro Bondielli, Alessandro Lenci and Lucia Passaro 53

Bias Mitigation in Misogynous Meme Recognition: A Preliminary Study
Gianmaria Balducci, Giulia Rizzi and Elisabetta Fersini 63

Building a Spoken Dialogue System for Supporting Blind People in Accessing Mathematical Expressions
Pier Felice Balestrucci, Luca Anselma, Cristian Bernareggi and Alessandro Mazzei 70

Contrastive Language–Image Pre-training for the Italian Language
Federico Bianchi, Giuseppe Attanasio, Raphael Pisoni, Silvia Terragni, Gabriele Sarti and Dario Balestri 78

Modelling and Publishing the Lexicon der indogermanischen Verbenas Linked Open Data
Valeria Irene Boano, Francesco Mambrini, Marco Passarotti and Riccardo Ginevra 86

How Green is Sentiment Analysis? Environmental Topics in Corpora at the University of Turin
Cristina Bosco, Muhammad Okky Ibrohim, Valerio Basile and Indra Budi 93

Ti blocco perché sei un trollazzo". Lexical Innovation in Contemporary Italian in a Large Twitter Corpus
Paolo Brasolin, Greta Hayley Franzini and Stefania Spina 100

Testing ChatGPT for Stability and Reasoning: A Case Study Using Italian Medical Specialty Tests
Silvia Casola, Tiziano Labruna, Alberto Lavelli and Bernardo Magnini 113

Assessing Language and Vision-Language Models on Event Plausibility
Maria Cassese, Alessandro Bondielli and Alessandro Lenci 120

GPT-based Language Models meet Emojitaliano: A Preliminary Assessment Test between Automation and Creativity
Francesca Chiusaroli, Tiberio Uricchio, Johanna Monti, Maria Laura Pierucci and Federico Sangati . . . 128

Linking Stance and Stereotypes About Migrants in Italian Fake News
Alessandra Teresa Cignarella, Simona Frenda, Tom Bourgeade, Cristina Bosco and Francesca D'Errico . . . 135

Interpretation of Generalization in Masked Language Models: An Investigation Straddling Quantifiers and Generics
Claudia Collacciani and Giulia Rambelli . . . 143

Highway to Hell. Towards a Universal Dependencies Treebank for Dante Alighieri's Comedy
Claudia Corbetta, Marco Passarotti, Flavio Massimiliano Cecchini and Giovanni Moretti . . . 154

Towards an Italian Corpus for Implicit Object Completion
Agnese Daffara and Elisabetta Jezek . . . 162

Linking the Corpus CLaSSES to the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin
Irene De Felice, Lucia Tamponi, Federica Iurescia and Marco Passarotti . . . 172

Processing Effort During Reading Texts in Young Adults: Text Simplification, Readability Assessment and Preliminary Eye-tracking Data
Maria De Martino . . . 179

An Experiment in Error Analysis of Real-time Speech Machine Translation Using the Example of the European Parliament's Innovation Partnership
Elisa Di Nuovo . . . 185

A Cognitive Linguistics Analysis of Phrasal Verbs Representation in Distributional Semantics
Melissa Donati and Carlo Strapparava . . . 196

How To Build Competitive Multi-gender Speech Translation Models For Controlling Speaker Gender Translation
Marco Gaido, Dennis Fucci, Matteo Negri and Luisa Bentivogli . . . 203

Hate Speech Detection in an Italian Incel Forum Using Bilingual Data for Pre-Training and Fine-Tuning
Paolo Gajo, Silvia Bernardini, Adriano Ferraresi and Alberto Barrón-Cedeño . . . 211

Linking the Dictionary of Medieval Latin in the Czech Lands to the LiLa Knowledge Base
Federica Gamba, Marco C. Passarotti and Paolo Ruffolo . . . 219

Check-IT!: A Corpus of Expert Fact-checked Claims for Italian
Jacopo Gili, Lucia Passaro and Tommaso Caselli . . . 227

End-to-end Dependency Parsing via Auto-regressive Large Language Model
Claudiu Daniel Hromei, Danilo Croce and Roberto Basili . . . 236

DisaggregHate It Corpus: A Disaggregated Italian Dataset of Hate Speech
Marco Madeddu, Simona Frenda, Mirko Lai, Viviana Patti and Valerio Basile . . . 243

Automatic Detection of Parkinson's Disease with Connected Speech Acoustic Features: Towards a Linguistically Interpretable Approach
Marta Maffia, Loredana Schettino and Vincenzo Norman Vitale . . . 251

Introducing Deep Learning with Data Augmentation and Corpus Construction for LIS
Manuela Marchisio, Alessandro Mazzei and Dario Sammaruga 259

XL-WA: a Gold Evaluation Benchmark for Word Alignment in 14 Language Pairs
Federico Martelli, Andrei Stefan Bejgu, Cesare Campagnano, Jaka Čibej, Rute Costa, Apolonija Gantar, Jelena Kallas, Svetla Peneva Koeva, Kristina Koppel, Simon Krek, Margit Langemets, Veronika Lipp, Sanni Nimb, Sussi Olsen, Bolette Sanford Pedersen, Valeria Quochi, Ana Salgado, László Simon, Carole Tiberius, Rafael-J Ureña-Ruiz and Roberto Navigli 272

Is Change the Only Constant? An Inquiry into Diachronic Semantic Shifts in Italian and Spanish
Matteo Melis, Anastasiia Salova and Roberto Zamparelli 281

Building Structured Synthetic Datasets: The Case of Blackbird Language Matrices (BLMs)
Paola Merlo, Giuseppe Samo, Vivi Nastase and Chunyang Jiang 292

Simplifying Administrative Texts for Italian L2 Readers with Controllable Transformers Models: A Data-driven Approach
Martina Miliani, Fernando Alva-Manchego and Alessandro Lenci 303

Multi-task Learning for German Text Readability Assessment
Salar Mohtaj, Vera Schmitt, Razieh Khamsehashari and Sebastian Möller 316

Is It Really That Simple? Prompting Large Language Models for Automatic Text Simplification in Italian
Debora Nozza and Giuseppe Attanasio 322

Unraveling Text Coherence from the Human Perspective: a Novel Dataset for Italian
Federica Papa, Luca Dini, Dominique Brunato and Felice Dell'Orletta 334

Lost in Labels: An Ongoing Quest to Optimize Text-to-Text Label Selection for Classification
Michele Papucci, Alessio Miaschi and Felice Dell'Orletta 342

Are All Languages Equal? Curriculum Learning over Different Languages
Giulia Pucci, Leonardo Ranaldi and Fabio Massimo Zanzotto 351

Prompting LLMs in Italian Language for Text-to-SQL Translation
Federico Ranaldi, Elena Sofia Ruzzetti, Leonardo Ranaldi, Davide Venditti, Cristina Giannone, Andrea Favalli, Raniero Romagnoli and Fabio Massimo Zanzotto 361

Beyond Obscuration and Visibility: Thoughts on the Different Strategies of Gender-Fair Language in Italian
Martina Rosola, Simona Frenda, Alessandra Teresa Cignarella, Matteo Pellegrini, Andrea Marra and Mara Floris 369

Blaze-IT: A Lightweight BERT Model for the Italian Language
Francesco Russo and Michele Filannino 379

Camoscio: An Italian Instruction-tuned LLaMA
Andrea Santilli and Emanuele Rodolà 385

Automatic Generation of Common Procurement Vocabulary Codes
Lucia Siciliani, Emanuele Tanzi, Pierpaolo Basile and Pasquale Lops 396

Inters8: A Corpus to Study Misogyny and Intersectionality on Twitter
Ivan Spada, Mirko Lai and Viviana Patti 403

Let's Keep an Eye on Russian: Testing Sensitivity to the Change in the Grammatical Number in Somatic Idioms with ruBERT
Elena Spaziani and Eliana Di Palma . . . 412

"That branch of the Lake of Como...": Developing a New Resource for the Analysis of I Promessi Sposi and its Historical Translations
Rachele Sprugnoli and Marco Sartor . . . 420

Qualitative Analysis of Persuasive Emotion Triggering in Online Content
Olga Uryupina . . . 427

When You Doubt, Abstain: A Study of Automated Fact-Checking in Italian under Domain Shift
Giovanni Valer, Alan Ramponi and Sara Tonelli . . . 433

On Incrementing Interpretability of Machine Learning Models from the Foundations: A Study on Syllabic Speech Units
Vincenzo Norman Vitale, Loredana Schettino and Francesco Cutugno . . . 441

Drug Name Recognition in the Cryptomarket Forum of Silk Road 2
Romane Werner, Thomas François and Sonja Bitzer . . . 448

Italian Crossword Generator: Enhancing Education through Interactive Word Puzzles
Kamyar Zeinalipour, Tommaso Iaquinta, Asya Zanollo, Giovanni Angelini, Leonardo Rigutini, Marco Maggini and Marco Gori . . . 455

Die Rätselrevolution: Automated German Crossword Solving
Andrea Zugarini, Thomas Röthenbacher, Kai Klede, Marco Ernandes, Bjoern M. Eskofier and Dario Zanca . . . 465

Towards a Multilingual System for Vaccine Hesitancy Using a Data Mixture Approach
Oscar Araque, María Felipa Ledesma Corniel and Kyriaki Kalimeri . . . 471

TAll: A New Shiny App of Text Analysis for All
Massimo Aria, Corrado Cuccurullo, Luca D'Aniello, Michelangelo Misuraca and Maria Spano 476

On the Impact of Language Adaptation for Large Language Models: A Case Study for the Italian Language Using Only Open Resources
Pierpaolo Basile, Pierluigi Cassotti, Marco Polignano, Lucia Siciliani and Giovanni Semeraro 480

Annotating Homeric Emotions by a Domain-Specific Language
Federico Boschetti, Laura Chilla, Maria Konstantinidou and John Pavlopoulos . . . 485

How Good is NLLB for Low-resource Languages? A Study on the Genoese Language
Davide Buscaldi and Paolo Rosso . . . 490

Debunker Assistant: A Support for Detecting Online Misinformation
Arthur Thomas Edward Capozzi Lupi, Alessandra Teresa Cignarella, Simona Frenda, Mirko Lai, Marco Antonio Stranisci and Alessandra Urbinati . . . 494

Towards a Multi-Level Annotation Format for the Interoperability of Automatic Term Extraction Corpora
Nicola Cirillo and Daniela Vellutino . . . 499

The PBSDS: A Dataset for the Detection of Pseudoprofound Bullshit
Evan D. DeFrancesco and Carlo Strapparava . . . 506

A Post-Modern Approach to Automatic Metaphor Identification
Dario Del Fante, Federico Manzella, Guido Sciavicco and Eduard Ionel Stan 510

Building a Corpus on Eating Disorders from TikTok: Challenges and Opportunities
Melissa Donati, Ludovica Polidori, Paola Vernillo and Gloria Gagliardi 515

A Preliminary Release of the Italian Parliamentary Corpus
Valentino Frasnelli and Alessio Palmero Aprosio 521

Extracting an Expectation-based Lexicon for UD Treebanks
Matteo Gay and Cristiano Chesi 525

An Analysis of Visually Grounded Instructions in Embodied AI Tasks
Marco Grazioso and Alessandro Suglia 530

Exploring Sentiments in Summarization: SentiTextRank, an Emotional Variant of TextRank
Md. Murad Hossain, Luca Anselma and Alessandro Mazzei 535

An Italian Verb Lexicon for Sentiment Inference
Manfred Klenner, Dylan Massey and Susanna Tron 540

The Inherence of Telicity: Unveiling Temporal Reasoning in Video Question Answering
Olga Loginova and Raffaella Bernardi 546

Identification of Multiword Expressions: Comparing the Performance of a Conditional Random Fields Model on Corpora of Written and Spoken Italian
Ilaria Manfredi and Lorenzo Gregori 551

Teasing LLMs Adapted to Italian
Leonardo Ranaldi, Giulia Pucci, Elena Sofia Ruzzetti, Fabio Massimo Zanzotto and André Freitas 557

Investigating Gender Bias in Large Language Models for the Italian Language
Elena Sofia Ruzzetti, Dario Onorati, Leonardo Ranaldi, Davide Venditti and Fabio Massimo Zanzotto 562

Towards a New Computational Lexicon for Italian: Building the Morphological Layer by Harmonizing and Merging Existing Resources
Flavia Sciolette, Simone Marchi and Emiliano Giovannetti 570

Textual Entailment with Natural Language Explanations: The Italian e-RTE-3 Dataset
Andrea Zaninello, Sofia Brenna and Bernardo Magnini 575

Preface to the CLiC-it 2023 Proceedings

Federico Boschetti[1,2], Gianluca E. Lebani[2], Bernardo Magnini[3] and Nicole Novielli[4]

[1]*CNR-Institute for Computational Linguistics "A. Zampolli", Pisa*
[2]*Ca' Foscari University of Venice*
[3]*Fondazione Bruno Kessler, Trento*
[4]*University of Bari "Aldo Moro"*

The ninth edition of the Italian Conference on Computational Linguistics (CLiC-it 2023) was held from 30th November to 2nd December 2023 at **Ca' Foscari University of Venice**, in the beautiful venue of the Auditorium Santa Margherita - Emanuele Severino. After the edition of 2020, which was organized in fully virtual mode due to the health emergency related to Covid-19, and CLiC-it 2021, which was held in hybrid mode, with CLiC-it 2023 we are back to a fully in-presence conference. Overall, almost 210 participants registered to the conference, confirming that the community is eager to meet in person and to enjoy both the scientific and social events together with the colleagues.

Concerning the scientific organization of the conference, there have been few important novelties proposed at CLiC-it 2023. First, we have introduced the **distinction between long and short papers**, with the aim of bringing CLiC-it closer to the standards of the major international conferences in the field. In total, we received 66 submissions for long papers and 20 submissions for short papers, for a total of 86 submissions. In addition we received 20 submissions for Research Communications, i.e., outstanding papers that have been already published during the last period in a major international CL conference or journal.

A second novelty of the conference concerns **the reviewing process**. Instead of having conference tracks, as in previous CLiC-it editions, we just listed a large number of topics to be used at submission time, and then submissions were assigned to area chairs (twelve program committee members) through a bidding mechanisms. This way we were able to achieve a better balance of papers for each area chair, while respecting their research interests. Then, assignment to reviewers was also managed globally, instead of separately for each area. We had a single pool of reviewers, which allowed to better monitoring the whole process, including good per paper distribution of senior and young reviewers.

During the reviewing process, each submission was reviewed by three independent reviewers in single-blind fashion. At the end of the process, **75 proposals were accepted** for presentation at the conference and for publication in the proceedings, resulting in an acceptance rate of 87.21%.

Out of the 75 accepted proposals, 26 were included in the program of CLiC-it 2023 as oral presentations and the remaining 59 were assigned to one of the three poster sessions of the conference. As usual, the criterion for assigning a proposal to an oral or a poster session was based on the contents and not on the quality of the proposal. Regardless of the format of presentation, long papers are limited to seven pages of content (one additional page w.r.t. the previous editions) plus unlimited pages of references in the proceedings, available as open access publication.

The program of CLiC-it 2023 is completed by **20 research communications** selected after the reviewing process. Research communications are not published in the proceedings, but four of them have been orally presented within dedicated sessions at the conference, while the remaining 16 were presented in one of the three poster sessions of the Conference.

In addition to the technical programme, this year the conference was opened by a welcome speech by Rodolfo Del Monte titled "Computational Linguistics in Venice: when it all began", and we were honoured to have as **invited speakers** such internationally recognized researchers as:

- **David Bamman** (UC Berkeley), with a keynote entitled "The Promise and Peril of Large Language Models for Cultural Analytics": *Much work at the intersection of NLP and cultural analytics/computational social science is focused on creating new algorithmic measuring devices for constructs we see encoded in text (including agency, respect, and power, to name a few). How does the paradigm shift of large language models change this? In this talk, I'll discuss the role of LLMs (such as ChatGPT, GPT-4 and open alternatives) for research in cultural analytics, both raising issues*

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
federico.boschetti@ilc.cnr.it (F. Boschetti); gianluca.lebani@unive.it (G. E. Lebani); magnini@fbk.eu (B. Magnini); nicole.novielli@uniba.it (N. Novielli)
https://www.ilc.cnr.it/people/federico-boschetti/ (F. Boschetti); https://www.unive.it/persone/gianluca.lebani (G. E. Lebani); https://collab.di.uniba.it/nicole/ (N. Novielli)

about the use of closed models for scholarly inquiry and charting the opportunity that such models present. The rise of large pre-trained language models has the potential to radically transform the space of cultural analytics by both reducing the need for large-scale training data for new tasks and lowering the technical barrier to entry, but need care in establishing the reliability of results.
- **Vera Demberg** (Saarland University), with a keynote entitled "Pragmatic processing in humans and language models": *Pragmatic processing concerns inferences that go beyond the literal meaning of a text or utterance. In my talk, I will go through different types of pragmatic inferences, including scalar implicatures, atypicality inferences and other tasks requiring theory of mind reasoning; for each of these, I will discuss recent work from our lab and others, regarding the ability of large language models, and of humans, to handle these phenomena.*

On the last morning of the Conference, Danilo Croce and Claudiu Daniel Hromei (University of Rome, Tor Vergata) gave a **tutorial** on the state-of-the-art Large Language Models and how to instruction-tuned in a sustainable way.

This year we received 7 candidate theses for the "**Emanuele Pianta Award for the Best Master Thesis**". This special prize for the best Master Thesis (Laurea Magistrale) in Computational Linguistics, submitted at an Italian University, is endorsed by AILC. The candidate theses have been evaluated by a jury composed by Pierpaolo Basile (University of Bari "Aldo Moro"), Gianluca E. Lebani (Ca' Foscari University of Venice) and Viviana Patti (University of Turin). The winner was awarded during the closing session of the conference by the members of the jury.

We thank **all the people and institutions** involved in the organization of the conference, all area chairs, reviewers, and all participants, who contributed to the success of the event. All area chairs and reviewers are named in the following pages. We are grateful to the Venice Centre for Digital and Public Humanities[1], that made CLiC-it 2023 possible by hosting the event and supporting us greatly in the processes of local organization, and to the organizations that endorsed our event: the Italian Association of Digital Humanities [2], the European Centre for Living Technology hosted by Ca' Foscari[3] and the Future Artificial Intelligence Research Foundation[4].

We would like to thank our **supporters**, who generously provided funds and services that are crucial for the realization of this event: Talia[5] (Platinum), Almawave[6] and Aptus.AI[7] (Gold), aecqua.tech and trnslated [8] (Bronze) and the Department of Linguistics and Comparative Cultural Studies of the Ca' Foscari University of Venice [9].

Finally, we want to thank very much the Associazione Italiana di Linguistica Computazionale (**AILC**), all the members of the Association Board and, in particular, the President Simonetta Montemagni and Elisabetta Fersini, who never let us alone in the sea of doubts and problems that organizing such an ever-changing event implies.

Venice, December 2023

Conference Chairs

- **Federico Boschetti**, CNR-Institute for Computational Linguistics "A. Zampolli" / Ca' Foscari University of Venice
- **Gianluca E. Lebani**, Ca' Foscari University of Venice
- **Bernardo Magnini**, Fondazione Bruno Kessler
- **Nicole Novielli**, University of Bari "Aldo Moro"

Local Organization Committee

- **Franz Fisher**, Ca' Foscari University of Venice
- **Cristina Procentese**, Ca' Foscari University of Venice
- **Matteo Radaelli**, Norwegian University of Science and Technology

Proceedings Chairs

- **Luca Molinari**, Ca' Foscari University of Venice / University of Warsaw
- **Andrea Zaninello**, Fondazione Bruno Kessler

Webmasters

- **Eleonora Ghizzota**, University of Bari "Aldo Moro"
- **Tiziano Labruna**, Fondazione Bruno Kessler

Publicity Chair

- **Sofia Brenna**, Fondazione Bruno Kessler

1 https://www.unive.it/vedph
2 http://www.aiucd.it//
3 https://www.unive.it/eclt
4 https://future-ai-research.it/
5 https://talia.cloud/
6 https://www.almawave.com/it/
7 https://www.aptus.ai/
8 https://translated.com/
9 https://www.unive.it/dslcc

Booklet

- **Alice Suozzi**, Ca' Foscari University of Venice

Area Chairs

- **Pierpaolo Basile**, University of Bari "Aldo Moro"
- **Valerio Basile**, University of Turin
- **Tommaso Caselli**, Rijksuniveristeit Groningen
- **Danilo Croce**, University of Rome Tor Vergata
- **Francesco Cutugno**, University of Naples Federico II
- **Elisabetta Fersini**, University of Milano-Bicocca
- **Gloria Gagliardi**, University of Bologna
- **Elisabetta Jezek**, University of Pavia
- **Alessandro Lenci**, University of Pisa
- **Rachele Sprugnoli**, University of Parma
- **Sara Tonelli**, Fondazione Bruno Kessler
- **Giulia Venturi**, CNR-Institute for Computational Linguistics "A. Zampolli"

Reviewers

Chiara Alzetta, Alessio Palmero Aprosio, Oscar Araque, Serena Auriemma, Francesco Barbieri, Mauro Bennici, Luisa Bentivogli, Andrea Bolioli, Helena Bonaldi, Giovanni Bonetta, Cristina Bosco, Sofia Brenna, Dominique Brunato, Davide Buscaldi, Elena Cabrio, Florian Cafiero, Jean-Baptiste Camps, Silvia Casola, Camilla Casula, Emmanuele Chersoni, Francesca Chiusaroli, Alessandra Teresa Cignarella, Fabio Ciotti, Claudia Roberta Combei, Gianpaolo Coro, Riccardo Del Gratta, Angelo Mario Del Grosso, Felice Dell'Orletta, Maria Di Maro, Marcello Ferro, Komal Florio, Greta Hayley Franzini, Simona Frenda, Francesca Frontini, Marco Gaido, Lorenzo Gregori, Claudiu Daniel Hromei, Anas Fahad Khan, Tiziano Labruna, Alberto Lavelli, Eleonora Litta, Simone Magnolini, Francesco Mambrini, Alessio Miaschi, Martina Miliani, Monica Monachini, Johanna Monti, Debora Nozza, Antonio Origlia, Lucia Passaro, Marco Carlo Passarotti, Viviana Patti, Matteo Pellegrini, Marco Polignano, Simone Paolo Ponzetto, Giulia Rambelli, Alan Ramponi, Irene Russo, Manuela Sanguinetti, Giovanni Semeraro, Lucia Siciliani, Claudia Soria, Manuela Speranza, Carlo Strapparava, Olga Uryupina, Serena Villata, Vincenzo Norman Vitale, Andrea Zaninello

When the Lab of CompLing was started at the University of Venice - Preface to the proceedings of the first workshop held in 1982

Rodolfo Delmonte[1]

[1]*Ca Foscari University of Venice*

Abstract
The chapters contained in the book - now out of print - and in its second edition published in 1988 with the title "Studi di Linguistica Computazionale"/Studies of Computational Linguistics, are some of the presentations held at the "Seminario Introduttivo alla Linguistica Computazionale"/Introductory Workshop in Computational Linguistics, on May 13/14/1982, at the University of Venice.

Keywords
what CL really is, Pisa contribution, Zampolli opinion, Venice contribution

Prefazione al libro

I contributi raccolti in questo volume sono alcuni degli interventi tenuti al Seminario Introduttivo alla Linguistica Computazionale, il 13-14 maggio 1982, all'Università di Venezia. Lo scopo dell'incontro era principalmente quello di presentare una panoramica dettagliata degli sviluppi di una disciplina, la linguistica computazionale, che in Italia ha per ora poco seguito. I motivi di questo stato di cose sono svariati, di tipo economico storico e sociale: tradizionalmente le Facoltà umanistiche infatti, non si servono di tecnologia per svolgere le propria ricerche, a differenza di quelle scientifiche. Il letterato, il filologo, il glottologo e il linguista in Italia solo sporadicamente si è rivolta all'elaboratore per avere un ausilio nei propri studi. All'estero è prassi comune che gli studiosi in campo umanistico, se non dispongono direttamente di un centro ricerche di linguistica computazionale, si siano rivolti all'elaboratore per verificare o convalidare ipotesi teoriche. Non vi è alcun dubbio però che gli elaboratori costano, e il livello di investimento nella ricerca in Italia è ben noto a tutti quelli che vi operano come basso. Ugualmente nota poi in Italia è la separazione che tuttora esiste tra ricerca umanistica attuta con carta, penna e intuito dallo scolaro, e ricerca scientifica che deve giocoforza affidarsi alla tecnologia. In più, storicamente, si aggiunge l'anatema crociano contro tutti quelli che per essere creativi non si affidino all'intuito e intelligenza individuale, ma magari vogliano fare lavoro d'equipe o "sacrilegio", utilizzare macchine. I risultati sono ovviamente quelli che dicevamo: benché l'interesse per la LC sia cresciuto negli ultimi sei o sette anni, ciò è dovuto quasi unicamente all'opera di organizzazione

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
delmont@unive.it (R. Delmonte)
http://rondelmo.it/ (R. Delmonte)
0000-0003-0282-7661 (R. Delmonte)

e ramificazione compiuta dall'attuale Istituto di Linguistica Computazionale di Pisa, diretto dal prof. Antonio Zampolli.

Come ha anche lucidamente chiarito Zampolli in un suo intervento[1] e nella relazione presentata a Venezia, è bene distinguere perlomeno tra linguistica quantitativa e linguistica computazionale: nel primo caso infatti l'elaboratore viene utilizzato principalmente allo scopo di condurre analisi di testi di tipo statistico, probabilistico e quantitativo in genere. L'elaboratore funge in questo caso da ausilio al linguista, il quale è primariamente interessato agli aspetti stilistici, filologici, glottologici e letterari del testo/i in esame: l'elaboratore fornirà allora spogli elettronici dalla cui consultazione ragionata il linguista potrà rilevare gli elementi di suo interesse. Ovviamente, una volta trasferito uno o più testi su un supporto appropriato per la macchina, schede perforate o nastro magnetico, l'analisi potrà essere ripetuta variando eventualmente di volta in volta gli elementi che si vuole che l'elaboratore estragga per il linguista. Analisi di questo genere, anche se compiute su corpora estesi, ad esempio di 100mila occorrenze, non richiedono all'elaboratore più di 50 secondi di tempo macchina e un tempo variabile tra i 15 e i 30 minuti per stampare i risultati.

La LC invece ha come scopo l'analisi del linguaggio e non solamente della lingua, e l'elaboratore non è più semplicemente un ausilio, ma lo strumento di simulazione dei processi linguistici sottostanti a un qualche modello della produzione/comprensione della lingua, implementati dal linguista, che vorrà verificarne la bontà di funzionamento. Il programma utilizzato in questo caso non servirà a produrre spogli di corpora o testi, ma rappresenterà una grammatica, tradotta in un linguaggio di programmazione. Il linguista sarà allora interessato a studiare gli effetti prodotti da modificazione, introduzioni o cancellazione di regole della grammatica sulla interpretazione e descrizione del campione di lingua utilizzato.

Più in generale, il linguista computazionale sarà interessato non soltanto al raggiungimento di obbiettivi esplicativi e cioè di descrizione del linguaggio mediante regole e rappresentazioni strutturali, ma vorrà anche sviluppare strumenti adeguati alla generazione o produzione e alla comprensione e percezione del linguaggio. Se quindi le ricerche quantitative operano in ambito matematico e statistico, utilizzando un approccio tassonomico e descrittivo ai fenomeni linguistici studiati, quelle computazionali sono interessate alla elaborazione di modelli della competenza e della realizzazione linguistica, in altre parole alla grammatica e al meccanismo che la realizza o processore.

Come spiega Zampolli[1], il sogno di giungere velocemente alla comprensione dei meccanismi della lingua attraverso la traduzione meccanica svanì lentamente negli anni '60. L'attenzione degli operatori in quel tempo si distolse dai problemi di implementazione su elaboratore di algoritmi intesi a risolvere questioni linguistiche, per rivolgersi più direttamente al funzionamento del linguaggio: da problemi legati a tecniche di programmazione e linguaggi, tecniche di immagazzinamento dati e altri aspetti del software e hardware intesi ad influire direttamente sulla lingua in esame, per dirigersi allora al problema costituito dalla struttura del linguaggio e dalle sue regole sottostanti.

Benché l'alto livello di formalizzazione e di esplicitazione nella descrizione strutturale e delle regole proposta da Chomsky possa indurre il linguista a considerare la grammatica generativa trasformazionale(GGT) come un esempio di algoritmo linguistico per elaboratore: ma in realtà non è così, e lo ha ribadito lo stesso Chomsky[2]. La GGT è sì esplicita nei suoi formalismi, ma i suoi obbiettivi descrittivo/esplicativi non si adeguano se non in parte a quelli che ci si pone in ambito LC. Nella GGT manca infatti una qualsiasi preoccupazione per esplorare e studiare i meccanismi relativi alle operazione di produzione e comprensione del linguaggio che Chomsky relega alla cosiddetta esecuzione o realizzazione linguistica, che in quanto tale non può costituire un campo di ricerca scientificamente valido per un linguista. E se il modello di competenza proposto da Chomsky costituisce fondamentalmente un modello psicologico, esso è però anche intenzionalmente il più astratto e il più distante possibile dalla realtà della realizzazione linguistica. Le regole contenute nella grammatica infatti secondo i principi della GGT, non potranno mai essere desunte o indotte da corpora linguistici attraverso procedimenti di scoperta per quanto ampi essi possano essere, saranno pur sempre deficitari rispetto alla quantità di materiale linguistico che produrrà/comprenderà nella sua vita un parlante qualsiasi della lingua.

Applicazioni computazionali della GGT come quella della Joyce Friedman[**?**] o altre, discusse in particolare dalla Prodanof, in realtà sono ispirate solo in parte alle posizioni teoriche della GGT. Infatti, la teoria e il processore non necessariamente coincidono, ed è soltanto il secondo che servirà da verifica della prima, che essenzialmente è e resterebbe solamente una ipotesi di funzionamento del linguaggio. In particolare poi, le ultime teorie generative hanno sostituito alla centralità della sintassi con il suo componente trasformazionale, il lessico e le categorie funzionali come primitive. I lavori di Bresnan[**?**], Kaplan e Bresnan[3] e Gazdar[4] hanno relegato il ruolo della sintassi alla sottocategorizzazione e al funzionamento di una grammatica a struttura sintagmatica che proietta strutture sintattiche direttamente dall'analisi superficiale. In termini di analizzatore sintattico o "parser", non sarà quindi più necessario procedere a ritroso con le trasformazioni alla rovescia per recuperare una struttura profonda difficilmente individuabile. Un parser ATN, o quello deterministico del Parsifal saranno allora sufficienti con sole regole di tipo context-free a descrivere la struttura sintattica sottostante - ma di questo tratterà estesamente il saggio della Prodanof.

Tradurre e comprendere discorsi e testi sono invece attività che richiedono ad un processore abilità alquanto diverse da quelle necessarie per la verifica di una teoria e del modello che essa rappresenta. Infatti il linguista computazionale dovrà simulare il comportamento linguistico di un parlante/ascoltatore in situazioni reali. In questo caso non sarà sufficiente l'informazione lessicale, morfologica, sintattica e semantica ma si dovrà utilizzare anche quella pragmatica; e le regole non potranno essere solo quelle contestuali, o context-sensitive, ma saranno del testo o del discorso, cotestuali o transfrastiche, in modo da catturare i processi di inferenza che dall'enunciato risalgono alla conoscenza o enciclopedia del parlante/ascoltatore. Solo così la codifica-decodifica del messaggio o produzione-comprensione dell'enunciato potrà realizzarsi efficacemente. Le ricerche in questo campo vanno da quelle di Petoefi[5] e Van Dijk[6] a quelle documentate in Conte[7], Parisi[8] e Castelfranchi e Parisi[9], nonché all'ambito di ricerca definito con il termine Intelligenza Artificiale, a cui accenna la relazione di Ferrari, all'interno di lavori orientati a compiere analisi automatiche del contenuto semantico del linguaggio.

Ed è proprio di questo campo più vicino alla realizzazione linguistica che si interessa l'ultimo saggio, che non vuole né può essere una panoramica dei lavori svolti in questa area di ricerca, in quanto le soluzioni adottate per l'inglese mal si adattano all'italiano. Il saggio propone un modello e la sua simulazione in un processore basato su teorie fonetico-fonologiche e prosodiche, o di quella parte più standardizzata di variabili implicate nella fase di realizzazione linguistica rappresentata dalla produzione del parlato. In pratica, il processore simula le operazioni di codifica e decodifica compiute da un parlante nel leggere un testo ad alta voce. In questo senso esso è uno strumento adeguato alla produzione di voce sintetica attraverso macchine comandate da elaboratori. I primi lavori in questo ambito sono di provenienza inglese e cioè di Holmes e Mattingly[10], e Mattingly[11] che propongono un approccio segmentale di tipo fonematico; i lavori del gruppo di Padova, composto da Francini, Debiasi e Spinabelli [12], propongono una soluzione al problema delle unità minime con i difoni, che sono di numero superiore ai fonemi ma risolvono interamente il problema delle transizioni tra due suoni contigui. Sempre nell'ambito della sintesi della voce, si registrano poi i lavori americani, in particolare quelli di Allen[13] e di Umeda[14], nonché il sistema di Klatt[15], basato di nuovo su difoni. Tutti questi sistemi di sintesi contengono un processore della lingua basato su teorie fonetico-fonologiche: ad esempio Allen utilizza le teorie chomskiane per il suo modello, e per prevedere la posizione dell'accento ha costruito un dizionario di 12mila morfi che assieme ad un algoritmo morfologico gli permettono di recuperare la struttura sillabica della parola in esame, quindi il suo riconoscimento e l'assegnazione dell'accento di parola per regole. Anche il nostro modello, come vedremo, prevede la localizzazione dell'accento di parola, utilizzando un lessico esiguo ma solo per le eccezioni alle regole.

RINGRAZIAMENTI

Le persone da ringraziare in questi casi sono ovviamente tantissime: innanzitutto chi ha permesso materialmente che avesse luogo l'incontro e cioè il prof. Volpato, direttore del CEDOSTA, il prof. Castellani, direttore del Centro di Calcolo e il prof. Csillaghy, direttore del Centro Linguistico dell'Università di Venezia. Quindi chi mi ha spronato ad organizzare l'incontro e cioè il prof. Zampolli, direttore dell'Istituto di LC di Pisa, e tutti gli amici, Riccardo Zipoli, Alberto Tomasin, Irina Prodanof, Giacomo Ferrari, Luciano Canepari, e gli altri che hanno seguito gli interventi. Per ultimo chi più di ogni altro ha contribuito di persona, Giorgio Vercellin. Il libro non sarebbe potuto uscire in questa forma senza il contributo del Centro Linguistico e l'aiuto dell'amico Marco Vio.

References

[1] A. Zampolli, Trattamento automatico di dati linguistici e linguistica quantitativa, in: Linguaggi e Formalizzazione - SLI", Bulzoni, Roma, 1979, pp. 349—-370.
[2] N. Chomsky, Current Issues in Linguistic Theory, Prentice-Hall, Englewood Cliffs N.J., 1964, pp. 50–118.
[3] R. K. . J. Bresnan, Lexical Functional Grammar: a Formal Syntax for Grammatical Representation, ????
[4] G. Gazdar, Phrase Structure Grammar, ????
[5] J. S. Petoefi, Text representation and lexicon as semantic network, in: Linguaggi e Formalizzazione - SLI", Bulzoni, Roma, 1979, pp. 573—-589.
[6] T. A. V. Dijk, Text and Context, Longman, London, 1977.
[7] M. E. Conte, La linguistica testuale, Feltrinelli, Milano, 1977.

Figure 1: Copyright @ 1983 by C.L.E.S.P. s.r.l. Via del Santo, 57/7 – 35123 Padova – Italy All rights reserved

[8] D. Parisi, Studi per un modello del linguaggio, volume 89, Quaderni della Ricerca Scientifica - CNR, Roma, 1975.
[9] C. C. . D. Parisi, Linguaggio conoscenza e scopi, Il Mulino, Bologna, 1980.
[10] J. N. H. . I. G. M. . J. N. Shearme, Speech synthesis by rule, Language and Speech 7 (1964) 127–143.
[11] I. G. Mattingly, Synthesis by rule of prosodic features, Language and Speech 9 (1966) 1–23.
[12] G. I. F. . G. B. D. . R. D. Spinabelli, Study of a system of minimal speech-reproducing units for italian speech, JASA 43 (1968) 1282–1286.
[13] J. Allen, Synthesis of speech from unrestricted text, in: Proceedings of the IEEE - "Man

Machine Communication by Voice", volume 64, 1976, pp. 433—-442. doi:10.1109/PROC.1976.10152.
[14] N. Umeda, Consonant duration in american english, JASA 61 (1977) 846–858.
[15] D. H. Klatt, Structure of a phonological component for a synthesis-by-rule program, ASSPT - IEEE Transactions Acoustic Speech Signal Process 24 (1976) 391–398.

Legal Summarization: to each Court its own model

Flavia Achena[1], David Preti[1], Davide Venditti[2], Leonardo Ranaldi[2], Cristina Giannone[1], Fabio Massimo Zanzotto[2], Andrea Favalli[1] and Raniero Romagnoli[1]

[1]*Cognitive and AI R&D, Almawave S.p.A.*

[2]*Università di Roma Tor Vergata*

Abstract

In the Italian Civil Law System, easily accessing legal judgments through *massime* is crucial. In this work, we compare extractive summarization models to produce massime in two Italian courts: the *Constitutional Court* and the *Supreme Court*. The aim of our study is to assess the effectiveness and efficiency of these models in summarizing the decisions of the two courts. Through a comprehensive analysis of two large datasets, we evaluate the quality of the summaries generated by each model and their ability to capture the key legal principles and linguistic features present in the courts' decisions.

Keywords

Legal Text Analysis, BERT-based Summarization, Legal NLP,

1. Introduction

In civil and common law systems, accessing legal judgments to retrieve legal decisions is crucial when lawyers have to defend clients, prosecutors have to build cases, and judges have to draw decisions. To ensure widespread information on the decisions of the courts, in Italy, for this purpose, a specific body drawn up *massime*.

These *massime* present, in a short but detailed way, a legal principle present in judgments. Hence, justice professionals can read these *massime* instead of the complete legal decisions.

The process of analyzing judgments and extracting relevant sentences can be significantly simplified through the use of pre-trained models [1, 2], which serve as versatile universal sentence/text encoders, capable of addressing various downstream tasks, including summarization [3]. These models consistently outperform other approaches, especially after fine-tuning or domain-adaptation [4]. However, despite the success of pre-trained transformers in other summarization tasks, the task of producing *massime* is challenging for current extractive and abstractive summarization systems. Unlike standard summaries, *massime* must follow rigorous specifications in some courts. Extractive and abstractive summarization datasets and relative systems, in contrast, aim to reduce the size of a text while preserving its overall meaning. However, this approach differs from the specific requirements of creating *massime*.

Additionally, legal texts are often extensive, further increasing the summarization task's complexity. Identifying the portions of the text that contain the relevant information to be reported in the *massime* becomes challenging due to their length[5].

Legal document summarization has seen rapid progress in recent years, and several approaches[6, 7] have been proposed to manage this kind of data, ranging from fine-tuned Transformer models on legal domain, Reinforcement Learning to Generative Models[8].

In this paper, we compare extractive summarization models to produce *massime* in two different contexts: the Constitutional Court (Corte Costituzionale) and the Supreme Court (Corte di Cassazione). We discuss similarities and differences about *massime* and how the kind of court impacts the data in terms of their availability and privacy management. Then, we propose two models tailored to the specific type of courts, discussing the approaches we implemented to circumvent the issue related to lengthy documents. Results of the experiments confirm that producing *massime* is a real challenge even for dedicated systems. Hence, these systems should be designed as facilitators in a human-in-the-loop environment [9].

2. Different courts, Different Judgments, and Different *massime*

The data for the Italian legal domain have some peculiarities that require careful consideration. Different courts, such as the Constitutional and Supreme Court, produce different judgments, leading to different *massime*. Moreover, within the same court, there can be judgments with varying numbers of related *massime*, ranging from one to five or even more. In addition, the availability of data depends on the presence or absence of sensitive informa-

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
[first-initial].[surname]@almawave.it (F. Achena); [name].[surname]@uniroma2.it (D. Venditti)

tion within the judgments, so the legal courts provide access to the data in different ways.

Both Constitutional and Supreme Court courts share that producing a *massima* requires a relevant cognitive task carried out by the "Massimario Office" body, as it involves identifying the principle of law present in the judgment and satisfying some precise criteria for its writing. The following are the characteristics of the two Italian Courts under consideration (Sec. 2.1), the nature of their judgments and *massime*, how a *massima* is structured (Sec. 2.3), and, finally, a comparative analysis of the two types of corpora that can be derived from these two courts (Sec. 3) is provided.

2.1. Two Italian Courts: Constitutional Court and Supreme Court

The Italian Constitutional Court has the primary responsibility to assess the constitutionality of the acts and laws of the State and the Regions. Among other functions, it assesses charges against the President of the Republic in accordance with constitutional provisions. The Court examines the admissibility of abrogative referendums. To ensure impartiality and independence, the Constitutional Court is composed of 15 lawyers, chosen from among judges, law professors, or lawyers with at least 20 years of experience.

The Supreme Court - also known as the "Corte di Cassazione" - is the highest authority in the Italian judicial system. It serves as the court of final appeal and has two main functions. Firstly, it resolves judicial conflicts to determine which judge has jurisdiction over a case. Secondly, it has a *nomophylactic* function, ensuring that the law is interpreted uniformly. Within the Court, there is the "Massimario Office", responsible for identifying nomophylactic judgments and producing concise summaries called *massime*. These summaries contain the legal principles from the Court's judgments, not just a summary of the cases themselves. The primary objective of the Massimario Office is to disseminate legal knowledge and facilitate comprehension of past court decisions. To accomplish this, the office updates its collection by incorporating new judgments, ensuring access to the most current precedents. This results in a large number of judgments and massime so in the vision of making the judicial system more efficient by digitising court proceedings, providing automatized support to the processes can reduce the time and effort required to analyze and summarize them.

2.2. Availability of judgments and *massime* in the two Courts

A fundamental element that affects data availability concerns personal information and privacy. In cases where judgments contain sensitive personal information, access to such data is restricted due to privacy protection laws.

The Italian Constitutional Court, since it is central to the defense of the Constitution, prioritizes the availability of data on its proceedings and decisions. The Court must ensure the integrity and adherence to constitutional principles within the legal system. As a result, inquiries made to the Court generally focus on broad issues that do not involve specific individuals. Consequently, judgments do not contain any personal information and are not subject to privacy-related restrictions. Data of the Italian Constitutional Court are thus open and accessible through its portal[1].

On the other hand, the Supreme Court deals with cases that may involve specific physical or juridic people, which requires compliance with privacy regulations. Consequently, access to its data must be restricted. Information on the proceedings and decisions of the Supreme Court is only accessible through the Italgiure platform[2], which is exclusively available to professionals and legal practitioners. Data cannot be shared, and accesses are controlled and logged. Currently, the dataset selected for the Supreme Court cannot be made public because it would require an expensive anonymization process to ensure privacy.

2.3. The shape of massime

Each legal judgment (also called decision), despite its individuality in terms of case and subject matter, has a shared overall structure. This structure comprises the following key components:

- Heading/Epigrafe: It is the initial part containing the indication of the members of the court, the details of the initiating document, the reporting judge, and the attorneys heard by the Court.
- Statement of Facts: Summarizes the relevant facts of the case, often introduced by "considered in fact and in law.".
- Reasons: It is the section where the Court provides an explanation or argumentation for the conclusions reached in the judgment. This section typically presents the legal principles, factual analysis, and logical reasoning that support the Court's decision.
- Ratio Decidendi: Establishes the binding legal principle or rule derived from the court's decision.

[1] https://dati.cortecostituzionale.it
[2] https://www.italgiure.giustizia.it/

- Disposition: Concludes the decision with the final ruling and any related orders or remedies. It is often introduced by *"P.Q.M."*[3]: It contains the determination of the judges.

Similar to the decisions, the creation of summaries of legal principles, commonly known as *massime*, follows well-defined summarization criteria. As outlined in [10], these *massime* must contain explicit legal references and embody the fundamental principles of law. This detailed approach ensures the effective spread of legal knowledge. *Massime* must meet the following requirements:

- Faithfulness to the decision.
- Conciseness in stating the legal principle.
- Clarity and precision of the stated principle.

Hence, *massima* represents the expression of the legal principle and must not be considered a summary of the decision.

Figure 1: The plot illustrates an increase in "similarity" between the *massima* and pronunciation after year 2000 (with a similarity threshold of 90%) in the Constitutional Court.

3. The datasets of judgments and *massime*

3.1. Analysis of the *massime* of the two Courts

To better understand how to develop a system for *massime* generation, we analyzed the correlation between the judgments and the *massime* of both courts as they have different roles and consequently deliver different judgments.

For the Supreme Court, we selected a subset of judgments, from 2010 to 2013, to build our dataset useful for the extractive summarization task. During our analysis, we noticed that some decisions may include a *massima* without any text or expressed with an abbreviation, such as *"CONFORME A CASSAZIONE ASN: ..."*. For these cases, we interpret them as references to previous *massime*, but decline to use these specific examples. We started by selecting only the judgments corresponding to at least one *massima*. Indeed, we observed that while most legal judgments of the Supreme Court are tied to a single *massima*, there are a sizable amount of cases in which multiple massime refers to the same judgment (see Tab. 1). Details about how we handled such cases are discussed in the next subsection.

In addition to analyzing judgments from the Supreme Court, we also conducted a systematic analysis of Italian Constitutional Court judgments from 1956 to 2021. We aligned sentences in massime with sentences in the judgments in order to understand how sentences in massime are different from those in the judgments (see Figure 1).

According to our analysis, massime of the Constitutional Court became more extractive after 2000. Indeed, in that period, it seems that Constitutional Court Judges forced the "Massimario Office" to avoid changing the text extracted from judgments because even a small change of a single word could significantly alter the overall concept expressed in the judgment. As a result, since 2000, the process of producing massime become an extractive summarization task guided by a topic presented in the last part of the judgment.

3.2. Producing *massime* as a classification task

Summarization is an inherently abstractive task. However, it can be treated as an extractive classification task once the target summary (i.e., a *massima*) is used to select the *relevant* or *irrelevant* sentences from the starting document (i.e., a judgment).

3.2.1. Supreme Court Extractive Data-set

As mentioned before, the first step of the extractive model used to deal with the Supreme Court dataset (see Sec. 2.1) consists in rephrasing a generic *abstractive* summarization dataset into something suitable for a (classical) classification model. This is achieved via the introduction of a *Oracle* meta-model[11].

For each pair (document, summary), all the sentences forming the set with the highest F1 Rouge [12] combination $R_1 + R_2$ concerning the summary are selected and annotated as *relevant*, while all the others are automatically identified as *irrelevant*. This automatically frames

[3] for these reasons

massima per judgment	Supreme Court	Constitutional Court
1	65%	36%
2	24%	26%
3	6%	24%
4	3%	11%
5+	2%	3%

Table 1
The fraction of distinct *massima* per judgment is displayed for both Courts under investigation.

the dataset into a binary classification perspective

$$(document, summary) \overset{\text{Oracle}}{\overbrace{\rightarrow}} (sentence, category)$$

with $category = 0, 1$. As shown in Tab. 1, when a judgment is related with more than one massima, the *Oracle* model acts independently on each judgment-maxima couple, and then the annotated sentences are merged together without repetitions. This is done because otherwise, it can most likely happen that in a multiple *massima* scenario, the same sentence in a judgment is related only with one *massima*, ending up with the same sentence annotated with opposite categories.

Starting from a dataset corresponding to 12000 couples of $(massime, judgments)$ we decided to keep only the data corresponding to at an Oracle rouge of $R_1 + R_2 \geq 0.55$, reducing our training data almost by half (6.849). We observed that, given the nature of the judgment, the number of relevant sentences in any judgment is a very small portion, inevitably producing a highly unbalanced dataset toward the *irrelevant* sentences.

3.2.2. Constitutional Court Extractive Data-set

Given the analysis of the *massime* and the related judgments of the Constitutional Court, we decided to define the task of producing *massime* as the classification task of selecting the appropriate sentences of the judgment given a target topic. The classification dataset is then built as follows, starting from the judgments and the related *massime*. For each judgment, we extracted the points of its operative part (*punti del dispositivo*). For each point, we selected the correlated *massima*. Then, we divided the judgments into sentences and produced a set of triples:

$$(sentence, point, in\_massima)$$

where $sentence$ is a sentence of the judgment, $point$ is a point of its operative part, and $in\_massima$ is True if the sentence overlaps for more than 90% with a sentence in the massima related to the $point$.

For our experiments, we extracted a subset of 40,000 data points from this expanded dataset. The selection process ensured a balanced distribution between positive and negative examples, maintaining a 50/50 ratio. It is important to note that the specific details and steps of the method used to derive the larger dataset from the original 14,316 rows are not provided in this paper. However, this method facilitated a focused analysis of the textual components, shedding light on the connections between phrases, device points, and the formation of massime.

4. Models

Our main challenge was identifying the most relevant parts of pronouncements to assist the massima producer in crafting legal maxims.

As mentioned before, extractive summarization models treat the task of automatic summarization generation as a straightforward sentence classification task. In this vision, the summary of a given document emerges by the concatenation of all the most relevant document *fragments* (i.e., sentences or sub-sentences) classified by the model, this could effectively provide the Massimario with the essential subparts of pronouncements for massima construction.

Both models proposed in the current work are essentially based on a deep encoder which maps the fragments to a vector representation in a high dimensional space subsequently classified into two classes: *relevant* sentences (i.e., candidates for the summary) or *irrelevant* sentences (i.e., not containing relevant information for the summary).

4.1. Supreme Court Model

Data-sets with very long documents (as the one introduced in Sec. 2.1) are usually difficult to handle using a BERT-based [2] transformer encoder. The well known self-attention (introduced in [13]) which characterizes most of the transformer networks is plagued by a fast scaling of computational and memory requirements with the input sequence. Instead of proceeding with a more memory-efficient attention implementation (for instance, see [14]), we decided to act on data and restrict the context length. In this perspective, we introduced a fixed

Court	Prec	Rec	F_1	R_1	R_2	R_3	$\bar{R}_1$	$\bar{R}_2$	$\bar{R}_3$	$\tilde{R}_1$	$\tilde{R}_2$	$\tilde{R}_3$
Supreme	0.40	0.31	0.35	0.47	0.32	0.28	0.64	0.52	0.50	1.97(8)	3.69(35)	4.45(54)
Constitutional	0.53	0.80	0.52	0.32	0.29	0.24	-	-	-	-	-	-

Table 2
Classification and Coverage results of the two models considered. Normalized values with respect to the *Oracle* coverage $\bar{R}_n = R_n/R_n^{\text{Oracle}}$ and random baseline $\tilde{R}_n = R_n/R_n^{\text{Random}}$ where sentences are extracted with the same frequency as in the train set.

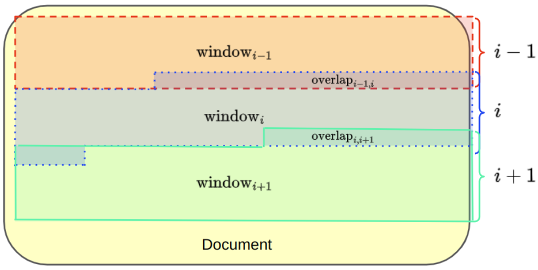


Figure 2: Sketch of the sliding window procedure with overlap. The shaded area in different colors corresponds to different windows (i.e., contexts).

length *sliding window* similar to the implementation described in [15]).

Our goal was to optimize the context length and mitigate context-truncation effects. To achieve this, we defined the context window based on word-pieces and introduced the possibility of overlapping windows up to a maximum number of sentences. The latter, while still under investigation, offers an intriguing tool to probe the context effects on the model. Even in the simplest implementation, with only one overlapping sentence (see Fig. 2), it is interesting to see the effect of a preceding or subsequent context on the same sentence.

The model used, with the aforementioned modification in the document pre-processing, is based on the one proposed in [16, 3] referred to as *BERTSUM*[4]. It is worth mentioning that while in their works, the predicted probability is used only as a ranking score and a fixed number of sentences are extracted neglecting the actual probabilities, we select the relevant sentences accordingly to a cutoff parameter. The latter is a necessary introduction since we observed that, in our data-set, it is common to have documents with a clear separation between sentences that are very likely to be extracted compared to others whose predicted probability is minimal. Therefore, fixing the number of extracted sentences introduces a strong bias toward the extraction of irrelevant sentences.

We observe that the model does not seem to reach high performances both in terms of absolute and relative scores, normalized with the *Oracle* Rouge scores, which are the maximum scores such a model can achieve (see Tab. 2). This is partially due to the violent class unbalance present in the dataset, even if marginally mitigated by the introduction of a weighted loss, with weights inversely proportional to the category frequency in the train set. As a baseline comparison, we decided to include the scores normalized with the one of a random classifier to assess that no random classifications are being performed.

4.2. Constitutional Court Model

The challenge lay in selecting the most useful subparts of legal judgments for the purpose of the massima producer. We sought to leverage BERT[2] to provide the best subparts of pronouncements to aid in massima construction. However, we soon realized that the task was exceptionally complex, requiring the ability to summarize and generalize the text in a unique manner.

To address the above multifaceted challenge, we focused on using BERT to assist us in identifying the most relevant subparts of legal judgment. Through this approach, we aimed to equip the massima producer with essential tools for constructing the maxim more effectively. While our efforts resulted in the development of a tool to assist the massima, producer, we must acknowledge that the results achieved with BERT were not as remarkable as initially hoped. The complexity of the problem, combining the tasks of summarization and generalization in a unique manner, presented formidable hurdles.

Nevertheless, we view this endeavor as a stepping stone toward understanding and tackling the intricacies of legal text processing. Our tool, despite its limitations, serves as a valuable resource for the Massimario, aiding them in the maxim construction process. We recognize that further research and advancements in natural language processing will be crucial in making substantial strides in this domain.

Even in this case, results are interesting but not yet satisfactory (see Tab. 2). Indeed, R1, R2, and R3 are 0.32,

[4] sometimes named *BertSumExt* in literature.

0.29, and 0.24 respectively. This suggests that the task of producing massime is indeed a challenging task.

5. Conclusions

In conclusion, while we did not achieve outstanding results, our efforts shed light on the intricacies of this challenging problem. During our analysis, we also noticed notable differences between the two courts, which further emphasizes the complexity of generating accurate *massime*. We find it particularly intriguing to explore the factors contributing to these variations and understand how they impact the summarization process. Despite the challenges, we remain committed to refining our approach and exploring innovative techniques. Recent advances in the field further motivate us to seek a proper solution that addresses data privacy concerns and significantly improves the task of summarization in the legal field for the Italian language.

Acknowledgments

This work was conducted within the *DATALAKE Giustizia project*; we acknowledge the partners and the scientific committee for their support.

References

[1] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, 2018. arXiv:1802.05365.

[2] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[3] Y. Liu, M. Lapata, Text summarization with pretrained encoders, 2019. URL: https://arxiv.org/abs/1908.08345. doi:10.48550/ARXIV.1908.08345.

[4] X. Jin, D. Zhang, H. Zhu, W. Xiao, S.-W. Li, X. Wei, A. Arnold, X. Ren, Lifelong pretraining: Continually adapting language models to emerging corpora, in: Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models, Association for Computational Linguistics, virtual+Dublin, 2022, pp. 1–16. URL: https://aclanthology.org/2022.bigscience-1.1. doi:10.18653/v1/2022.bigscience-1.1.

[5] E. Bauer, D. Stammbach, N. Gu, E. Ash, Legal extractive summarization of u.s. court opinions, 2023.

[6] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, I. Androutsopoulos, LEGAL-BERT: The muppets straight out of law school, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 2898–2904. URL: https://aclanthology.org/2020.findings-emnlp.261. doi:10.18653/v1/2020.findings-emnlp.261.

[7] I. Chalkidis, A. Jana, D. Hartung, M. Bommarito, I. Androutsopoulos, D. Katz, N. Aletras, LexGLUE: A benchmark dataset for legal language understanding in English, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 4310–4330. URL: https://aclanthology.org/2022.acl-long.297. doi:10.18653/v1/2022.acl-long.297.

[8] M. Cherubini, F. Romano, A. Bolioli, N. De Francesco, I. Benedetto, Summarization di testi giuridici: una sperimentazione con gpt-3, Rivista Italiana di Informatica e Diritto (2023). doi:10.32091/RIID0103.

[9] F. M. Zanzotto, Viewpoint: Human-in-the-loop artificial intelligence, Journal of Artificial Intelligence Research 64 (2019) 243–252. URL: https://doi.org/10.1613%2Fjair.1.11345. doi:10.1613/jair.1.11345.

[10] F. Costantini, P. D'Ovidio, Sintesi dei criteri della massimazione civile e penale, https://www.cortedicassazione.it/cassazioneresources/resources/cms/documents/SINTESI_CRITERI_DELLA_MASSIMAZIONE_CIVILE_E_PENALE.pdf, 2023.

[11] C.-Y. Lin, E. Hovy, The potential and limitations of automatic sentence extraction for summarization, in: Proceedings of the HLT-NAACL 03 Text Summarization Workshop, 2003, pp. 73–80. URL: https://aclanthology.org/W03-0510.

[12] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017. arXiv:1706.03762.

[14] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The long-document transformer, 2020. arXiv:2004.05150.

[15] Q. Grail, J. Perez, E. Gaussier, Globalizing BERT-

based transformer architectures for long document summarization, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, 2021, pp. 1792–1810. URL: https://aclanthology.org/2021.eacl-main.154. doi:10.18653/v1/2021.eacl-main.154.

[16] Y. Liu, Fine-tune bert for extractive summarization, 2019. URL: https://arxiv.org/abs/1903.10318. doi:10.48550/ARXIV.1903.10318.

CorpusCompass: A Tool for Data Extraction and Dataset Generation in Corpus Linguistics

Muhadj Adnan[1], Nicolo' Brandizzi[2]

[1]Arabic Linguistics, University of Bayreuth, Bayreuth, 95447, Germany, DE
[2]Department of Computer, Automation and Management Engineering, Sapienza University of Rome, 00185, Italy, IT

Abstract
As the need for effective tools in Corpus Linguistics continues to grow, particularly for under-resourced languages and nonstandard annotation tasks, specialized software has become essential for processing and analyzing large and complex datasets. This paper introduces *CorpusCompass*, a new open source tool for data extraction and dataset creation, which offers a number of functionalities for researchers interested in analyzing corpora. The tool can derive structured datasets from text annotated with custom annotation schemes, while also checking for errors and consistency. By defining custom variables of interest and annotation rules, researchers can tailor the tool to their specific needs, making it particularly valuable for unique linguistic research domains. When used in conjunction with statistical analysis or visualization tools, *CorpusCompass* helps researchers to gain insights into the factors that are affecting language use. In this paper, we introduce the tool and give a real-world example in the field of language variation.

Keywords
Corpus Linguistics, Under-resourced languages, Nonstandard annotation tasks, Data Extraction, Dataset generation, Statistical analysis, Language variation, Sociolinguistics

1. Introduction

In recent years, there has been a growing interest in studying language variation in under-resourced languages. Mair [1] identifies a lack of resources for spoken data in corpus linguistics and emphasizes the need for more computational tools for different languages and varieties. The process of creating and analyzing a spoken language corpus is complex, posing a range of challenges for researchers in the field of Corpus Linguistics.

In this context, we identify six main steps for this process, each presenting its own set of practical and technical challenges, see Figure 1. Step (i) entails *sourcing and recording data*, along with the associated metadata, which provides essential contextual information about the recordings. Following this, step (ii) involves *transcribing* the spoken data to convert it into a text-based format, allowing for more straightforward analysis. The third step (iii) involves *annotating* the data with linguistic features, such as phonological or morphological information. Subsequently, step (iv) requires *data preprocessing* to clean and organize the data, preparing it for the fifth step, (v), which is *data analysis*. This stage allows researchers to derive insights from the corpus by examining patterns and connections within the data. Lastly, the final step, (vi), involves *publishing and sharing* the corpus with the wider research community, promoting collaboration and further research based on the spoken language data.

Annotating can be a time-consuming and error-prone process, especially when working with large corpora. Errors in a manually annotated corpus can potentially affect the evaluation. Additionally, poor quality annotation in the corpus can lead to misleading results in a linguistic analysis. This is where *CorpusCompass* comes into play.

In this paper, we provide a detailed overview of *CorpusCompass*[1], including its design, implementation, and functionalities. The tool is based on *Jupyter Notebook* and is designed to help Corpus Linguistics researchers focusing on language variation to create a structured dataset from their previously annotated corpus/corpora and a list of variables of interest (see Section 3.1). The tool is coded in Python and can be run in an interactive manner using *Google Colab*. Once run, it generates a structured dataset, i.e. a systematically organized collection of data, that includes linguistic variables and potentially relevant metadata (see Section 3.2).

On the one hand, the dataset enables corpus exploration, assisting in discovering patterns that can inform the creation of new hypotheses or the dismissal of initial assumptions (see Al-Wer et al. [2], pp. 37-38). On the other hand, it facilitates performing statistical analyses using established methods through platforms like Rbrul

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
muhadj.adnan@uni-bayreuth.de (M. Adnan); brandizzi@diag.uniroma1.it (N. Brandizzi)
https://nicofirst1.github.io/ (N. Brandizzi)
0009-0001-5174-3897 (M. Adnan); 0000-0002-3191-6623 (N. Brandizzi)

[1]The code is available on GitHub, for the URL, please visit the website https://www.corpuscompass.com/. Please note that the code for *CorpusCompass* is constantly evolving and, in this paper, we refer to version 1.0.0.

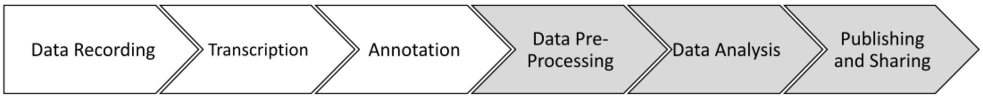


Figure 1: Practical and technical challenges when creating spoken language corpora in six steps: from sourcing and recording data (and metadata), to transcribing, annotating, and marking up datasets. The last three steps (data preprocessing, data analysis, and publishing and sharing) are highlighted, as these are the areas where *CorpusCompass* provides support and assistance to researchers.

[3], SPSS [4], and R [5].

To exemplify the practical application of *CorpusCompass* within a research context, Section 4 explores a case study on inter-generational linguistic variation among Iraqi speakers living in Germany, and highlights various potential uses of the tool.

For linguists without any or limited programming skills, *CorpusCompass* has the potential of saving time (by automating repetitive tasks) and improving the accuracy of their research. It is intended to bridge the gap between researchers and advanced statistical analysis by facilitating the connection between them, as well as addressing research questions that require the use of manually annotated data. Moreover, *CorpusCompass* was developed by researchers in Linguistics and Computer Science to advance Corpus Linguistics tools and promote interdisciplinary collaboration.

2. Related Work

Many text annotation tools have been developed primarily in the context of Artificial Intelligence and Natural Language Processing. However, such tools typically do not address specific needs, focusing more on the annotation of informational content rather than linguistic properties of text (phonological, grammatical, lexical, etc.). For this reason, linguists typically work with tools that have been developed specifically for the purpose of annotating linguistic corpora. In the following, we will briefly outline the importance of these tools and demonstrate how *CorpusCompass* complements and extends their functionality.

There are several tools available for researchers in Corpus Linguistics (see Neves and Ševa [6], Berberich and Kleiber [7]). *AntConc* [8], *Monoconc Pro* [9], and *WordSmith* [10] remain popular tools due to their wide range of functions, including KWIC (key word in context) concordancers, collocates, word frequencies, keywords and other corpus analysis features.

AntConc is a powerful corpus analysis tool, but it has certain drawbacks. One major limitation is the lack of a feature to create structured datasets from the corpus or various corpora, which is essential for statistical analysis pipelines as well as to share data. This is particularly problematic for researchers working with complex data involving various phonological, morphological and lexical variables as well as different speakers, as detailed and structured extraction is necessary. Sociolinguistic studies, in particular, require flexibility in handling multiple speakers and their background information. With *CorpusCompass*, we aim to address this gap in functionality.

WordSmith and *Monoconc Pro* share these limitations with *AntConc*, but they also have an additional drawback: they are not freely available. This lack of open access can be a significant barrier for corpus linguistics practitioners who may not have the financial resources to purchase these tools.

Other useful modern tools are typically focused on the annotation process rather than the analysis. For example, INCEpTION[2] [12] is a cloud-based platform that enables researchers to create and share linguistic annotations in a collaborative environment, for various languages. Another example is FLAT [13], a web-based linguistic annotation tool that revolves around the FoLiA format, a customizable XML-based format for linguistic annotation. However, for state-of-the-art analysis and error-checking of annotations produced by such tools, we would prefer to not rely on their built-in capabilities, but instead make use of common data science methodology such as statistical analysis in *R* or visualization with *Gephi*. This requires exporting the annotations from the typically XML-like formats that these tools use to a tabular data format used in data science such as CSV or TSV files. This is a core functionality of *CorpusCompass*, addressing a gap in the existing tools.

Biber et al. [14], Gries [15], and Weisser [16] suggest that learning programming and developing one's own analytical tools can overcome limitations in existing corpus tools.

Biber et al. [14] suggest that this would allow corpus linguists to perform faster and more accurate analyses, and the ability to tailor the output to suit the particular research requirements. Furthermore, according to Gries [15], utilising a pre-existing tool lets the researcher become dependent of the company or individual developing them, whereas programming allows them to have control over their research needs. Therefore, corpus lin-

[2]The software was previously known as *WebAnno* [11].

Figure 2: Example of a subset of the corpus dealing with linguistic variation, highlighting the annotations matching the standard *regex* pattern in green. Note the different anonymized speakers based on age: young (A,BSH); old (S, SUH). Image made with https://regex101.com/.

guists have clear benefits from learning a programming language, both in terms of the flexibility to develop tools for specialized tasks, as well as providing them with an understanding of the issues faced by tool developers creating general purpose tools.

As seen in the overview above, each tool addresses unique needs, thus implementing specific functionalities. In comparison, *CorpusCompass* tackles a complementary set of challenges. It is implemented as a *Jupyter Notebook* and can be run in an interactive manner using *Google Colab*, which not only makes it more accessible to users but also gives beginners the possibility to get familiar with programming.

3. Methodology

The modular code structure of *CorpusCompass* includes a file handling module for managing JSON and CSV files, an annotation parsing module that extracts annotations using regular expressions (refer to Section 3.1), a dataset construction module for creating a structured CSV dataset, and a logging module that displays relevant information such as program status and annotation details. *CorpusCompass* provides helpful functions for string manipulation, data cleaning, and error handling, simplifying the data analysis process. These functions enhance dataset accuracy by mitigating errors and inconsistencies during the data preparation phase. In the following, we describe the pipeline of *CorpusCompass* and demonstrate how it simplifies the process of extracting valuable insights from spoken language data.

3.1. Defining Variables

Defining variables of interest is a crucial step in using *CorpusCompass*, as it allows researchers to tailor the tool to what they aim to investigate. In the field of Corpus Linguistics, variables are often used to study language variation and how it is affected by various factors such as speaker demographics (e.g. age, sex, education), linguistic context (e.g. dialect, register), social context (e.g. audience, situation), or properties of a construction (e.g. morphemes, idioms). By defining both independent and dependent variables in their structured dataset, researchers have maximum freedom in the exploration of their corpora and the creation of their unique datasets.

Regular Expressions Based on their research objectives, researchers may use automated annotation tools or choose to manually annotate data in more complex linguistic situations (as described in Section 4). This leads to a broad range of annotation rules. To accommodate this variety, *CorpusCompass* employs regular expressions (*regex*) [17] for accurate extraction of annotations from the corpus. *Regex* allows the user to define text patterns, and is useful for tasks such as input validation, text modifications, and data extraction.

Figure 2 shows four paragraphs taken from our corpus[3], where annotations are highlighted in green. For the sake of simplicity, we kept only the dependent variable that are at the basis of our analysis in Section 4.2.

[3]Strict phonetic transcription was not followed due to the focus on pre-selected specific features, as adhering to it for the extensive 22-hour audio recordings would have been time-consuming.

Appendix A reports the full list of variables used for the study. It is important to note that in our complete annotated files, we typically have multiple annotations per word.

3.2. Generating a Structured Dataset

After the variables and *regex* rules have been defined, researchers can run *CorpusCompass* to generate a structured dataset. The dataset construction module automatically performs several steps, including cleaning and preprocessing the extracted annotations, grouping them by speaker and file, and writing them to a CSV file. Six output files are created, including five CSVs (dataset, annotation_info, missed_annotations, unk_variables, and binary_dataset) and one JSON file (corpus_stats).

The *dataset* file is a structured dataset based on the defined variables, with each row corresponding to a token in the annotated corpus and each column representing a variable category. This organized representation of annotations facilitates the analysis. The *annotation_info* file contains information about the annotations included in the dataset file, such as the token itself, the number of times it appears in the dataset, and the number of times it appears for each speaker. This information can be useful for identifying patterns in the data, such as the most common tokens or the distribution of annotations across speakers.

The *missed_annotations* file tracks tokens that were previously annotated but not consistently annotated in subsequent instances[4]. The file's purpose is to identify tokens that were once deemed important but not annotated consistently. Furthermore, in projects with multiple annotators, inter-annotator disagreement is a known challenge [18]. During the annotation process, it is possible for researchers to come up with new variables that were not previously specified in the JSON. However, researchers may forget to add these variables to the file, leading to inconsistencies in the dataset. The *unk_variables* file contains a list of variables that were not specified in the JSON. Finally, the *corpus_stats* file provides an overview of the corpus by reporting key statistics. Access to these statistics enables researchers to better understand the size and structure of their corpus, and can also provide valuable information for reproducibility purposes.

For a comprehensive description and additional information regarding the CSV files, please see Appendix B.

[4]The file might contain false positives since *CorpusCompass* does not differentiate between different meanings of the same token.

4. Analysing Linguistic Variation Using *CorpusCompass*

Following the annotation of the data with linguistic features, we used *CorpusCompass* for preprocessing in order to prepare the data for the analysis process. This transition from data preprocessing to analysis was facilitated by the integration of the tool into our workflow, significantly enhancing the efficiency and effectiveness of our exploration process.

In our case study, we examine Arabic-speaking communities, specifically Iraqis and Syrians, residing in Germany since 2014, following standard sociolinguistic variationist research practices. The participants are Iraqi and Syrian Arabic native speakers. We select phonological, morphological, and lexical variables for statistical analysis, with age as a key independent variable influencing linguistic variation. The study aims to examine inter-generational differences within the two groups and explore the extent to which a koiné (common variety) results from dialect and language contact in the migration context between the Syrian and Iraqi participants.

4.1. The Corpus

The corpus utilized in this study, of which a sample is illustrated in Figure 2, represents only half of our entire dataset and has been phonetically transcribed[5] using the International Phonetic Alphabet (IPA) and annotated by a single person in *Notepad++*. While having one annotator can be a common case in the field, mostly due to resource limitation, it can also be prone to errors and inconsistencies. The analyzed corpus is comprised of 2,101 paragraphs and encompasses 114,550 words and 654,431 characters. It features 24 speakers in total, 14 of whom are Iraqis, considered speakers of interest for our analysis. The dataset contains 35 variables, with 25 being dependent variables and 10 being independent variables. These variables are represented by 69 distinct values, 53 of which correspond to dependent variable values, and 16 to independent variable values.

In total, the corpus contains 3,366 unique annotations, with 13,641 annotated tokens. Given the substantial size of the dataset and the numerous variables involved, organizing the data in a structured manner is crucial for efficient analysis.

For comprehensive details regarding the corpus collection process, transcription methodology, annotation procedures, and the specific tool employed, please refer to Appendix C.

[5]Supported by one assistant during transcription of the recorded data.

Research Question In the following section, we use *CorpusCompass* to answer two research questions focusing on Iraqi speakers: (i) does age influence the usage of religious expressions? (ii) are young speakers more subject to German borrowings while speaking Arabic? These questions will guide our exploration of potential correlations within the dataset, with the understanding that the current analysis serves as a simplified demonstration. However, it should be noted that the dataset is well-suited for rigorous statistical analysis, including techniques such as regression analysis.

Dependent Variables For the purpose of this example, we have selected two categories of dependent variables to investigate: (i) religious expressions[6], represented by the label RELIG; (ii) the influence of German language, represented by multiple labels such as G-DL for daily life, G-EDU for education, and G-JOB for working contexts. Since we are interested in the general use of German words, we generalize the labels to GERM.

Independent Variable The independent variable chosen for analysis in the corpus is *age*, which is an important factor to consider in language variation. The speakers are divided into two categories, young (21-26 years) and old (46-55 years).

4.2. Analyzing and Sharing the Data

This section discusses two types of analysis: error checking and data analysis. Error checking is the process of identifying and fixing errors or inconsistencies in the data, while data analysis involves using statistical and visualization techniques to extract insights and draw conclusions from the data.

Error Checking *CorpusCompass* can identify any errors or inconsistencies during the annotation process and generate separate CSV files that provide information on annotated and non-annotated tokens. Thanks to the generated file, we were able to find circa 400 (3% of all the annotations) ill-formatted annotations (e.g. "[$G-JOB.biriif["), 1,305 missed annotations of which 205 where considered correctly identified and more than 9 unknown variables, i.e. dependent variables that are present in the corpus but not specified beforehand by the user.

Data Analysis By importing *the binary_dataset* in Excel [21], we determined the cross-tabulation (pair-wise frequency) of dependent and independent variables. Table 1 presents these frequencies along with the total number of words spoken by young and old speakers. By normalizing the frequencies and estimating the proportions (old vs. young), we observe the following:

Table 1
Frequency of dependent variables (*GERM* and *RELIG*) across age groups (Old and Young) along with the total number of words spoken by each group.

Age Group	GERM	RELIG	Words
Old	221	357	42,483
Young	505	175	39,406

$$RELIG_p = \frac{RELIG_{old}}{RELIG_{yng}} \cdot \frac{Words_{yng}}{Words_{old}} = 1.89$$

This indicates that old speakers use 189% times more religious phrases than young speakers. Applying the same method for GERM, we obtain:

$$GERM_p = \frac{GERM_{old}}{GERM_{yng}} \cdot \frac{Words_{yng}}{Words_{old}} = 0.40$$

Correspondingly, old speakers use 40% of the amount of German borrowings compared to young speakers. To assess the significance of our findings, we conducted standard t-test analyses with DATAtab [22]. When comparing the proportion of older speakers using religious expressions to that of younger speakers using the same expressions, the result was the following:

$$t(10695) = -7.91,\ p < .001$$

In contrast, the analysis of German borrowings between older and younger populations yielded:

$$t(10695) = 10.59,\ p < .001$$

The *p-value* suggests that the dependent variables (German borrowings and religious expressions) play a role in the language variation exhibited by young and old Iraqi migrants residing in Germany and requires further investigation. Ultimately, these findings validate our initial research question and demonstrate the value of structured data in facilitating robust statistical analyses.

5. Limitations

CorpusCompass, while offering numerous features, is not without its limitations. In this section, we outline some of the primary constraints of the tool, alongside potential future developments.

[6] Jaradat [19] and Piamenta [20] describe religious phrases, such as *Inshallaah* (God willing), *alhamdulillaah* (Praise be to God), *Allah ysallimak* (may God protect you) etc. as "Allah expressions". They include an explicit or implicit reference to Allah, which is literally translated as "the God".

The user interface, especially the integration of *regex* and *Jupyter Notebook* might pose a challenge to linguists, particularly those hesitant to engage with programming. This could be solved by developing the tool into a more user-friendly application. Furthermore, we showcased the functionality of *CorpusCompass* in addressing a specific research question. While the tool has already been applied to address other research questions [23] and on another corpus[7], the extend of its usability remains a point of investigation. It is essential to assess its performance on multiple corpora to enhance its robustness and confirm its applicability for diverse research contexts.

Additionally, while the tool identifies errors, as delineated in Section 4.2, the manual correction process can be tedious and time-consuming. Looking forward, future iterations of *CorpusCompass* might integrate an automatic error correction feature that suggests possible corrections, allowing users to either accept or decline them. Another area for consideration is the tool's reliance on the CSV format, which might present compatibility issues with other linguistic tools. Transitioning to more standardized formats, such as XML, upcoming versions could address this limitation.

In summary, there are numerous opportunities to refine and enhance *CorpusCompass*. By addressing its current constraints, introducing new functionalities, and emphasizing user-centric enhancements, this tool has the potential to become an even more invaluable asset in Corpus Linguistics.

6. Conclusion

Creating and analyzing a corpus is a complex task that requires a range of technical and practical skills. In this paper, we have explored the challenges involved in these steps and introduced *CorpusCompass* as an innovative solution. The tool's aim is to simplify data extraction and dataset generation, facilitating the identification of significant features and syntactic errors in the annotations. This contributes to advancing the overall replicability of studies within the field of Corpus Linguistics. As *CorpusCompass* is implemented as a *Jupyter Notebook*, it also serves as an accessible introduction to programming for researchers who wish to expand their skill set and gain more control over their analytical processes.

Additionally, we have presented a real-world example of how *CorpusCompass* can be applied in the field of language variation by using a subset of our corpus of Arabic varieties spoken by migrants in Germany, representing an under-resourced language. The example shows how the generated dataset can be used in conjunction with existing analysis tools to answer unique research questions. With *CorpusCompass*, we aim to contribute to the development of tools for spoken language corpora. The existence of this tool and its accessibility to researchers without a background in programming will lead to more quantitative studies that analyse such corpora. The tool exemplifies interdisciplinary collaboration and emphasizes the importance of linguistics researchers working with experts from computer science and engineering. This collaboration results in the development of flexible corpus tools applicable to a wide range of research studies.

Sharing is Caring It is essential to highlight that structured datasets are crucial for sharing data in linguistic research, particularly in connection to research data management practices and platforms such as *Figshare* [24]. Organized data facilitates sharing and reusability among researchers, enabling more extensive collaborations and the creation of larger datasets. Furthermore, structured datasets allow researchers to replicate and verify research findings, promoting transparency and accountability in the scientific community. Therefore, creating a structured dataset is not only essential for internal analysis but also for the advancement of the field and the dissemination of knowledge.

[7]The corpus is focused on Nigerian Arabic and has been kindly provided by Prof. Dr. Jonathan Owens.

Acknowledgments

We extend our sincere appreciation to Dr. Jelke Bloem (University of Amsterdam) and Prof. Dr. Jonathan Owens (University of Bayreuth) for their invaluable feedback.

The development of *CorpusCompass* was undertaken as part of the project *Modernity, Migration, and Minorities: Three Case Studies of Arabic in Contact* at the University of Bayreuth and funded by the Deutsche Forschungsgemeinschaft (DFG), No. 429257272.

References

[1] C. Mair, Erfolgsgeschichte korpuslinguistik?, in: Korpuslinguistik, De Gruyter, 2018, p. 5–26. URL: http://dx.doi.org/10.1515/9783110538649-002. doi:10.1515/9783110538649-002.

[2] E. Al-Wer, U. Horesh, B. Herin, R. De Jong, Arabic Sociolinguistics, Cambridge University Press, 2022. doi:10.1017/9781316863060.

[3] D. E. Johnson, Getting off the GoldVarb standard: Introducing rbrul for mixed-effects variable rule analysis, Language and Linguistics Compass 3 (2009) 359–383. URL: https://doi.org/10.1111/j.1749-818x.2008.00108.x. doi:10.1111/j.1749-818x.2008.00108.x.

[4] H. Norman, C. H. Hull, SPSS: Statistical package for the social sciences, McGraw-Hill Book Company, 1975.
[5] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2021. URL: https://www.R-project.org/.
[6] M. Neves, J. Ševa, An extensive review of tools for manual annotation of documents, Briefings in Bioinformatics 22 (2019) 146–163. URL: https://doi.org/10.1093/bib/bbz130. doi:10.1093/bib/bbz130.
[7] K. Berberich, I. Kleiber, Tools for corpus linguistics, https://corpus-analysis.com/ (Mar. 2023), 2020.
[8] L. Anthony, Antconc: A learner and classroom friendly, multi-platform corpus analysis toolkit, proceedings of IWLeL (2004) 7–13.
[9] A. Svedkauskaite, Monoconc pro 2.0 and the corpus of spoken professional american english: Resources from athelstan, Style 38 (2004) 127–133.
[10] S. Mike, Wordsmith tools version 8, 2023.
[11] R. E. De Castilho, E. Mújdricza-Maydt, S. M. Yimam, S. Hartmann, I. Gurevych, A. Frank, C. Biemann, A web-based tool for the integrated annotation of semantic and syntactic structures, in: Proceedings of the workshop on language technology resources and tools for digital humanities (LT4DH), 2016, pp. 76–84.
[12] J.-C. Klie, M. Bugert, B. Boullosa, R. E. de Castilho, I. Gurevych, The inception platform: Machine-assisted and knowledge-oriented interactive annotation, in: Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, Association for Computational Linguistics, 2018, pp. 5–9. URL: http://tubiblio.ulb.tu-darmstadt.de/106270/, event Title: The 27th International Conference on Computational Linguistics (COLING 2018).
[13] M. van Gompel, M. Reynaert, Folia: A practical xml format for linguistic annotation–a descriptive and comparative study, Computational Linguistics in the Netherlands Journal 3 (2013) 63–81.
[14] D. Biber, S. Conrad, R. Reppen, Corpus linguistics: Investigating language structure and use, Cambridge University Press, Shaftesbury Road ,Cambridge, CB2 8BS , United Kingdom, 1998.
[15] S. T. Gries, What is corpus linguistics?, Language and linguistics compass 3 (2009) 1225–1241.
[16] M. Weisser, Essential programming for linguistics, Edinburgh University Press, The Tun - Holyrood Road. 12 (2f) Jackson's Entry. Edinburgh EH8 8PJ. UK., 2009.
[17] A. V. Aho, Algorithms for finding patterns in strings, in: J. VAN LEEUWEN (Ed.), Algorithms and Complexity, Handbook of Theoretical Computer Science, Elsevier, Amsterdam, 1990, pp. 255–300. URL: https://www.sciencedirect.com/science/article/pii/B9780444880710500102. doi:https://doi.org/10.1016/B978-0-444-88071-0.50010-2.
[18] Y. Oortwijn, T. Ossenkoppele, A. Betti, Interrater disagreement resolution: A systematic procedure to reach consensus in annotation tasks, in: Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval), Association for Computational Linguistics, Online, 2021, pp. 131–141. URL: https://aclanthology.org/2021.humeval-1.15.
[19] A. A. Jaradat, The linguistic variants of allah expressions in jordanian arabic, Cross-Cultural Communication 10 (2014) 61–68.
[20] M. Piamenta, The Muslim conception of God and human welfare: As reflected in everyday Arabic speech, Brill Archive, Leiden, The Netherlands, 1983. URL: https://brill.com/view/title/1464. doi:https://doi.org/10.1163/9789004661820.
[21] C. Microsoft, Microsoft excel, https://office.microsoft.com/excel (Apr. 2023), 2023. URL: https://office.microsoft.com/excel.
[22] T. DATAtab, Datatab: Online statistics calculator, https://datatab.net/ (Apr. 2023), 2023. URL: https://datatab.net/.
[23] M. Adnan, J. Owens, Imperfect Verbal Prefixes as Discourse Markers, volume XXXV, Perspectives on Arabic Linguistic, Benjamins, Amsterdam, To Appear.
[24] T. Figshare, Figshare: the open research repository platform, 2023. URL: https://figshare.com/.
[25] L. Milroy, M. Gordon, Sociolinguistics: Method and interpretation, Wiley, 2003. URL: https://doi.org/10.1002/9780470758359. doi:10.1002/9780470758359.
[26] S. A. Tagliamonte, Analysing sociolinguistic variation, Cambridge University Press, 2006.
[27] P. Boersma, D. Weenink, Praat: Doing phonetics by computer, Ear and Hearing 32 (2011). URL: https://journals.lww.com/ear-hearing/Fulltext/2011/03000/Praat__Doing_Phonetics_by_Computer.12.aspx.

Speakers	Independent variables
`{ "A" : ["female","young"], "BSH" : ["male","young"], "S" : ["female","old"], "SUH" : ["female", "old"] }`	`{ "Gender":["male","female"], "Age":["old","young"] }`

Dependent variables

```
{
  "German Context" : [ "G-SCHOOL",  "G-JOB"],
  "Religious Phrases" : "RELIG"
}
```

Figure 3: Content of variables and speaker JSON files.

A. Variables

The study presented in Section 4.1 revolves around the variables and speaker information detailed in Table 3. The employed syntax adheres to the JSON specification, providing considerable flexibility in the examination of linguistic variables.

There are two main types of variables: independent and dependent variables. Independent variables, also known as input variables, are the factors that the researcher manipulates or controls in a study. In contrast, dependent variables, also known as output variables, are the outcomes or responses being measured. The dependent variables are affected by the independent variables.

The speakers involved in the example are represented by four anonymized aliases. For each individual, two attributes (i.e. independent variables), namely *age* and *gender*, are taken into consideration. These attributes encompass the categories of *male* and *female*, as well as *old* and *young*. Finally, the corpus is annotated with two dependent variables: *german context*, which can be categorized as either G-SCHOOL or G-JOB, and *religious phrases*, identified by the annotation RELIG.

B. Generated Files

Additional details regarding the files generated by *CorpusCompass* are provided in the following. In order to maintain simplicity and avoid overwhelming the reader with excessive information, we conducted calculations on a subset corpus, as illustrated in Figure 2. This facilitates a visual association between the output of *CorpusCompass* and the input data. Conversely, the files containing missed annotations and unknown variables are presented for the entire corpus, which serves as the foundation for the analysis in Section 4.2. This distinction is necessary as the subset corpus exhibits no errors or missed annotations.

Dataset The *dataset* file encompasses a structured dataset that reflects the defined variables, where each row corresponds to a token within the annotated corpus, and each column represents a distinct variable category (see Figure 4). Additionally, a *binary_dataset* file contains a one-hot encoded version of the dataset, specifically designed for seamless integration with statistical models and machine learning pipelines, without necessitating additional preprocessing steps.

Within the provided CSV example, Figure 4, *token* refers to the individual tokens in the corpus, *German Context* indicates the context of the text (either G-JOB or G-SCHOOL), *Religious Phrases* denotes the presence of religious phrases (annotated with RELIG), *age* and *gender* specify the respective attributes of the speaker, *speaker* identifies the speaker's anonymized alias, *interlocutor/s* denotes the interlocutor(s) in the conversation, *file* points to the file path where the token was spotted (truncated in the example), and *context* provides additional contextual information.

For instance, the token *kindarbifleega* has G-JOB as the German context, no religious phrases, a young female speaker (age and gender), identified as *A*, and an anonymized speaker-interlocutor combination of *BSH, S, SUH*. The corresponding context is *A uw huwwa yixtiṣir kindarbifleega...* .

token	German Context	Religious Phrases	Age	Gender	speaker	interlocutor/s	file	context
kindarbifleega	G-JOB		young	female	A	BSH,S,SUH	Path/2/file	A uw huwwa yixtiṣir kindarbifleega w il kraankinbifleega w il aḷtinbifleega , uw, mirtaaħa bii
kraankinbifleega	G-JOB		young	female	A	BSH,S,SUH	Path/2/file	A uw huwwa yixtiṣir kindarbifleega w il kraankinbifleega w il aḷtinbifleega , uw, mirtaaħa bii [$RELIG.l ħamdi l
aḷtinbifleega	G-JOB		young	female	A	BSH,S,SUH	Path/2/file	uw huwwa yixtiṣir kindarbifleega w il kraankinbifleega w il aḷtinbifleega , uw, mirtaaħa bii l ħamdi l laa , ʃwayya ṣaʕub,
l ħamdi l laa		RELIG	young	female	A	BSH,S,SUH	Path/2/file	w il kraankinbifleega w il aḷtinbifleega , uw, mirtaaħa bii l ħamdi l laa , ʃwayya ṣaʕub, wa laakin [$RELIG.ħamdi l
ħamdi l laah		RELIG	young	female	A	BSH,S,SUH	Path/2/file	bii l ħamdi l laa , ʃwayya ṣaʕub, wa laakin ħamdi l laah raad la diraasa w taʕab ʃwayya, [$RELIG.in
in ʃaaʔa ḷḷaah		RELIG	young	female	A	BSH,S,SUH	Path/2/file	ħamdi l laah raad la diraasa w taʕab ʃwayya, in ʃaaʔa ḷḷaah , w qabilha sawweet koors bi-zwaay , bi-aaynz uw
koors	G-SCHOOL		young	female	A	BSH,S,SUH	Path/2/file	w taʕab ʃwayya, in ʃaaʔa ḷḷaah , w qabilha sawweet koors bi-zwaay , bi-aaynz uw bi-zwaay dirasit biruufʃuula , aah, santeen, uw
bi-zwaay	G-SCHOOL		young	female	A	BSH,S,SUH	Path/2/file	taʕab ʃwayya, in ʃaaʔa ḷḷaah , w qabilha sawweet koors bi-zwaay , bi-aaynz uw bi-zwaay dirasit biruufʃuula , aah, santeen, uw ð
bi-aaynz	G-SCHOOL		young	female	A	BSH,S,SUH	Path/2/file	ʃwayya, in ʃaaʔa ḷḷaah , w qabilha sawweet koors bi-zwaay , bi-aaynz uw bi-zwaay dirasit biruufʃuula , aah, santeen, uw ðalleet r
bi-zwaay	G-SCHOOL		young	female	A	BSH,S,SUH	Path/2/file	ʃaaʔa ḷḷaah], w qabilha sawweet koors bi-zwaay , bi-aaynz uw bi-zwaay dirasit biruufʃuula , aah, santeen, uw ðalleet mistimirra y
biruufʃuula	G-SCHOOL		young	female	A	BSH,S,SUH	Path/2/file	w qabilha sawweet koors bi-zwaay , bi-aaynz uw bi-zwaay dirasit biruufʃuula , aah, santeen, uw ðalleet mistimirra yaʕni dixalt ih
koorsaat	G-SCHOOL		young	female	A	BSH,S,SUH	Path/2/file	xumuṣ ṭaʕaʃ, uw balliʃat raħlat id diraasa w il koorsaat w madrasa w il aaxri wa hassa awsbildung , w
awsbildung	G-JOB		young	female	A	BSH,S,SUH	Path/2/file	il koorsaat w madrasa w il aaxri wa hassa awsbildung , w in ʃaa ḷḷa ib alfeen tlaaθa w ʕiʃriin
in ʃaa ḷḷa		RELIG	young	female	A	BSH,S,SUH	Path/2/file	w madrasa w il aaxri wa hassa awsbildung , w in ʃaa ḷḷa ib alfeen tlaaθa w ʕiʃriin ʃahr id
in ʃaa ḷḷa		RELIG	young	female	A	BSH,S,SUH	Path/2/file	tlaaθa w ʕiʃriin ʃahr id daʕaʃ atxarraj min naa, in ʃaa ḷḷa
ħamdu li llaah		RELIG	young	male	BSH	A,S,SUH	Path/2/file	biyya aʃtuɣuḷ igullu gʕud bi l maktab bass hiič ħamdu li llaah la ma, fa faytabildung ma yṣiir, leen
faytabildung	G-JOB		young	male	BSH	A,S,SUH	Path/2/file	maktab bass hiič ħamdu li llaah la ma, fa faytabildung ma yṣiir, leen aani ma ʕindi awsbildung ma ʕindi
awsbildung	G-JOB		young	male	BSH	A,S,SUH	Path/2/file	ma, fa faytabildung ma yṣiir, leen aani ma ʕindi awsbildung ma ʕindi ma ʕindi biriif ihnaa gittilhum aani ʕindi
biriif	G-JOB		young	male	BSH	A,S,SUH	Path/2/file	leen aani ma ʕindi awsbildung ma ʕindi ma ʕindi biriif ihnaa gittilhum aani ʕindi jaamiʕa maθalan kaða w ij
waḷḷa		RELIG	old	female	S	A,BSH,SUH	Path/2/file	ma biiha ʕumur, bi đ đabuṭ, aani agullu lhum waḷḷa marraat igulluu li yaḷḷa ṃaaṃa dursi da ħaawli ti,
waḷḷa		RELIG	old	female	S	A,BSH,SUH	Path/2/file	yaʕni la tinsiiha, aa, yaʕni ħatta gitti lhum ii waḷḷa aani mfakkira innu aani aaxið il bi-aaynz uw baʕdeen
bi-aaynz	G-SCHOOL		old	female	S	A,BSH,SUH	Path/2/file	lhum ii waḷḷa aani mfakkira innu aani aaxið il bi-aaynz uw baʕdeen aruuħ asawwi oosbildung
oosbildung	G-JOB		old	female	S	A,BSH,SUH	Path/2/file	innu aani aaxið il bi-aaynz uw baʕdeen aruuħ asawwi oosbildung
waḷḷaahi		RELIG	old	female	SUH	A,BSH,S	Path/2/file	SUH haay il qiṣṣa uw ma biiha eh ħaaliyan waḷḷaahi aani činit aʃtiɣiḷ ferkawfarin baay (...) b bekeraay ib
ferkawfarin	G-JOB		old	female	SUH	A,BSH,S	Path/2/file	uw ma biiha eh ħaaliyan waḷḷaahi aani činit aʃtiɣiḷ ferkawfarin baay (...) b bekeraay ib erlangin miθil ma gittlič,
baay	G-JOB		old	female	SUH	A,BSH,S	Path/2/file	ma biiha eh ħaaliyan waḷḷaahi aani činit aʃtiɣiḷ ferkawfarin baay (...) b bekeraay ib erlangin miθil ma gittlič, bass,
bekeraay	G-JOB		old	female	SUH	A,BSH,S	Path/2/file	ħaaliyan waḷḷaahi aani činit aʃtiɣiḷ ferkawfarin baay (...) b bekeraay ib erlangin miθil ma gittlič, bass, aa, ijatti koroona
ma ʃaaʔa ḷḷaah		RELIG	old	female	SUH	A,BSH,S	Path/2/file	ib erlangin miθil ma gittlič, bass, aa, ijatti koroona ma ʃaaʔa ḷḷaah ṭayyirat il awwal w it taali, fa,

Figure 4: Dataset file generated from example corpus in Figure 2.

token	annotated	not annotated	not_annotated_interest	A not annotated	A annotated	BSH not annotated	BSH annotated	S not annotated	S annotated	SUH not annotated	SUH annotated	total
kindarbifleega	1	0	0	0	1	0	0	0	0	0	0	1
kraankinbifleega	1	0	0	0	1	0	0	0	0	0	0	1
aḷtinbifleega	1	0	0	0	1	0	0	0	0	0	0	1
l ħamdi l laa	1	0	0	0	1	0	0	0	0	0	0	1
ħamdi l laah	1	0	0	0	1	0	0	0	0	0	0	1
in ʃaaʔa ḷḷaah	1	0	0	0	1	0	0	0	0	0	0	1
koors	1	0	0	0	1	0	0	0	0	0	0	1
bi-zwaay	2	0	0	0	2	0	0	0	0	0	0	2
bi-aaynz	2	0	0	0	1	0	0	0	1	0	0	2
biruufʃuula	1	0	0	0	1	0	0	0	0	0	0	1
koorsaat	1	0	0	0	1	0	0	0	0	0	0	1
awsbildung	2	0	0	0	1	0	1	0	0	0	0	2
in ʃaa ḷḷa	2	0	0	0	2	0	0	0	0	0	0	2
ħamdu li llaah	1	0	0	0	0	0	1	0	0	0	0	1
faytabildung	1	0	0	0	0	0	1	0	0	0	0	1
biriif	1	0	0	0	0	0	1	0	0	0	0	1
waḷḷa	2	0	0	0	0	0	0	0	2	0	0	2
oosbildung	1	0	0	0	0	0	0	0	1	0	0	1
waḷḷaahi	1	0	0	0	0	0	0	0	0	0	1	1
ferkawfarin	1	0	0	0	0	0	0	0	0	0	1	1
baay	1	0	0	0	0	0	0	0	0	0	1	1
bekeraay	1	0	0	0	0	0	0	0	0	0	1	1
ma ʃaaʔa ḷḷaah	1	0	0	0	0	0	0	0	0	0	1	1

Figure 5: Annotation information file generated from example corpus in Figure 2.

Annotation Information The *annotation_info* file contains information about the annotations included in the *dataset* file, such as the token itself, the number of times it appears in the dataset, and the number of times it appears for each speaker, Figure 5.

Missed Annotations The *missed_annotations* file tracks tokens that were previously annotated but not consistently annotated in subsequent instances. It contains the token and its context, determined by a user-defined n-gram size. Figure 6 reports an example taken from the from corpus in Section 4.1.

File	token	context 1	context 2	context 3	context 4
Path/2/file	haaði	huwwa yaaxuð malaabis iħađđirhin uw haaði, ʃaayilhin, aani aruuħ [$IA.yam] aa jiddu uw biibi			
Path/2/file	kinna	la ma [$GQ.agʕud] ma ʕindi ixtiḷaaṭaat min kinna bi l ʕiraaq yaʕni, ma aħibb li [$DEM-HA.hal] yoom			
Path/2/file	gaam	aani ṣaarat [$IA.aku] , waaħid ʃaxiṣ ihnaana, gaam [$CK.yiħči] uw [$DEM-I-END.haaði]	ʕad, leeʃ ii, leeʃ aani [$CK.aħči] aani b ṣaraaħa, gaam [$CK.yiħči] uw [$DEM-I-END.haaði], fa zaʕalit		
Path/2/file	maʕa	uw zaʕal uw [$GQ.gittla] zaʕal zaʕal, aani maʕa l ʕilim ahaa, aani m aʕurfa yaʕni uw [$GQ.galloola]	ma ṣaarat, wa la, wa la sumaʕit yaʕni, maʕa l ʕilim aani... ey, ey, [$IA.zeena] musaalima uw	-SCHOOL.koors] kulla ma axðaw [$G.G-GER.bii-aaynz] maʕa l ʕilim banaat [$IA.WIYYA.wiyyaana] Saarlhum tlaθ	
Path/2/file	mu	li iħna inħibbha yaʕni daaʔiman ey waḷḷa ii, mu l kull, ii (laughing) ṣaħiiħ, bass ihna, ihna	[$IA.WIYYA.wiyya] kull in naas, il ʃaraaʔiħ il muʃtamaʕ akθar, kull ʃii bii bii ħaʃaakum	ʃii bii bii ħaʃaakum iz [$IA.zeen] uw [$IA.aku] l mu [$IA.zeen] w il, ee yaʕni ixt... bass aani aʕruf i	xallii iħna b ɣaaba, inxallii yixtiliṭ ib, yaʕni b muʃtamaʕ uw inti ma tʕurfiin, il [$DEM-HAAY.haay] il

Figure 6: Subset of missed annotation file generated from corpus in Section 4.1.

variable	speaker	context	token	tag
KQ	SUH	l leel sawweetha laffeetha uu θaani yoom gaʕadit min wakit iṣ ṣubiħ, ðabbeetha ʕa naar axaðitha ħaare ilhum axaðit	wakit	[$KQ.wakit]
QK	A	hamm ib nafs il waqit innu ma ansa luɣati l ʕarabiyya, wiya , ey mumkin	waqit	[$QK.waqit]
QK	BSH	li l kull, wugaf bass ili, leen aani ʕindi waqit maħduud,	waqit	[$QK.waqit]
QK	DUN	awakkilhum uu aɣayyirilhum, aaxuð waqit akθar waqti wiyya binti , adarrisha uu ahtamm biiha uu anayyimha	waqti	[$QK.waqti]
QK	DUN	beet hammeen ey, ikuun waqti malyaan, ey	waqti	[$QK.waqti]
RAISE	A	baʕad santeen uu atxarraj bass aani haaða l qisim ħabbeeta uu huwwa jidiid	atxarraj	[$RAISE.atxarraj]
GA	A	iħna gaaʕdiin il diktoora gaaʕda tiʃraħ	gaaʕda tiʃraħ	[$GA.gaaʕda tiʃraħ]
SUF-NO-H	A-G	hiyya il killa iṭ ṭarafeen	killa	[$SUF-NO-H.killa]
D-DH	S-G	daaʔiman ifaadaat bi ḍ diwal fa kill ma yruuħ yaaxudna maʕaa b ayy dawla fa itʔaqlamna	yaaxudna	[$D-DH.OTH-DIAL.yaaxudna]

Figure 7: Subset of unknown variables file generated from corpus in Section 4.1.

Unknown Variables The *unk_variables* file is designed point to a list of variables that were not specified in the JSON file. This file includes information about the speakers, the context, and the file it was taken from, making it easy for researchers to identify and correct any inconsistencies in the dataset, Figure 7.

Descriptive Statistics Knowing basic descriptive statistics is fundamental in language research. The *corpus_stats* JSON file provides an overview of the corpus by reporting key statistics. The file contains four types of information: (i) word-related information such as the number of paragraphs[8], words, and characters; (ii) variable information, including the number of dependent and independent variables and their values; (iii) speaker-wise information, such as the total number of speakers, speakers of interest, and words spoken per speaker; and (iv) annotation-wise information, such as the number of unique annotations and annotated tokens, see Table 2.

C. Corpus

The study is based on 20 sociolinguistic individual interviews (circa 60 minutes each) conducted in Bayreuth and Nuremberg, located in Bavaria, Germany. Additionally, two group conversations were recorded with the same speakers (90 minutes per interview), where four Iraqi and Syrian speakers were paired together.

The individuals in each group come from the same dialect area and almost all come from the same circle of friends/family. In order to minimize possible influences on the interview conversations, the interviews were conducted by two assistants who are native speakers of the respective varieties.

Since the sociolinguistic interview is used as a basic tool in the study of sociolinguistic variation and it is the most common method for collecting sociolinguistic data [25], the research data were collected using this method. The goal was to move from general and impersonal questions to more specific and personal questions. Questions on selected topics encouraged respondents to narratively talk about their personal experiences (e.g., life in Germany/home country, refugee experience, friends/family, fears and concerns). Thus, the speaker's natural language could be elicited [26]. After data collection, phonetic transcription of the recordings was performed using the transcription program *Praat* [27]. Demographic information, as well as details about the respondents' backgrounds and environments, were also collected. In addition, questionnaires were employed to gather data on the interviewees' language contact behavior with speakers of other languages and language varieties.

[8] In our example, each paragraph is a turn-taking component.

Table 2
Corpus statistics file generated from example corpus in Figure 2.


```
"paragraphs": 4,
"speakers_of_interest": 4,
"all_speakers": 4,
"dependent_variables": 2,
"independent_variables": 6,
"variables": 8,
"variables_values": 18,
"dependent_variable_values": 10,
"independent_variable_values": 8,
```

```
"words": 333,
"characters": 1897,
"unique_annotations": 22,
"annotated_tokens": 26,
"speaker_num_words": {
    "A": 107,
    "SUH": 75,
    "S": 81,
    "BSH": 90
    }
```

Harnessing Il Manifesto Newspaper Archive for Knowledge Base Creation: Techniques and Findings in the MeMa Project

Robert J. Alexander[1], Matteo Bartocci[2], Oriana Persico[3] and Guido Vetere[1,4]

[1]*Isagog S.R.L., Via Faà di Bruno 54, 00195 Roma, Italy*
[2]*Il Manifesto Soc. Coop., Via Angelo Bargoni 8, 00153 Roma, Italy*
[3]*Human Ecosystems Relazioni S.R.L, via Umberto Guarnieri 15, 00177 Roma, Italy*
[4]*Università Guglielmo Marconi, Via Plinio 44, 00193 Roma, Italy*

Abstract
English.The historical archive of the newspaper "il Manifesto" is a valuable asset protected by the Italian Ministry of Cultural Heritage. The *MeMa* project aims to create an "intelligent archive" using AI principles, fostering collaboration and transparency. The platform, built around Apache Jena and open linguistic technologies, addresses the newspaper community's specific needs. This paper presents the platform's architecture, knowledge base construction process, and future directions, emphasizing journalism enhancements through AI while respecting "Il Manifesto"'s principles. **Italiano.**L'archivio storico del quotidiano "il Manifesto" è tutelato dal Ministero dei Beni Culturali. Il progetto *MeMa* mira a creare un "archivio intelligente" basato su una intelligenza artificiale che favorisce la collaborazione e la trasparenza. La piattaforma, costruita attorno ad Apache Jena e tecnologie linguistiche aperte, risponde alle esigenze specifiche della comunità del giornale. Questo contributo presenta l'architettura della piattaforma, il processo di costruzione della base di conoscenza e le direzioni future, discutendo il potenziamento del giornalismo attraverso l'intelligenza artificiale nel rispetto dei principi de "Il Manifesto".

Keywords
AI in journalism, Open linguistic technologies, Knowledge graphs, Newspaper community

1. Introduction

The historical archive of the newspaper "il Manifesto" is an asset protected by the Italian Ministry of Cultural Heritage as of particular interest [1]. The archive includes a paper collection starting from 1971, and a digitized collection starting from the 1990s. The resource is now entrusted to the "Nuovo Manifesto Società Cooperativa Editrice", which publishes the newspaper and its digital editions since 2013. The cooperative is committed to maintain and improve the archive, as well as to guarantee free access and digital consultation facilities to anyone interested in it [2]. The digital archive, produced in different phases over the years, reflects the historical and technological evolution of the publishing sector. The database initially included 10,013 digitized files containing about 160,000 articles, with few gaps in the years 1985-1986 and 1994-2002. Il Manifesto considers an "intelligent archive" to be the cornerstone of its digital strategy, and for this reason seeks to align it with new technologies with appropriate investments in research and development. The *MeMa* (Memoria Manifesta) project started in 2020 by a partnership with Salvatore Iaconesi [3] and Oriana Persico, with the aim of developing new archive infrastructure based on Artificial Intelligence. This would be a "Community AI" [1] based on the principles of openness, transparency, collaboration and non-extractiveness, thus being able to establish productive relationships between the archive, the editorial staff, the user communities and society in general [2].
When, in 2023, the project was resumed, the new board decided to continue the original plan by making it evolve in the direction of Linked Open Data, and taking advantage of the latest advances in language and knowledge technologies. The idea was to build a standards-based Knowledge Graph (KG) using editorial metadata and structured information extracted from article text. By itself, this idea is by no means new [3] [4] [5]. Also, there are commercial platforms that have been offering solutions for the newspaper industry some years now, such as Neo4j [6] or Ontotext [7]. However, we realized that the success of the project depended significantly on how the platform would adapt to the way content is produced, extracted, organised, enriched and experienced by the professional and user communities gathered around

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
bob@isagog.com (R. J. Alexander); bartocci@ilmanifesto.it (M. Bartocci); oriana.persico@he-r.it (O. Persico); g.vetere@isagog.com (G. Vetere)
0000-0002-6703-7276 (G. Vetere)

[1]Legislative Decree 42/2004 with provision of the Regional Director for the Cultural Heritage of Lazio (24/2013, 12 March 2013)
[2]https://archiviopubblico.ilmanifesto.it/
[3]Salvatore Iaconesi (Livorno 1973, Reggio Calabria 2022) has been an engineer, artist, hacker and interaction designer

the newspaper. Rather than forcing these habits to an out-of-the-box commercial platform, we opted to tailor a specific solution. Moreover, as a sociotechnical platform, *MeMa* should be open to user curation and contribution (e.g. from readers, archivists, and journalists), collaboratively contributing to the evolution of the AI, including correcting the inevitable errors of current NLP technologies. Hence, we started designing a custom platform around a core open graph database, namely Apache Jena [4] and a selection of open linguistic technologies suitable for the Italian language. The solution falls into the broad area of Enterprise Knowledge Graphs [8] which are gaining momentum as "rational counterparts" of generative linguistic technologies based on neural models [9]. This work is a first account of what emerged in the first months of analysis, design and development of the solution, and a discussion of our plans to meet the socio-technical requirements we have analyzed so far. Our contribution is a "reality check" of the use of knowledge and language technologies applied to complex texts produced by an Italian publishing community over more than 40 years of work. In general, our research concerns the interaction between digital systems and human beings to make their contents fully transparent and accessible to different user communities. From a linguistic point of view, relevant aspects include the specificity of the texts produced over a wide period of time, characterized by a specific idiolect but also by diachronic variations.

This paper is organized as follows. In Section 2, we present an architectural overview of the platform under development. Section 3 delves into the process of constructing the knowledge base, detailing the steps involved in gathering and organizing the relevant information. In Section 4, we discuss challenges and ideas about the future directions. Note that automatic content generation is not included in the journalism enhancements driven by AI, as intended by "Il Manifesto".

2. System Overview

MeMa's software architecture comprises several components that work together to handle a graph database with indexed attributes, enabling efficient ingestion, analysis, and semantic querying. The key components of this architecture include:

1. Knowledge Graph: The core of the system is a graph database of the RDF (Resource Description Framework) family with inference capabilities, based on Apache Jena, the Pellet OWL reasoner, the search engine Lucene, and custom components, where a number of KG attributes are indexed and embedded to optimize search and retrieval operations.
2. NLP Service: A REST service that provides an abstraction layer over various NLP functionalities to support the system's operations. It wraps capabilities such as text analysis, entity recognition, topic analysis, semantic similarity, and other NLP tasks based on open source transformers [10]. This service collaborates with the ingestion process to extract valuable insights from the content being ingested.
3. Ingestion Processor: A batch process that is responsible for ingesting content into the KG. This process integrates different sources, analyzes texts to extract relevant information using the NLP service, and produces RDF sources to feed the KG according to the *MeMa* ontology.
4. Query and Update Service: A REST service that is responsible for handling queries and update operations on the KG. It integrates similarity searches and SPARQL queries to retrieve relevant graph entities. This service leverages the indexed attributes to optimize query performance and speed up retrieval operations, and the NLP Service to transform user's queries and evaluate response ranking.

This software architecture employs a services and API-based approach, enabling functional evolution, flexible deployment, and seamless scalability. The service architecture is an abstraction of a general functionality that can be applied to a variety of scenarios. Based on this design, we have developed custom application services that can be used in a front-end designed for the editorial staff of the newspaper.

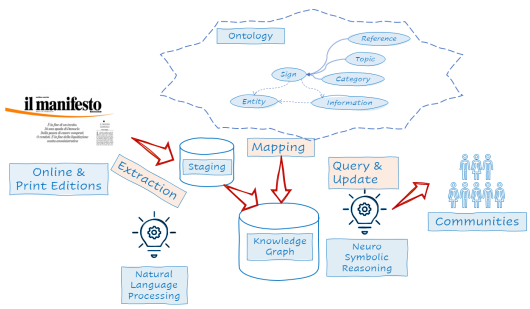


Figure 1: Architectural overview

"Il Manifesto" has a print edition and an online edition, each managed by its own Content Management System (CMS). The two editions largely coincide, however each one may contain articles not present in the other. As a result, the same article (with slight variations) may be available in two different repositories. When consolidating all editorial content into one Knowledge Base, we had to harmonize and integrate the contents from both CMSs.

[4]https://jena.apache.org/

3. The Knowledge Base

Modeling editorial content in a KG requires the adoption of a suitable ontology. Although editorial content modeling has already been studied and tested [11], we did not identify a simple, well-established model that suited our needs. In particular, we aimed to represent how agents interpret specific tokens as referring to entities based on established conventions or procedures. In other words, we were interested in semiotics. At the best of our knowledge, even comprehensive conceptualizations, like the CIDOC Conceptual Reference Model [12], which include linguistic and symbolic objects, do not provide modeling primitives to represent interpretation processes. This is why we decided to develop our own conceptualization, which we will illustrate in the following section. Mappings to existing conceptual frameworks, such as schema.org[5], are preserved as annotations.

3.1. The *MeMa* Ontology

The *MeMa* ontology focuses on the way entities are mentioned, rather than on the characterization of those entities, which is mostly left to external sources. As such, the *MeMa* ontology adopts a semiotic perspective [13] in the line of [14] and [15]. The structure of our ontology is sketched as follows:

- Class: Sign
  An immaterial entity that stands to someone (or something) for some other entity as the outcome of an interpretation
  - Subclass: Category
    A sign standing for a class of entities
  - Subclass: Reference
    A sign standing for a single (even collective) entity
  - Subclass: Topic
    A sign standing for a focus of interest in a larger context
- Class: Information
  An immaterial thing that conveys interconnected signs
  - Subclass: Text
    A textual information object
  - Subclass: Sentence
    Part of a text
  - Subclass: Token
    Part of a sentence
- Class: Entity
  A spatio-temporal thing
  - Subclass: Agent
    An entity that has the capacity to initiate or perform actions
  - Subclass: Location
    An identified portion of space
  - Subclass: Event
    An entity that unfolds in time
  - Subclass: Object
    An entity that unfolds in space

A key feature of this ontology is the distinction of Reference and Token, where the latter instantiates the former [6]. As a Sign, a Reference is based on an interpretation process, whether human or automated, e.g., for *DBpedia Spotlight*, interpreting the string "Aristotle" as the name of the philosopher from Stagira. Sign instances support properties (*interpretation records*) that keep track of these processes. A Token, on the other hand, is a portion of Text, e.g. the string "Aristotle" that appears in a document at a given offset, which may trigger the processes mentioned above. In this way, the semantic qualification of the text is provided with the means to trace the underlying interpretation, be it automatic or human. This is essential for ensuring the traceability and accountability of the knowledge base's content.

3.2. Handling Metadata

Extracting knowledge from newspaper articles essentially consists of working on the both metadata and text in a consistent way. This process has currently generated about 650.000 stored articles and grows roughly by 1000 new articles a month.

[5] https://schema.org/

[6] This aligns with Peirce's distinction of *type* and *token*

According to our ontology, assertions about articles are based on two types of properties, which we call *editorial* and *semantic*. The former includes attributes such as *publication date* or *author*, the latter are generically intended to characterize the content, including standard categorization (*sports*, *business*, etc.), references to people, places and other named entities, and arbitrary classifiers which are typically encoded in freely invented wording. However, this distinction is neither fully aligned with the structure of the legacy metadata schemes, nor fully reflected in how metadata are actually produced. For historical and organizational reasons, in fact, the online and print editions are metadated separately, with different schemes and guidelines. Looking into it, we realized that integrating them could not be done by simply mapping schemes to our ontology, but instead required a thoughtful analysis of the actual data. We carried out qualitative and quantitative analyses which led us to devise an adequate treatment of the metadata content. Here is a summary of the historical archive scheme:

- **ARGOMENTO** (subject) is fed with labels with no semantic relationship amongst them. The raw count for these labels is 792.000 with **4023 distinguished values (0.51%)**, which comprise synonyms, typos, abbreviations, and other variants.
- **CATEGORIA** (category) field, on the other hand, is used with a prevalence of editorial tags (*front page*, *editorial* etc) but again we often encounter values that also belong to the ARGOMENTO field. The raw count usage for CATEGORIA is 828.805, with **1358 different values (0.16%)**, which also comprise synonyms, typos, abbreviations, and other variants.
- **LOCALITA** (location) accommodates editor's or archivist description of what geopolitical entities are involved. They might not be mentioned literally in the article. We observed redundant tagging where many broader geopolitical concepts, which could be inferred, are explicitly stated somewhat arbitrarily (e.g., CUTRO, CR, Italia). Whenever we successfully link a geopolitical mention to GeoNames, this redundancy becomes unnecessary, as GeoNames allows for full hierarchical navigation.
- **RIFERIMENTI** (references) is used as a placeholder for a variety of annotations, which also overlap other fields. Most often, these are short summaries which should facilitate keyword based retrieval. We currently count 949248 occurrences of these annotations, **679760 of which are unique (71,6%)**, thus qualifying by far as the most informative facet.

Overall, the frequency distribution of all these properties exhibits long tails with low frequencies typical of a lack of annotation guidelines and tools. In particular, the RIFERIMENTI field appears to be very heterogeneous, as it mixes editorial tags (e.g. *breve*, *cronaca*), named entities and content summaries. As a result of this analysis, we decided to ignore the formal meaning (if any) of the legacy metadata schema and instead focus on the annotation content. In particular, with respect to our ontology, we want to distinguish among *classifiers* (Sign) and *descriptions* (Information). To this end, we use:

- Two handcrafted tagsets, for editorial marks and standard topics respectively, obtained by clearing and deduplicating the contents of ARGOMENTO, CATEGORIA and the most recurrent RIFERIMENTI
- A lemmatizer for out of tagset values
- A rule-based classifier for multi-word RIFERIMENTI values, which discriminates *descriptions* from multi-word topics

Classifiers are instantiated as either as Category or Topic, and suitably linked to the article, while descriptive summaries are kept as data properties, whose content is indexed. We plan to add a vector representation of summaries to include them in semantic similarity searches and/or clustering.

3.3. Knowledge Extraction

Besides annotated metadata, *MeMa* analyzes the full article text. At the current stage, we only perform entity recognition and linking. There are no limits to the kind of entities that can be mentioned in a newspaper article. However, there are limits to the kinds that can be efficiently retrieved by standard NLP pipelines. One of the richest known inventories [16], includes up to 18 categories, but as a matter of facts the available recognizers for the Italian language, e.g. Spacy [17] and Stanza [18] are limited to just a few of them, such as **PER**(son), **LOC**(alization), and **ORG**(anization). We currently use a combination of Stanford's Stanza [18] (in particular: `tokenize, mwt, pos, lemma, depparse`, and `ner` processors), DBPedia Spotlight [19], GeoNames [7], along with a number of custom processing functions. We choose Stanza because of the state-of-the-art performances on Italian benchmarks[8]. We evaluated the NER performance on our sources by randomly choosing 30 articles, manually annotating their content, and matching the pipeline outcome. Results presented in Table 2 align with the current state of the art [20].

For the **PER** class we also adopt a simple co-referencing matching based on the fact that within an article we mostly find a fully named instance of the person and subsequently only the first or last names. Along with the span, we therefore generate a Person co-refernce ID. We then proceed to the grounding attempt against the DBpedia API which we invoke via its Spotlight function. We have found no added precision/recall by giving it more textual context. For both the grounded and the

[7] https://www.geonames.org/

[8] Stanza's performance on NER Corpora https://stanfordnlp.github.io/stanza/ner_models.html

annotation	occurrences
breve (*short*)	5324
cronaca (*news*)	1860
analisi (*analisys*)	901
programma (*program*)	732
scheda (*form*)	691
crisi (*crisis*)	688
scenario (*scenario*)	671
le lettere di oggi (*today's letters*)	662
storia (*history*)	648
ritratto (*portrait*)	575
campagna elettorale (*election campaign*)	564
reazioni (*reactions*)	544
famiglia incertezza e preocupazioni (sic) (*family uncertainty and worries*)	1
oggi sciopero marcia globale per il clima (*global climate march strike today*)	1
giorgio forti, alessandro stoppoloni, christian picucci (proper names)	1

Table 1
An excerpt of both recurrent and unique values of RIFERIMENTI

Type	Precision	Recall	F1 Score
PER	0.9117	0.9612	0.9280
LOC	0.9194	0.8703	0.8763
ORG	0.8071	0.8213	0.7847
Overall	0.8816	0.8868	0.8657

Table 2
Average Precision, Recall, and F1 Score per Type and Overall

ungrounded **PER**sons, we then store the span of surface, a fuzzy score of the match with DBpedia's entity to accommodate typos and variations which are especially common with the Italian rendition of foreign names and the reference to the current article. We therefore have the spans where the surface of the person was mentioned and the grounded/ungrounded reference to the article in a separate collection. A similar process is performed for the **LOC**ation named entities against the GeoNames resource. Linking to the GeoNames resource gives us a wealth of added information amongst which geolocalization and administrative and geographical data. Also for **LOC** we store the spans within the article's and the mentions in their dedicated collection. We also tried using DBpedia Spotlight for **ORG**anizations but the results were not satisfactory. One of the causes may be the lack of precision at the **NER** stage. Also, there are often false positive groundings given that there are several organizations with namesakes or placenames. We didn't conduct a comprehensive analysis of the entity linking performance; however, an initial examination revealed that roughly 10% of the total links were incorrect. Finally, the last stages of our pipeline transforms the staging data into corresponding RDF data (Turtle format). We therefore generate article individuals with metadata from both the historical and the digital corpora leveraging the reconciliation when possible and we also generate individuals, topics and all of their cross-linked mentions. The resulting knowledge base is currently expressed with approximately 12.5 million triples, and loaded into Apache Jena Fuseki to be used as a SPARQL endpoint.

4. Challenges and Ideas

Newspaper articles pose several interpretative challenges [21]. The reporting of events, with their participants and their contextual characterization, are the most relevant parts of their content. Metonymy, regular polysemy and presupposition, even combined, stand out as prominent linguistic phenomena. Take for instance the headline: "Di Maio al Colle, ma non da Mattarella" (≈*"Di Maio at the Colle, but not meeting with Mattarella"*) [9]. "Di Maio" and "Mattarella" can be plainly identified as person mentions and linked to their corresponding individuals (Italian politicians). But what about "Colle"? Even if it were identified as a place (the Quirinal hill in Rome) it is clear that, contextually, the token intends to signify the institutional function of the presidency of the Italian Republic. Also, the people mentioned in the sentence represent their public roles at the time the article was written, rather than any identified human being. This kind of metonymic use of language makes classification of named entities more difficult [22]. As for the news

[9]https://ilmanifesto.it/di-maio-al-colle-ma-non-da-mattarella

in question, note that apparently there is no mention of any event, but presumably something happened. Event mining is also a long-standing challenge of NLP, as well as reasoning about implicature and presupposition [23]. These tasks are usually approached with ML methods [24]. In particular, supervised learning strategies have been implemented in recent years, but they are limited in that they require specific annotated corpora and training procedures [25].
The use of Large Language Models (LLM) with generative transformers are now regarded to as the way forward for the challenges mentioned above [26], so we plan to replace part of the existing NLP machinery with a suitable implementation of these techniques. At present, we are piloting publicly available services. For instance, if suitably promped, GPT4 [27] produces the following analysis of the headline mentioned d above:

1. Di Maio: *This refers to a person, likely Luigi Di Maio, an Italian politician. He could be the subject of the discussion or the context in which the sentence is used.*
2. at the Colle: *Refers to "Colle" or "Quirinale", which is the official residence of the President of the Italian Republic, located in Rome. Therefore, "Di Maio at the Colle" indicates that Luigi Di Maio is at or going to the Quirinale Palace.*
3. Mattarella: *Refers to Sergio Mattarella, who was the President of the Italian Republic as of September 2021. He might be mentioned because he is the incumbent president at the time the sentence is spoken.*

Llama2 [28] gives similar results [10]:

1. Di Maio: *This refers to Luigi Di Maio, an Italian politician who served as the Minister of Foreign Affairs and International Cooperation from 2018 to 2019.*
2. Colle: *This is a reference to the Quirinal Hill, a hill in Rome, Italy, where the President of Italy resides.*
3. Mattarella: *This refers to Sergio Mattarella, the President of Italy from 2015 to 2022.*

In both cases, entities are correctly identified and connected to relevant background knowledge, where their respective professional role are also highlighted. When it comes to implicatures, GPT4 is pretty inventive:

So, the sentence could mean that Luigi Di Maio is going to or present at the Quirinale, but he is not receiving instructions or direction directly from Sergio Mattarella. It could be used in a political or governmental context to express a situation where Di Maio is acting independently of the President of the Republic.

Llama2 seems to be less imaginative:

Therefore, the entities mentioned in the phrase are two politicians (Luigi Di Maio and Sergio Mattarella) and a geographic location (Quirinal Hill)

These examples show how, using LLMs appropriately, events can also be found in nominal constructions (such as the headline in question), and their participants, along with some other contextual element, can be reliably identified even with little superficial evidence. The LLMs generative ability of "connecting the dots" seem to be particularly effective when dealing with journalistic jargon, which is actually full of elliptical constructions. As for lexical units other than entities and events, framing complex notions such as *not receiving instructions* in a Knowledge Graph may raise ontological challenges, e.g. in this case that of representing negative facts. The "ontological cut-of" operated in the design phase, i.e. the way in which linguistic and logical (conceptual) expressiveness is arranged, plays here a crucial role. Our ontology is such that only basic patterns (e.g. *participation in action*) are ingested into the KG as logic assertions (i.e. triples), while blurry concepts (e.g. *receiving instructions*) are kept at the lexical level. Lexical concepts can be mapped to onto-lexical resources and interleaved by semantic relationships, as well as associated to distributional embeddings. In any case, the "ontological cut-of" requires the division of KG's reasoning into logical and linguistic inference procedures and the integration of their results, which is at the core of our future developments. The current prototype does not include semantic relationships and deep linguistic inference, but we do evaluate semantic similarity based on embeddings of textual fragments (e.g. headlines and summaries), e.g. when re-ranking KG queries results.
To improve knowledge extraction, we are in the process of experimenting LLMs generative models. It is already clear, however, that for giant models available only through remote services, such as those of the OpenAI family, the feasibility of these experiments could be problematic, since the stability of their behaviour seems to be questionable [29]. Also, the use of remote services would not comply with Il Manifesto's digital strategy, due to unwanted bindings to external business entities. Therefore, we are focusing on the use of on-premise open LLMs, trading some functionality for dependability, freedom, control, and cost effectiveness. At the time of writing, although the use of open models such as LLama2 seems promising, we have identified some hallucinations, for example the person "Matteo Meloni", erroneously identified as reference for "Meloni" in the context of "governo Meloni", who looks like a disturbing hybridization of the current Italian Prime Minister and his Deputy. How to deal with invented entities and fancy judgments is a general concern for the productive use of these new NLP methods.Our approach will be to involve editors, archivists and readers in reviewing and amending AI results.

[10] We are using the 13B parameters deployed on a virtual host

5. Conclusion

The construction of *MeMa*'s KG is an opportunity to discuss the state of the art perspective of NLP in the context of a real Italian content production environment. The KG will be made available later this year through a SPARQL endpoint and a dataset collection. At the current stage, our experience shows the potential, but also the limits, of NLP technologies applied to a large corpus of newspaper articles extended over a relevant time interval, which are characterized by a sophisticated use of the Italian language. In general, structured knowledge extraction can be achieved with various levels of granularity by integrating NLP processors, such as named entities recognizers, event recognizers and role labelers, keyword and topic extractors. Pre-trained multilingual LLM-based generative transformers will probably replace the supervised methods that have dominated the technology of these processors the last decade, considerably easing the task of extracting qualified semantic information. However, the new neural technologies do not seem free from errors, mainly due to the kind of inventive linguistic generation that may produce. Giving the user community the ability to "educate" AI, i.e. monitor and correct its results, remains the main route for us. Transparent logical structures such as Knowledge Graphs offer the best support for this type of activity. How information automatically extracted from text can be conceptualized and critically scrutinized by user communities will have a profound impact on the harmonization of AI in human ecosystems.

References

[1] S. Iaconesi, O. Persico, When my child is ai: Learning and experiencing through ai outside the school - the experiences of a community ai, QTimes Journal of Education Anno XIII (2021).

[2] S. Iaconesi, The illustrated principles of nuovo abitare, Medium, 2021. Https://medium.com/@salvatoreiaconesi/the-illustrated-principles-of-nuovo-abitare-f5c0af6d69a5.

[3] M. Lakshika, H. Caldera, Knowledge graphs representation for event-related e-news articles, Machine Learning and Knowledge Extraction 3 (2021) 802–818. URL: https://www.mdpi.com/2504-4990/3/4/40. doi:10.3390/make3040040.

[4] A. L. Opdahl, T. Al-Moslmi, D.-T. Dang-Nguyen, M. Gallofré Ocaña, B. Tessem, C. Veres, Semantic knowledge graphs for the news: A review, ACM Comput. Surv. 55 (2022). URL: https://doi.org/10.1145/3543508. doi:10.1145/3543508.

[5] A. Berven, O. A. Christensen, S. Moldeklev, A. L. Opdahl, K. J. Villanger, A knowledge-graph platform for newsrooms, Computers in Industry 123 (2020) 103321. URL: https://www.sciencedirect.com/science/article/pii/S0166361520305558. doi:https://doi.org/10.1016/j.compind.2020.103321.

[6] T. Bratanic, Making sense of news, the knowledge graph way, Neo4j Developer Blog, 2021. Published online on Feb 2, 2021.

[7] M. Yankova, Journalism in the age of open data, Ontotext Blog, 2016. Last accessed on July 2023.

[8] J. Z. Pan, G. Vetere, J. M. Gomez-Perez, H. Wu (Eds.), Exploiting Linked Data and Knowledge Graphs in Large Organisations, Springer, Cham, 2017. doi:10.1007/978-3-319-45654-6.

[9] P. Hitzler, M. K. Sarker, Neuro-symbolic artificial intelligence: The state of the art (2022).

[10] Hugging face, the ai community building the future, 2023. Https://huggingface.co/ (Accessed: 2023-07-20).

[11] N. Fernández, D. Fuentes, L. Sánchez, J. A. Fisteus, The news ontology: Design and applications, Expert Systems with Applications 37 (2010) 8694–8704. URL: https://www.sciencedirect.com/science/article/pii/S0957417410005592. doi:https://doi.org/10.1016/j.eswa.2010.06.055.

[12] C. Bekiari, G. Bruseker, M. Doerr, C.-E. Ore, S. Stead, A. Velios, Definition of the CIDOC Conceptual Reference Model v7.1.1, The CIDOC Conceptual Reference Model Special Interest Group, 2021. Release Date: June 9, 2021.

[13] C. S. Peirce, Collected Papers of Charles Sanders Peirce, Harvard University Press, Cambridge, MA, 1931-1958.

[14] J. Sowa, Ontology, metadata, and semiotics, Conceptual Structures: Logical, Linguistic, and Computational Issues (2000) 55–81.

[15] A. Gangemi, semiotics.owl: A content ontology pattern that encodes a basic semiotic theory, Linked Open Vocabularies, 2007. Https://lov.linkeddata.es/dataset/lov/vocabs/semiotics.

[16] R. Weischedel, S. Pradhan, L. Ramshaw, M. Palmer, N. Xue, M. Marcus, E. Hovy, R. Belvin, R. MacIntyre, R. Grishman, et al., Ontonotes release 5.0 ldc2013t19, in: Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC), 2011, pp. 2525–2530. Https://catalog.ldc.upenn.edu/LDC2013T19.

[17] M. Honnibal, I. Montani, spacy 2: Industrial-strength natural language processing in python, 2020. URL: https://spacy.io.

[18] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, C. D. Manning, Stanza: A Python natural language processing toolkit for many human languages, in: Proceedings of the 58th Annual Meeting of the Association for Computational

Linguistics: System Demonstrations, Association for Computational Linguistics, 2020, pp. 272–277. Https://www.aclweb.org/anthology/2020.acl-demos.34.

[19] P. N. Mendes, M. Jakob, A. García-Silva, C. Bizer, DBpedia Spotlight: Shedding Light on the Web of Documents, in: Proceedings of the 7th International Conference on Semantic Systems, ACM, 2011, pp. 101–108. URL: https://dbpedia.org/spotlight.

[20] S. Vajjala, R. Balasubramaniam, What do we really know about state of the art ner?, in: Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), European Language Resources Association (ELRA), Marseille, 2022, pp. 5983–5993. Conference held on 20-25 June 2022.

[21] T. A. van Dijk, News as Discourse, Lawrence Erlbaum Associates, 1988.

[22] K. Markert, M. Nissim, Semeval-2007 task 08: Metonymy resolution at Semeval-2007, in: Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), Association for Computational Linguistics, 2007, pp. 36–41.

[23] P. Jeretic, A. Warstadt, S. Bhooshan, A. Williams, Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8690–8705. doi:10.18653/v1/2020.acl-main.768, https://aclanthology.org/2020.acl-main.768.

[24] Q. Li, J. Li, J. Sheng, S. Cui, J. Wu, Y. Hei, H. Peng, S. Guo, L. Wang, A. Beheshti, P. S. Yu, A survey on deep learning event extraction: Approaches and applications, IEEE Transactions on Neural Networks and Learning Systems 14 (2022) November 2022. doi:10.1109/TNNLS.2022.xxxxxxx.

[25] K. A. Mathews, M. Strube, A large harvested corpus of location metonymy, in: International Conference on Language Resources and Evaluation, 2020.

[26] S. Wang, X. Sun, X. Li, R. Ouyang, F. Wu, T. Zhang, J. Li, G. Wang, Gpt-ner: Named entity recognition via large language models, 2023. arXiv:2304.10428.

[27] OpenAI, Gpt-4 technical report, 2023. arXiv:2303.08774.

[28] H. Touvron, al., Llama 2: Open foundation and fine-tuned chat models, 2023. arXiv:2307.09288.

[29] L. Chen, M. Zaharia, J. Zou, How is chatgpt's behavior changing over time?, 2023. arXiv:2307.09009.

Unmasking the Wordsmith: Revealing Author Identity through Reader Reviews

Chiara Alzetta[1], Felice Dell'Orletta[1], Chiara Fazzone[1], Alessio Miaschi[1] and Giulia Venturi[1]

[1]ItaliaNLP Lab, CNR, Istituto di Linguistica Computazionale 'A.Zampolli', Pisa, Italy

Abstract
Traditional genre-based approaches for book recommendations face challenges due to the vague definition of genres. To overcome this, we propose a novel task called *Book Author Prediction*, where we predict the author of a book based on user-generated reviews' writing style. To this aim, we first introduce the 'Literary Voices Corpus' (LVC), a dataset of Italian book reviews, and use it to train and test machine learning models. Our study contributes valuable insights for developing user-centric systems that recommend leisure readings based on individual readers' interests and writing styles.

Keywords
Book Author Prediction, Italian reviews, stylistic analysis, user-generated book reviews

1. Introduction and Background

Reading for pleasure is currently experiencing a significant decline, as evidenced by surveys indicating that leisure reading has reached an unprecedented low[1]. Book recommender systems have been proposed as a valuable tool to promote the practice of reading for pleasure [1]. These systems provide personalized suggestions and aid users in navigating the vast array of available literary works [2]. Their integration into e-commerce services has long been explored, as it benefits both sellers and consumers [3].

Typically integrated with online platforms, book recommender systems rely on the history of users to predict their future interests and provide recommendations based on the literary genre or authors that users have previously engaged with. While recommending the other books by an author that the reader enjoyed is trivial, suggesting books belonging to the same genre remains a complex area of study, particularly concerning literary novels [4]. This is mostly due to the fact that the notion of *genre* represents a quite heterogeneous object of study due to multiple factors [5]. In fact, the same book can be assigned to more than one literary genre either on the same reading platform or across diverse platforms. Accordingly, various approaches have been proposed to automatically identify literary genres using book content [6, 7, 8], titles or summaries [9], and even cover designs [10]. Nevertheless, these models often face challenges when book content is inaccessible due to licensing restrictions.

Consequently, an alternative and promising line of research on book recommender systems involves leveraging user reviews as a valuable source of information for generating recommendations. Analyzing reviews allows for a unique perspective on books from the viewpoint of their readers, without requiring access to their content.

Reviews offer valuable insights into readers' opinions and preferences, and they have been effectively utilized to predict trends in the book market [11, 12, 13, 14, 15]. There are few attempts to exploit user reviews also for literary genre identification. These include [16] and [17] for English and Portuguese book reviews respectively. We have also contributed to this line of research by focusing on Italian book reviews [18]. In our previous work, we demonstrated how book reviews published by amateur readers on two social reading platforms, namely Amazon and Goodreads, can be exploited to automatically identify the genre of the reviewed book.

Building upon our prior investigations, our current research aims to explore whether the writing style of user-generated reviews, analyzed in terms of lexical and (morpho-)syntactic characteristics, can serve as a reliable source of information also to predict the author of a reviewed book. We started from the assumption that the vague definition of literary genres might make recommendations based on related authors more effective than genre-based approaches. To this end, inspired by the literature on Authorship Attribution [19], we introduced a novel task named *Book Author Prediction*. We tackle the problem as a supervised classification task, where the objective is to predict the author of a given book from a set of potential candidates. It is important to note that, unlike the traditional Authorship Attribution task, our information source consists of user-generated reviews

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
chiara.alzetta@ilc.cnr.it (C. Alzetta); felice.dellorletta@ilc.cnr.it (F. Dell'Orletta); chiara.fazzone@ilc.cnr.it (C. Fazzone); alessio.miaschi@ilc.cnr.it (A. Miaschi); giulia.venturi@ilc.cnr.it (G. Venturi)

[1]See https://www.istat.it/it/archivio/284591, https://literacytrust.org.uk/research-services/annual-literacy-survey/

rather than the books authored by the novelists themselves. This distinction adds a layer of complexity to the task, making it particularly challenging and novel in its approach. As a crucial step towards this objective, we introduce a novel dataset of Amazon[2] and Goodreads[3] book reviews, the 'Literary Voices Corpus' (LVC). The dataset successfully served in diverse experimental settings we explored in this work aimed at training and testing pre-trained and traditional machine learning models, that use different configurations of lexical and (morpho-)syntactic features, to accomplish the new prediction task.

The work presented in this study falls within the context of collective efforts to foster the habit of reading and enlarge the readership across different target audiences[4]. Among these initiatives, LettERE (Letture pER TE) is a project that aims to encourage and promote the practice of reading by creating a reading recommendation system that provides personalised recommendations tailored to the reader's language skills and interests (see Acknowledgements). In this regard, the research presented in this paper contributes significantly to the LettERE project's objectives by showing that user-generated reviews can be effectively used to identify readers sharing common interests and ultimately provide personalised book recommendations.

The remainder of the paper is organised as follows. Section 2 presents LVC, the novel collection of Italian book reviews referring to the books of six popular authors. Section 3 introduces the Book Author Prediction task and details the methodology and models exploited in this work to address it. Section 4 presents the results of our experiments. Finally, Section 5 offers conclusions and outlines potential future research directions.

2. The Literary Voices Corpus

We performed our experiments on the 'Literary Voices Corpus' (LVC), which encompasses a collection of book reviews in Italian published on two leading platforms for Digital Social Reading (DSR), Amazon Books and Goodreads and covering the work of several authors of fiction novels.[5] This corpus is a spin-off of the 'A Good Review' corpus, which we introduced in [18]. The LVC corpus is aimed at being representative of two different approaches to writing book reviews, a diversity specific to the peculiarities of the two platforms. In fact, while Goodreads gathers a large community of amateur readers to exchange opinions and reading recommendations, Amazon has a marked commercial vocation and treats books mainly as a consumer good. Goodreads reviews are typically exploited to predict the orientation of the book market [11, 13], to map reading preferences across various communities of users [20], as well as to analyze the linguistic style adopted by readers to describe their reading experiences [21, 22]. Conversely, reviews posted on Amazon Books have mostly been investigated within marketing and buyers' behaviour studies, often relying on sentiment analysis [23, 24, 25].

When building LVC, we first chose popular novelists in order to acquire a diverse but rich collection of reviews from amateur readers. These are J.K. Rowling, Stephen King, J.R.R. Tolkien, Jane Austen, Sarah J. Maas, and Dan Brown.[6]

Since *literary genre* is not a monolithic notion [4], the books of these authors traverse multiple genres. For example, King's repertoire encompasses horror, thriller, and science-fiction, while Maas's fantasy novels also incorporate a substantial element of romance. Then, we extracted the reviews for their respective books from the 'A Good Review' corpus and we integrated the set with new books if necessary using the ISBN number of a book to unambiguously identify it on Amazon and Goodreads and to collect its reviews written in Italian. This was done to reach a minimum of 1,100 reviews per novelist from Goodreads and 800 reviews from Amazon. While we successfully obtained the desired number of reviews for most authors, we encountered challenges for Austen and Maas on Amazon. Nonetheless, the number of reviews collected for these authors can still be considered reasonably comparable to the desired amount. The statistics of the final LVC dataset are reported in Table 1.

As can be noted, the two portions of the dataset (i.e., Amazon and Goodreads) are quite different in terms of the length of a single review. This difference arises in part from the lower number of reviews collected from Amazon, but mostly from the comparatively greater length of Goodreads reviews in terms of sentences and tokens. Thus, achieving a balanced number of reviews across authors does not correspond to an equal number of tokens. Furthermore, we notice a tendency to produce longer reviews among the readers of certain authors, such as King, Maas, or Austen, on both platforms. This represents one of the first general characterization of the diversity across literary voices we collected.

3. Book Author Prediction

The novel task of Book Author Prediction consists of predicting the author of a book from the readers' reviews. We explored the performance on the task of a suite of machine learning algorithms that vary with re-

[2] https://www.amazon.it
[3] https://www.goodreads.com
[4] See for instance: https://www.regione.toscana.it/-/un-patto-per-la-lettura.
[5] The LVC corpus is freely available under request for research purposes.
[6] The complete list of books whose reviews in Italian have been included in LVC can be found in Appendix A.

	Rowling	King	Tolkien	Austen	Maas	Brown	All
				Goodreads			
Books	6	8	7	7	6	7	41
Reviews	1,100	1,100	1,100	1,100	1,100	1,100	6,600
Sentences Total	5,951	7,479	6,224	6,914	11,447	5,151	43,166
Tokens Total	155,653	202,027	180,680	214,921	302,687	129,684	1,185,652
Avg Sentences per Review	5.41	6.80	5.65	6.28	10.40	4.68	6.54
Avg Tokens per Review	141.50	183.66	164.25	195.38	275.17	117.89	179.64
				Amazon			
Books	6	8	6	7	5	7	39
Reviews	800	800	800	749	653	800	4,602
Sentences Total	1,712	3,525	2,695	2,326	3,961	2,422	16,641
Tokens Total	21,899	69,078	48,275	40,875	81,668	40,719	302,514
Avg Sentences per Review	2.14	4.40	3.36	3.10	6.06	3.03	3.61
Avg Tokens per Review	27.37	86.34	60.34	54.57	125.06	50.89	65.73

Table 1
Literary Voices Corpus statistics.

Raw text
Number of sentences and tokens
Average tokens per sentence and average characters per token
Vocabulary Richness
Type/Token Ratio for words and lemmas (first 100/200 tokens)
Morphosyntactic information
Distibution of UD POS
Lexical density
Inflectional morphology
Distribution of lexical verbs and auxiliaries for inflectional categories (tense, mood, person, number)
Verbal Predicate Structure
Distribution of verbal heads and verbal roots
Average verb arity and distribution of verbs by arity
Global and Local Parsed Tree Structures
Average depth of the whole syntactic trees
Average length of dependency links and of the longest link
Average length of prepositional chains and distribution by depth
Average clause length
Relative order of elements
Distribution of subjects and objects in post- and pre-verbal position
Syntactic Relations
Distribution of dependency relations
Use of Subordination
Distribution of subordinate and principal clauses
Average length of subordination chains and distribution by depth
Distribution of subordinates in post- and pre-principal clause position

Table 2
Linguistic features acquired from book reviews.

spect to the architecture and features used for training (see Section 3.1). The models leverage a wide spectrum of text properties acquired from the reviews of increasing informativeness, which range from n-grams of words to stylistic features (Section 3.2), up to contextual sentence representations of Neural Language Models. For all models, we adopted a 5-fold cross-validation approach for training and testing. The train and test sets always contain reviews of different books, thus increasing the complexity of the classification tasks. Note that, considering the high discriminative power of proper nouns in this classification scenario, we performed the linguistic analysis of reviews and sanitized the text [26] by masking all tokens marked as proper nouns (POS = PROPN).

3.1. Models

Linear Support Vector Machine We define two LinearSVM models, referred to as 'Profiling' and 'Ngrams' models. The former takes the set of linguistic characteristics described in Sec. 3.2. Ngrams exploits lexical information since it uses as input feature a simple contiguous sequence of *n* words acquired from the reviews (i.e. n-grams, with *n* equal to 1, 2, and 3).

Neural Language Model We relied on the Italian pretrained version of the BERT model (12 layers, 768 hidden units) [27][7], which was pretrained using the Italian Wikipedia and the Italian portion of the OPUS corpus [28], a multilingual collection of translated open source documents available on the Internet, and fine-tuned on the Book Author Classification task.

LinearSVM + NLM We combined the previous models into a classifier based on LinearSVM and trained using the internal representations of the BERT model fine-tuned on the author classification tasks. We refer to this model as *SVM (BERT)*. *SVM (BERT+Profiling)* is an additional LinearSVM model trained using both the fine-tuned representations produced by BERT and Profiling-UD features. The BERT representations used as input features of the SVM model were computed by averaging the embeddings of all the tokens in each review.

Baselines We compared the performance of the above models against a random uniform classifier, i.e. a model that uniformly generates random predictions for each author.

[7]https://huggingface.co/dbmdz/bert-base-italian-cased

Model	Rowling	King	Tolkien	Austen	Maas	Brown	All
				Goodreads			
Baseline	0.19	0.15	0.16	0.18	0.15	0.16	0.16
Profiling	0.21	0.18	0.26	0.27	0.40	0.25	0.26
Ngrams	0.42	0.36	0.46	0.51	0.46	0.44	0.44
BERT	**0.69**	**0.70**	**0.72**	**0.79**	**0.73**	**0.74**	**0.73**
SVM (BERT)	0.44	0.51	0.55	0.58	0.57	0.56	0.54
SVM (BERT + Profiling)	0.46	0.50	0.51	0.54	0.56	0.57	0.52
Average	0.44	0.45	0.50	**0.54**	**0.54**	0.51	0.50
				Amazon			
Baseline	0.16	0.15	0.17	0.16	0.16	0.14	0.16
Profiling	0.38	0.18	0.27	0.17	0.32	0.22	0.26
Ngrams	0.44	0.35	0.40	0.38	0.58	0.39	0.42
BERT	**0.57**	**0.60**	**0.56**	**0.64**	**0.72**	**0.61**	**0.61**
SVM (BERT)	0.39	0.40	0.45	0.45	0.63	0.43	0.46
SVM (BERT+Profiling)	0.41	0.42	0.39	0.46	0.56	0.36	0.43
Average	0.44	0.39	0.41	0.42	**0.56**	0.40	0.44

Table 3
Results of book author prediction on Goodreads and Amazon reviews.

3.2. Linguistic Features

To model the linguistic properties of the reviews, we relied on a set of 150 linguistic features. These features correspond to specific aspects of the document structure and were derived using Profiling-UD [29], a web-based tool conceived to linguistically profile multilingual texts by relying on the Universal Dependencies (UD) formalism [30]. The features encompass 9 dimensions of document structure, which are detailed in Table 2. They range from morpho-syntactic and inflectional properties to more complex aspects of sentence structure, such as the depth of the syntactic tree. Other features pertain to the structure of sub-trees and include the order of subjects and objects in relation to the verb, as well as the use of subordination.

4. Results

Table 3 presents the classification accuracies for the task of Book Author Prediction. Notably, all models outperformed the random uniform baseline on both Amazon and Goodreads. Upon closer examination of the models, we notice that lexical information has more discriminative power than linguistic properties in the task. As proof, consider the global and author-level scores obtained by the *Profiling* model compared to the *Ngram* and, most notably, the *BERT* models. Interestingly, using the fine-tuned BERT representations as input features for the SVM classifier (*SVM (BERT)*) yielded lower results than simply using pre-trained BERT, and the results are comparable – or lower – when combining contextualized representations with linguistic features (*SVM (BERT+Profiling)*).

Comparing the two platforms, Goodreads reviews exhibit on average higher accuracy scores overall. This is possibly due to a typical trait of commercial platforms like Amazon, whose reviews frequently encompass aspects beyond the book's content, such as parcel delivery or the edition's book cover. These topics cause the reviews to be quite standardised, thus more difficult to discriminate. Conversely, Goodreads reviews primarily focus on the book's content possibly containing a larger amount of stylistic elements which help the automatic classification. This trend holds also when classifying individual authors, except Rowling for the *Profiling* and *Ngrams* models.

When looking at the results obtained for individual authors, Sara J. Maas turned out to be the most accurately predicted author on both platforms, considering the average scores across all models. However, upon closer inspection of the results obtained with the top-performing model (*BERT*), we observe that while Maas remains the most accurately identified author in Amazon reviews, the reviews of Jane Austen's books exhibit the highest level of distinctiveness on Goodreads.

4.1. Discussion

To take a closer look at the classification results, Fig. 1 reports the confusion matrices with the percentage of the predictions made by all models in the Book Author Prediction task. This complements the classification results by showing which authors are more confusing and which are the most wrongly classified ones.

In general, we observe that as the model performance improves, the matrices become less sparse, regardless of the platform. This means that when the correct author is predicted most of the time, the erroneous predictions are distributed quite evenly among all possible authors.

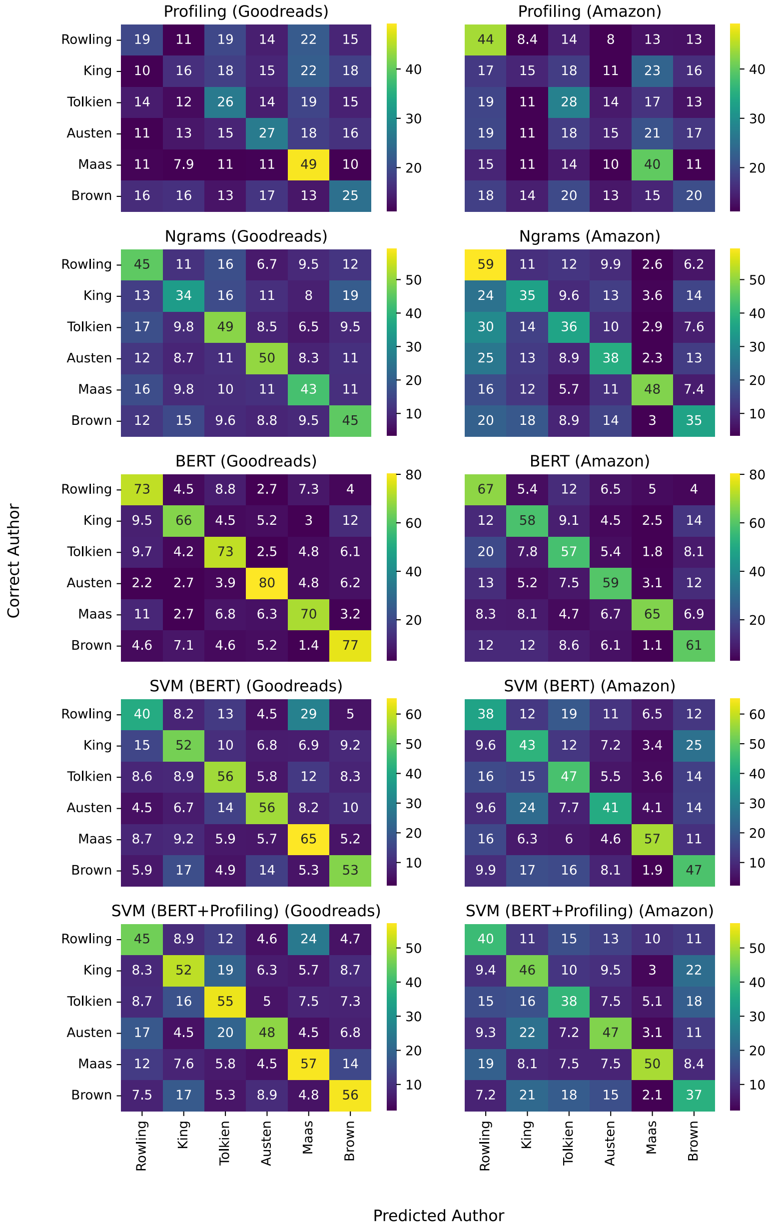


Figure 1: Confusion matrices of the classification task for all models: cells report the percentage of reviews automatically assigned to an author by each classification model (column) with respect to their actual author (row).

Consider, for instance, the matrices obtained from the analysis of *BERT* and compare them with the matrices referring to the *Profiling* and *Ngrams* models, which yield the most sparse matrices.

Notable differences arise in the distribution of predicted authors across the two platforms. For instance, when considering the *Profiling* model applied to Goodreads reviews, we observe that Maas is the most frequently predicted author, leading to other authors' books being frequently misclassified as Maas's works. Notably, the reviews of *It* by King and of the fourth book from the Harry Potter saga by Rowling are often incorrectly assigned to Maas. The content of these books, at the crossroads between the fantasy and horror genres, may contribute to the model confusion. However, the most influencing factor to the *Profiling* model predictions appears to be the review length. On Goodreads, reviews of King's and Rowling's books that are longer than 150 tokens are wrongly classified as referring to Maas in over 40% of cases. On Amazon, we observe an opposite tendency, but for a different author: when a review has less than 10 tokens, the model assigns the review to Rowling in around 60% of cases.

The analysis of the feature rankings[8] produced by the classifiers trained on both Amazon and Goodreads reviews confirms the importance of review length for the *Profiling* model. Indeed, features that capture structural properties are particularly relevant for the model: the use of subordination (*subordinate_dist*) is crucial for classifying Rowling's and King's reviews on Goodreads, as they exhibit respectively the lowest and highest use of subordinate clauses. Conversely, on Amazon, the average number of verb dependents (*verb_edges*) and the distribution of function words (namely, conjunctions, auxiliary verbs and determiners) are discriminative for Rowling, Tolkien, and Maas.

For what concerns the *Ngram* model, the feature ranking consists of the n-grams employed by the model ordered by relevance for book author classification purposes on Amazon and on Goodreads. Quite expectedly, the analysis of the top 100 most relevant n-grams reveals that, on Amazon, parcel delivery is a highly referenced topic (e.g. '*tempi previsti*', expected timing, and '*ben confezionato*', well packaged), especially among the readers of Tolkien and Rowling, which have the most similar n-gram rankings (Spearman correlation score = 0.235, $p < 0.05$). The two authors are the most frequently confused by the model, especially for what concerns the reviews of Tolkien's 'The Hobbit' and 'The Silmarillion', wrongly classified as referring to Rowling's books. Indeed, it is possible that the two authors attract a similar readership interested in books involving intricate mythologies, and that feature multi-dimensional characters with strengths, flaws, and internal struggles. Such closeness between the Amazon reviews of these authors is captured also by the BERT model which, although performing better than other models on the task, seems quite confused by the reviews of the same Tolkien books.

On Goodreads reviews, where parcel delivery is not relevant, the most impactful n-grams tend to revolve around book appreciation (e.g., *'ho apprezzato'*, I appreciated; *'lettura piacevole'*, pleasant reading; *'non mi aspettavo'*, I did not expect) or plot (*'il maghetto'*, the little wizard; *'signore di'*, lord of; *'chiesa'*, church; *'di epoca'*, historical; *'drago'*, dragon; *'di vampiri'*, of vampires). Therefore, it is not surprising to see that King's reviews are most frequently misclassified as referring to Brown's work, also by the BERT model. Both authors, despite their differences, are known for building suspense and tension in their narratives and incorporating detailed historical settings and psychological aspects into their work.

The classification of Goodreads review performed by the *SVM (BERT)* and *SVM (BERT + Profiling)* models highlight author commonalities that did not emerge so strongly with other models. The reviews of Rowling's books, for instance, are frequently wrongly classified as referring to Maas's work. Both authors are known for their contributions to popular literature, particularly in the genres of fantasy and young adult fiction, which attract a readership interested in exploring themes of personal growth and self-discovery through the characters' coming-of-age journeys.

Overall, no particular author appears to be systematically confused by all models. This finding is particularly interesting from our perspective since it shows that using user-generated reviews as an information source allows to successfully address the Book Author Prediction task. It suggests that books authored by different novelists attract readers who are interested in similar topics and also adopt similar communication strategies in their writing. It also implies that the proposed methodology could have a positive impact on the development of user-centric book recommender systems.

5. Conclusions

This paper has explored an innovative approach that leverages user reviews as a source of information for Book Author Prediction. Building upon our prior work, we introduced a novel dataset of Amazon and Goodreads book reviews, LVC, which has been used for training and evaluating machine learning models addressing the novel book author prediction task.

Our findings highlight the challenging nature of predicting the author of a novel from a reader's review. However, the analysis of erroneous predictions pointed us to cases of books sharing a similar readership. This observation supports the intuition that user-generated reviews can effectively serve as a basis for personalized book recommendations. By analyzing reviews, we gained insights into readers' preferences beyond the writing style of the book's author, opening up new avenues for more tailored and user-centric recommendations.

Moving forward, this research could be expanded by investigating the impact of exploiting user judgments as an additional feature for classification. Furthermore, the sentiment expressed by readers about a book, whether positive or negative, could be leveraged to validate and fine-tune personalized recommendations.

[8] See Appendix B and C.

Acknowledgments

We thank the "Letture pER TE" (LettERE) project (2022-2024) funded by Regione Toscana (Progetti Congiunti di Alta Formazione – POR FSE 2014-2020 Investimenti a favore della crescita e dell'occupazione) in collaboration with M.E.T.A. Srl company.

References

[1] H. Alharthi, D. Inkpen, S. Szpakowicz, Authorship identification for literary book recommendations, in: Proceedings of the 27th International Conference on Computational Linguistics (COLING), ACL, 2018, pp. 390–400.

[2] H. Alharthi, D. Inkpen, S. Szpakowicz, A survey of book recommender systems, Journal of Intelligent Information Systems 51 (2018) 139–160.

[3] J. B. Schafer, J. Konstan, J. Riedl, Recommender systems in e-commerce, in: Proceedings of the 1st ACM conference on Electronic commerce, 1999, pp. 158–166.

[4] J.-M. Schaeffer, Qu'est-ce qu'un genre littéraire?, Seuil, 1989.

[5] D. Biber, S. Conrad, Genre, Register, Style, Cambridge University Press, 2009.

[6] L. Shamir, UDAT: Compound quantitative analysis of text using machine learning, Digital Scholarship in the Humanities 36 (2020) 187–208.

[7] Rahul, Ayush, D. Agarwal, D. Vijay, Genre classification using character networks, in: Proceedings of the 5th International Conference on Intelligent Computing and Control Systems (ICICCS), IEEE, 2021, pp. 216–222.

[8] J. Worsham, J. Kalita, Genre identification and the compositional effect of genre in literature, in: Proceedings of the 27th International Conference on Computational Linguistics (COLING), ACL, 2018, pp. 1963–1973.

[9] E. Ozsarfati, E. Sahin, C. J. Saul, A. Yilmaz, Book genre classification based on titles with comparative machine learning algorithms, in: Proceedings of 2019 4th International Conference on Computer and Communication Systems (ICCCS), IEEE, 2019, pp. 14–20.

[10] P. Buczkowski, A. Sobkowicz, M. Kozlowski, Deep learning approaches towards book covers classification, in: Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods (ICPRAM), SCITEPRESS-Science and Technology Publications, 2018, pp. 309–316.

[11] K. Wang, X. Liu, Y. Han, Exploring Goodreads reviews for book impact assessment, Journal of Informetrics 13 (2019) 874–886.

[12] G. Aerts, T. Smits, P. W. Verlegh, How online consumer reviews are influenced by the language and valence of prior reviews: A construal level perspective, Computers in Human Behavior 75 (2017) 855–864.

[13] S. K. Maity, A. Panigrahi, A. Mukherjee, Analyzing social book reading behavior on Goodreads and how it predicts Amazon best sellers, Influence and Behavior Analysis in Social Networks and Social Media (2019) 211–235.

[14] S. Dimitrov, F. Zamal, A. Piper, D. Ruths, Goodreads versus Amazon: the effect of decoupling book reviewing and book selling, in: Proceedings of International AAAI Conference on Web and Social Media (ICWSM), volume 9, 2015, pp. 602–605.

[15] M. Thelwall, Reader and author gender and genre in Goodreads, Journal of Librarianship and Information Science 51 (2019) 403–430.

[16] M. Saraswat, Leveraging genre classification with rnn for book recommendation, International Journal of Information Technology (2022) 1–6.

[17] C. Scofield, M. O. Silva, L. de Melo-Gomes, M. M. Moro, Book genre classification based on reviews of portuguese-language literature, in: Proceedings of the International Conference on Computational Processing of the Portuguese Language (PROPOR), 2022, pp. 188–197.

[18] C. Alzetta, F. Dell'Orletta, A. Miaschi, E. Prat, G. Venturi, Tell me how you write and I'll tell you what you read: a study on the writing style of book reviews, Journal of Documentation Forthcoming (2023).

[19] E. Stamatatos, A survey of modern authorship attribution methods, Journal of the American Society for Information Science and Technology 60 (2009) 538–556.

[20] K. Bourrier, M. Thelwall, The social lives of books: Reading victorian literature on Goodreads, Journal of Cultural Analytics 5 (2020) 12049.

[21] B. Driscoll, D. Rehberg Sedo, Faraway, so close: Seeing the intimacy in Goodreads reviews, Qualitative Inquiry 25 (2019) 248–259.

[22] L. Nuttall, C. Harrison, Wolfing down the twilight series: Metaphors for reading in online reviews, Contemporary media stylistics (2020) 35–60.

[23] K. Kaur, T. Singh, Impact of online consumer reviews on Amazon books sales: Empirical evidence from india, Journal of Theoretical and Applied Electronic Commerce Research 16 (2021) 2793–2807.

[24] F. Chiavetta, G. L. Bosco, G. Pilato, A lexicon-based approach for sentiment classification of Amazon books reviews in Italian language, in: International Conference on Web Information Systems and Technologies (WEBIST), volume 3, Scitepress, 2016, pp. 159–170.

[25] K. Srujan, S. Nikhil, H. Raghav Rao, K. Karthik, B. Harish, H. Keerthi Kumar, Classification of Amazon book reviews based on sentiment analysis, in: Information Systems Design and Intelligent Applications, Springer, 2018, pp. 401–411.
[26] V. Vasudevan, A. John, A review on text sanitization, International Journal of Computer Applications 95 (2014).
[27] T. Wolf, L. Debut, V. Sanh, alii, Transformers: State-of-the-art natural language processing, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), ACL, 2020, pp. 38–45.
[28] J. Tiedemann, L. Nygaard, The OPUS corpus - parallel and free, in: Proceedings of the Conference on Language Resources and Evaluation (LREC), ELRA, 2004.
[29] D. Brunato, A. Cimino, F. Dell'Orletta, G. Venturi, S. Montemagni, Profiling-UD: a tool for linguistic profiling of texts, in: Proceedings of the Conference on Language Resources and Evaluation (LREC), ELRA, 2020, pp. 7147–7153.
[30] M. C. De Marneffe, C. D. Manning, J. Nivre, D. Zeman, Universal dependencies, Computational linguistics 47 (2021) 255–308.

A. Books of the Literary Voices Corpus

Author	Book
Jane Austen	Emma
	Lady Susan
	Mansfield Park
	Northanger Abbey
	Persuasion
	Persuasion
	Pride and Prejudice
	Sense and Sensibility
Dan Brown	Angels and Demons
	Deception Point
	Digital Fortress
	Inferno
	Origin
	The Da Vinci Code
	The Lost Symbol
Sarah J. Maas	A Court of Mist and Fury
	A Court of Frost and Starlight
	A Court of Wing and Ruin
	A Court of Silver Flames
	Throne of Glass
J.K. Rowling	Harry Potter and the Chamber of Secrets
	Harry Potter and the Goblet of Fire
	Harry Potter and the Half Blood Prince
	Harry Potter and the Order of the Phoenix
	Harry Potter and the Prisoner of Azkaban
	Harry Potter and the Sorcerer's Stone
J.R.R. Tolkien	The Fellowship of the Ring
	The Children of Húrin
	The Hobbit
	The Return of the King
	The Silmarillion
	The Two Towers
Stephen King	Salem's Lot
	Carrie
	Doctor Sleep
	It
	Misery
	Mr. Mercedes
	Pet Sematary
	The Shining

Table 4
List of the books present in the LVC.

B. Feature ranking Profiling Model (Goodreads)

	Dan Brown		J.K. Rowling		J.R.R. Tolkien	
	Feature	Avg	Feature	Avg	Feature	Avg
1	ttr_form_chunks_100	0.22	subordinate_post	67.91	ttr_lemma_chunks_100	0.26
2	ttr_lemma_chunks_100	0.20	subordinate_dist_1	58.66	ttr_form_chunks_100	0.30
3	upos_dist_AUX	4.46	subordinate_pre	11.46	aux_tense_dist_Pres	71.32
4	avg_prepositional_chain_len	0.93	dep_dist_orphan	0.00	ttr_form_chunks_200	0.15
5	dep_dist_aux	2.54	verbs_form_dist_Part	26.69	ttr_lemma_chunks_200	0.13
6	upos_dist_DET	12.87	ttr_form_chunks_100	0.24	n_prepositional_chains	7.29
7	dep_dist_cop	1.57	avg_prepositional_chain_len	0.86	n_tokens	164.25
8	prep_dist_2	10.46	upos_dist_ADP	11.04	upos_dist_AUX	4.84
9	prep_dist_1	68.65	subordinate_dist_2	15.22	upos_dist_ADP	12.02
10	dep_dist_det	12.06	dep_dist_mark	2.69	dep_dist_orphan	0.00
11	ttr_form_chunks_200	0.10	upos_dist_SCONJ	1.54	upos_dist_DET	14.35
12	ttr_lemma_chunks_200	0.09	ttr_lemma_chunks_100	0.21	aux_mood_dist_Ind	73.27
13	dep_dist_flat:name	1.38	verb_edges_dist_1	15.34	dep_dist_aux	2.60
14	avg_verb_edges	2.39	aux_tense_dist_Pres	67.71	aux_tense_dist_Imp	5.79
15	prep_dist_3	1.02	avg_subordinate_chain_len	1.08	dep_dist_case	10.55
16	dep_dist_cc	3.42	verb_edges_dist_2	25.27	dep_dist_cop	1.90
17	dep_dist_flat:foreign	0.08	dep_dist_case	9.77	dep_dist_mark	2.64
18	upos_dist_NUM	0.77	verb_edges_dist_3	23.72	verbs_form_dist_Part	28.82
19	upos_dist_PROPN	4.68	verbs_form_dist_Fin	38.53	dep_dist_flat:name	0.62
20	upos_dist_CCONJ	3.43	upos_dist_AUX	4.97	aux_num_pers_dist_Sing+3	52.14

	Jane Austen		Sarah J. Maas		Stephen King	
	Feature	Avg	Feature	Avg	Feature	Avg
1	ttr_lemma_chunks_200	0.15	ttr_form_chunks_200	0.25	upos_dist_CCONJ	3.29
2	ttr_form_chunks_200	0.18	ttr_lemma_chunks_200	0.22	dep_dist_cc	3.29
3	upos_dist_CCONJ	3.88	verbs_form_dist_Fin	38.74	avg_prepositional_chain_len	0.98
4	verbal_head_per_sent	3.35	verbs_form_dist_Part	30.77	ttr_form_chunks_200	0.17
5	avg_prepositional_chain_len	0.99	verb_edges_dist_2	27.51	ttr_lemma_chunks_200	0.15
6	n_tokens	195.38	verb_edges_dist_1	12.55	prep_dist_2	10.28
7	dep_dist_cc	3.87	verb_edges_dist_3	26.86	prep_dist_1	74.51
8	verbs_form_dist_Fin	38.72	verbs_form_dist_Inf	21.37	subordinate_post	76.19
9	tokens_per_sent	29.92	aux_tense_dist_Past	5.30	dep_dist_orphan	0.00
10	ttr_lemma_chunks_100	0.30	avg_prepositional_chain_len	0.95	prep_dist_3	0.89
11	prep_dist_1	72.48	n_prepositional_chains	9.46	subordinate_dist_1	66.18
12	ttr_form_chunks_100	0.34	verb_edges_dist_4	16.36	tokens_per_sent	26.85
13	n_sentences	6.29	aux_tense_dist_Pres	75.14	n_tokens	183.66
14	dep_dist_advmod	7.51	prep_dist_1	74.15	aux_tense_dist_Pres	72.50
15	prep_dist_2	10.92	verbal_head_per_sent	3.60	ttr_lemma_chunks_100	0.31
16	verb_edges_dist_2	27.55	aux_form_dist_Part	5.03	avg_verb_edges	2.54
17	verb_edges_dist_3	26.57	prep_dist_2	9.36	subordinate_pre	12.63
18	upos_dist_ADV	8.00	n_tokens	275.17	verbal_head_per_sent	3.22
19	upos_dist_AUX	4.79	aux_mood_dist_Ind	78.48	dep_dist_case	10.52
20	dep_dist_case	10.32	verb_edges_dist_5	6.62	upos_dist_ADP	12.03

Table 5
Top 20 ranked features by the Profiling model for the classification of each author on Goodreads. Average values of the linguistic features are also reported (columns *Avg*).

C. Feature ranking Profiling Model (Amazon)

	Dan Brown		J.K. Rowling		J.R.R. Tolkien	
	Feature	**Avg**	**Feature**	**Avg**	**Feature**	**Avg**
1	avg_subordinate_chain_len	0.98	ttr_form_chunks_200	0.01	upos_dist_AUX	4.57
2	dep_dist_cc	3.47	ttr_lemma_chunks_200	0.01	dep_dist_det	12.67
3	upos_dist_AUX	4.06	upos_dist_CCONJ	2.67	dep_dist_aux	2.23
4	upos_dist_CCONJ	3.46	dep_dist_cc	2.67	upos_dist_ADV	6.84
5	aux_tense_dist_Pres	62.35	upos_dist_AUX	3.99	ttr_lemma_chunks_100	0.07
6	dep_dist_aux	2.18	ttr_form_chunks_100	0.03	ttr_form_chunks_100	0.08
7	dep_dist_cop	1.63	dep_dist_aux	2.11	dep_dist_cop	1.95
8	subordinate_dist_2	12.89	verb_edges_dist_3	16.65	upos_dist_DET	13.35
9	lexical_density	0.57	verb_edges_dist_2	25.40	dep_dist_root	10.92
10	verbs_form_dist_Part	33.30	ttr_lemma_chunks_100	0.03	dep_dist_advmod	6.41
11	ttr_lemma_chunks_200	0.01	n_tokens	27.37	verb_edges_dist_2	29.13
12	upos_dist_DET	12.06	dep_dist_cop	1.57	verb_edges_dist_3	21.28
13	subordinate_dist_3	2.54	verb_edges_dist_4	6.61	ttr_lemma_chunks_200	0.02
14	aux_form_dist_Fin	63.30	verbs_form_dist_Inf	11.28	aux_tense_dist_Pres	64.68
15	subordinate_dist_1	61.25	lexical_density	0.65	verb_edges_dist_4	11.55
16	verbs_form_dist_Fin	34.92	aux_form_dist_Part	2.23	avg_verb_edges	2.14
17	upos_dist_PUNCT	10.96	verb_edges_dist_1	18.43	verbs_form_dist_Part	35.10
18	ttr_form_chunks_200	0.01	avg_verb_edges	1.59	dep_dist_case	10.77
19	verb_edges_dist_3	23.54	verbs_form_dist_Fin	21.86	ttr_form_chunks_200	0.02
20	dep_dist_case	10.13	aux_tense_dist_Past	2.27	verb_edges_dist_1	18.95

	Jane Austen		Sarah J. Maas		Stephen King	
	Feature	**Avg**	**Feature**	**Avg**	**Feature**	**Avg**
1	aux_tense_dist_Pres	59.15	verbs_form_dist_Part	33.77	dep_dist_det	11.73
2	upos_dist_AUX	4.30	ttr_lemma_chunks_200	0.09	dep_dist_cc	3.12
3	avg_verb_edges	2.13	ttr_form_chunks_200	0.10	upos_dist_CCONJ	3.19
4	upos_dist_DET	12.40	verbs_form_dist_Fin	33.45	upos_dist_DET	12.63
5	dep_dist_case	9.13	verb_edges_dist_2	30.79	ttr_form_chunks_200	0.05
6	upos_dist_ADP	10.77	lexical_density	0.54	ttr_lemma_chunks_200	0.05
7	dep_dist_cc	3.73	verb_edges_dist_1	15.47	avg_verb_edges	2.27
8	dep_dist_aux	2.36	verbs_form_dist_Inf	20.95	verbs_form_dist_Part	32.46
9	verbs_form_dist_Part	28.94	verb_edges_dist_3	23.55	aux_tense_dist_Pres	63.26
10	avg_subordinate_chain_len	0.90	dep_dist_flat	0.00	verbs_form_dist_Fin	34.79
11	dep_dist_det	11.54	upos_dist_DET	13.45	verbs_form_dist_Inf	19.15
12	upos_dist_CCONJ	3.73	upos_dist_NUM	0.55	lexical_density	0.55
13	dep_dist_cop	1.67	verb_edges_dist_4	12.00	ttr_lemma_chunks_100	0.12
14	verbs_form_dist_Fin	34.52	dep_dist_nummod	0.53	verb_edges_dist_1	15.93
15	aux_mood_dist_Ind	59.79	upos_dist_ADP	10.91	dep_dist_root	11.04
16	ttr_form_chunks_200	0.02	dep_dist_flat:name	0.31	upos_dist_AUX	4.45
17	ttr_lemma_chunks_200	0.02	verb_edges_dist_0	2.18	dep_dist_aux	2.44
18	aux_form_dist_Fin	61.29	avg_subordinate_chain_len	1.02	principal_proposition_dist	39.50
19	subordinate_dist_2	11.26	dep_dist_det	12.28	dep_dist_det:poss	0.68
20	aux_tense_dist_Imp	3.92	verbs_form_dist_Ger	2.63	dep_dist_flat:foreign	0.03

Table 6
Top 20 ranked features by the Profiling model for the classification of each author on Amazon. Average values of the linguistic features are also reported (columns *Avg*).

Integrated Gradients as Proxy of Disagreement in Hateful Content

Alessandro Astorino[1], Giulia Rizzi[1,2] and Elisabetta Fersini[1,*]

[1]University of Milano-Bicocca, Milan, Italy
[2]Universitat Politècnica de València, Valencia, Spain

Abstract
Online platforms have increasingly become hotspots to spread not only opinions but also hate speech, posing substantial obstacles to developing constructive and inclusive online communities. In this paper, we propose a novel approach that leverages the integrated gradients of pre-trained language models to automatically predict both hate speech and the potential disagreement that can arise from readers. The integrated gradient attributions are used to shed light on the model's decision-making process attributing importance scores to individual tokens and enabling the identification of crucial factors contributing to disagreement and hate speech classifications. The integrated gradients' straightforwardness allows for the recognition of fundamental causes of disagreements and hate speech content. By adopting an interpretable approach, we bridge the gap between model predictions and human comprehension. Our experimental results highlight the effectiveness of our approach, outperforming traditional BERT models and state-of-the-art methods in both prediction tasks.

Keywords
Learning with Disagreement, Integrated Gradients, Hateful Content

1. Introduction

In the modern era, human beings are constantly subject to absorbing content of various kinds generated and shared on the web. To ensure the sustainability of continuously produced information and promote individual and societal well-being in the context of online content is important to recognize where hate content can harm from a personal perspective. Different individuals, according to their cultural beliefs and backgrounds, may be more or less susceptible to potentially offensive content. It is, therefore, necessary to safeguard the perceptions of different individuals by defining Natural Language Processing (NLP) models that are able to capture and model different perceptions. How to deal with disagreement, in particular related to hate speech detection problems, is a topic that has attracted increasing interest during the last few years [1, 2, 3, 4]. Although a good number of approaches able to deal with disagreement in hate speech detection problems have been proposed [5, 6, 7, 8], only a few of them have been focused on really modelling perspectivism.

Recognizing potential disagreements within hateful content, especially in identifying controversial elements, is of paramount importance for multiple reasons. When the possibility of disagreement arises in hateful texts shared on social media platforms (e.g. Twitter), it becomes critical to have a service that recognizes if that text written in that manner causes disagreement and works as a filter for these texts based on a personal perspective

Moreover, a detoxification strategy could be implemented to notify the authors of user-generated texts, cautioning them about the potential perception of their content as hateful by certain readers, and suggesting revisions for the original message. Identifying disagreements within hateful sentences and determining the associated disagreement-related elements can significantly contribute to the creation of reliable benchmarks. Primarily, for contents prone to disagreements, specific annotation policies can be implemented (e.g., involving more annotators, excluding samples requiring annotation from the dataset, etc.). Additionally, annotators could be provided with targeted cues to focus on particular constituents that may be perceived differently by readers (e.g., underlining words, hashtags, or emojis identified as disagreement-related elements warranting careful evaluation).

In this paper, we try to connect hate speech and disagreement by determining which hateful constituents can contribute more to predicting disagreement. In particular, we combine pre-trained language models and integrated gradients providing the following main contributions:

- a *filtering strategy* of textual constituents that contributes remarkably to explain hateful messages;
- a *unified model* that, considering the prediction of the hateful contents and the selected explanations,

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
*Corresponding author.
a.astorino2@campus.unimib.it (A. Astorino); g.rizzi10@campus.unimib.it (G. Rizzi); elisabetta.fersini@unimib.it (E. Fersini)
0000-0002-0619-0760 (G. Rizzi); 0000-0002-8987-100X (E. Fersini)

Dataset	Language	Types	Training Size	Task	Annotators	Pool Ann.	% Full Agreement
HS-Brexit [9]	En	Tweets	1,120	Hate Speech	6	6	69%
ArMis [10]	Ar	Tweets	943	Misogyny and sexism detection	3	3	86%
ConvAbuse [11]	En	User-agent dialogues	4,050	Abusive Language detection	2-7	7	65%
MD-Agreement [12]	En	Tweets	10,753	Offensiveness detection	5	>800	42%

Table 1
Datasets characteristics.

predicts if disagreement could arise when reading such contents;

The rest of the paper is organized as follows. In Section 2 an overview of the state of the art is provided. The adopted datasets are described in Section 3. In Section 4 the proposed approach is detailed. The results achieved by the proposed approaches are reported in Section 5. Finally, conclusions and future research directions are drawn in Section 6.

2. Related Work

The fast rise of social media and online communication platforms has changed the way people communicate, exchange information, and express their ideas, while simultaneously increasing the spread of hate speech. Hateful content includes a wide range of various forms of offensive, abusive, and discriminatory language targeted at individuals or groups based on their race, religion, ethnicity, gender, or other protected characteristics. The propagation of hate speech online has major implications, perpetuating discrimination, stoking antagonism, and instigating violence, necessitating the urgent need for effective anti-hate speech solutions. Over the years, significant progress has been made in developing automatic hate content detection systems that leverage advancements in Natural Language Processing (NLP), machine learning, and deep learning techniques. In this section, we highlight some of the state-of-the-art approaches and methodologies employed in hate speech detection. The dominant approach for hate speech detection is represented by supervised learning [13, 14]. In particular, the approaches based on Language Models (LM) [15, 16, 17] have shown promising results in capturing contextual information and semantic relationships, leading to improved classification performance.

One of the key challenges in hate speech detection is the ability to make sense of the context in which the offensive language is used. Researchers have explored context-aware models [18, 19] that consider the surrounding text or conversation to make more accurate predictions. This can exploit speaker attributes, or discourse patterns to better grasp the intended meaning and differentiate between hate speech and non-hateful expressions. In recent years, hate speech detection has extended to encompass multimodal data analysis to keep up with the increasing usage of images and videos in online communication. Combining textual information with visual cues from images and videos has shown promise for improving the accuracy and granularity of hate speech identification systems [20, 14]. An increasing number of datasets are collecting multimodal examples of hate content ranging from memes [20, 21] to advertisements [22] and videos [23].

The latest datasets are addressing the problem of hate speech under the Learning with Disagreements paradigm reporting information both on the *hard label* (usually obtained through majority voting) and on the *soft label* (with all the annotators' labels or a confidence level attached to the labels). The inclusion of different perspectives allows us to address the subjectivity of the task by representing the multiple perceptions of the annotators with different points of view and understanding [24]. The information that represents annotators' disagreement is not only used to improve the quality of the dataset [25] but also in the training process by weighting the samples according to their disagreement values [26] or by directly training from disagreement, without considering any aggregates label [27, 28].

3. Dataset

The four benchmark datasets provided by SemEval 2023 task 11 related to Learning With Disagreements [29] have been considered in order to address the problem of predicting disagreement in hateful content. The datasets have different characteristics for what concerns language, type, and goal as summarized in Table 1. All the datasets have been adapted by the challenge organizer to share a common structure for what concern the textual input and the hard and soft labels (additional dataset-specific attribute are present). Since in this work, the disagreement prediction is addressed as a binary task, an *agreement label* has been derived from the *soft label*. This is because taking the levels of disagreement into account requires knowledge of the number of annotators, which is not taken into account at this time since the objective is to distinguish agreement and disagreement and not the various levels of disagreement. In particular, the agreement label is set equal to $(+)$ when there is a 100% agreement between the annotators, regardless of the value of the hard label, while equal to $(-)$ in all the other cases.

4. Proposed Approach

The proposed approach aims at addressing the tasks of predicting both disagreement and hate speech while maintaining the method fully interpretable through the adoption of integrated gradients. Integrated gradients are used to shed light on the model's decision-making process attributing importance scores to individual tokens and enabling the identification of crucial factors contributing to the model's decision.

In particular, the proposed approach is composed of four main steps:

1. **Fine-tuning of a pre-trained LM:** the multilingual BERT (m-BERT) has been fine-tuned to distinguish hateful content from non-hateful ones. The textual input (i.e. the tweet or the conversation depending on the dataset) has been given as input to the m-BERT model with a final sigmoid layer. Additionally, to overcome the datasets' class imbalance, in the training phase, the loss function has been penalized accordingly to the class distribution. The optimal decision threshold has been determined according to the Youden's J statistics [30]. The statistics, which is a linear combination of sensitivity and specificity, is maximized by evaluating several cut-offs.
2. **Estimation of the attribution score:** the attribution score for each textual constituent has been estimated using the integrated gradients presented in [31] on the fine-tuned model. This attribution score assumes values from -1 to 1, 1 means that that token has a high contribution to the prediction of the model and -1 the opposite. A visual representation of the integrated gradient on two available samples is reported in Figure 1. On one hand, each attribution score allows us to identify those tokens that contribute more to the final prediction, and on the other hand, those compositions of tokens characterized by divergent values make the content controversial potentially leading to disagreement. The variability and the magnitude of attribution values within a text are subsequently exploited to detect a potential disagreement.
3. **Filtering constituents:** the integrated gradient's attribution scores have been used to filter out those tokens that do not bring a significant contribution to explain the target label. In particular, let t_{im} be the i-th token within a text m and s_{im} the corresponding attribution score. The token t_{im} is considered significant and maintained for the subsequent disagreement model if $s_{im} \geq \tau$, otherwise the token is removed from the original input text. In our case study, τ is a specific threshold estimated according to a grid search approach.
4. **Extraction of latent representations:** the tokens considered significant according to the previous step are used to extract the corresponding latent representation of the filtered sentence from the fine-tuned m-BERT model.
5. **Creation of the disagreement input space:** the latent representation obtained at the previous step is used according to the following strategies:
    - *Filtered Embeddings:* the embedding of the filtered sentence is obtained by fine-tuned model on hate and used to train the subsequent disagreement model.
    - *Predicted Label:* the Boolean labels predicted by the model fine-tuned to distinguish hateful from non-hateful messages are included in the input space for training the disagreement model.
    - *Distribution values:* the distribution probability obtained through the sigmoid layer of the fine-tuned models has been alternatively considered.
6. **Training of the disagreement model:** the derived input space (latent representation of the selected token, concatenated with the predicted label or probability distribution) is given as input to a trivial Neural Network with the following structure to predict disagreement labels:
    - *Input layer*: layer that reflects the shape of the input, with Relu as activation function and dropout of 0.7;
    - *Hidden layer*: layer that halves the size of the input with Relu and dropout of 0.7;
    - *Output layer*: one output neuron with a sigmoid function to predict the final agreement/disagreement.

The entire proposed approach is synthesized in Figure 2.

5. Experimental Results

In this section, the results obtained by the proposed approach are reported. We measured Precision (P), Recall (R) and F-Measure (F), distinguishing between hateful $(+)$ and not hateful $(-)$ labels and reporting also the

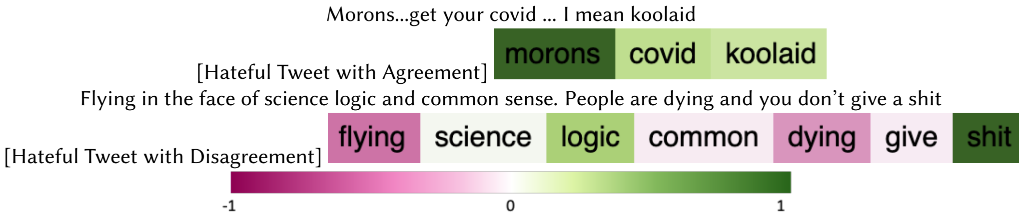


Figure 1: Visual representation of the integrated gradients on sentences from the MD-Agreement dataset. Positive values are represented with the green colour, negative values are associated with the pink colour, while the white colour is used for attribution values equal to zero.

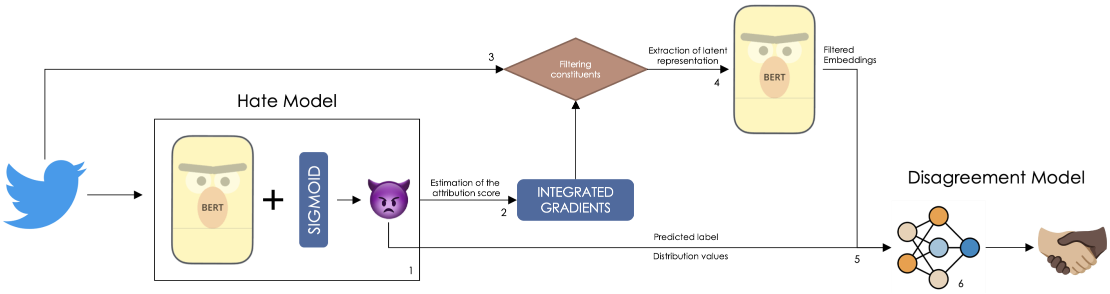


Figure 2: Proposed Approach

Macro F-Measure. We show in Table 2 the performance achieved by the fine-tuned model on the hate speech detection task. The achieved results denote good prediction capability, especially for the negative class (non-hateful). This behaviour is mainly due to the unbalanced nature of the datasets and in some cases to the limited number of instances available.

Now, we report in Table 3 the performance on the disagreement prediction, distinguishing however between agreement $(+)$ and disagreement $(-)$. The results of the proposed method are shown according to the input space previously described. In particular, we report:

- **m-BERT**: a baseline m-BERT model fine-tuned according to the disagreement label;
- **NN + Filt**: a neural network that takes as input the embedding representation of the sentence composed of the tokens selected according to the attribution scores and trained on the disagreement label. This configuration corresponds to the one described in step 5(a);
- **NN + Pred**: a neural network that takes as input the embedding representation of the sentence composed of the tokens selected according to the attribution scores with an additional Boolean feature denoting the label predicted by the fine-tuned model on the hate. This configuration corresponds to the one described in step 5(b);
- **NN + Dist**: a neural network that takes as input the embedding representation of the sentence composed of the tokens selected according to the attribution scores with two additional features denoting the probability distribution associated with the labels predicted by the fine-tuned model on the hate. This configuration corresponds to the one described in step 5(c);

In order to understand whether the proposed approaches obtain significant results compared with m-BERT, a McNemar Test has been performed. In particular, the McNemar Test has been adopted to perform a pairwise comparison between the m-BERT predictions and each of the proposed strategies according to a confidence level equal to 0.95. If a given model outperforms m-BERT and its error distribution is different compared to m-BERT, then the corresponding F1-Score is marked with a wildcard symbol $(*)$ in Table 3.

It can be easily noted that, in the majority of the considered datasets, all of the proposed approaches significantly outperform the considered baseline m-BERT. It is also interesting to highlight that, considering the datasets are even more unbalanced and with a very limited number of samples, the proposed approach NN-Dist tends to achieve more balanced performance between the two labels than the other methods. The McNemar test confirms that the NN-Dist strategy is not only the best-performing one but also that the predictions are different with respect

Dataset	P+	R+	F+	P−	R−	F−	Macro F
HS-brexit	0.37	0.78	0.58	0.97	0.84	0.90	0.70
ArMIS	0.55	0.76	0.64	0.75	0.54	0.63	0.63
ConvAbuse	0.77	0.50	0.60	0.90	0.97	0.93	0.77
MD-Agreement	0.73	0.56	0.63	0.80	0.90	0.85	0.74

Table 2
Model performance on the hate speech detection task on the test set.

Dataset	Approach	P+	R+	F+	P−	R−	F−	Macro F
HS-Brexit	m-BERT	0.85	0.69	0.76	0.51	0.73	0.60	0.68
	NN + Filt	0.69	0.86	0.83	0.62	0.50	0.55	0.69
	NN + Pred	0.79	0.86	0.83	0.62	0.50	0.55	0.69
	NN + Dist	0.84	0.78	0.81	0.57	0.67	<u>0.62</u>	**0.71**
ArMIS	m-Bert	0.60	0.27	0.37	0.32	0.65	0.43	0.40
	NN + Filt	0.64	0.93	0.76	0.50	0.11	0.18	0.47*
	NN + Pred	0.66	0.84	0.73	0.46	0.25	0.32	0.53*
	NN + Dist	0.67	0.75	0.71	0.47	0.38	<u>0.42</u>	**0.56***
ConvAbuse	m-BERT	0.87	0.99	0.93	0.33	0.03	0.05	0.49
	NN + Filt	0.94	0.59	0.73	0.23	0.76	0.25	0.54*
	NN + Pred	0.92	0.67	0.78	0.24	0.65	0.35	0.56*
	NN +Dist	0.94	0.70	0.80	0.27	0.72	<u>0.40</u>	**0.60***
MD-Agreement	m-BERT	0.43	0.34	0.38	0.58	0.68	0.63	0.50
	NN + Filt	0.47	0.71	0.57	0.67	0.43	0.53	0.55*
	NN + Pred	0.53	0.52	0.52	0.66	0.66	0.66	0.59*
	NN + Dist	0.54	0.52	0.53	0.66	0.68	<u>0.67</u>	**0.60***

Table 3
Comparison of the different approaches on the test set for disagreement detection. **Bold** denotes the best approach according to the F1-Score, while <u>underline</u> represents the best approach according to the disagreement label. (*) denotes that model outperforms M-BERT and obtains results that are statistically different.

to the ones given by m-BERT. This implies that the performance of the proposed approach could be considered statistically significant.

An additional remark concerns the relationship that exists between the disagreement prediction model and the model able to predict hateful content. The performances of the proposed models are strictly related to the recognition capabilities of the model fine-tuned to distinguish hateful content from non-hateful ones. Improving the recognition capabilities of the hateful model is expected to increase the recognition potential of the proposed disagreement models.

For what concerns the errors of the most promising approach, i.e., NN-Dist, we can highlight that on the HS-Brexit dataset, most of the misclassifications are due to the absence of relevant information. In particular, in 70% of the misclassified samples, there are references to users and links that have been omitted, making the understanding of the context even more complex. Regarding the ArMis dataset, most of the errors are related to the implicit language used to express hateful content against women (no explicit insults or sexist expressions are used, but more subtle misogynous samples are reported). In ConvAbuse, the misclassification of the proposed approach is mainly due to the reduced number of tokens of the text. In fact, 40% of the original text contains less than 3 tokens, making difficult the prediction of disagreement. Finally, in MD-Agreement the error rate is quite higher (42.79%) compared to the other datasets. In this scenario, the misclassified samples are almost balanced between the two classes, (i.e., 0.45% for the agreement and 55% for the disagreement class). The main reason behind the high classification error can be found in the different arguments covered by the dataset. This suggests that disagreement is not only related to different beliefs or backgrounds but also to specific discussed topics.

6. Conclusions and Future works

The proposed paper introduces a novel approach for detecting disagreement in hateful content. The method leverages integrated gradients from pre-trained language models to predict both hate speech and potential disagreement arising from different readers. The approach is evaluated on four benchmark datasets related to Learning

With Disagreements, and the results show that the proposed method outperforms the baseline m-BERT model in disagreement prediction tasks. One of the proposed strategies, namely NN + Dist, performs particularly well and achieves statistically significant improvements compared to a baseline model based on m-BERT. Overall, the proposed approach demonstrates the potential to predict disagreement in hateful content compared to bert. Future work could focus on exploring the applicability of the proposed approach to other languages and expanding the scope to include multimodal data analysis, considering the increasing use of images and videos in online communication.

Acknowledgments

The work of Elisabetta Fersini has been partially funded by the European Union – NextGenerationEU under the National Research Centre For HPC, Big Data and Quantum Computing - Spoke 9 - Digital Society and Smart Cities (PNRR-MUR), and by MUR under the grant "Dipartimenti di Eccellenza 2023-2027" of the Department of Informatics, Systems and Communication of the University of Milano-Bicocca, Italy.

References

[1] A. Uma, T. Fornaciari, A. Dumitrache, T. Miller, J. Chamberlain, B. Plank, E. Simpson, M. Poesio, Semeval-2021 task 12: Learning with disagreements, in: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), 2021, pp. 338–347.

[2] P. Kralj Novak, T. Scantamburlo, A. Pelicon, M. Cinelli, I. Mozetič, F. Zollo, Handling disagreement in hate speech modelling, in: International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Springer, 2022, pp. 681–695.

[3] P. Fortuna, M. Domínguez, L. Wanner, Z. Talat, Directions for nlp practices applied to online hate speech detection, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 11794–11805.

[4] E. Leonardelli, G. Abercrombie, D. Almanea, V. Basile, T. Fornaciari, B. Plank, V. Rieser, A. Uma, M. Poesio, SemEval-2023 task 11: Learning with disagreements (LeWiDi), in: Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 2304–2318.

[5] D. Grötzinger, S. Heuschkel, M. Drews, CICL_DMS at SemEval-2023 task 11: Learning with disagreements (le-wi-di), in: Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1030–1036. URL: https://aclanthology.org/2023.semeval-1.141.

[6] M. Sullivan, M. Yasin, C. L. Jacobs, University at buffalo at semeval-2023 task 11: Masda–modelling annotator sensibilities through disaggregation, in: Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), 2023, pp. 978–985.

[7] S. Shahriar, T. Solorio, SafeWebUH at SemEval-2023 task 11: Learning annotator disagreement in derogatory text: Comparison of direct training vs aggregation, in: Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 94–100. URL: https://aclanthology.org/2023.semeval-1.12.

[8] E. Gajewska, eevvgg at SemEval-2023 task 11: Offensive language classification with rater-based information, in: Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 171–176. URL: https://aclanthology.org/2023.semeval-1.24.

[9] S. Akhtar, V. Basile, V. Patti, Whose opinions matter? perspective-aware models to identify opinions of hate speech victims in abusive language detection, 2021. arXiv:2106.15896.

[10] D. Almanea, M. Poesio, ArMIS - the Arabic misogyny and sexism corpus with annotator subjective disagreements, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 2282–2291. URL: https://aclanthology.org/2022.lrec-1.244.

[11] A. Cercas Curry, G. Abercrombie, V. Rieser, ConvAbuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational AI, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 7388–7403. URL: https://aclanthology.org/2021.emnlp-main.587. doi:10.18653/v1/2021.emnlp-main.587.

[12] E. Leonardelli, S. Menini, A. Palmero Aprosio, M. Guerini, S. Tonelli, Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 10528–10539. URL: https://aclanthology.org/2021.emnlp-main.822. doi:10.

18653/v1/2021.emnlp-main.822.

[13] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, V. Patti, Resources and benchmark corpora for hate speech detection: a systematic review, Language Resources and Evaluation 55 (2021) 477–523.

[14] A. Chhabra, D. K. Vishwakarma, A literature survey on multimodal and multilingual automatic hate speech identification, Multimedia Systems (2023) 1–28.

[15] M. Mozafari, R. Farahbakhsh, N. Crespi, Hate speech detection and racial bias mitigation in social media based on bert model, PloS one 15 (2020) e0237861.

[16] H. S. Alatawi, A. M. Alhothali, K. M. Moria, Detecting white supremacist hate speech using domain specific word embedding with deep learning and bert, IEEE Access 9 (2021) 106363–106374.

[17] H. Saleh, A. Alhothali, K. Moria, Detection of hate speech using bert and hate speech word embedding with deep model, Applied Artificial Intelligence 37 (2023) 2166719.

[18] M. Fernandez, H. Alani, Contextual semantics for radicalisation detection on twitter (2018).

[19] M. Bilal, A. Khan, S. Jan, S. Musa, Context-aware deep learning model for detection of roman urdu hate speech on social media platform, IEEE Access 10 (2022) 121133–121151. doi:10.1109/ACCESS.2022.3216375.

[20] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, D. Testuggine, The hateful memes challenge: Detecting hate speech in multimodal memes, Advances in neural information processing systems 33 (2020) 2611–2624.

[21] E. Fersini, F. Gasparini, G. Rizzi, A. Saibene, B. Chulvi, P. Rosso, A. Lees, J. Sorensen, SemEval-2022 task 5: Multimedia automatic misogyny identification, in: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), Association for Computational Linguistics, Seattle, United States, 2022, pp. 533–549. URL: https://aclanthology.org/2022.semeval-1.74. doi:10.18653/v1/2022.semeval-1.74.

[22] F. Gasparini, I. Erba, E. Fersini, S. Corchs, et al., Multimodal classification of sexist advertisements., in: ICETE (1), 2018, pp. 565–572.

[23] M. Das, R. Raj, P. Saha, B. Mathew, M. Gupta, A. Mukherjee, Hatemm: A multi-modal dataset for hate video classification, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 17, 2023, pp. 1014–1023.

[24] A. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, Learning from disagreement: A survey, The Journal of Artificial Intelligence Research 72 (2021) 1385–1470. doi:https://doi.org/10.1613/jair.1.12752.

[25] B. Beigman Klebanov, E. Beigman, From annotator agreement to noise models, Computational Linguistics 35 (2009) 495–503.

[26] A. Dumitrache, F. Mediagroep, L. Aroyo, C. Welty, A crowdsourced frame disambiguation corpus with ambiguity, in: Proceedings of NAACL-HLT, 2019, pp. 2164–2170.

[27] A. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, Learning from disagreement: A survey, Journal of Artificial Intelligence Research 72 (2021) 1385–1470.

[28] T. Fornaciari, A. Uma, S. Paun, B. Plank, D. Hovy, M. Poesio, et al., Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2021.

[29] E. Leonardelli, A. Uma, G. Abercrombie, D. Almanea, V. Basile, T. Fornaciari, B. Plank, V. Rieser, M. Poesio, Semeval-2023 task 11: Learning with disagreements (lewidi), 2023. arXiv:2304.14803.

[30] W. J. Youden, Index for rating diagnostic tests, Cancer 3 (1950) 32–35.

[31] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: International conference on machine learning, PMLR, 2017, pp. 3319–3328.

Challenging specialized transformers on zero-shot classification

Serena Auriemma[1], Mauro Madeddu[1], Martina Miliani[1], Alessando Bondielli[1], Alessandro Lenci[1] and Lucia Passaro[3]

[1]Dipartimento di Filologia, Letteratura e Linguistica, Università di Pisa, Via Santa Maria, Pisa, 56126, Italy
[3]Dipartimento di Informatica, Università di Pisa, Largo B. Pontecorvo, 3 Pisa, 56127, Italy

Abstract
This paper investigates the feasibility of employing basic prompting systems for domain-specific language models. The study focuses on bureaucratic language and uses the recently introduced BureauBERTo model for experimentation. The experiments reveal that while further pre-trained models exhibit reduced robustness concerning general knowledge, they display greater adaptability in modeling domain-specific tasks, even under a zero-shot paradigm. This demonstrates the potential of leveraging simple prompting systems in specialized contexts, providing valuable insights both for research and industry.

Keywords
Domain Adaptation, Transformers, Prompting, Zero-shot, Italian Bureaucratic Language, Public Administration

1. Introduction

Pre-trained Language Models (PLMs) have had a significant impact on Natural Language Processing (NLP), and the pre-train and fine-tune paradigm has become the predominant approach for applying effective models on a wide variety of downstream tasks [1, 2, 3, *inter alia*].

However, one of the main concerns when working with PLMs is the paucity of annotated data, especially for specific domains, required to fine-tune the additional classification layer on top of these models for downstream tasks, such as classification. Recently, prompt-based tuning has started to affirm as a promising way to perform similar tasks, significantly reducing the need for annotated data. This approach has been proven to be very effective with Large Language Models (LLMs) [4]. However, it is often the case that LLMs are not available for low-resource languages, and that their performance drastically decreases when they are challenged on specific domains. Hence, we decided to test a domain-specific model, BureauBERTo [5], a LM further pre-trained on Italian bureaucratic texts (e.g., administrative acts, banking and insurance documents), in a zero-shot scenario exploiting the prompt-based tuning technique.

Since BureauBERTo has shown to be particularly accurate in the fill mask task [5],[1] where the model had to predict both random and in-domain masked words, we wanted to further inspect the domain lexical knowledge acquired by this model during the domain adaptation. We aimed at leveraging this knowledge to implement two classification tasks in the PA domain, modeled as prompt-based classification. Thus, we challenged the model to predict both the topics of PA texts, and the type of generic and PA-related named entities occurring in sentences extracted from administrative documents.

We conducted two prompting experiments for each task. We first adopted the Italian name of the classification classes as label words, then we associated in-domain terms to each class. We also compared BureauBERTo with an Italian generic PLM, UmBERTo (Section 3).

Our findings show that in a zero-shot classification scenario when the label words of each class are shallowly related to the content of the text or to the entity type fed to the model in the prompt template, both the generic and the domain-specialized models perform poorly in the classification task. However, when the classes are represented by multiple word labels semantically related to the text/entity to be classified, the PLMs improve their performance by a wide margin. This gaining is particularly evident in the domain-adapted model BureauBERTo, which outperformed UmBERTo in both prompt-document classification and prompt-entity typing tasks, suggesting that the domain linguistic knowledge acquired by this model during the additional pre-training phase could be particularly useful in a prompt-based tuning scenario where the model is much more reliant on its word knowledge, compared to when the same task is accomplished via

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
serena.auriemma@phd.unipi.it (S. Auriemma)
0009-0006-6846-5826 (S. Auriemma); 0009-0002-7844-3963 (M. Madeddu); 0000-0003-1124-9955 (M. Miliani); 0000-0003-3426-6643 (A. Bondielli); 0000-0001-5790-4308 (A. Lenci); 0000-0003-4934-5344 (L. Passaro)

CEUR Workshop Proceedings (CEUR-WS.org)

[1]See Appendix B for the plot of the model results in the fill-mask task

fine-tuning.

2. Related work

PLMs have proven to be effective in NLP tasks related to specific domains, whether they were trained from scratch [6, 7], or further pre-trained on domain data [8, 9, 10] with a Masked Language Modeling (MLM). More recently, the MLM training objective has been leveraged to solve various NLP tasks reformulated as a sort of cloze task, allowing the PLM to directly solve it without any or with very few labelled examples. One of the first works in this direction was proposed by [11], who performed zero-shot learning using pre-trained LMs without fine-tuning on a dataset of training examples. Within similar conditions, but using the larger GPT-3, [4] achieved near state-of-the-art results for some SuperGLUE [12] tasks. [13] showed that competitive performance with those of GPT-3 can be achieved with much smaller models like the 220M parameters ALBERT, by performing some gradient-based fine-tuning of the model using the labeled examples on a cloze task. Since then, prompt-based learning has gained attention as a simple way to perform, among other tasks, zero-shot classification [14]. However, it's essential to note that the performance of prompt-based learning techniques scales with model size [15]. Consequently, general purpose Large Language Models (LLMs) with billions of parameters are typically used in prompt-learning experiments, even for specialized domains such as the legal one [16]. In contrast, for the biomedical and clinical domains, [17] showed that smaller specialized models like BioBERT [8] and Clinical BERT [18] outperform GPT-2 and T5 in a few-shot prompt based classification of medical texts. The authors hypothesize that the advantage of the BERT-based models is possibly due both to their domain adaptation and to their bidirectional MLM training objective, which is more similar to the prompt template format than those of auto-regressive and sequence-to-sequence models like GPT-2 and T5. [19] reported a similar finding even for the much larger GPT-3 over BioBERT. Nevertheless, these approaches are constrained by the model input size, which limits the length of the conditioning input context and can significantly affect performance [19].

Although prompt-based classification with specialized models has been explored for the medical and clinical domains, to the best of our knowledge, this is the first work that focuses on applying prompts to the Italian administrative language and in a zero-shot classification scenario. Additionally, a notable challenge in prompt-based approaches lies in their sensitivity to variations in prompt templates and verbalizers [20, 21, 22]. We conducted experiments using different verbalizers, i.e., a generic verbalizer and a custom verbalizer using domain-specific terms, to investigate how domain-related word labels affect the model's performance in different classification tasks.

3. Models

For our experiments, we decided to compare the performance of two PLMs, namely UmBERTo and BureauBERTo. UmBERTo[2] is a RoBERTa-based language model trained on the Italian section of the OSCAR corpus,[3] that has been shown to perform well on administrative data [23] compared to other generic PLMs of the same size (110M parameters). BureauBERTo[4] [5] is a domain-adapted model obtained by further pre-training UmBERTo on Italian PA, banking, and insurance documents.

4. Experimental settings

4.1. Prompting

Prompt-based classification requires a specific template to reformulate the original classification task as a cloze-task, where the text to be classified is fed to the model followed by a prompt sentence, such as "This `<text>` is about [MASK]". In this way, the model has to predict the probability that a certain word is filled in the "[MASK]" token. The mapping from the label candidate word to a specific class is gained through the *verbalizer* [13], which represents the original class names as a set of label words, greatly influencing the model performance in the task [24]. Hence, we decided to conduct our prompt-based classification experiments in two settings, using a standard and a custom *verbalizer* to better understand the correlation between the lexical knowledge of PLMs and the use of domain-related terms as the set of word labels in the prompt *verbalizer*.

The first verbalizer, i.e., the *base-verbalizer* simply uses the Italian name of the classification classes as label words (e.g., *Ambiente - "Environment"* is the label word for the class AMBIENTE - "ENVIRONMENT"), while the second verbalizer is a *manual verbalizer* that we constructed by adding some synonyms of the class name and some related PA terms as label words for each class, to better depict the document classes and the entity types (in this case the label words for the class AMBIENTE - "ENVIRONMENT" are: *ambiente - "environment", natura - "nature", territorio - "territory", flora - "flora",etc.*).

[2]https://github.com/musixmatchresearch/umberto
[3]https://oscar-corpus.com
[4]https://huggingface.co/colinglab/BureauBERTo

4.2. Datesets

We evaluate the models in two tasks on two different datasets. For the **prompt document classification**, we used a subset of the ATTO corpus [23], which is a collection of administrative documents annotated with labels denoting topics. We filtered this dataset keeping only those instances (2,811) that were annotated with a single topic label.

For the **prompt entity typing** task, we used the PA-corpus of [25], a collection of 460 PA-documents with token-level annotations of Named Entities denoting both general entities, such as persons, locations, organizations, and domain-specific entities, like legislative norms, acts, and PA-related organizations.

4.3. Evaluation metrics

We evaluated the performance of the models with common classification metrics.

4.4. Prompt entity typing

We modeled the NER task introduced by [25] as an entity typing task. Entity typing can be considered a subtask of NER and focuses on entity classification. In other words, systems assign a label to an already extracted entity. This task is often formulated to challenge systems at retrieving sub-categories organized in a hierarchical structure (e.g., an entity corresponding to a person may be specified as director, major, lawyer, etc.) As in [25], we asked models to identify only coarse-grained entities: generic ones, such as persons (PER), locations (LOC), and organizations (ORG); and related to the administrative domain: law references (LAW), administrative acts (ACT), and PA organizations (OPA).

We prompted the models by giving as input a sentence and an entity occurring in it, asking to predict the entity type in place of a masking token. The resulting template is: `<text>. In questa frase, <entity> è un esempio di <mask>.`[5]

As anticipated, we verbalized the entities in two ways. In the first experiment, we provided an Italian translation of the entity or a single word representing the entity class. In the second experiment, we expanded most of the label words by including synonyms and other terms related to the various classes.[6]

4.5. Prompt document classification

For the recognition of the topics in PA documents, we designed the following template to model the document classification task as a masked language modeling problem: `<text>.Questo documento parla di <mask>.`[7]

Thus, PLMs are challenged to infer the topic of the document by predicting the most appropriate label word to represent the masked token in the prompt, following the document text. Since the ATTO corpus contains only short documents of a maximum of 600 tokens, by setting the tokenizer's truncation at 512 tokens[8], we were able to feed the models the entire document in almost all cases. Like with the prompt entity typing, we perform the prompt-based classification twice. In the first experiment, we used the *base verbalizer*, where each class is linked to one or few label words that correspond to the names of the classes in the original annotation of the ATTO corpus. For the second experiment, we use the *manual verbalizer*, which contains, in addition to the label words of the *base verbalizer*, a collection of domain terms manually selected as PA representative topic labels for each class. The complete list of the label words used in both verbalizers is shown in Table 1.[9]

5. Results and discussion

Table 2 shows the results of prompting applied to the entity typing task.

In the first experiment, where a single class label is used (see Sec. 4.4), UmBERTo almost doubled the results obtained by BureauBERTo for F1 Micro (0.404 vs. 0.263) and Macro Average (0.335 vs. 0.201). Surprisingly, for a domain entity like ACT, BureauBERTo missed all the entities, whereas UmBERTo obtained a low but higher score (0.140). For the LAW entity, UmBERTo overpasses BureauBERTo, as well. We may suppose that this is due to the fact that UmBERTo was trained on Common-Crawl, which also contains legal and administrative texts in its Italian section. Very high results are obtained by UmBERTo for PER entities, reaching 0.827 in our zero-shot scenario. On the contrary, both models obtain very low results for LOC, OPA, and ORG. These two latter classes are very similar to each other: ORG refers to organizations in general, comprising firms and associations, whereas OPA can be considered as a subclass of ORG, and refers to organizations within the Public Administration, such as municipal departments. Such overlapping may impact on classification.

For what concerns the second experiment, we added to the prompt also highly distinctive words for each class. In this case, we notice a better ability of BureauBERTo to recognize domain-specific entities such as ACT, LAW,

[5] In English: `<text>. In this sentence, <entity> is an example of <mask>.`

[6] Both verbalizers for entity typing are in Appendix A.

[7] In English: `<text>. This document is about <mask>.`

[8] 512 is the maximum number of tokens that these Transformers models can receive as input.

[9] See Appendix A for the English translation of the label words for document classification.

Table 1
The Table shows the label words adopted in the experiments of prompt document classification.

Class	Basic Labels	+In-domain Lexicon
Ambiente	ambiente	ambiente, natura, territorio, flora, fauna, animali, clima, inquinamento, rifiuti, igiene, caccia, pesca, verde, ecologia, agricoltura, acque
Avvocatura	avvocatura	avvocatura, avvocati, giustizia, legale, ricorso, giudici, Tribunale, Corte, Appello, Assise, notifica, atti, Albo, Pretorio, protocollo
Bandi-Contratti	bandi, contratti	bandi, contratti, bando, contratto, gara, appalto, assunzione, liquidazione
Commercio-Attività-economiche	commercio, attività, economiche	commercio, economia, attività, economica, beni, commerciare, vendite, acquisti, commercianti, confesercenti
Cultura-Turismo-sport	cultura, turismo, sport	cultura, turismo, sport, culturale, turisti, musei, arte, cinema, vacanze, spettacolo, scuola, manifestazioni
Demografico	demografico	demografico, popolazione, abitanti, residenti, censimento, anagrafe, residenza, domicilio, cittadinanza, leva
Edilizia	edilizia	edilizia, costruzioni, cantiere, ristrutturazione, planimetrie, residenziale
Personale	personale	personale, risorse, umane, assunzioni, lavoro, part-time
Pubblica-Istruzione	istruzione	istruzione, istituto, scolatisco, scuola, insegnante, formazione, educazione
Servizi-Informativi	servizi, informazioni	servizi, informazioni, informativi
Servizio-Finanziario	finanza	finanza, euro, finanziario, contabilità, contabile, copertura, rimborsi, pagamenti, versamenti, bilancio, spese, sanzioni, multe, tributi, retribuzioni, emolumenti
Sociale	sociale	sociale, leva, militare, disabili, protezione, civile, invalidi
Urbanistica	urbanistica	urbanistica, trasporti, trasporto, traffico, circolazione, veicoli, viabilità, viaria

and OPA. However, despite the general improvement in recognizing such classes, we notice that it performs worse than UmBERTo for traditional entities. This experiment based on the comparison of general-purpose language models and domain-adapted ones has yielded compelling insights. Generally, both types of models demonstrate enhanced performance when enriched with domain-specific terms within their prompts. However, it is evident that the domain-adapted model outperforms the general-purpose model, exhibiting an improvement of more than twofold (0.516 vs 0.368 for Macro Average F1 score). This significant boost in performance suggests that the domain-adapted model is likely to be more attuned and proficient in leveraging domain-specific terminology.

Nevertheless, it is important to acknowledge that domain-specific terms may wield less influence over generic entities such as PER. With the in-domain lexicon added to the verbalizer, UmBERTo fails to recognize any PER entity. By looking at the confusion matrix for UmBERTo, we observed that the model identifies almost all the people's names as ORG entities. Thus, we carried out an ablation study by deleting the in-domain terms added for the PER entity class, i.e. *generalità - "particulars"* and *nominativo - "name"*.

The results in Table 3 show that the performance of UmBERTo increases not only for the PER entities but that the ablation improves the F1-score of the ORG class as well. Whereas UmBERTo reaches the highest performances for overall F1 Micro Avg, the deletion of in-domain lexicon from the verbalizer seems to penalize BureauBERTo in the recognition of PER entities. Following the trend observed in UmBERTo, the ablation impacts the model's ability to properly recognize the other classes. Despite this, the adapted model still obtained higher results on the in-domain entity classes: ACT, LAW, and OPA further solidifying the advantages of domain-adapted models in specialized contexts. Finally, it is worth noting that we observed a high variability of results according to different prompts and verbalizer configurations, as shown in the ablation study. In fact, deleting the in-domain lexicon related to one of the entity classes affected the performance achieved by the models on all the others, due to wrong classifications (e.g., people names confused with location addresses or company names). Therefore, future investigations into prompt tuning are necessary and can lead to further interesting insights.

Regarding the prompt document classification experiments, whose results are summarized in table 4, we observed a similar trend. When only one or few word labels are used to represent a topic class, both the generic and the domain-specialized models obtained a rather low accuracy (0.22 vs. 0.09) and Macro Average F1 scores (0.16 vs. 0.06). In this case, UmBERTo outperformed BureauBERTo in almost all classes, with the exception of Cultura, Turismo e Sport - 'Culture, tourism, and sports', Demografico - 'Demographics', and Personale - 'Personnel'. Looking into the details of the scores obtained by UmBERTo in its most recognizable classes (Pubblica istruzione - 'Public Education', Edilizia - 'Constructions' and Urbanistica - 'Urban planning'), we speculate that the single-word labels used to define these classes provided a sufficient cue to enable the model to appropriately recognize these topics. This is in line with the fact that the UmBERTo pre-training corpus included texts extracted from Italian municipalities' web pages, which often refer to such topics.

On the other hand, in the second experiment, where we manually added to the prompt *verbalizer* a set of salient PA-related terms to depict the document topics at a finer-grained level, we observed a significant improvement in the overall performance of both models. The benefits of a custom-made set of domain-related terms are particularly evident for the specialized model BureuBERTo, which reached a better accuracy (0.60 vs. 0.54) and Weighted Average F1-score (0.57 vs. 0.51) than UmBERTo. It appears that the model adapted to the domain may possess heightened sensitivity, enabling it to effectively capitalize on the contextual cues offered by domain-specific terms. However, by performing a class-wise comparison between the two experimental settings, we observed that for some classes that shared a common domain lexicon, such as Pubblica Istruzione - 'Public Education' and Cultura, Turismo e Sport - 'Culture, Tourism, and Sports', or Servizi finanziari - 'Financial services' and Bandi e Contratti - 'Tenders and Contracts' the models' classification could have been influenced in favor of one of the two classes, due to their topic descriptor lexical overlap. These findings confirm the necessity of further inquiry into the effect of lexical specificity on prompt-based classifications, especially for domain-adapted models.

6. Conclusion and future work

In this paper, we propose a zero-shot prompt tuning classification approach for solving two tasks related to the Italian PA domain: the classification of documents according to their topic and the recognition of the entity types occurring in administrative sentences.

We compared the results obtained in these two tasks by the PA-specialized model BureauBERTo with those of the domain-agnostic model UmBERTo. Our findings show that by enriching with domain terms the set of word labels encoded in the prompt *verbalizer* both models demonstrated enhanced performances. Moreover, BureauBERTo exhibited an improvement over UmBERTo of +0.06 Weighted Average F1 score in the document classification (0.51 vs. 0.57) and of more than twofold in the entity typing task (0.516 vs. 0.368 for Macro Average F1 score), meaning that the domain adapted model is more proficient in leveraging domain-specific terminology.

These results underscore the importance of tailoring language models to specific domains to unlock their full potential and address the nuanced challenges posed by diverse subject matters. However, it is also worth mentioning that we noticed a high variability in the task results according to different prompting and different label words. In particular, when the label words adopted to depict a certain topic class are, within the domain context, semantically related to the label words of another class, the models' classification output seems to be biased in favor of one of the two classes.

In conclusion, our study underscores the critical need for a thorough exploration of prompt engineering, particularly in the context of the entity typing task. This imperative arises not only from the potential to augment the predictive capabilities of models, but also from the need to consolidate the knowledge related to general entity classes. Notably, the Public Administration (PA) domain exhibits distinctive characteristics, both in terms of referencing entity names within documents and employing domain-specific terminology. Notably, the identified patterns within the PA domain deviate from the broader, general-purpose Italian style, indicating the necessity for tailored, domain-specific prompt experimentation.

This investigative effort shed the linguistic intricacies that exert an impact on Transformer model performance. Our findings, as revealed in the ablation study on entity linking, emphasize the pivotal importance of delving into the interplay among different entity classes present in datasets. A nuanced analysis of how these classes interact and potentially overlap is indispensable for honing the model's ability to distinguish between them in a domain-specific context.

To conclude, this leads us to surmise as a future direction for our work a further inspection of how domain-adapted PLMs encode in their embedding the semantics of domain-related terms and how this information relates to their performance in prompt-based tasks.

Table 2
Performance comparison of UmBERTo and BureauBERTo on the entity typing task. We grouped together generic entities (LOC, ORG, PER) and domain-related entities (ACT, LAW, OPA). In the upper part of the table are the results of the first experiment, with a unique word as a label. In the bottom part, we report the results for the second experiment where we used multiple labels for each entity class. In bold the best results for each experiment. The best overall results are underlined.

Model	Measure	LOC	ORG	PER	ACT	LAW	OPA	MicAvg	MacAvg
		Basic Labels							
UmBERTo	P	0.7	0.181	0.836	0.4	0.455	0.818	0.462	0.565
	R	0.045	0.372	0.818	0.085	0.618	0.107	0.36	0.341
	F1	**0.085**	<u>**0.244**</u>	<u>**0.827**</u>	**0.140**	**0.524**	**0.189**	**0.404**	**0.335**
BureauBERTo	P	0.5	0.115	0.774	0	0.294	1	0.323	0.447
	R	0.013	0.223	0.526	0	0.447	0.024	0.221	0.205
	F1	0.025	0.152	0.626	0	0.355	0.047	0.263	0.201
		+In-domain Lexicon							
UmBERTo	P	0.767	0.11	0	0.63	0.6	0.364	0.421	0.412
	R	0.445	0.234	0	0.309	0.756	0.476	0.368	0.370
	F1	<u>**0.563**</u>	0.15	0	0.414	0.669	0.412	0.393	0.368
BureauBERTo	P	0.814	0.178	0.45	0.521	0.727	0.797	0.534	0.581
	R	0.368	0.245	0.555	0.404	0.756	0.607	0.492	0.489
	F1	0.507	**0.206**	**0.497**	<u>**0.455**</u>	<u>**0.741**</u>	<u>**0.689**</u>	<u>**0.512**</u>	<u>**0.516**</u>

Table 3
Ablation study conducted on BureauBERTo and UmBERTo on entity typing task. In bold are the best results for each entity class.

Model	Measure	LOC	ORG	PER	ACT	LAW	OPA	MicAvg	MacAvg
UmBERTo	P	0.792	0.294	0.482	0.634	0.728	0.536	0.569	0.578
	R	0.368	0.266	0.577	0.277	0.805	0.536	0.482	0.471
	F1	**0.502**	**0.279**	**0.525**	0.385	0.764	0.536	**0.522**	0.499
BureauBERTo	P	0.746	0.280	0.385	0.629	0.793	0.750	0.573	0.597
	R	0.303	0.223	0.453	0.415	0.780	0.571	0.456	0.458
	F1	0.431	0.249	0.416	**0.500**	**0.787**	**0.649**	0.508	**0.505**

Acknowledgments

This research has been funded by the Project "ABI2LE (Ability to Learning)", Regione Toscana (POR Fesr 2014-2020); by PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI", funded by the European Commission under the NextGeneration EU programme; and partially supported by: TAILOR, a project funded by EU Horizon 2020 research and innovation programme under GA No 952215.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[2] J. Howard, S. Ruder, Universal language model fine-tuning for text classification, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 328–339. URL: https://aclanthology.org/P18-1031. doi:10.18653/v1/P18-1031.

[3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19

Table 4
Performance comparison of UmBERTo and BureauBERTo on the document classification task. On the left side of the table are the results of the first experiment, where we employed basic label words. On the right side are the results for the second experiment where we used multiple labels for each class. In bold the best results for each experiment. The best overall results are underlined. C-A-E refers to Commercio-Attività-Economiche, whereas C-T-S stands for Cultura-Turismo-Sport.

	Basic Labels						+In-domain Lexicon					
	P		R		F1		P		R		F1	
Class	UmB	BB	UmB	BB	UmB	BB	UmB	BB	UmB	BB	UmB	BB
Ambiente	1.00	0.91	0.04	0.03	**0.08**	0.07	0.79	0.86	0.37	0.38	0.51	**0.52**
Avvocatura	0.00	0.00	0.00	0.00	0.00	0.00	0.50	1.00	0.02	0.03	0.03	**0.07**
Bandi-contratti	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.11	0.36	0.31	0.09	**0.16**
C-A-E	0.00	0.00	0.12	0.00	**0.01**	0.00	0.50	1.00	0.38	0.25	**0.43**	0.40
C-T-S	0.62	0.44	0.07	0.12	0.13	**0.19**	0.61	0.41	0.21	0.40	0.31	**0.40**
Demografico	0.00	0.06	0.00	0.98	0.00	**0.12**	0.73	0.64	0.43	0.28	**0.54**	0.39
Edilizia	0.56	1.00	0.31	0.03	**0.40**	0.06	0.88	0.74	0.11	0.29	0.20	**0.41**
Personale	0.10	0.45	0.51	0.22	0.16	**0.29**	0.38	0.70	0.71	0.45	0.49	**0.55**
Pubblica-istruzione	0.67	0.00	0.49	0.00	**0.57**	0.00	0.25	0.14	0.03	0.07	0.05	**0.09**
Servizi-informativi	0.02	0.00	0.53	0.00	**0.03**	0.00	0.06	0.02	0.47	0.27	**0.10**	0.03
Servizio-finanziario	0.58	0.00	0.19	0.00	**0.28**	0.00	0.89	0.54	0.49	0.90	0.63	**0.67**
Sociale	0.20	0.00	0.03	0.00	**0.06**	0.00	1.00	0.00	0.02	0.00	**0.04**	0.00
Urbanistica	0.45	0.50	0.30	0.00	**0.36**	0.00	0.62	0.83	0.98	0.97	0.76	**0.89**
Accuracy	-	-	.	.	**0.22**	0.09	-	-	-	-	0.54	**0.60**
Macro Avg	0.32	0.26	0.20	0.11	**0.16**	0.06	0.56	0.54	0.35	0.35	0.32	**0.35**
Weighted Avg	0.45	0.37	0.22	0.09	**0.24**	0.05	0.69	0.65	0.54	0.60	0.51	**0.57**

-1423. doi:10.18653/v1/N19-1423.

[4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[5] S. Auriemma, M. Madeddu, M. Miliani, A. Bondielli, L. C. Passaro, A. Lenci, BureauBERTo: adapting UmBERTo to the Italian bureaucratic language, in: F. Falchi, F. Giannotti, A. Monreale, C. Boldrini, S. Rinzivillo, S. Colantonio (Eds.), Proceedings of the Italia Intelligenza Artificiale - Thematic Workshops co-located with the 3rd CINI National Lab AIIS Conference on Artificial Intelligence (Ital IA 2023), volume 3486 of *CEUR Workshop Proceedings*, CEUR-WS.org, Pisa, Italy, 2023, pp. 240–248. URL: https://ceur-ws.org/Vol-3486/42.pdf.

[6] I. Beltagy, K. Lo, A. Cohan, Scibert: A pretrained language model for scientific text, arXiv preprint arXiv:1903.10676 (2019).

[7] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, ACM Transactions on Computing for Healthcare (HEALTH) 3 (2021) 1–23.

[8] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2020) 1234–1240.

[9] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, I. Androutsopoulos, Legal-bert: The muppets straight out of law school, arXiv preprint arXiv:2010.02559 (2020).

[10] D. Licari, G. Comandè, Italian-legal-bert: A pre-trained transformer language model for italian law, in: CEUR Workshop Proceedings (Ed.), The Knowledge Management for Law Workshop (KM4LAW), 2022.

[11] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners (2019).

[12] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, S. Bowman, Superglue: A stickier benchmark for general-purpose language understanding systems, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 32, Curran

Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/4496bf24afe7fab6f046bf4923da8de6-Paper.pdf.

[13] T. Schick, H. Schütze, It's not just size that matters: Small language models are also few-shot learners, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 2339–2352. URL: https://aclanthology.org/2021.naacl-main.185. doi:10.18653/v1/2021.naacl-main.185.

[14] R. Puri, B. Catanzaro, Zero-shot text classification with generative language models, Computing Research Repository (CoRR) abs/1912.10165 (2019). URL: http://arxiv.org/abs/1912.10165. arXiv:1912.10165.

[15] B. Lester, R. Al-Rfou, N. Constant, The power of scale for parameter-efficient prompt tuning, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 3045–3059.

[16] F. Yu, L. Quartey, F. Schilder, Legal prompting: Teaching a language model to think like a lawyer, arXiv preprint arXiv:2212.01326 (2022).

[17] S. Sivarajkumar, Y. Wang, Healthprompt: A zero-shot learning paradigm for clinical natural language processing., in: AMIA... Annual Symposium proceedings. AMIA Symposium, volume 2022, 2022, pp. 972–981.

[18] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, M. McDermott, Publicly available clinical bert embeddings, arXiv preprint arXiv:1904.03323 (2019).

[19] M. Moradi, K. Blagec, F. Haberl, M. Samwald, Gpt-3 models are poor few-shot learners in the biomedical domain, arXiv preprint arXiv:2109.02555 (2021).

[20] E. J. Hu, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al., Lora: Low-rank adaptation of large language models, in: International Conference on Learning Representations, 2021.

[21] Y. Lu, M. Bartolo, A. Moore, S. Riedel, P. Stenetorp, Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 8086–8098.

[22] D. Trautmann, A. Petrova, F. Schilder, Legal prompt engineering for multilingual legal judgement prediction, arXiv preprint arXiv:2212.02199 (2022).

[23] S. Auriemma, M. Miliani, A. Bondielli, L. C. Passaro, A. Lenci, Evaluating pre-trained transformers on italian administrative texts, in: Proceedings of 1st Workshop AIxPA (co-located with AIxIA 2022), 2022.

[24] T. Gao, A. Fisch, D. Chen, Making pre-trained language models better few-shot learners, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 3816–3830. URL: https://aclanthology.org/2021.acl-long.295. doi:10.18653/v1/2021.acl-long.295.

[25] L. C. Passaro, A. Lenci, A. Gabbolini, Informed PA: A NER for the italian public administration domain, in: R. B. andMalvina Nissim, G. Satta (Eds.), Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017), Rome, Italy, December 11-13, 2017, volume 2006 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2017. URL: https://ceur-ws.org/Vol-2006/paper048.pdf.

A. Label Words

Table 5 shows the verbalizer for entity typing. Table 6 contains the English version of the verbalizer adopted for the document classification (see Table 1 for the Italian version).

Table 5
The Table shows the label words adopted in the experiments of prompt entity typing.

Class	Basic Labels	+In-domain Lexicon
PER	persona	persona (person), generalità (particulars), nominativo (name)
LOC	luogo	luogo (place), località (locality)
ORG	organizzazione	organizzazione (organization), azienda (firm), società (corporation), associazione (association), compagnia (company)
LAW	legge	legge (law), norma (rule), decreto (decree), legislativo (legislative)
ACT	atto	atto (act), delibera (resolution), determina (decision), deliberazione (deliberation), regolamento (regulation)
OPA	ufficio	ufficio (office)

Table 6
The Table shows the label adopted in two experiments related to Document Classification. This is an English translation of Table 1. Although some Italian words are translated as multi words word labels can be represented as single words only.

Class	Basic Labels	+In-domain Lexicon
Environment	environment	environment, nature, land, flora, fauna, animals, climate, pollution, waste, hygiene, hunting, fishing, green, ecology, agriculture, water
Advocacy	advocacy	advocacy, attorneys, justice, legal, appeal, judges, courthouse, court, appello, assise, notification, acts, albo, pretorio, protocol
Tenders-Contracts	tenders, contracts	tenders, contracts, notice, contract, tender, hiring, liquidation
Trade-Economic-Activities	trade, economic, activities	trade, economy, business, economic, goods, trade, sales, purchases, merchants, confesercenti
Culture-Turism-Sport	culture, turism, sport	culture, tourism, sports, cultural, tourists, museums, art, cinema, vacations, entertainment, school, events
Demographic	demographic	demographics, population, inhabitants, residents, census, registry, residence, domicile, citizenship, conscription
Building	building	building, construction, yard, renovation, planimetry, residential
Personnel	personnel	personnel, resources, human, hiring, work, part-time
Education	education	education, institute, school, teacher, training, education
Information-Services	services, information	services, information, informative
Financial-Services	finance	finance, euro, financial, accounting, accountant, coverage, refunds, payments, disbursements, budget, expenses, penalties, fines, taxes, wages, emoluments
Welfare	welfare	welfare, conscription, military, disabled, protection, civilian, disability
Urban-Planning	urban planning	urban planning, transportation, transports, traffic, circulation, vehicles, roadway

B. Fill-mask results

Preliminary experiments on a fill-mask task (Fig.1) showed that BureauBERTo outperformed UmBERTo when predicting masked words on Public Administration documents [5]. This motivated us to evaluate BureauBERTo domain-specific knowledge in an unsupervised setting in prompt-based zero-shot classification tasks.

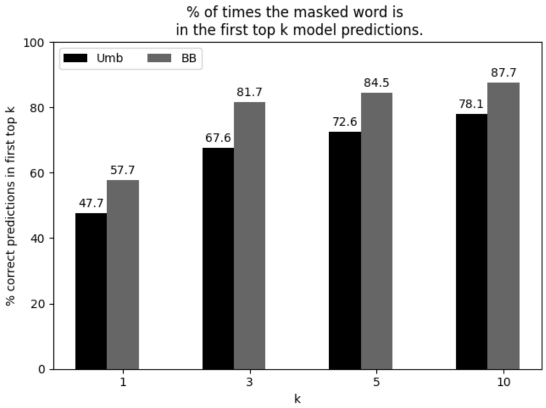


Figure 1: Results of a fill-mask task experiment in which [5] masked domain-specific words in sentences from the ATTO corpus (PA domain). Percentages indicate the number of times the masked word was in the model's top k predictions.

Bias Mitigation in Misogynous Meme Recognition: A Preliminary Study

Gianmaria Balducci[1,2], Giulia Rizzi[1,3] and Elisabetta Fersini[1,*]

[1]University of Milano-Bicocca, Milan, Italy
[2]PMI Reboot S.r.l., Milan, Italy
[3]Universitat Politècnica de València, Valencia, Spain

Abstract
In this paper, we address the problem of automatic misogynous meme recognition by dealing with potentially biased elements that could lead to unfair models. In particular, a bias estimation technique is proposed to identify those textual and visual elements that unintendedly affect the model prediction, together with a naive bias mitigation strategy. The proposed approach is able to achieve good recognition performance characterized by promising generalization capabilities.

Keywords
Bias Mitigation, Bias Estimation, Misogyny Identification, Meme

1. Introduction

In the context of social media, memes have become popular as a means of expressing irony or opinions on various topics. However, these memes can also perpetuate discriminatory behaviours towards certain groups and minorities. Misogyny, in particular, has gained attention as a form of ***hateful language*** conveyed through memes in various ways, such as female stereotyping, shaming, objectification, and violence. While misogyny recognition mechanisms have been widely investigated focusing on textual sources (i.e., tweets) [1, 2, 3, 4], misogynous identification in multimodal settings, and in particular on memes, is still in its infancy. In [5], a few naive unimodal and multimodal approaches have been investigated to understand the contribution of textual and visual cues. Further investigations from the same authors [6] have introduced a multimodal approach that considers both visual (in the form of captioning) and textual information to distinguish between misogynous and non-misogynous memes. Recently, the performance of multiple pre-trained and trained from scratch models have been compared to verify if domain-specific pre-training could help to improve the recognition performance [7].

Independently on the textual, visual or multimodal sources, several authors highlighted how the classification models may be subject to ***bias*** that could affect the real performance of the models [8, 9] in a real setting. Most of the investigations propose a few bias estimation metrics and related mitigation policies that are based on a fixed set of seed words to quantify and minimize the bias at the dataset or model level. When dealing with misogynous memes recognition, metrics to estimate the bias and techniques to mitigate it are still missing.

To this purpose, we provide the following main contributions:

- a *candidate biased elements identification* in a multi-modal setting, focusing on both textual and visual constituents of a meme;
- a mitigation strategy at training time, named *Masking Mitigation*, that masks the candidate biased elements to reduce the distortion introduced by their presence.

The rest of the paper is organized as follows. In Section 2 a summary of the state of the art is reported. In Section 3 the candidate biased element identification strategy is detailed. In Section 4 the proposed mitigation strategy is presented. In Section 5 the experimental results are reported. In Section 6 conclusion are reported.

2. Related work

The majority of works on hate content detection focus on tweets, while, only in recent years, they have started to address multimodal content such as memes. For instance, the approach proposed in [5] aims to counter the phenomenon of memes that can convey sexist messages ranging from stereotyping women to shaming, objectification, and violence, investigating both unimodal and multimodal approaches to understand the contribution of textual and visual cues. In [10], the authors indicate how the visual mode may be much more informative

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
*Corresponding author.
g.balducci1@campus.unimib.it (G. Balducci); g.rizzi10@campus.unimib.it (G. Rizzi); elisabetta.fersini@unimib.it (E. Fersini)
0000-0002-0619-0760 (G. Rizzi); 0000-0002-8987-100X (E. Fersini)

for detecting hate speech than the linguistic mode in memes. More recently, two benchmark datasets have been proposed to facilitate the investigation related to misogynous meme detection. The first benchmark presented in [11] is composed of 800 memes from the most popular social media platforms. The dataset has been labelled through a crowdsourcing platform, involving 60 subjects, in order to collect three evaluations for each instance. Each instance, labelled according to misogyny, aggressiveness and irony, has been labelled by three annotators from the crowd and three expert labellers. A more recent benchmark has been collected for *MAMI* shared task at SemEval 2022 [12]. The dataset, composed of 10.000 memes for training and 1.000 memes for testing, allowed to approach: (i) the identification of misogynistic memes, and (ii) the recognition of the type of misogyny among potential overlapping categories. For the MAMI challenge, most of the participants [13, 14, 15, 16] exploited pre-trained models and ensemble strategies.

Regarding the potential bias that the models could inherit from the training dataset, most of the investigations focus only on a unimodal setting and more precisely on the textual component [17, 18, 19]. In particular, special attention has been devoted to *identity terms*, i.e. those terms frequently associated with hateful expressions in the dataset referred to a specific target (e.g., woman, wife, girlfriend, etc...). It has been demonstrated that such identity terms lead the models to biased implicit associations between such terms and a given class label, finally originating unfair predictions. In order to counteract the potential bias, several mitigation strategies have been proposed in the literature. One of the most widely used strategies is data augmentation [4, 20, 21], which consists in adding data containing examples of non-toxic comments that bring back those identity terms that have the most disproportionate distribution in the dataset. Alternative solutions are focused on mitigating directly the models by means of specific objective functions [22, 23] or optimization strategies [24, 25, 26]. Although the above-mentioned strategies represent a fundamental step towards bias mitigation, they are defined for unimodal settings. Bias estimation and mitigation for multimodal perspective are still missing for misogynous meme identification.

3. Bias Estimation

In order to understand if a given misogyny identification model is biased, three main steps are performed: (i) Candidate Biased Elements Estimation, which allows us to identify specific textual or visual elements that could lead a model to unfair predictions, (ii) the creation of a Synthetic Dataset with specific characteristics that allow evaluating models behaviours in challenging examples, and (iii) the definition of a metric to quantify how a model could be biased from such elements. The proposed method has been evaluated on the MAMI Dataset [12] consisting of 10.000 memes for training and 1.000 memes for testing. The MAMI test set will later be referred to as *raw*.

3.1. Candidate Bias Elements Estimation

As highlighted in the literature, classification models may be affected by bias: the presence of specific elements can lead the model to an erroneous behaviour by predicting a specific label due to the presence of such elements. This distortion in the investigated data-derived models can be in fact caused by an imbalance distribution, in relation to the prediction label, of specific terms or visual elements strongly associated with a given class label. Those *candidate biased elements* can be distinguished in *candidate biased terms*, which are related to the superimposed text of a meme, and *candidate biased tags*, which are concerned with the objects that describe the scene of a meme. We exploit a novel estimation for identifying candidate biased elements [26] that overcomes the limitations of the Polarized Weirdness Index (PWI) [27], which is unbounded and does not consider the context in which the elements appear, and extended the estimation process to address more than one modality.

Given a multimodal dataset $\mathcal{D}$, e is a visual or textual element belonging to the set $\mathcal{T}$ that comprises all the terms and tags of D. A bias score $S(e)$ can be estimated for each element e according to the following formula:

$$\text{S(e)} = \frac{1}{|\mathcal{M}|}\sum_{m=1}^{|\mathcal{M}|} P(c^{+} \mid T_m) - P(c^{+} \mid \{T_m - e\}) \quad (1)$$

where $\mathcal{M}$ is the set of memes containing e, c^{+} represents the misogynous label and T_m denotes the set of terms and tags in a given meme m. $P(c^{+} \mid T_m)$ represents the probability of a meme m of being associated with the misogynous label, given the terms and tags T_m within the meme itself, and, analogously, $P(c^{+} \mid \{T_m - e\})$ denotes the probability of a meme m of being associated with the misogynous label c^{+}, given the text (tags) present in the instance (meme), excluding the evaluated element e except for the term (tag) in analysis. The proposed bias score ranges into the interval $[-1; +1]$. The higher positive the score, the more likely the element would induce bias towards the positive class (misogynous). On the other hand, the lower negative the score, the more likely the element would be associated with the negative class (not misogynous). Terms and tags with a score close to zero, are considered neutral with respect to a given label.

We report in Tables 1 and 2 the set of biased terms and biased tags identified on the MAMI training dataset. As we can see, the set of candidate biased terms with the highest score for the misogynous class is composed of words that are typically associated with some specific misogyny categories like *dishwasher* and *chick* for stereotype and *whore* for objectification. The remaining tokens are websites that have been used to collect only misogynous memes. A few terms identified as convey potential bias are related to the seed words used to collect the dataset (e.g. *whore*), confirming the ability of the proposed approach to capture the bias introduced in the dataset-creation phase (*Selection Bias*). On the other hand, the presence of other terms (e.g. *chloroform*) highlights the ability of the proposed approach to generalize with respect to the dataset creation process and include elements that may induce bias due to their unintended unbalanced distribution. Concerning the set of terms with the highest negative bias score for the not misogynous class, it is composed of words that are very general and commonly used in a variety of popular memes. An analogous consideration can be drawn for the candidate biased tags.

Candidate Biased Terms			
Misogynous		Not Misogynous	
Term	Score	Term	Score
demotivational	0.39	mcdonald	-0.26
dishwasher	0.38	ambulance	-0.24
promotion	0.35	communism	-0.23
whore	0.35	anti	-0.21
chick	0.34	valentine	-0.20
motivate	0.33	developer	-0.20
chloroform	0.30	template	-0.20
blond	0.30	weak	-0.19
diy	0.30	zipmeme	-0.18
belong	0.28	identify	-0.17

Table 1
Top-10 candidate biased terms.

3.2. Synthetic Dataset

In order to measure the bias of the models when making predictions, a *synthetic dataset* has been created with specific characteristics that can effectively help to highlight the bias of the models given the presence of the candidate biased elements.
In particular, let E_t^+ and E_o^+ be respectively the set of all the biased candidate terms and tags with a positive score, which qualifies elements that are expected to introduce the bias towards the misogynous class. Also, let E_t^- and E_o^- be respectively the set of all the biased candidate terms and tags with a negative score, which qualifies elements that are expected to introduce the bias towards the not misogynous class. Given a specific element $e_t^+ \in E_t^+$ and $e_o^+ \in E_o^+$, we collected misogynous and not misogynous memes according to the following criteria:

- a not misogynous meme is part of the synthetic dataset if it contains e_t^+ (or e_o^+) and it does not contain any biased candidate terms (or tags) with a negative score. This is to evaluate the impact of e_t^+ (or e_o^+) in introducing a bias towards the misogynous class in not misogynous memes;
- a misogynous meme is part of the synthetic dataset if it contains e_t^+ (or e_o^+) and it does not contain any other element in E_t^+ (or E_o^+). This is to verify if the model, given the presence of e_t^+ (or e_o^+), is able to perform well on misogynous memes.

Candidate Biased Tags			
Misogynous		Not Misogynous	
Tag	Score	Tag	Score
Woman	0.11	Penguin	-0.27
Earring	0.11	Cat	-0.26
Lip	0.11	Whisker	-0.23
Strap	0.11	Beak	-0.18
Tire	0.10	Gun	-0.17
Eyebrow	0.10	Dog	-0.16
Girl	0.09	Toy	-0.15
Teeth	0.08	Paw	-0.15
Short	0.08	Animal	-0.14
Dress	0.08	Bear	-0.14

Table 2
Top-10 candidate biased tags.

An analogous procedure has been adopted to create misogynous and not misogynous memes according to the candidate biased terms and tags with a negative score. The synthetic test set will later be recalled as *synt*.

3.3. Multimodal Bias Estimation (MBE)

In order to measure if a given model is affected by bias we introduce the **Multimodal Bias Estimation** (MBE) metric, which combines the area under the curve (AUC_{raw}) estimated on a test set belonging to the original MAMI test set and the area under curve estimated on the test set belonging to the synthetic dataset (AUC_{synt}):

$$MBE = \frac{1}{2}AUC_{raw} + \frac{1}{2}AUC_{synt} \qquad (2)$$

where AUC_{synt} is computed as reported in Equation 3. $\mathcal{M}_t$ represents the subgroup of memes identified by the presence of a biased term t, T is the subset of selected

$$AUC_{synt} = \frac{1}{2}\frac{\sum_{t\in T} AUC_{\text{Subgroup}}(\mathcal{M}_t) + \sum_{t\in T} AUC_{BPSN}(\mathcal{M}_t) + \sum_{t\in T} AUC_{BNSP}(\mathcal{M}_t)}{|T|} + \frac{1}{2}\frac{\sum_{i\in I} AUC_{\text{Subgroup}}(\mathcal{M}_i) + \sum_{i\in I} AUC_{BPSN}(\mathcal{M}_i) + \sum_{i\in I} AUC_{BNSP}(\mathcal{M}_i)}{|I|} \tag{3}$$

woman	cat	desk	chair	man	car	bicycle
0.9	0.3	0.8	0.43	0.87	0.13	0.0

woman	cat	desk	chair	man	car	bicycle	MASK
0.0	0.3	0.8	0.43	0.87	0.13	0.0	1.0

Figure 1: Visual Masking

biased terms. $\mathcal{M}_i$ denotes the subgroup of memes identified by the presence of a biased tag i and I denotes the subset of selected biased tags.

AUC_{synt} is a three per-element AUC-based measure, which considers both the biased terms and the biased tags, composed of the following estimations:

- $AUC_{Subgroup}(\cdot)$, estimated on the subset of the synthetic dataset identified by the presence of a biased element;
- $AUC_{BPSN}(\cdot)$, computed on the background-positive subgroup-negative subset that corresponds to the subset of misogynous memes identified by the absence of the biased element and the not misogynous memes containing the biased element;
- $AUC_{BNSP}(\cdot)$, computed on the background-negative subgroup-positive subset that corresponds to the subset of not misogynous memes identified by the absence of the biased element and the misogynous memes containing the biased element.

The *MBE* metric, which ranges into the interval $[0, 1]$, estimates the ability of the models on performing a good prediction on the raw test data and simultaneously achieving a significant performance on memes that, by construction, can lead to a biased prediction.

4. Debiasing Strategy

Several baseline models have been initially considered for distinguishing between misogynous and not misogynous memes. We trained SVM, KNN, Naive Bayes, Decision Tree, and Multi-layer Perception independently on each unimodal representation of the memes. In particular, the following modalities have been considered as (separate) input space:

- **textual component**, that is the transcription of the text contained within the meme (obtained with OCR) embedded through the Universal Sentence Encoder (USE) [28].
- **visual component**, expressed by the objects identified within the meme (*object tags*) by the Scene Graph Generation method [29] and represented through a n-dimensional vector that denotes if a given meme contains one or more predefined objects with the corresponding probabilities.

The classifiers have been combined, accordingly to each modality (e.g. visual or textual), through a Bayesian Model Averaging (BMA) [30] ensemble paradigm. BMA has been employed also for creating a multimodal ensemble that considers all the predictions provided by the above-mentioned models trained on each representation independently.

4.1. Mitigation Strategy

Bias mitigation is adopted in both unimodal and multi-modal contexts. In an unimodal setting, only the considered modality is mitigated. In a multi-modal scenario, all the models based on visual and textual components that compose the ensemble are mitigated. In order to debias the model at training time (and inference time), a **Masking Mitigation** is proposed. In particular, for what concerns the textual component, each biased term is masked according to the class label that they affect more (see Table 1). Any given biased term, estimated using to the strategy presented in section 3, is masked in the training dataset according to the class towards they induce bias. In particular, if a candidate biased term induces a bias towards the misogynous label, then it is replaced with a positive mask [POS-MASK] in misogynous memes. On the contrary, if a candidate biased term induces a bias towards the not misogynous label, then it is replaced with a negative mask [NEG-MASK] in not misogynous memes. An example is reported in the following.

`Original Text:` *When you can't afford a new* ***dishwasher*** *so you...*

`Masked Text:` *When you can't afford a new [POS-MASK] so you...*

Regarding the visual component, when a candidate biased tag is present, the probability value of that tag is set equal to 0 and a new feature indicating the presence of the masking is added to the original n-dimensional vector. A toy example is reported in Figure 1.

5. Experimental Results

We report in this section the results of the proposed mitigation strategy, comparing the performance with several approaches. In particular, we report AUC_{raw}, AUC_{synt} and MBE related to each model enclosed in the ensemble, i.e., Support Vector Machines (SVM), K-Nearest Neighbour (KNN), Naive Bayes (NB), Decision Tree (DT), and Multi-layer Perception (MLP) together with their Bayesian Model Averaging (BMA). We also show the performance of the proposed Masking Mitigation on BMA (BMA-MM). Finally, we report a baseline debiasing technique available in the state of the art. In particular, we used REPAIR [31] as a benchmark mitigation model. It computes a weight w_i for each sample based on its proportional loss contribution with respect to a reference model and resamples the original training dataset according to several strategies. In particular, given a weight w_i for each meme i, it keeps $p = 50\%$ examples with the largest weight w_i from each class.

We show in Tables 3-5, the comparison between all the considered models, distinguished according to the modalities used to perform the training and the corresponding mitigation phase. A T-test has been performed to compute the statistical equality with a pairwise analysis between the best-performing approach (BMA) against the compared mitigation strategies, i.e. BMA-MM and REPAIR.

A few considerations can be derived from Table 3, where the models have been trained using the textual component only: (1) training on the textual component only lead all the models to obtain good results on both raw and $synt$ test sets, (2) BMA is able to achieve remarkable results compared with the baselines, (3) the proposed Masking Mitigation strategy (BMA-MM) significantly outperforms all the baseline models and the original BMA, but also the REPAIR strategy. BMA-MM is able to maintain good recognition performance on the raw test set, still improving significantly the generalization capabilities on the controversial memes available in the $synt$ test set.

Textual Component Only			
Model	AUC_{raw}	AUC_{synt}	MBE
SVM	0.7202	0.7801	0.7501
KNN	0.7173	0.7041	0.7107
NB	0.7010	0.7687	0.7348
DT	0.6301	0.7475	0.6880
MLP	0.7257	0.7521	0.7389
BMA	**0.7326**	0.7841	0.7583
REPAIR	0.6775	0.6811	0.6793
BMA-MM	0.7325	**0.8052**	**0.7689***

Table 3
Model performance using the textual component only. **Bold** denotes the best result, while (*) reflects that the mitigated model outperforms the best non-mitigated approach (BMA) and the improvement is statistically significant.

Visual Component Only			
Model	AUC_{raw}	AUC_{synt}	MBE
SVM	0.6808	0.5918	0.6363
KNN	0.6623	0.5942	0.6283
NB	0.6635	0.5773	0.6204
DT	0.6499	0.5888	0.6194
MLP	**0.6912**	0.6047	0.6480
BMA	0.6870	0.5990	0.6430
REPAIR	0.6651	0.5922	0.6286
BMA-MM	0.6655	**0.6416**	**0.6535***

Table 4
Model performance using the visual component only. **Bold** denotes the best MBE, while (*) reflects that the mitigated model outperforms the best non-mitigated approach (BMA) and the improvement is statistically significant.

For what concerns Table 4, where the models have been trained using the visual component only, the considerations are a bit different. As demonstrated in other state-of-the-art studies [26], the visual component is less impactful on the recognition capabilities than the textual one. We hypothesize that the reduced contribution of the pictorial component is mainly due to conceptualization issues to relate a given object to a an abstract concept (e.g. dishwasher). However, also in this case, BMA is able to achieve better results than the baselines and BMA-MM is still able to significantly outperform the original BMA and REPAIR.

Multimodal Components			
Model	AUC_{raw}	AUC_{synt}	MBE
SVM	0.7632	0.7794	0.7713
KNN	0.7590	0.7277	0.7433
NB	0.7326	0.7794	0.7560
DT	0.7006	0.7483	0.7245
MLP	0.7690	0.7374	0.7532
BMA	**0.7802**	0.7908	0.7855
REPAIR	0.7360	0.6982	0.7171
BMA-MM	0.7676	**0.8306**	**0.7991***

Table 5
Model performance using the multimodal components. **Bold** denotes the best result, while (*) reflects that the mitigated model outperforms the best non-mitigated approach (BMA) and the improvement is statistically significant.

Regarding the performance of the multimodal settings reported in Table 5, we can assert that not only the proposed mitigation strategy significantly outperforms all the other configurations presented above, but it is also able to achieve a very promising compromise between raw and $synt$ samples that facilitate the adoption of the BMA-MM in a real setting.

6. Conclusions

This paper addressed the problem of mitigating misogynous meme detection. In particular, a candidate biased element estimation and a corresponding mitigation strategy is proposed to perform fair prediction in a real setting. The proposed approach, validated on a benchmark dataset, achieved remarkable results both in terms of prediction and generalization capabilities, reducing the bias in a significant way.

Acknowledgments

The work of Elisabetta Fersini has been partially funded by the European Union – NextGenerationEU under the National Research Centre For HPC, Big Data and Quantum Computing - Spoke 9 - Digital Society and Smart Cities (PNRR-MUR), and by MUR under the grant “Dipartimenti di Eccellenza 2023-2027" of the Department of Informatics, Systems and Communication of the University of Milano-Bicocca, Italy.

References

[1] M. Anzovino, E. Fersini, P. Rosso, Automatic identification and classification of misogynistic language on twitter, in: International Conference on Applications of Natural Language to Information Systems, Springer, 2018, pp. 57–64.

[2] M. A. Bashar, R. Nayak, N. Suzor, Regularising lstm classifier by transfer learning for detecting misogynistic tweets with small training set, Knowledge and Information Systems 62 (2020) 4029–4054.

[3] H. T. Ta, A. B. S. Rahman, L. Najjar, A. Gelbukh, Transfer learning from multilingual deberta for sexism identification, in: CEUR Workshop Proceedings, volume 3202, CEUR-WS, 2022.

[4] R. Calderón-Suarez, R. M. Ortega-Mendoza, M. Montes-Y-Gómez, C. Toxqui-Quitl, M. A. Márquez-Vera, Enhancing the detection of misogynistic content in social media by transferring knowledge from song phrases, IEEE Access 11 (2023) 13179–13190.

[5] E. Fersini, F. Gasparini, S. Corchs, Detecting sexist MEME on the Web: A study on textual and visual cues, in: 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), 2019, pp. 226–231.

[6] E. Fersini, G. Rizzi, A. Saibene, F. Gasparini, Misogynous meme recognition: A preliminary study, in: International Conference of the Italian Association for Artificial Intelligence, Springer, 2021.

[7] S. Singh, A. Haridasan, R. Mooney, “female astronaut: Because sandwiches won’t make themselves up there”: Towards multimodal misogyny detection in memes, in: The 7th Workshop on Online Abuse and Harms (WOAH), 2023, pp. 150–159.

[8] R. Song, F. Giunchiglia, Y. Li, L. Shi, H. Xu, Measuring and mitigating language model biases in abusive language detection, Information Processing & Management 60 (2023) 103277.

[9] T. Shen, J. Li, M. R. Bouadjenek, Z. Mai, S. Sanner, Towards understanding and mitigating unintended biases in language model-driven conversational recommendation, Information Processing & Management 60 (2023) 103139.

[10] B. O. Sabat, C. C. Ferrer, X. G. i Nieto, Hate speech in pixels: Detection of offensive memes towards automatic moderation, 2019. arXiv:1910.02334.

[11] F. Gasparini, G. Rizzi, A. Saibene, E. Fersini, Benchmark dataset of memes with text transcriptions for automatic detection of multi-modal misogynistic content, Data in brief 44 (2022) 108526.

[12] E. Fersini, F. Gasparini, G. Rizzi, A. Saibene, B. Chulvi, P. Rosso, A. Lees, J. Sorensen, SemEval-2022 task 5: Multimedia automatic misogyny identification, in: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), Association for Computational Linguistics, Seattle, United States, 2022, pp. 533–549.

[13] Z. Zhou, H. Zhao, J. Dong, N. Ding, X. Liu, K. Zhang,

DD-TIG at semeval-2022 task 5: Investigating the relationships between multimodal and unimodal information in misogynous memes detection and classification, in: The 16th International Workshop on Semantic Evaluation, 2022.
[14] L. Chen, H. W. Chou, RIT boston at semeval-2022 task 5: Multimedia misogyny detection by using coherent visual and language features from CLIP model and data-centric AI principle, in: The 16th International Workshop on Semantic Evaluation, 2022.
[15] S. Hakimov, G. S. Cheema, R. Ewerth, TIB-VA at semeval-2022 task 5: A multimodal architecture for the detection and classification of misogynous memes, in: The 16th International Workshop on Semantic Evaluation, 2022.
[16] J. M. ZHI, Z. Mengyuan, M. Yuan, D. Hu, X. Du, L. Jiang, Y. Mo, X. Shi, PAIC at semeval-2022 task 5: Multi-modal misogynous detection in MEMES with multi-task learning and multi-model fusion, in: The 16th International Workshop on Semantic Evaluation, 2022.
[17] D. Nozza, C. Volpetti, E. Fersini, Unintended bias in misogyny detection, in: IEEE/WIC/ACM International Conference on Web Intelligence, WI '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 149–155. URL: https://doi.org/10.1145/3350546.3352512. doi:10.1145/3350546.3352512.
[18] N. Zueva, M. Kabirova, P. Kalaidin, Reducing unintended identity bias in russian hate speech detection, in: Proceedings of the Fourth Workshop on Online Abuse and Harms, 2020, pp. 65–69.
[19] F. R. Nascimento, G. D. Cavalcanti, M. Da Costa-Abreu, Unintended bias evaluation: An analysis of hate speech detection and gender bias mitigation on social media using ensemble learning, Expert Systems with Applications 201 (2022) 117032.
[20] R. Zmigrod, S. J. Mielke, H. Wallach, R. Cotterell, Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 1651–1661.
[21] K. Guo, R. Ma, S. Luo, Y. Wang, Coco at semeval-2023 task 10: Explainable detection of online sexism, in: Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), 2023, pp. 469–476.
[22] M. Xia, A. Field, Y. Tsvetkov, Demoting racial bias in hate speech detection, in: Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media, 2020, pp. 7–14.
[23] R. Sridhar, D. Yang, Explaining toxic text via knowledge enhanced text generation, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 811–826. URL: https://aclanthology.org/2022.naacl-main.59. doi:10.18653/v1/2022.naacl-main.59.
[24] V. Perrone, M. Donini, M. B. Zafar, R. Schmucker, K. Kenthapadi, C. Archambeau, Fair bayesian optimization, in: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 2021, pp. 854–863.
[25] S. Sikdar, F. Lemmerich, M. Strohmaier, Getfair: Generalized fairness tuning of classification models, in: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022, pp. 289–299.
[26] G. Rizzi, F. Gasparini, A. Saibene, P. Rosso, E. Fersini, Recognizing misogynous memes: Biased models and tricky archetypes, Information Processing & Management 60 (2023) 103474.
[27] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, V. Patti, Resources and benchmark corpora for hate speech detection: a systematic review, Language Resources and Evaluation 55 (2021) 477–523.
[28] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. St. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, B. Strope, R. Kurzweil, Universal Sentence Encoder for English, in: Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations, 2018, pp. 169–174.
[29] X. Han, J. Yang, H. Hu, L. Zhang, J. Gao, P. Zhang, Image scene graph generation (sgg) benchmark, 2021. arXiv:2107.12604.
[30] E. Fersini, E. Messina, F. A. Pozzi, Sentiment analysis: Bayesian Ensemble Learning, Decision Support Systems 68 (2014) 26–38.
[31] Y. Li, N. Vasconcelos, Repair: Removing representation bias by dataset resampling, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 9572–9581.

Building a Spoken Dialogue System for Supporting Blind People in Accessing Mathematical Expressions

Pier Felice Balestrucci[1], Luca Anselma[1], Cristian Bernareggi[2] and Alessandro Mazzei[1]

[1]Dipartimento di Informatica, Università degli Studi di Torino, Italy
[2]Laboratorio Polin - Dipartimento di Matematica, Università degli Studi di Torino, Italy

Abstract
English. Mathematical expressions are complex hierarchical structures of symbols that are usually accessed by visual inspection. These expressions are seldom rendered with natural language since users are not usually required to read them aloud. People with a visual impairment generally use LaTeX with *screen readers* to acquire mathematical expressions. However, LaTeX can be verbose, slow to listen, and difficult to learn. This work proposes a way to make mathematical expressions easier to be accessed by people with disabilities by exploiting their hierarchical structures. We describe and evaluate a dialogue system to vocally navigate mathematical expressions in English. In contrast with standard screen readers, the vocal interaction allows people to query the system about sub-parts of the expressions.

Italiano. Le espressioni matematiche sono complesse strutture gerarchiche di simboli generalmente esplorate visivamente. Queste espressioni raramente vengono rappresentate tramite linguaggio naturale, perché agli utenti di solito non è richiesto di leggerle ad alta voce. Le persone con disabilità visiva di solito usano il LaTeX con uno *screen reader* per ascoltare le espressioni matematiche. Tuttavia, il LaTeX può essere verboso, lento da ascoltare e difficile da imparare. Questo lavoro propone una alternativa per rendere più accessibili le espressioni matematiche alle persone con disabilità sfruttando la loro struttura gerarchica. Descriviamo ed valutiamo quindi, un sistema di dialogo per navigare vocalmente le espressioni matematiche in inglese. A differenza dei normali screen reader, l'interazione vocale consente alle persone anche di interrogare il sistema sugli elementi che compongono le espressioni.

Keywords
dialogue system, natural language processing, visually impaired people

1. Introduction

In the last few years, improvements in speech technologies have enabled the use of Natural Language Processing in various application domains. Among these, the field of assistive technologies stands out as one with the most potential. Specifically, spoken dialogue systems (SDSs) can enhance the quality of life for numerous users with special needs.

In this paper we study the task of accessing mathematics for blind people. Traditionally, people with a visual impairment, in order to acquire mathematics, use LaTeX with *screen readers*, that are dedicated text-to-speech software. However, LaTeX is verbose, slow to listen, and not well known in young people. Indeed, LaTeX is designed as a typographical language, that is a technical language to provide typographical details rather than a compact description of a mathematical content. Moreover, mathematical expressions can range over multiple dimensions (e.g. fractions are arranged on two levels) and the linear encoding of LaTeX can produce a very long sequence with distant relations among symbols closely connected.

To tackle these issues, in this paper we propose a SDS based on the main idea to use standard natural language to interact with mathematical contents. For instance by using a SDS people with a visual impairment do not need to use a keyboard and a mouse to use their device. They can activate vocally the SDS to interact directly. Furthermore, this approach can also address other disabilities, including motor impairments. The SDS uses *mathematical sentences*, that are natural language sentences containing the semantics of a mathematical expression [1, 2]. There are many advantages in using mathematical sentences with respect to LaTeX. First, mathematical sentences are shorter. Second, they are not semantically ambiguous as some LaTeX expressions (e.g. $f(x)$ could also stand for $f \cdot (x)$). Third, they do not require knowledge of LaTeX, so that they can be used by a wider class of users (e.g. children). In particular, the SDS converts the LaTeX encoding of the mathematical expression in a semantic language, called Content Math Markup Language (CMML)[1]. From CMML, English mathematical sentences are generated according to good practices for spoken mathematics [3], through a standard Natural Language Generation (NLG)

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
pierfelice.balestrucci@unito.it (P. F. Balestrucci); luca.anselma@unito.it (L. Anselma); cristian.bernareggi@gmail.com (C. Bernareggi); alessandro.mazzei@unito.it (A. Mazzei)

[1]https://www.w3.org/TR/MathML3/chapter4.html

architecture, i.e. a sentence planner and a realizer.

The interaction design of the SDS is quite straightforward: the SDS pronounces the mathematical sentences and the user can interrupt it to "navigate" the expression (e.g. ask for repetition of its parts). So, the dialogue manager component coordinates the recognition of a repetition command with the generation of (a subpart of) the mathematical sentence.

We performed a user-based evaluation on the effectiveness and the usability of the SDS. The system implementation and evaluation were conducted with the involvement of visually impaired experts. The results show that the developed system has a good impact both on the comprehension of mathematical expressions and on the user experience, constituting a promising approach for helping people with visual impairments.

The paper is structured as follows: In Section 2, we report related work. In Section 3, we describe the main components of the SDS. In Section 4, we describe a human-based evaluation of the SDS and in Section 5, we closes the paper with some consideration and future work. Finally, Section 6 discusses some limitations of our SDS.

2. Related Work

2.1. Access to mathematics by people with visual impariments

Many different solutions have been investigated to enable people with visual impariments to access mathematical expressions. They can be divided up into two main categories: systems to read mathematics and systems to type and simplify maths expressions. The former category includes applications to read LaTeX in PDF files [4, 5, 6, 7], maths in web pages through speech rendering of MathML [8] or MathJax [9, 10], to read source LaTeX documents [11, 12] and maths in R Markdown [13]. These solutions propose and evaluate reading models based on sequential reading or hierarchical reading of maths expressions based on keyboard interaction. To the best of our knowledge, to date no studies have investigated a dialogue system to facilitate reading and exploration on maths expressions through speech input and output.

The latter category includes specialized applications that are designed to facilitate students with visual impairments to simplify expressions in LaTeX format [14] or in a multimodal work environment specifically designed for inclusive classes [15, 16] and to work with bi-dimensional mathematical procedures as arithmetic operations [17].

2.2. Speech-to-text solutions for entering maths expressions

This section introduces the solutions which have been investigated to write mathematics through speech input. TalkMaths [18] [19] is a prototype application which translates a limited set of arithmetic, algebraic and trigonometric expressions from spoken English into LaTeX or MathML. It adopts Dragon Naturally Speaking[2] (DNS) as speech recognition system. The translation rules are defined only for English and the recognition implements a dictation model based on pauses, which slow down the dictation process [20].

Mathifier [21] is an open source software module which converts a subset of mathematical expressions from English into LaTeX. It combines a dictionary, a language model and an acoustic model to recognize mathematical English utterances. It is based on Sphinx-4 [22] to recognize speech. This project has not been maintained and updated regularly.

CamMath [20] is a proof of concept prototype application designed to prove the advantages of continuous speech over discrete utterance of mathematical expressions in English.

Metroplex MathTalk[3] is a commercial application that provides speech input of arithmetic, algebra, calculus and statistics in English.

EquatIO[4] enables dictation of simple maths expressions in English in MS Word and in GSuite applications.

Even though these applications have been designed to enable speech recognition of mathematical expressions, none of them has addressed the needs of people with visual impairments by combining speech input and speech output.

3. Spoken Dialogue System

In this section we describe the main components of the SDS. As most rule-based SDSs [23], the information flow follows a path initiated by the user, who starts the interaction with a request to read a specific mathematical expression (Fig. 1). The SDS pronounces the mathematical sentence. The user listens to the produced sentence, and possibly interrupts the SDS asking for clarification. At this point the SDS answers to the request and the dialogue goes on. In Fig. 1 we report the architecture of the SDS, which is based on three main components, that are the response generator (described in Section 3.1), the language understanding and the dialogue management (described both in Section 3.2).

[2] https://www.nuance.com/it
[3] www.metroplexvoice.com/
[4] https://www.texthelp.com/en-us/products/equatio/

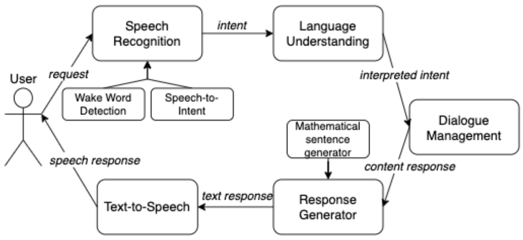


Figure 1: Dialogue System Architecture

3.1. Response Generator and Text-to-Speech

The generation side of the SDS follows the traditional NLG architecture composed of document planning, sentence planning and realizer [24]. In particular, for the mathematical sentence generation, we follow the pipeline described in Mazzei et al. [1, 2]. Note that, in contrast with Mazzei et al. [1], we model the grammar of the mathematical sentence for English rather than Italian for this novel SDS. Moreover, we propose an SDS, whereas in Mazzei et al. [1] we proposed a pure NLG system, where the interaction was limited to listening to the generated sentence.

The document planning consists of an encoding of the mathematical expression from LaTeX into a semantically unambiguous format, i.e. CMML, through the LatexML tool [25], specifying some heuristics defined by the specific mathematical sub-domain (e.g. algebra).

Let us consider the following formula as an example: "$A \times B = \{(x, y) \mid x \in A, y \in B\}$" and its LaTeX representation: "A \times B = { (x, y) \mid x \in A, y \in B}". The mathematical formula will be converted in the following CMML format:

```
<apply>
<eq/>
    <apply>
    <cartesianproduct/>
        <ci>A</ci>
        <ci>B</ci>
    </apply>
    <apply>
    <conditional-set/>
        <apply>
        <pair/>
            <ci>x</ci>
            <ci>y</ci>
        </apply>
        <apply>
        <and/>
            <apply>
            <in/>
                <ci>x</ci>
                <ci>A</ci>
            </apply>
            <apply>
            <in/>
                <ci>y</ci>
                <ci>B</ci>
            </apply>
        </apply>
    </apply>
</apply>
```

The CMML representation of the above formula is unambiguous: for each operator (empty tag–e.g. <eq/>), there is an opening and closing tag (i.e. <apply>...</apply>). Within these tags, we can find nested parts of the mathematical sentence such as other operators and the variables that compose the arguments of each operator.

For implementing the sentence planning, we model the syntactic structure of the English mathematical expressions by using as reference the linguistic constructions presented in Chang [3], a standard reference point for spoken mathematics in assistive technologies. With this linguistic reference for English, following the approach of Mazzei et al. [1], we divide the mathematical operators in categories characterized by the same arity, and for each one we define a syntactic template. For instance, the operators in Table 1 (e.g. $+$) are generally modeled with declarative sentence (e.g. $a\ plus\ b$), while elementary functions (e.g. sin) are modeled with a noun phrase (e.g. $sin\ of\ x$).

Symbol	Operator	English Form
$+$	plus	plus
$-$	minus	minus
$/$	divide	over
$*$	times	times
$[...]^{[..]}$	power	to
$\setminus$	settdiff	minus
$\times$	cartesianproduct	cross
$\cap$	intersect	the intersection of
$\cup$	union	the union of ... and ...

Table 1
Algebraic, Arithmetic and Set Operators

Note that within a category there are still different ways to compose a sentence. Taking as an example the minus operator, it can appear as: (1) $minus\ 5$, (2) $a\ minus\ b$ (a noun phrase and a declarative phrase, respectively). The difference between these two sentences depends on how the operator is used. In (1) $minus$ is used to define a negative value, in (2) $minus$ is used as the difference between two variables. From a realisation point, we need to distinguish two different forms within the category of algebraic, arithmetic and set operators: (1) unary form (e.g.

minus 5): a so-called adjective phrase must be defined where the operator (e.g. *minus*) works as an adjective; (2) binary form (e.g. *a minus b*): a declarative structure must be defined where the operator (e.g. *minus*) works as the parent node of the expression.

The realisation phase uses SimpleNLG [26], which is a Java library for morphological realization and linearization in English. Note that we added special symbols into SimpleNLG lexicon to produce both parentheses and pauses. Finally, we control the pronunciation, volume and pitch by using the Speech Synthesis Markup Language (SSML)[5]. As the last step, the mathematical sentence is pronounced through a Text-to-Speech technology.

We experiment with two different vocal synthesizers: the commercial AWS Polly[6] and the open source eSpeak[7] (both support SSML). AWS Polly is based on an advanced deep learning technology and has a human-kind voice, which is clear and highly user-adjustable, whereas eSpeak is based on a so-called formant synthesis method and produces a robotic voice. It is worth noting that eSpeak is a very familiar voice for visually impaired people. The users can choose which vocalizer to use.

3.2. Speech Recognition, Language Understanding and Dialogue Management

The speech recognition module in Fig. 1 is composed of two different sub modules: Wake-up word detection and Speech-to-Intent. A wake-up word [27] is a keyword that triggers the speech-to-intent module. The classification is binary and happens in real time. For this task we used Porcupine (v.2.1) [28] which has good results in comparison to other commercial systems. A Speech-to-Intent system is able to recognize a user's intent in a very specific context. The system works on a small vocabulary of terms and classifies each user's request. For this task we used Rhino (v.2.1) [29] which allowed us to detect in real time the user's vocal commands for our mathematical context.

The intent is translated into a request which is interpreted by the language understanding module with regular expressions, which matches the intent produced by the speech recognition with a domain specific speech act. These speech acts are: (1) repetition, (2) query and (3) resuming. The repetition intent lets the user ask the repetition (from a subpart) of the mathematical sentence (e.g. "Repeat from the first integer"). In this case the system searches within the content response what the user wants to be repeated. The query intent lets the user ask the system to repeat a specific part of the mathematical sentence (e.g. "What is the limit of the second integer?"). In this case, the specific operator in the user's request will be searched. If matched, the system produces an answer with the requested part of the operator. Finally, the resuming intent lets the user resume the interaction after an interruption (e.g. "Go on").

The main algorithm of this module is represented below:

```
dialogue():
1.  Say "I'm starting to say the sentence"
2.  Activate the Wake-up Word Detection
    system
3.  For each word within the sentence:
4.      Say the word
5.      If the Wake-up Word Detection
        system has detected "Hey stop":
6.          Say "Ok, I'm listening to you"
7.          Activate the Request
            Recognition system
8.          Fulfill the request
9.  Say "I've finished reading this
    sentence, but I'm still here for you"
10. Activate the Request Recognition
    system
```

In lines 1 and 9 the system announces to the user the start and the end of the expression. After the activation of the wake-up word detection system, for each word the system will pronounce it. If the user says the wake-up word, the system will stop and put itself into a listening phase waiting for a request. Completed the expression, the System can continue to answer user's requests.
The "fulfill the request" method in line 8 works as follows:

```
fulfill_request(intent, request):
1.  Parsify the request
2.  If the intent is "Query":
3.      Search for an operator in the
        request
4.      Match the request with the
        operator arguments
5.      If the match is successful:
6.          Answer the question
7.          Otherwise, utter an
            apology message
8.  If the intent is "RepetitionFrom":
9.      Search within the sentence
        uttered what the user
        wants to hear repeated
10.     If the search is successful:
11.         Repeat from that point
12.         Otherwise, utter an
            apology message
13. If the intent is "Resume":
14.     Resume the synthesis of the
        mathematical expression
```

[5] https://www.w3.org/TR/speech-synthesis11/
[6] https://aws.amazon.com/it/polly/
[7] http://espeak.sourceforge.net/

If the intent has been interpreted as repetition, the system will search within the content response what the user wants to be repeated. If the user's intent has been interpreted as query, the specific operator in the user's request will be searched. If matched, the system will answer the requested part of the operator. The dialogue management module is in charge of making decisions. If the interpreted request does not correspond to any of the supported possibilities, it asks the user to repeat the request. Otherwise, the dialogue management searches the content of the response for the user and sends it to the response generator module.

4. Evaluation

To evaluate the SDS, we conducted two experiments involving visually impaired users with the approval of the University's ethical committee. For the first experimentation we tested the effectiveness and solidity of the English generation system in a similar way as Mazzei et al. [1] and for the second experiment the effectiveness and usability of the SDS. In Experiment 1 we recruited two blind native Italian speakers proficient in English with an excellent maths knowledge that participated freely and without compensation. We retrieved from a calculus textbook [30] 10 mathematical expressions of different length and difficulty (Table 2 and Table 3), and we represented them in CMML. The difficulty of a formula is related to the number of parentheses and the number of nodes in its CMML representation. Then, we used SDS to synthesize the expressions. Using the same experimental setting of Mazzei et al. [1], we obtained 25 audios (10 easy ones generated using different synthesizers, i.e. eSpeak and Polly, and 15 difficult ones generated with different strategies for generating pauses, e.g. for parentheses). We uploaded the audio files to Youtube and provided them to the users along with a questionnaire[8] (on Google Form, because it is accessible) containing profiling questions and the request to write down the expressions in the audio files in an unambiguous notation.

We evaluated the written expressions with two metrics: Exact Match (EM) and SPICE [31]. EM is 1 if the original CMML and the one obtained by the user are the same, and 0 otherwise. SPICE is obtained by calculating the F-score of the overlapping between the original CMML tree and the one obtained from the user. The overlapping is measured by decomposing the CMML trees in typed elementary substructures, which are operands, operators and their relations. Experiment 1 results (Table 4) show that the generation system seems to be effective since the users obtained a good understanding of the expressions.

[8] https://docs.google.com/forms/d/e/1FAIpQLSc93-1NAIWy_SXQaTJH2tAnHKg9PHSysurOSzf3xbSvtKG7Ig/formResponse?pli=1

LaTeX Formula	Nodes
$\sqrt[n]{x} = x^{1/n}$	10
$x > b \Longrightarrow \lvert f(x) \rvert < M$	10
$g^{-1}(y) = f^{-1}((y-b)/a)$	13
$\int_b^c a \, \mathrm{d}x = a(c-b)$	14
$A \times B = \{(x,y) \mid x \in A, y \in B\}$	15

Table 2
Easy expressions - Few parenthesises and nodes - Experiment 1

LaTeX Formula	Nodes
$\lim \left(1 + \frac{1}{n}\right)^n = e$	10
$\int \frac{1}{\sqrt{m^2 - x^2}} \mathrm{d}x = \arcsin \frac{x}{m} + c$	20
$y = f(a) + \frac{f(b) - f(a)}{b - a}(x - a)$	21
$\sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k$	28
$\lim_{x \to x_0} \left\{ \frac{f(x) - f(x_0)}{x - x_0} - f'(x_0) \right\} = 0$	31

Table 3
Difficult expressions - More parenthesises and nodes - Experiment 1

User	Metrics	Tot. (25)	Easy (10)	Difficult (15)
1	EM	0.92	1.00	0.87
	SPICE	0.98	1.00	0.97
2	EM	1.00	1.00	1.00
	SPICE	1.00	1.00	1.00
avg	EM	0.96	1.00	0.93
	SPICE	0.99	1.00	0.99

Table 4
EM and SPICE - Experiment 1

In Experiment 2 we recruited five blind native Italian speakers that declared a good maths knowledge and proficiency in English; however, one of the users dropped out because of their low competency. The users participated freely and without compensation. In Experiment 2 the users connected via Google Meet to a client running the SDS. This modality has been decided on the basis of COVID-19 restrictions still in force at that time. A facilitator established the Google Meet connection, presented the experimental protocol (see the instructions presented to the users in Fig. 2), and observed the user interactions while remaining neutral. After a short time (about 30 minutes) when they could practice with the SDS, the users were presented with 3 easy and 3 difficult expressions chosen among the 10 expressions of Experiment 1 (Table 5 and Table 6). The SDS used Polly for the speech synthesis. The users interacted autonomously with the SDS and they could interrupt the system and ask questions. Finally, the users were asked to write down the expressions questionnaire similarly to Experiment 1.

```
First of all, thank you for taking part in this
experimentation.
During today’s test you will have to interact
with a voice assistant who will pronounce six
mathematical phrases to you. Your job is to write
them in any understandable formalism,
even in English, and send me your answers by
the Google Meet chat or by email.
The voice assistant is able to receive
various commands when interrupted by saying “Hey stop!”.
I recommend you spell these words well and use a high
volume of voice. The commands are:
 1. Repeat from anywhere in the sentence.
 For example, consider the expression “a + b”:
 if “Say again from plus” is said, the
 assistant will repeat “+ b”. The keyword
 is “Say again from” and you can
 repeat a variable or an operator.
 2. Specify an operator.  For example consider
 the expression “a + b + c” you can ask
 “What is the first sum?” or “What is the
 left argument of second sum”. The keyword is
 “What is” and you can ask for information about
 a variable or an operator. Regarding the
 operator you can ask how you have already
 heard to repeat the argument in its case,
 specifying whether left or right, lower or upper.
 3. In the event that you accidentally interrupted
 the assistant or do not remember the question you
 wanted to ask him, just say “Go on”.
If you try to stop the assistant without success,
please repeat “Hey stop!” several times. From the moment
it stops you have about 1 minute maximum to make the
request. Let’s take a few examples now,
if you don’t have any questions!
```

Figure 2: Instructions for testers - Experiment 2

LaTeX Formula	Nodes
$\sqrt[n]{x} = x^{1/n}$	10
$g^{-1}(y) = f^{-1}\left((y-b)/a\right)$	13
$\int_b^c a \, \mathrm{d}x = a(c-b)$	14

Table 5
Easy expressions - Few parentheses and nodes - Experiment 2

LaTeX Formula	Nodes
$\int \frac{1}{\sqrt{m^2-x^2}} \mathrm{d}x = \arcsin\frac{x}{m} + c$	20
$y = f(a) + \frac{f(b)-f(a)}{b-a}(x-a)$	21
$\lim_{x\to x_0}\left\{\frac{f(x)-f(x_0)}{x-x_0} - f'(x_0)\right\} = 0$	31

Table 6
Difficult expressions - More parenthesises and nodes - Experiment 2

As in Experiment 1, we used EM and SPICE as evaluation measures. Moreover, the users also compiled the User Experience Questionnaire (UEQ) [32] to evaluate their experience in terms of attractiveness, perspicuity, efficiency, dependability, stimulation and novelty[9].

The scores for Experiment 2 are worse than the ones of Experiment 1 both on EM and on SPICE (cf. Table 7). This could be explained by the complexity of the setting, because users had to learn how to use a new tool in a short time, whereas the users in Experiment 1 were familiar with the linear fruition of a Youtube audio. We observed that the performance of the users improved over the time of the experiment, which can be due to an acquired familiarity of the tool.

In Table 8 we present the results of the UEQ along different attributes. For each attribute we report the score of the attribute on a scale between -3 and $+3$ and compare it with a benchmark provided by UEQ [32] that includes a dataset with 468 products evaluated by 21, 175 users. It is worth noting that the SDS scored high over stimulation, novelty and attractiveness and fair over perspicuity, efficiency and dependability. The scores over efficiency

[9]https://www.ueq-online.org/

User	Metrics	Tot. (6)	Easy (3)	Difficult (3)
1	EM	0.50	0.67	0.33
	SPICE	0.86	0.89	0.87
2	EM	1.00	1.00	1.00
	SPICE	1.00	1.00	1.00
3	EM	0.33	0.33	0.67
	SPICE	0.66	0.98	0.82
4	EM	0.33	0.00	0.67
	SPICE	0.80	0.95	0.89
avg	EM	0.54	0.50	0.67
	SPICE	0.83	0.95	0.89

Table 7
EM and SPICE - Experiment 2

Attribute	Score	Comparison to benchmark
Attractiveness	1.83	between 75% and 90%
Perspicuity	1.28	between 50% and 75%
Efficiency	1.33	between 50% and 75%
Dependability	1.33	between 50% and 75%
Stimulation	1.75	above 90%
Novelty	1.92	above 90%

Table 8
UEQ Benchmark

and novelty are consistent with the hypothesis that users would benefit from a longer training time to become proficient with this new tool, that they however deem stimulating and attractive. These preliminary experiments seem to be promising, nevertheless it would be beneficial to enlarge the pool of users. However, it is known in accessibility studies [33] that involving visually impaired people in experiments is significantly hard and several studies tend to engage only sighted people.

5. Conclusion

In this paper we described a SDS designed for allowing visually impaired people to access mathematical expressions. In Experiment 1 we focused on the understanding of the mathematical sentences generator for English (i.e. using EM and SPICE measures), replicating the good results obtained for Italian in Mazzei et al. [1, 2]. In Experiment 2, we tested the complete SDS allowing user to ask for repetition. With respect to expressions understanding, the results of this experimentation are less encouraging than Experiment 1, but we speculate that this is a consequence of the complexity of the experimental setting due to the necessity of online interaction. However, the UEQ showed that the users really appreciated the interaction with the SDS. In the future we want to improve the SDS by adding new intents. Moreover, we want to define a new SDS designed for diagrams and other visual structures, creating accurate descriptions and making them navigable.

6. Limitations

The SDS developed in this paper has two main limitations. The design interaction is limited to repetition request concerning a subpart of the expression. A better interaction could consider the possibility to ask for mathematical clarification on the role of a subpart (e.g. “what is x?”). The evaluation has two aspects that could be improved: 1. the limited number of testers, and 2. no native English speakers participated.

References

[1] A. Mazzei, M. Monticone, C. Bernareggi, Using NLG for speech synthesis of mathematical sentences, in: Proceedings of the 12th International Conference on Natural Language Generation, Association for Computational Linguistics, Tokyo, Japan, 2019, pp. 463–472. URL: https://aclanthology.org/W19-8658. doi:10.18653/v1/W19-8658.

[2] A. Mazzei, M. Monticone, C. Bernareggi, Evaluating speech synthesis on mathematical sentences, in: Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13-15, 2019, 2019, pp. 1–7. URL: http://ceur-ws.org/Vol-2481/paper46.pdf.

[3] L. A. Chang, Handbook for Spoken Mathematics, The Regent of the University of California, 1983.

[4] T. Armano, A. Capietto, S. Coriasco, N. Murru, A. Ruighi, E. Taranto, An automatized method based on latex for the realization of accessible pdf documents containing formulae, in: International Conference on Computers Helping People with Special Needs, Springer, 2018, pp. 583–589.

[5] D. Ahmetovic, T. Armano, C. Bernareggi, A. Capietto, S. Coriasco, D. Boris, K. Alexandr, N. Murru, et al., Automatic tagging of formulae in pdf documents and assistive technologies for visually impaired people: the latex package axessibility 3.0, in: ICCHP 2020 17th International Conference on Computers Helping People with Special Needs, volume 1, ICCHP, 2020, pp. 69–73.

[6] D. Ahmetovic, T. Armano, C. Bernareggi, M. Berra, A. Capietto, S. Coriasco, N. Murru, A. Ruighi, E. Taranto, Axessibility: A latex package for mathematical formulae accessibility in pdf documents, in: Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility, 2018, pp. 352–354.

[7] D. Ahmetovic, T. Armano, C. Bernareggi, A. Capietto, S. Coriasco, N. Murru, et al., Axessibility 2.0: creating tagged pdf documents with accessible formulae, Ars Texnica (2019) 138–145.

[8] N. Soiffer, Browser-independent accessible math, in: Proceedings of the 12th International Web for All Conference, 2015, pp. 1–3.

[9] D. Cervone, P. Krautzberger, V. Sorge, Employing semantic analysis for enhanced accessibility features in mathjax, in: 2016 13th IEEE Annual Consumer Communications & Networking Conference (CCNC), IEEE, 2016, pp. 1129–1134.

[10] V. Sorge, C. Chen, T. Raman, D. Tseng, Towards making mathematics a first class citizen in general screen readers, in: Proceedings of the 11th Web for All Conference, 2014, pp. 1–10.

[11] R. Chauhan, I. Murray, R. Koul, Audio rendering

of mathematical expressions for blind students: a comparative study between mathml and latex, in: 2019 IEEE International Conference on Engineering, Technology and Education (TALE), IEEE, 2019, pp. 1–5.

[12] A. Bansal, M. Balakrishnan, V. Sorge, Comprehensive accessibility of equations by visually impaired, ACM SIGACCESS Access. Comput. 126 (2020) 1. URL: https://doi.org/10.1145/3386280.3386281. doi:10.1145/3386280.3386281.

[13] J. Seo, S. McCurry, A. Team, Latex is not easy: Creating accessible scientific documents with r markdown, Journal on Technology and Persons with Disabilities 7 (2019) 157–171.

[14] S. Arooj, S. Zulfiqar, M. Qasim Hunain, S. Shahid, A. Karim, Web-alap: A web-based latex editor for blind individuals, in: The 22nd International ACM SIGACCESS Conference on Computers and Accessibility, volume 28, 2020, pp. 1–6.

[15] V. Sorge, Supporting visual impaired learners in editing mathematics, in: Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility, 2016, pp. 323–324.

[16] C. Bernareggi, Non-sequential mathematical notations in the lambda system, in: ICCHP, Springer, 2010, pp. 389–395.

[17] A. Gerino, N. Alabastro, C. Bernareggi, D. Ahmetovic, S. Mascetti, Mathmelodies: inclusive design of a didactic game to practice mathematics, in: International Conference on Computers for Handicapped Persons, Springer, 2014, pp. 564–571.

[18] A. Wigmore, G. Hunter, E. Pflügel, J. Denholm-Price, V. Binelli, Using automatic speech recognition to dictate mathematical expressions: The development of the "talkmaths" application at kingston university., Journal of Computers in Mathematics and Science Teaching 28 (2009) 177–189.

[19] A. M. Wigmore, E. Pflugel, G. J. Hunter, J. Denholm-Price, M. Colbert, Talkmaths better! evaluating and improving an intelligent interface for creating and editing mathematical text, in: 6th International Conference on Intelligent Environments, IEEE, 2010, pp. 307–310.

[20] C. Elliott, J. Bilmes, Computer based mathematics using continuous speech recognition, Vocal Interaction in Assistive Technologies, Games and More (2007).

[21] S. N. Batlouni, H. S. Karaki, F. A. Zaraket, F. N. Karameh, Mathifier—speech recognition of math equations, in: 18th International Conference on Electronics, Circuits, and Systems, IEEE, 2011, pp. 301–304.

[22] W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, J. Woelfel, Sphinx-4: A flexible open source framework for speech recognition, 2004.

[23] K. Jokinen, M. McTear, Spoken Dialogue Systems, Synthesis lectures on human language technologies, Morgan & Claypool Publishers, 2010. URL: https://books.google.it/books?id=ualwulnD020C.

[24] E. Reiter, R. Dale, Building Natural Language Generation Systems, Natural Language Processing, Cambridge University Press, 2000. URL: http://prp.contentdirections.com/mr/cupress.jsp/doi=10.2277/052102451X. doi:DOI:10.2277/052102451X.

[25] B. Miller, Latexml: A LaTeX to XML converter, https://math.nist.gov/~BMiller/LaTeXML, 2007.

[26] A. Gatt, E. Reiter, SimpleNLG: A realisation engine for practical applications, in: Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009), 2009, pp. 90–93.

[27] Y. Wang, Wake word detection and its applications, Johns Hopkins University, 2021.

[28] Picovoice, Benchmarking a Wake Word Detection Engine, https://picovoice.ai/blog/benchmarking-a-wake-word-detection-engine/, 2018.

[29] Picovoice, Picovoice Console — Rhino Speech-to-Intent Engine, https://picovoice.ai/docs/quick-start/console-rhino/, 2018.

[30] L. Pandolfi, ANALISI MATEMATICA 1, Dipartimento di Scienze Matematiche "Giuseppe Luigi Lagrange", Politecnico di Torino, 2013.

[31] P. Anderson, B. Fernando, M. Johnson, S. Gould, SPICE: semantic propositional image caption evaluation, CoRR, abs/1607.08822, 2016.

[32] M. Schrepp, User experience questionnaire handbook. all you need to know to apply the ueq successfully in your project, https://www.ueq-online.org/, 2015.

[33] E. Brulé, B. J. Tomlinson, O. Metatla, C. Jouffrais, M. Serrano, Review of quantitative empirical evaluations of technology for people with visual impairments, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 2020, pp. 1–14.

Contrastive Language–Image Pre-training for the Italian Language

Federico Bianchi[1], Giuseppe Attanasio[2], Raphael Pisoni[3], Silvia Terragni[4], Gabriele Sarti[5] and Dario Balestri[3]

[1] *Stanford University, California, USA*
[2] *Bocconi University, Milan, Italy*
[3] *Independent Researcher*
[4] *Telepathy Labs, Zürich, Switzerland*
[5] *University of Groningen, The Netherlands*

Abstract
Recently, multi-modal systems such as CLIP (Contrastive Language–Image Pre-training) were introduced to represent images and texts jointly in the same embedding space. These models are trained on massive amounts of image-caption pairs and show impressive performance on zero-shot image classification. However, their usage is limited to English due to their training data. Training the same model for different languages is non-trivial since the amount of natural data in those might not be sufficient, and automatic translations of original captions might not have sufficient quality, harming performance. In this paper, we present the first CLIP model for the Italian Language (CLIP-Italian), trained on more than 1.4 million image-text pairs. Results show that CLIP-Italian outperforms a multilingual CLIP model on image retrieval and zero-shot classification tasks for the Italian language.[1]

Sistemi multimodali come CLIP (Contrastive Language-Image Pre-training) sono stati proposti di recente al fine di ottenere rappresentazioni di immagini e testo in uno spazio latente condiviso. Questi modelli sono allenati su enormi quantità di immagini associate alle loro didascalie, e dimostrano abilità eccellenti nell'effettuare classificazioni "zero-shot". Ciononostante, il loro utilizzo è limitato all'inglese, la lingua utilizzata durante il loro addestramento. Ottenere modelli del genere per altre lingue non è cosa da poco, poiché la quantità di dati a disposizione per queste lingue potrebbe non essere sufficiente e la traduzione automatica delle didascalie inglesi originali potrebbe portare a risultati non soddisfacenti. In questo articolo presentiamo il primo modello CLIP per la lingua italiana (CLIP-Italian), addestrato con più di 1.4 milioni di immagini e rispettive didascalie. I risultati riportati dimostrano l'efficacia di CLIP-Italian per l'estrazione e la classificazione zero-shot in italiano, ottenendo risultati migliori di un modello CLIP multilingue.

Keywords
clip, italian, contrastive, language, image, pretraining, multimodal

1. Introduction

The recent interest in combining different source domains to incorporate broader context in the training process has led to a surge in multi-modal models spanning modalities like text and vision [1] or text and speech [2]. A multi-modal architecture learns by jointly optimizing its parameters on two or more input domains (e.g., images, texts, tabular data, or audio signals), with a cost function that may vary depending on the task.

Contrastive Language–Image Pre-training (CLIP) [1] is a multi-modal model for joint learning image and text representations. CLIP learns to pair visual concepts with descriptions in natural language by leveraging a contrastive loss that pushes images and their respective captions closer in an embedding space. CLIP is trained on a large-scale dataset of images and their corresponding captions. The dataset used in the seminal paper contains 400 million images collected from the web. In recent years, there have been many successful domain-specific implementations of CLIP [3, 4, 5, 6, 7, *inter alia*].

While the model shows impressive zero-shot performance across various supervised tasks, its capabilities are bounded to the language the model is trained in, i.e., English. Despite the ongoing efforts on training multilingual variants of CLIP, different works have shown that multilingual models often do not achieve the same level of performance as language-specific ones [8, 9, 10].

In this paper, we describe how to fine-tune a specialized version of CLIP in a language different than English, i.e., Italian. We dub this model CLIP-Italian.[1] Crucially, we collect for the task a dataset of 1.4M high-quality text-image pairs for Italian, the largest collection of this kind to date. We release our best-performing checkpoint, the modeling and training code, a CometML report with training longs and metrics, and a live demo to showcase

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
giuseppe.attanasio3@unibocconi.it (G. Attanasio)

CEUR Workshop Proceedings (CEUR-WS.org)

[1]While Italian was selected for this study, the approach presented in this paper can be generalized to other languages and domains without loss of generality.

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

```
# using BERT and ViT to encode raw images and texts
encoded_images = clip.image_encoder(images)
encoded_texts = clip.text_encoder(images)

# nomalize the projections
embedded_images = l2_normalization(clip.image_projection(encoded_images))
embedded_texts = l2_normalization(clip.text_projection(encoded_texts))

logits = np.dot(embedded_images, embedded_texts.T) * logit_scale

labels = np.arange(n)  # correct image-text match is on the main diagonal
loss_images = cross_entropy_loss(logits, labels, axis=0)
loss_texts = cross_entropy_loss(logits, labels, axis=1)
loss = (loss_images + loss_texts) / 2
```

Figure 1: Numpy-like pseudo code that describes the CLIP-Italian loss.

CLIP-Italian capabilities and streamline testing.[2]

Contributions. We create the largest publicly available multi-modal dataset for the Italian language. We use this dataset to train and release the first CLIP image-text model for the Italian language. We show that this model performs better than its multilingual counterpart in two well-established multi-modal tasks: image retrieval and zero-shot image classification. Moreover, we release the model checkpoint, code, and an online demo to showcase CLIP-Italian capabilities.

2. Contrastive Language–Image Pre-training

CLIP is trained to put images and captions in close positions in the vector space. Therefore, the model is taught to associate visual concepts and their natural language descriptions.

CLIP's architecture consists of two distinct encoders, one for images and one for texts. At training time, all images and texts in a mini-batch are each projected to a 512-dimensional space. Next, vector similarities are computed for each image-text pair, and cross-entropy loss is applied. Finally, the average loss along the image and text dimensions is used to update model parameters. The loss is used to align the two 512-dimensional projection spaces. Figure 1 briefly summarizes how the contrastive loss is computed in CLIP. We refer the reader to [1] for additional details.

After training, CLIP can be used without further training for a variety of different tasks. Since images and texts are embedded in the same space, CLIP embeddings can be used for zero-shot text-based image retrieval and zero-shot image classification by looking at the similarities between available texts and images.

CLIP-Italian differs from the original CLIP in that encoders are not trained from scratch. We continue training from checkpoints of other pre-trained models. This approach allows us to leverage pre-training knowledge of existing models and remap it to new lexical items to create CLIP-Italian. We extensively cover training details in Section 4.

3. Datasets

We describe the four data sources we used to train our CLIP-Italian model.

- WIT [11] is a multilingual image-caption dataset collected from Wikipedia. We pre-process and extract the Italian subset, selecting the Reference Description captions as captions of interest. While several possible captions are available, we select those described as the most topical and highest-quality captions in the original paper.
- MSCOCO-IT [12].[3] The captions of this dataset come from the original MSCOCO dataset [13] and are translated with Microsoft Translator. The 2017 MSCOCO training set contains more than 100K images. More than one caption is available for each image.
- Conceptual Captions (CC) [14].[4] In this dataset, there are more than 3 million image-caption pairs, collected from the web. All images with available URLs were downloaded, and their captions

[2] Model: https://huggingface.co/clip-italian/clip-italian, Logs: https://www.comet.ml/g8a9/clip-italian/reports/clip-italian-training-metrics, Demo: https://huggingface.co/spaces/clip-italian/clip-italian-demo

[3] https://github.com/crux82/mscoco-it

[4] https://github.com/google-research-datasets/conceptual-captions

Dataset	Ratio	Captions
WIT	38%	525,950
MSCOCO-IT	8%	116,195
CC	52%	712,890
ILPOST	2%	29,055
Total		1,384,090

Table 1
A summary of datasets used in this work with the number of captions collected per dataset.

were translated to Italian using DeepL.[5], totaling roughly 710K captions.

- La Foto del Giorno (ILPOST).[6] This image-caption dataset is collected from *Il Post*, a prominent Italian online newspaper. Starting from early 2011, every day, the editors at Il Post have selected several images picturing the most salient events in the world. Each photo comes along with an original Italian caption. The resulting collection contains almost 30K pairs of images-captions.

3.1. Translations

We used automatic translation to augment the training set due to the low amount of captioned images for Italian compared to the original CLIP training dataset. Instead of relying on open-source translators, we use the proprietary DeepL API to obtain readily available high-quality English captions. While this choice aims to minimize the noise in translated data, we know about the bias (e.g., gender and age) that translation systems introduce during translations [15]. Some of the captions are available in Figure 2.

To assess the translation quality, three native Italian speakers among the authors inspected a sample of 100 translations alongside their original English sources, rating translations with scores between 1 and 4. We adopt the following categorization for the provided scores: 1, the sentence has lost its meaning, or it is not possible to understand it; 2, it is possible to get the idea, but there is something wrong; 3, good, however, a native speaker might complain about some parts of the translation; 4, correct translation.

The average score was 3.78, suggesting that the translations were good on average. We also computed an inter-rater agreement with Gwet's AC1 using ordinal weighting, obtaining a value of 0.858. This value suggests a strong agreement between annotators.

[5]https://www.deepl.com/
[6]https://www.ilpost.it/foto-del-giorno/

3.2. Data Cleaning

Many of the captions in WIT describe encyclopedic facts (e.g., "Roberto Baggio in 1994"). We believe these descriptions will not be helpful in learning a good mapping between images and captions, as most of the information in the description is factual knowledge. To prevent polluting the data with overly specific factual captions, we used Part-Of-Speech (POS) tagging using spacy[7] on the text and removed all the captions that were composed for the 80% or more by proper nouns (around 10% of the total captions for WIT). This simple solution allowed us to retain much of the dataset without introducing noise. Captions like "Dora Riparia", "Anna Maria Mozzoni", "Joey Ramone Place", "Kim Rhodes", "Ralph George Hawtrey" which are proper nouns (`PROPN`) have been removed. For the dataset ILPOST, we used `langdetect`[8] to filter non-Italian captions, resulting in only 2% captions being removed.

4. Training

Our CLIP-Italian model is based on previous pre-trained state-of-the-art models for both the vision and textual parts. We use Vision Transformer (ViT) [16] and BERT-inspired [17] text encoder. We limit the sequence length to 96 tokens and use a local batch size of 128 for each of the 8 TPU cores we used. For the optimization procedure, we used the AdaBelief optimizer [18] with Adaptive Gradient Clipping (AGC) and a Cosine Annealing Schedule [19]. We run training for a maximum of 15 epochs, evaluate at the end of each epoch, and release the checkpoint with the best validation loss.

Data Augmentation Following standard practices in computer vision, we applied several augmentations to the available images. In particular, we used random affine transformations, perspective changes, occasional equalization, and random changes to brightness, contrast, saturation, and hue. Importantly, we made sure to keep hue augmentations limited to allow the model to learn color definitions.

Projection Layers Warmup Since pre-trained checkpoints were used as starting points for both the vision and the text encoders, we found it helpful to warm-up projection layers. To do so, we first train the entire network using frozen vision and text encoders until loss convergence. After this first phase, the rest of the model is unfrozen to perform end-to-end training. We always pick the model with the best evaluation loss, evaluating every 15 epochs.

[7]https://spacy.io/models/it#it_core_news_lg
[8]https://github.com/Mimino666/langdetect

English Caption	Italian Caption
an endless cargo of tanks on a train pulled down tracks in an empty dry landscape	un carico infinito di carri armati su un treno trascinato lungo i binari in un paesaggio secco e vuoto
person walking down the aisle	persona che cammina lungo la navata
popular rides at night at the county fair	giostre popolari di notte alla fiera della contea

Table 2
Examples of automatically translated captions from the Conceptual Captions dataset.

Starting Checkpoints We used an Italian BERT checkpoint[9] as text encoder and the original CLIP vision encoder.[10]

Logits Scaling Both images and texts are then projected to 512-dimensional vectors to which we apply the loss defined in CLIP using logit scaling equal to 20. We empirically observed that logit scaling has a strong positive impact on model performance, suggesting that the embeddings have similar Euclidian norms and that scaling their dot similarities helped the cross entropy.

5. Quantitative Evaluation

To our knowledge, CLIP-Italian is the first multi-modal system explicitly trained for the Italian language. Hence, to provide meaningful comparisons, we compare its performance to an available multilingual CLIP[11] model trained with multilingual knowledge distillation [20].

5.1. Image Retrieval

The image retrieval task is as follows. Given a caption, the task is to retrieve the correct image from a set of available images, where the correct image is the one that is described by the caption. This search can be done by embedding the caption and the images and selecting the nearest neighbors to the caption embedding. We use the MSCOCO-IT validation dataset left out for this purpose during the training procedure, containing a total of 2,000 image-caption pairs.

Metric We compare models on the standard Mean Reciprocal Rank (MRR) retrieval metric. The metric computes the rank assigned to each image to be retrieved (r, where $r = 1$ is best), takes its reciprocal, and averages it across all the dataset samples ($MRR = 1/|D| \cdot \sum_i^{|D|} 1/r_i$). We consider only the first k retrieved images for each sample's contribution. If the target image is not within them, we approximate $1/r_i$ to 0 (MRR@k).

Measure (↑)	CLIP-Italian	mCLIP
MRR@1	**0.3797**	0.2874
MRR@5	**0.5039**	0.3957
MRR@10	**0.5204**	0.4129

Table 3
Results on MSCOCO image retrieval task. Best result in bold.

Table 3 reports the results for the image retrieval task, in terms of MRR@k, where $k \in \{1, 5, 10\}$. CLIP-Italian outscores mCLIP across the board.

5.2. Zero-shot Classification

The zero-shot image classification task replicates the experiment run by Radford et al. [1] on ImageNet. We first used DeepL to translate the image labels in ImageNet automatically. Then we prepend all test set labels with determiners and translate them (e.g., *a cat* is translated into "un gatto") and then prepended with the text "una foto di" (a photo of) as in "una foto di un gatto" (a photo of a cat) to obtain the final caption. This procedure is simpler than the one adopted by Radford et al. [1], where different templates are tested and averaged. Given an input image and the so-generated captions, we generate the embeddings (both for the image and all captions) and compute the similarities, assessing whether the correct image label corresponds to the closest caption in the embedding space.

Metric We compare models on the standard accuracy. Similarly to MRR@k, we consider a "hit" if the predicted class is within the top k retrieved and a "miss" otherwise. Similarly to the image retrieval task, we compute accuracy at k (Accuracy@k) with $k \in \{1, 5, 10\}$.

Table 4 reports the results for the zero-shot classification task. CLIP-Italian outscores mCLIP across the board.

5.3. Discussion

Our results across two tasks confirm that CLIP-Italian is very competitive and outperforms mCLIP on the two

[9] https://huggingface.co/dbmdz/bert-base-italian-xxl-cased
[10] https://huggingface.co/openai/clip-vit-base-patch32
[11] https://huggingface.co/sentence-transformers/clip-ViT-B-32-multilingual-v1

Measure (↑)	CLIP-Italian	mCLIP
Accuracy@1	**22.11**	20.15
Accuracy@5	**43.69**	36.75
Accuracy@10	**52.55**	42.91

Table 4
Results on ImageNet-1000 classification task. Best result in bold.

Figure 2: Result of the query "due cani sulla neve" (eng: two dogs on the snow) on Unsplash25K.

tasks we have been testing. Note that the performance for zero-shot ImageNet classification of the CLIP-Italian model (trained on 1.4 million image-text pairs) are lower than those shown in Radford et al. [1] (trained on 400 million image-text pairs). However, considering that our results align with those obtained by mCLIP, we think that the quality of the translated image labels most probably impacted the final scores.

6. Qualitative Evaluation

We examine some examples related to the image retrieval task on the Unsplash25K dataset.[12] Figure 2 shows the results of the query "due cani sulla neve" (two dogs on the snow), the model correctly finds the image, combining

[12]https://github.com/unsplash/datasets

Figure 3: Result of the query "una coppia al tramonto" (eng: a couple at the sunset) on Unsplash25K.

the concept of "snow" and the one of "two dogs".[13] We anecdotally find moderate numeracy capabilities during empirical evaluation, with sufficient ability to identify up to three distinct or repeated elements inside images, with a steep drop in coherence when more than three elements are present. Given the likely low number of training points depicting more than three subjects in a scene, we impute this finding to implicit bias in the training set. Figure 3 shows a similar performance for "una coppia al tramonto" (a couple during sunset), where the model could identify two people with sunlight in the background. A similar query, but with a mountain as a background, can be found in Figure 4. Despite the overall good performances, the model is inevitably subject to limitations and biases. For example, Figure 5 shows an image of a tiny hedgehog retrieved using the query "un topolino" (a tiny mouse). We leave a more thorough exploration of biases and stereotypes learned by the CLIP-Italian model to future work.

7. Conclusions

This paper presents the first CLIP model for the Italian language, trained on 1.4 million image-text pairs. The model shows promising zero-shot performance in two well-established tasks, suggesting many possible future applications.

Acknowledgments

This work was possible thanks to Hugging Face and Google which provided the computational resources to train CLIP-Italian. This project has also in part received

[13]Note, however, that compositional understanding in CLIP is limited, see [21]

Figure 4: Result of the query "una coppia in montagna" (eng: a couple in the mountains) on Unsplash25K.

Figure 5: Result of the query "un topolino" (eng: a small mouse) on the Unsplash25K dataset.

funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (No. 949944, INTEGRATOR). We thank Sri Lakshmi for her help on the project.

Ethical Considerations and Limitations

Large-scale models are difficult and costly to train, and important considerations have to be taken into account when developing them [22, 23]. We computed the cost of the different experiments we ran, and we estimated a total of 2,688$ for each TPU used. This result comes from the hourly cost of the TPU (8$) for 14 days; Note that we had access to a second TPU VM for part of the project and that, in this estimate, we are ignoring storage and data transfer costs. Strubell et al. [22] describe how these models can have a substantial environmental impact. As described by Bianchi and Hovy [24], these computational needs are quickly becoming unfeasible for many universities.

Moreover, recent evidence has shown that large-scale multimodal vision and language models exhibit biases in portraying several sociodemographic groups [25, 26, 27, 28, *inter alia*]. Moreover, the datasets on which these models have been trained on often contain harmful content [29]. As we build on pretrained vision and language models, we cannot exclude the presence of such biases. However, we want to point out that our vision and language models were pretrained on different language data. Hence, "concepts" in embedding spaces are not aligned. While we cannot exclude that models pick up biases from *our* training data, starting from unaligned embedding spaces can reduce the risk of unwanted biased associations.

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 8748–8763. URL: http://proceedings.mlr.press/v139/radford21a.html.

[2] S. Schneider, A. Baevski, R. Collobert, M. Auli, wav2vec: Unsupervised Pre-Training for Speech Recognition, in: Proc. Interspeech 2019, 2019, pp. 3465–3469. doi:10.21437/Interspeech.2019-1873.

[3] P. J. Chia, G. Attanasio, F. Bianchi, S. Terragni, A. R. Magalhães, D. Gonçalves, C. Greco, J. Tagliabue, Contrastive language and vision learning of general fashion concepts, Scientific Reports 12 (2022). URL: https://api.semanticscholar.org/CorpusID:253387447.

[4] E. Tiu, E. Talius, P. Patel, C. P. Langlotz, A. Y. Ng, P. Rajpurkar, Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning, Nature Biomedical Engineering 6 (2022) 1399–1406.

[5] S. Zhang, Y. Xu, N. Usuyama, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, C. Wong, et al., Large-scale domain-specific pretraining for biomedical vision-language processing, arXiv preprint arXiv:2303.00915 (2023).

[6] G. Chen, L. Hou, Y. Chen, W. Dai, L. Shang, X. Jiang, Q. Liu, J. Pan, W. Wang, mclip: Multilingual clip via cross-lingual transfer, in: Proceedings of the 61st Annual Meeting of the Association for Compu-

tational Linguistics (Volume 1: Long Papers), 2023, pp. 13028–13043.

[7] Z. Huang, F. Bianchi, M. Yuksekgonul, T. J. Montine, J. Y. Zou, A visual–language foundation model for pathology image analysis using medical twitter, Nature Medicine 29 (2023) 2307 – 2316. URL: https://api.semanticscholar.org/CorpusID:260970273.

[8] D. Nozza, F. Bianchi, D. Hovy, What the MASK? making sense of language-specific bert models, arXiv preprint arXiv:2003.02912 (2020).

[9] P. Rust, J. Pfeiffer, I. Vulić, S. Ruder, I. Gurevych, How good is your tokenizer? on the monolingual performance of multilingual language models, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 3118–3135. URL: https://aclanthology.org/2021.acl-long.243. doi:10.18653/v1/2021.acl-long.243.

[10] G. Sarti, M. Nissim, It5: Large-scale text-to-text pretraining for italian language understanding and generation, ArXiv abs/2203.03759 (2022).

[11] K. Srinivasan, K. Raman, J. Chen, M. Bendersky, M. Najork, Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning, arXiv preprint arXiv:2103.01913 (2021).

[12] A. Scaiella, D. Croce, R. Basili, Large scale datasets for image and video captioning in italian, IJCoL. Italian Journal of Computational Linguistics 5 (2019) 49–60.

[13] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: European conference on computer vision, Springer, 2014, pp. 740–755.

[14] P. Sharma, N. Ding, S. Goodman, R. Soricut, Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 2556–2565. URL: https://www.aclweb.org/anthology/P18-1238. doi:10.18653/v1/P18-1238.

[15] D. Hovy, F. Bianchi, T. Fornaciari, “you sound just like your father” commercial machine translation systems include stylistic biases, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 1686–1690. URL: https://www.aclweb.org/anthology/2020.acl-main.154. doi:10.18653/v1/2020.acl-main.154.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2021. URL: https://openreview.net/forum?id=YicbFdNTTy.

[17] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://www.aclweb.org/anthology/N19-1423. doi:10.18653/v1/N19-1423.

[18] J. Zhuang, T. Tang, Y. Ding, S. Tatikonda, N. Dvornek, X. Papademetris, J. S. Duncan, Adabelief optimizer: Adapting stepsizes by the belief in observed gradients, arXiv preprint arXiv:2010.07468 (2020).

[19] I. Loshchilov, F. Hutter, SGDR: stochastic gradient descent with warm restarts, in: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings, OpenReview.net, 2017. URL: https://openreview.net/forum?id=Skq89Scxx.

[20] N. Reimers, I. Gurevych, Making monolingual sentence embeddings multilingual using knowledge distillation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 4512–4525. URL: https://www.aclweb.org/anthology/2020.emnlp-main.365. doi:10.18653/v1/2020.emnlp-main.365.

[21] M. Yuksekgonul, F. Bianchi, P. Kalluri, D. Jurafsky, J. Zou, When and why vision-language models behave like bags-of-words, and what to do about it?, in: The Eleventh International Conference on Learning Representations, 2022.

[22] E. Strubell, A. Ganesh, A. McCallum, Energy and policy considerations for deep learning in NLP, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 3645–3650. URL: https://www.aclweb.org/anthology/P19-1355. doi:10.18653/v1/P19-1355.

[23] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big?, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021, pp. 610–623.

[24] F. Bianchi, D. Hovy, On the gap between adoption and understanding in NLP, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 3895–3901. URL: https://aclanthology.org/2021.findings-acl.340. doi:10.18653/v1/2021.findings-acl.340.
[25] J. Wang, Y. Liu, X. Wang, Are gender-neutral queries really gender-neutral? mitigating gender bias in image search, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 1995–2008. URL: https://aclanthology.org/2021.emnlp-main.151. doi:10.18653/v1/2021.emnlp-main.151.
[26] R. Wolfe, A. Caliskan, American== white in multimodal language-and-image ai, in: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, 2022, pp. 800–812.
[27] A. S. Luccioni, C. Akiki, M. Mitchell, Y. Jernite, Stable bias: Analyzing societal representations in diffusion models, arXiv preprint arXiv:2303.11408 (2023).
[28] F. Bianchi, P. Kalluri, E. Durmus, F. Ladhak, M. Cheng, D. Nozza, T. Hashimoto, D. Jurafsky, J. Zou, A. Caliskan, Easily accessible text-to-image generation amplifies demographic stereotypes at large scale, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 2023, pp. 1493–1504.
[29] A. Birhane, V. U. Prabhu, E. Kahembwe, Multimodal datasets: misogyny, pornography, and malignant stereotypes, arXiv preprint arXiv:2110.01963 (2021).

Modelling and Publishing the “Lexicon der indogermanischen Verben” as Linked Open Data

Valeria Irene Boano[1], Francesco Mambrini[1], Marco Passarotti[1] and Riccardo Ginevra[1]

[1]*Università Cattolica del Sacro Cuore, Milan, Italy*

Abstract

This paper describes the modelling and publication of part of the etymological information in the *Lexicon der indogermanischen Verben*, an etymological dictionary of verbs attested in ancient Indo-European languages, as Linguistic Linked Open Data. The lexicon has been made interoperable with a set of lexical and textual linguistic resources for Latin in the Lila Knowledge Base.

Keywords

Linked Open Data, Ontolex Lemon, Etymology, Etymological dictionary

1. Introduction

Over the past decades, several linguistic resources for historical languages have become available in digital format. This has given scholars the chance to access and exploit them in a quicker and deeper way.

In particular, several different linguistic resources are available for Latin today. They consist of textual corpora, such as the LASLA corpus[1] for Classical Latin and the *Index Thomisticus* Treebank [1] for Medieval Latin, and lexical resources, like the Lewis and Short dictionary [2] and the Logeion metadictionary[2].

Due to the centuries-long lexicographic tradition for the Latin language, its lexical resources comprise a number of etymological dictionaries. A dictionary is defined as etymological when it contains information about the etymology of its entries, that is about their origin and historical development: for Indo-European (IE) languages, etymological dictionaries often put their entries in relation with reconstructed Proto-Indo-European (PIE) roots, minimal lexical units to which the dictionary's entry and further related formations may be traced back. It is also often explained by which morphological processes the attested word has been formed from the root.

Even if the available resources (for Latin and beyond) provide a huge amount of linguistic information, at the present day, their full exploitation is still hindered by their isolation. In fact, most resources can be accessed only individually, and cannot interact.

Isolation of resources is an issue because each resource can really reach its potential only when it is made interoperable with other (types of) resources. Today, interoperability between linguistic resources can be obtained by describing and publishing their data according to the principles of the Linked Open Data paradigm [3]. As a consequence, in recent years, the amount of linguistic resources published as Linked Open Data has been raised substantially, as witnessed by the growing size of the Linguistic Linked Open Data Cloud (LLOD Cloud)[3]. In particular, the LiLa Knowledge Base[4] represents a successful example of Linked Open Data (LOD) principles applied to linguistic resources for Latin.

Although LiLa currently makes quite a number of linguistic resources for Latin interoperable, there is still a large set of digitized materials to interlink in the Knowledge Base. Among them is an important etymological dictionary, the *Lexicon der indogermanischen Verben* (LIV) [4]. This paper describes the process of transforming information contained in this dictionary into a LLOD resource, linked to LiLa. Section 2 describes the LiLa architecture and the LIV structure. Section 3 details the modelling of the resource, and the linking process. Section 4 describes some possible examples of exploitation and interaction. Finally, Section 5 discusses conclusions and sketches the future work.

2. The LiLa Knowledge Base and the *Lexicon der indogermanischen Verben*

2.1. The LiLa Knowledge Base

The LiLa Knowledge Base (KB) [5] is a linguistic hub for Latin, containing FAIR [6] linguistic resources, published as LOD. As usual in LLOD, structural interoperability

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

valeriairene.boano01@icatt.it (V. I. Boano); francesco.mambrini@unicatt.it (F. Mambrini); marco.passarotti@unicatt.it (M. Passarotti); riccardo.ginevra@unicatt.it (R. Ginevra)

[1]http://web.philo.ulg.ac.be/lasla/.
[2]https://logeion.uchicago.edu/.
[3]https://linguistic-lod.org.
[4]https://lila-erc.eu.

between resources is based on the Resource Description Framework (RDF) [7], which is the data model used for the Semantic Web [8]. Conceptual interoperability [9] is achieved by using common ontologies built and adopted by the LLOD community, such as the Ontolex Lemon model[5] and the OLiA ontology[6] [10].

LiLa is built around the so-called Lemma Bank, which contains a set of more than 200k Latin lemmas, taken from the database of the morphological analyser LEMLAT [11] and constantly extended. Each lemma of the Lemma Bank is a gateway between the different linguistic resources linked to the Knowledge Base, starting from the assumption that words (indexed by their lemmas) can be used as the point of contact between textual resources (which are made of occurrences of words), lexical resources (which describe words) and NLP tools (which process words).

Each entry in the LiLa Lemma Bank is an instance of `ontolex:Form`[7]. In particular, the `lila:Lemma`[8] is a form that can be linked to an `ontolex:LexicalEntry`[9] via the property `ontolex:canonicalForm`[10], which identifies the canonical form used to represent a lexical entry. Every other realization of a word is linked to the lexical entry via the property `ontolex:lexicalForm`[11]. Each lemma and each form may also be described with other properties, which give information, for example, about phonetic representation, Part of Speech (POS) tagging and other grammatical features.

Lexical resources are connected to LiLa by linking the `ontolex:LexicalEntry` of each resource to the `lila:Lemma` via the property `ontolex:canonicalForm`. Once a linguistic resource is linked to the KB via the Lemma Bank, all the interoperable resources can be queried together, using a SPARQL endpoint[12] also through a user-friendly interface[13].

The textual resources connected to LiLa so far include more than 3,5M words from Latin texts of different eras, such as the LASLA corpus[14], the *Index Thomisticus* Treebank [12], containing works of Thomas Aquinas, and *UDante*, a Universal Dependencies[15] treebank for Dante Alighieri's Latin works [13]. The lexical resources of LiLa include a derivational lexicon, *Word Formation Latin* [14], a manually checked subset of the Latin WordNet connected to a valency lexicon [15], the *Etymological dictionary of Latin and other Italic Languages* [16, 17] and a resource of principal parts of Latin words, *PrinParLat* [18].

2.2. The *Lexicon der indogermanischen Verben*

The *Lexicon der indogermanischen Verben*, also known as LIV, is an etymological dictionary of verbs attested in ancient Indo-European languages. After the first edition, curated by Helmut Rix [19] and published by Reichert Verlag in 1998, a second edition was published in 2001 with additions and corrections by Martin Kümmel and Helmut Rix [4].

The LIV is the main reference work for Proto-Indo-European verbal roots and contains three types of information:

- **Reconstructed Proto-Indo-European verbal roots**, which coincide with the entries of the dictionary and are provided with their presumed lexical meaning and their phonological structure. For each root, the corresponding index in the *Indogermanisches etymologisches Wörterbuch* [20] is specified as well.
- **Reconstructed Proto-Indo-European primary verbal stems**, which are either root formations or are formed by adding to the roots primary affixes that mainly express categories of aspect and actionality. The meaning of the stem is usually not specified.
- **Word forms that are historically attested in ancient IE languages**, which show how the Proto-Indo-European stems evolved in the various daughter languages. Each attested form is provided with its lexical meaning in the respective language. At the end of certain entries are sometimes listed innovative verbal stems that may ultimately be traced back to Proto-Indo-European roots, but are unlikely to directly reflect Proto-Indo-European verbal stems, having been created according to language-specific productive patterns.

The original data used during the linking process consists of a spreadsheet containing information from the LIV extracted and structured by Thomas Olander, with the collaboration of Simon Poulsen and Anders Richardt Jørgensen, and shared with us by the authors. For each of the 7,888 LIV entries the spreadsheet records its root, stem and attested forms.

The LIV is copyrighted by the Reichert Verlag, and LiLa is not authorized to reproduce the full content of the dictionary. The publisher has however agreed to allow us to model the basic etymological relations between

[5] https://www.w3.org/2016/05/ontolex/.
[6] https://acoli-repo.github.io/olia/.
[7] http://www.w3.org/ns/lemon/ontolex#Form.
[8] https://lila-erc.eu/lodview/ontologies/lila/Lemma.
[9] http://www.w3.org/ns/lemon/ontolex#LexicalEntry.
[10] http://www.w3.org/ns/lemon/ontolex#canonicalForm.
[11] http://www.w3.org/ns/lemon/ontolex#lexicalForm.
[12] https://lila-erc.eu/sparql/.
[13] https://lila-erc.eu/query/.
[14] https://www.lasla.uliege.be/cms/c_8508894/fr/lasla.
[15] https://universaldependencies.org/.

the PIE roots, the stems and the Latin words and stems, provided that explicit bibliographical attribution is given to the linguistic reconstruction.

This is the information that we modelled to be linked to the LiLa Knowledge Base, as described in the following section.

3. Modelling and linking the LIV

Making linguistic resources interoperable means using a shared set of vocabularies for knowledge description, as defined in specialized ontologies, to represent the information contained in them. The process of linking, on the other hand, aims to connect this information to a wider network of data, so that a meaningful context is provided[16]. Within the network of LiLa, this step means that all entries of a lexical resource must make reference to the canonical forms of the Lemma Bank, as described above.

This section details how we modelled our target information from the LIV, how we applied such modelling to the publication of these data as LOD and how we linked them to the LiLa collection.

3.1. Modelling

In the lexical resources linked to LiLa, etymological information has been expressed using the `lemonEty` extension of the Ontolex-Lemon model [21]. This ontology was used in LiLa to represent loanwords from Greek [22] and for the *Etymological Dictionary of Latin and the other Italic Languages* [16, 17].

The set of classes and properties of `lemonEty` are suitable to express the etymological information of the LIV too, but, compared to the aforementioned dictionaries in LiLa, a more complex modelling and a series of extensions are also required.

The `lemonEty` ontology establishes etymological relations between instances of the Ontolex's class `LexicalEntry`. In particular, a special subclass called `Etymon` is reserved for lexical items of the source language that are introduced in order to explain the history of the entries in the target language.

Two core classes of `lemonEty` that are particularly important are `Etymology`[17], and `EtyLink`[18]. The former "reifies the whole process of etymological reconstruction as scientific hypothesis" [17, p. 22]. Etymological links, on the other hand, connect "linguistic elements" from the source language to the corresponding elements of the target.

In applying this model to the LIV data, it is crucial to define what the "linguistic elements" connected via etymological links are. The previously mentioned lexical resources rested on a simple model where the etymological links involved only Latin lexical entries and etymons from a source language, so that e.g. the Latin word *abacus*[19] was the target of a link that had its source in the Greek etymon *ábax* 'reckoning board'.

In the LIV, on the other hand, relations are established between:

- **Inflected forms** of a historical language (e.g. Latin). In the case of Latin, those forms are used in the LIV as placeholders for all forms derived from the same **stem**; so, for instance, the Latin 1st-person perfect *fidi* stands for all forms from the perfect stem of the verb *findo* 'to cleave, split';
- **the PIE stems**, to which the inflected forms and stems of Latin (and other languages) must be traced back;
- **the PIE root** that underlies the PIE stems.

In the case of Latin, thus, the LIV documents etymological relations between a PIE and a Latin stem (the latter represented by a Latin inflected form). While the PIE root (e.g. *b^he̯id-) and the Latin target lexical item (e.g. *findo*, inclusive of all its stems) can be conceptualised as lexical entries, the stems and the word forms must be described using concepts from other vocabularies.

For the Latin forms and stems we reused the individuals of the class `Stem`[20] provided by *PrinParLat*, a lexical resource listing all Latin "principal parts". Principal parts are sets of inflected wordforms from which the content of all the other paradigm cells can be inferred[21].

For the perfect stem, *PrinParLat* already includes all forms linked to their `Stem` therein, which could thus be immediately reused. As for the present stems, however, the related forms were not available, and had therefore to be generated and linked to their `Stem` via the property `ontolex:lexicalForm`[22].

Some specific information provided by the dictionary that could not be represented with any of the available

[16]This is the fifth and final step in Berners-Lee's five-star rating system: see https://www.w3.org/DesignIssues/LinkedData.html.

[17]http://lari-datasets.ilc.cnr.it/lemonEty#Etymology.

[18]http://lari-datasets.ilc.cnr.it/lemonEty#EtyLink.

[19]http://lila-erc.eu/data/lexicalResources/IGVLL/id/LexicalEntry/abacus.

[20]https://lila-erc.eu/lodview/ontologies/prinparlat/Stem. This class will be defined as a sub-class of `morph:Morph`, when the module will be released.

[21]https://github.com/CIRCSE/PrinParLat.

[22]Note that, following the conventional definition of a form in *PrinParLat*, which is more granular than the one used by LiLa, we chose to recreate the form for the same paradigmatic slot that is traditionally chosen as lemma. *PrinParLat* creates a unique form for each graphical or spelling variant of a form; lemmas in the Lemma Bank, on the other hand, unify them under a single form. In many cases, this choice results in a duplicated form for the 1st-person singular present indicative verb.

modules required the creation of *ad hoc* classes. In particular, some Latin stems that trace back to PIE roots, but are unlikely to directly reflect a PIE stem, are classified by the LIV as *Neubildungen*, that is 'innovations', since they have been created according to language-specific productive patterns. These innovations cannot be traced back to a PIE stem, so that no etymological link can be created. We therefore created a specific class `Innovation`, which contains all those innovative stems.

Moreover, some Latin entries are defined by the dictionary as 'remodelings' (*Umbildungen*): their stems may be traced back to PIE stems, but have been reshaped following language-specific productive patterns (e.g. Latin *fodio* has lost the first syllable of the PIE reduplicated stem **bʰé-bʰodʰh₂/bʰdʰh₂-*). The remodeled Latin stems are now defined as instances of the new class `Remodeling`.

3.2. Linking

LIV provides etymological information for other IE languages in addition to Latin. Since, however, LiLa is limited to Latin resources, we restricted our attention only on entries where a connection to Latin forms was explicitly mentioned.

In total, we identified 550 Latin forms linked to PIE roots, 354 of which corresponded to the main lemma of a verb; the remaining 196 were instead analysed as inflected forms.

The forms were analysed with the UDPipe pipeline[23], in order to perform the POS tagging of all forms and lemmatization of 196 inflected forms. The results were manually checked, which confirmed a good accuracy of 97% for POS-tagging (only 11 cases were incorrectly tagged), but much lower performances for lemmatization (87 out of 196, i.e. 44%).

For each of the remaining lemmas in the manually corrected set, we created a lexical entry in our new etymological resource. The canonical forms of these entries were identified by matching the lemma strings with the written representations[24] of the lemmas in the LiLa Lemma Bank. In 143 cases, manual disambiguation was needed, as the query returned more than one possible match. In one case, it was not possible to link the form (*tātōd*) to any lemma in the Lemma Bank: we decided against adding the invariable form to the Lemma Bank and instead created a LIV `LexicalEntry` *tātōd*, without connecting it to any lemma.

Moreover, a set of 11 entries required a special treatment. For a series of entries, like for instance $*pleh_1$- [4, p. 482], in fact, the LIV does not point to a single Latin word form, but rather to a whole lexical root that is analogous to LiLa's lexical base [5, p. 191]. This morpheme represents a lexical element that is neither a prefix nor a suffix and is shared by all members of a derivational family. Comparably, for instance, the Latin hyphenated form *-pleo* in the entry $*pleh_1$- is used in the LIV as a placeholder for all the possible Latin verbs that can be formed adding different preverbs to the same base (e.g. *compleo* 'to fill up', *depleo* 'to empty', *expleo* 'to fill up'...).

In those cases where the LIV uses this notation, we chose to create one lexical entry for each verb connected to the corresponding lexical base in LiLa (e.g. the 'base of *pleo*'[25]).

Once the lexical entries had been created, the correct stems in the *PrinParLat* resource were easily identified by leveraging the advantages of the LOD model. In fact, each LiLa's lemma is linked to the appropriate stems via an instance of the *PrinParLat* class of `Flexeme`[26]; the stems for a lexical entry are therefore easily recoverable once the LiLa lemma is known.

3.2.1. A LIV lexical entry linked to LiLa

The Figure 1 is taken from the LodLive[27] visualization in LiLa. It shows an example of how a LIV lexical entry (*glubo* 'to peel') was modelled and linked to the Lemma Bank.

On the left side of the figure is the LIV `LexicalEntry` *glubo*, which is linked to LiLa's lemma *glubo* via the property `canonicalForm`: this simple but crucial link allows us to connect the LIV etymological relations with the other resources of the Knowledge Base.

Then, the LIV `LexicalEntry` is connected via the property `lexicalRel`[28] to two *PrinParLat* Latin stems, the present stem *glub-* and the perfect stem *glups-*. Each stem is connected with a Latin form: the present form *glubo* is part of the LIV resource, and is thus linked to the present stem via the property `lexicalForm`; on the other hand, the perfect form *glupsi* is part of the *PrinParLat* resource, and is thus linked to the perfect stem via the property `consistsOf`.

We then link the Latin stems to their Proto-Indo-European ancestors. Each of the two Latin stems is in fact the `etyTarget` of an `EtyLink` (*Etymology link: pres glubo* and *Etymology link: perf glupsi*), which connects them to their `etySource`, that is the corresponding PIE stem (for the present **g/ĝléu̯bʰ-/g/ĝlubʰ-*, and for the perfect *?*g/ĝlḗu̯bʰ/g/ĝléu̯bʰ-s-*). These etymological links

[23] https://lindat.mff.cuni.cz/services/udpipe/.
[24] In Ontolex Lemon, written representations are the different graphical variants of a form. See http://www.w3.org/ns/lemon/ontolex#writtenRep.
[25] http://lila-erc.eu/data/id/base/107. Note that we excluded those verbs that are only documented in lexicons of Medieval Latin, such as the Du Cange dictionary [23].
[26] http://lila-erc.eu/ontologies/prinparlat/Flexeme.
[27] "LodLive project provides a demonstration of the use of Linked Data standards (RDF, SPARQL) to browse RDF resources". http://en.lodlive.it/.
[28] http://www.w3.org/ns/lemon/vartrans#lexicalRel.

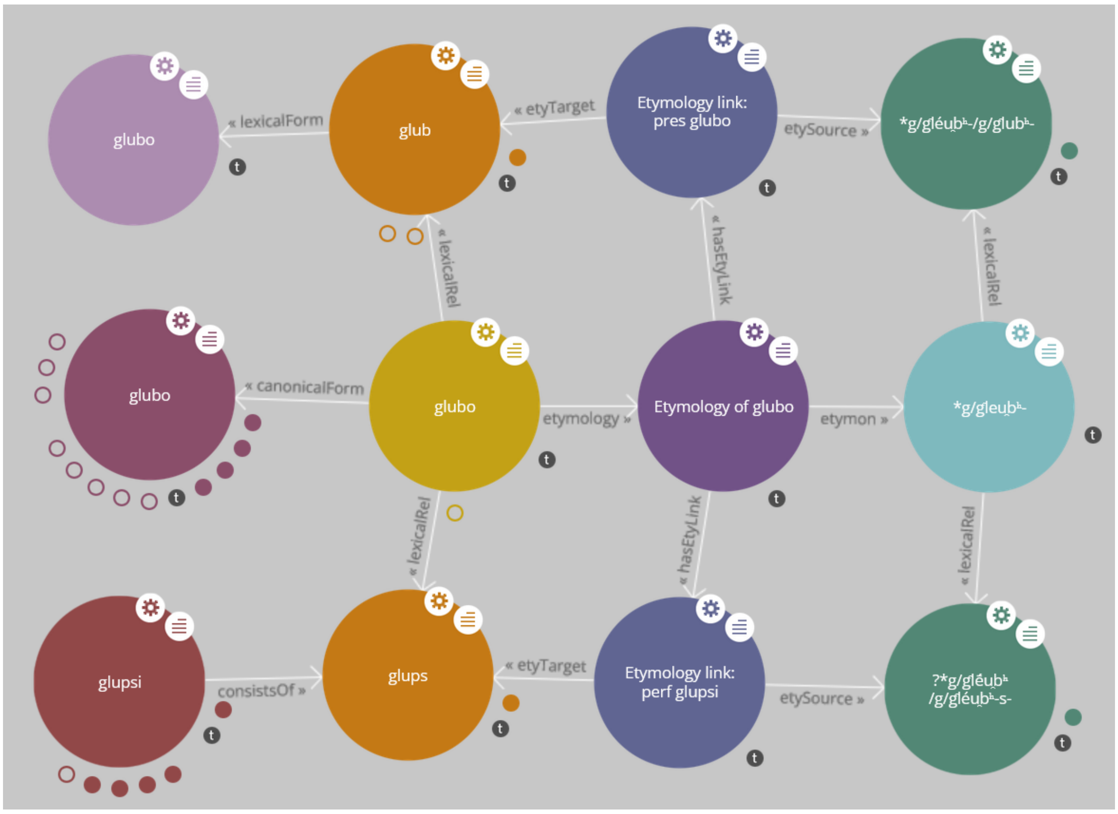


Figure 1: The linking of LIV etymological relations: the case of *glubo*.

reify the etymological relations that the LIV postulates between the stems, and constitute the bridge between Latin and PIE.

On the right side of the figure is the PIE symmetrical counterpart of the model. The PIE root **g/ĝléu̯b^h*- (which is an individual of the class `Etymon`, subclass of the class `LexicalEntry`) is linked to the two PIE stems via the property `lexicalRel`, in the same way as the `LexicalEntry` is linked to the Latin stems.

Finally, the generic etymological relation between the PIE root and the Latin lexical entry is reified by the `Etymology` class: this class establishes a link between them via the properties `etymon` and `etymology`, respectively. The `Etymology` is also connected with the two `EtyLink`, thanks to the property `hasEtyLink`, and thus constitutes a central crossroad between the LIV lexical items.

4. Querying the LIV data in LiLa

The modelling and linking work has, as shown above, benefited greatly from the advantages provided by the LOD paradigm. The re-use of the lemmas from the LiLa Lemma Bank as canonical forms for the LIV entries has allowed us to retrieve the stems from *PrinParLat*, as well as all words derived from a handful of selected lexical bases.

The exploration of the full set of words derived via the regular Latin word-formation rules from verbs of IE origin can be extended to all the entries in the LIV, beyond the 11 entries explicitly marked as bases in the dictionary. In fact, 342 regular entries of the LIV currently linked to a LiLa lemma are connected to a lexical base. Via this relation, we can access a set of 4,019 other verbs. Also, by leveraging the links in the LiLa network to textual resources, we can easily access the earliest occurrences in the corpora.

The full network of the resources linked to LiLa allows for even more advanced inquiries in historical linguistics

Entry URI	Lexical Base	Nr. of verbs
`liv:5`	Base of *facio*	256
`liv:147`	Base of *ago*	71
`liv:157`	Base of *fio*	67
`liv:252`	Base of *fero*	53
`liv:167`	Base of *capio*	51
`liv:260`	Base of *eo*	49

Table 1
LIV Entries connected to the most productive lexical bases in LiLa, with nr. of verbs (LIV entry excluded) connected to it.

and in the study of Latin lexicon. Table 1[29] reports the most productive bases connected to entries in the LIV, with the number of other verbs linked to each lexical base (note that the lemma of the LIV entry was excluded from the calculation).

As can be seen, the most productive words are some of the most common verbs belonging to the oldest IE substratum of Latin, like *facio* 'to do, make', *fero* 'to bring' or *capio* 'to seize, take'. Indeed, a joint query between the two dictionaries with IE etymologies, viz. LIV and [16], and the lexical bases in LiLa confirm that the bases that have at least one lemma that is traced back to PIE are considerably richer and more productive than those lexical families without any inherited lexemes. While the former have on average 23.70 members, the latter display only an average of 4.86 members. This fact can be easily explained considering that the latter group is mostly made up by loanwords, which are generally technical terms (especially from the Greek scientific or technical lexicon), and tend to be more specialised and less productive in terms of word formation.

The two dictionaries combined provide now information on PIE etymologies for 1,473 lemmas: 1,393 are connected to entries in the *Etymological Dictionary of Latin and the Other Italic Languages* [16, 17], 355 in the LIV. In particular, 275 lemmas are shared by the two resources: for these entries, it is therefore now possible to use LiLa to compare the approach to etymological reconstruction by the LIV and the dictionary by de Vaan.

5. Conclusion and Future Work

Linking a set of data from the LIV to LiLa enhances the Knowledge Base with etymological information about the processes that, starting from PIE roots, have led to the formation of the Latin word forms. Given the highly lexically-based nature of the architecture of the Knowledge Base, this makes the linking of the LIV an important achievement of the LiLa project.

The information provided by the LIV is now interoperable with that of the several other lexical resources currently interlinked through LiLa and can be queried together with the textual data provided by the Latin corpora published in the Knowledge Base.

In collecting and publishing as LOD the wealth of different digital resources for Latin built so far, an important challenge is to impact the scholarly community that has long been using the data provided today by these resources. The web-based interoperability among resources permitted by the LiLa Knowledge Base makes it possible to exploit such wealth of (meta)data like never before, in terms both of the quantity of the (meta)data under analysis and of the quality of the process leading to their retrieval. Interlinking through the Knowledge Base a set of data from the LIV, a reference lexical resource for the communities of Classicists and Historical Linguists, is expected to help overcome the challenge of making the use of LiLa a daily presence in the life of scholars who work in the fields of Classics and Historical Linguistics.

Finally, it is worth considering that the LIV provides etymological information not only about the Latin word forms, but, for each PIE root, it also reports a set of word forms which reflect the same root in several other IE daughter languages. The availability of this information allows for substantial research work to be performed in the near future. Indeed, by applying the principles of the Linked Data paradigm and reusing the same vocabularies adopted in LiLa to interlink the distributed linguistic resources for Latin, it is now possible to move one step further from the Latin language and aim to make interoperable word forms from several IE languages, by using the collection of PIE roots provided by the LIV as a kind of pivot resource to interlink them all.

6. Acknowledgments

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. We thank the Reichert Verlag publisher for granting us permission to use some data from the LIV. We thank Matteo Pellegrini for his valuable insights about PrinParLat.

References

[1] M. Passarotti, The project of the index thomisticus treebank, Digital classical philology. Ancient Greek and Latin in the digital revolution 10 (2019) 299–319.

[2] C. Lewis, C. Short, A Latin Dictionary. Founded on Andrews' edition of Freund's Latin dictionary, Clarendon Press, Oxford, 1879.

[29] The namespace `liv` in Tab. 1 refers to the URL http://lila-erc.eu/data/lexicalResources/LIV/id/LexicalEntry/.

[3] C. Chiarcos, P. Cimiano, T. Declerck, J. P. McCrae, Linguistic linked open data (LLOD). introduction and overview, in: Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and linking lexicons, terminologies and other language data, Association for Computational Linguistics, Pisa, Italy, 2013, pp. i – xi. URL: https://aclanthology.org/W13-5501.

[4] H. Rix, ed., LIV. Lexikon der indogermanischen Verben. Die Wurzeln und ihre Primärstammbildungen, 2nd ed., Reichert Verlag, Wiesbaden, 2001.

[5] M. Passarotti, F. Mambrini, G. Franzini, F. M. Cecchini, E. Litta, G. Moretti, P. Ruffolo, R. Sprugnoli, Interlinking through lemmas. the lexical collection of the lila knowledge base of linguistic resources for latin, Studi e Saggi Linguistici 58 (2020) 177–212.

[6] M. Wilkinson et al., The fair guiding principles for scientific data management and stewardship., Scientific Data 3 (2016). doi:https://doi.org/10.1038/sdata.2016.18.

[7] O. Lassila, R. R. Swick, Resource Description Framework (RDF) Model and Syntax Specification, 1998. URL: https://www.w3.org/TR/1999/REC-rdf-syntax-19990222/.

[8] T. Berners-Lee, The semantic web, Scientific American 284 (2001).

[9] N. Ide, J. Pustejovsky, 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, 2010.

[10] C. Chiarcos, M. Sukhareva, Olia – ontologies of linguistic annotation, Semantic Web 6 (2015) 379–386. doi:10.3233/SW-140167.

[11] M. Passarotti, M. Budassi, E. Litta, P. Ruffolo, The lemlat 3.0 package for morphological analysis of latin, in: Proceedings of the NoDaLiDa 2017 workshop on processing historical language, 2017, pp. 24–31.

[12] F. M. Cecchini, M. Passarotti, P. Marongiu, D. Zeman, Challenges in Converting the Index Thomisticus Treebank into Universal Dependencies, in: Proceedings of the Second Workshop on Universal Dependencies (UDW 2018), Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 27–36. URL: https://aclanthology.org/W18-6004. doi:10.18653/v1/W18-6004.

[13] F. Cecchini, R. Sprugnoli, G. Moretti, M. Passarotti, UDante: First Steps Towards the Universal Dependencies Treebank of Dante's Latin Works, in: Seventh Italian Conference on Computational Linguistics (CLiC-it 2020), Bologna, 2020.

[14] E. Litta, M. Passarotti, F. Mambrini, 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, Institute of Formal and Applied Linguistics, Charles University in Prague, Prague, Czech Republic, 2019, pp. 35–43. URL: https://ufal.mff.cuni.cz/derimo2019/pdf-files/derimo2019.pdf.

[15] F. Mambrini, M. Passarotti, E. Litta, G. Moretti, Interlinking Valency Frames and WordNet Synsets in the LiLa Knowledge Base of Linguistic Resources for Latin, in: M. Alam, P. Groth, V. de Boer, T. Pellegrini, H. J. Pandit, E. Montiel, V. Rodríguez Doncel, B. McGillivray, A. Meroño-Peñuela (Eds.), Further with Knowledge Graphs. Studies on the Semantic Web 53, IOS Press, Amsterdam, 2021. URL: https://ebooks.iospress.nl/doi/10.3233/SSW210032. doi:10.3233/SSW210032.

[16] M. de Vaan, Etymological Dictionary of Latin and the Other Italic Languages, Brill, Leiden and Boston, 2008.

[17] F. Mambrini, M. Passarotti, Representing Etymology in the LiLa Knowledge Base of Linguistic Resources for Latin, in: Proceedings of the Globalex Workshop on Linked Lexicography. LREC 2020 Workshop, European Language Resources Association (ELRA), Paris, 2020, pp. 20–28. URL: https://lrec2020.lrec-conf.org/media/proceedings/Workshops/Books/GLOBALEX2020book.pdf. doi:10.5281/zenodo.3862156.

[18] M. Pellegrini, Flexemes in theory and in practice, Morphology (2023) 1–35.

[19] H. Rix, ed., LIV. Lexikon der indogermanischen Verben. Die Wurzeln und ihre Primärstammbildungen, Reichert Verlag, Wiesbaden, 1998.

[20] J. Pokorny, Indogermanisches etymologisches Wörterbuch (IEW), Francke Verlag, 1959.

[21] A. Khan, Towards the representation of etymological data on the semantic web, Information 9 (2018).

[22] G. Franzini, F. Zampedri, M. Passarotti, F. Mambrini, G. Moretti, Græcissâre: Ancient Greek Loanwords in the LiLa Knowledge Base of Linguistic Resources for Latin, in: Proceedings of the Seventh Italian Conference on Computational Linguistics. Bologna, Italy, March 1-3, 2021, CEUR-WS.org, Bologna, 2020, pp. 1–6. URL: http://ceur-ws.org/Vol-2769/paper_06.pdf.

[23] C. Du Cange, Bénédictins de Saint-Maur, P. Carpentier, L. Henschel, L. Favre, Glossarium Mediae et Infimae Latinitatis, Léopold Favre, Niort, 1883-1887.

How Green is Sentiment Analysis? Environmental Topics in Corpora at the University of Turin

Cristina Bosco[1,4], Muhammad Okky Ibrohim[2], Valerio Basile[1] and Indra Budi[3]

[1]*Dipartimento di Informatica - Università degli Studi di Torino, Italy*
[2]*langing.ai, Indonesia*
[3]*Faculty of Computer Science - Universitas Indonesia, Indonesia*
[4]*CRISIS - Centro di Ricerca Interuniversitario sui cambiamenti Socio-ecologici e la transizione alla Sostenibilità, Italy*

Abstract

Despite the unanimous recognition of the plight associated with environmental phenomena and the proliferation of the discourse about it, there is still little work on these issues in the field of NLP. This paper provides a report on the activities we are carrying on at the University of Turin in the application of Sentiment Analysis to environmental topics. In pursuit of the goal of developing resources and tools specifically designed for addressing the complexity of the ongoing environmental debate, we are currently focused on exploring the language used for green issues and defining some annotation schemes that can describe them at different granularity.

Keywords

environment, corpora, sentiment analysis

1. Introduction

It has become increasingly common to apply Sentiment Analysis (SA) and text classification to issues with social impact about which people debate. On the one hand, studying a socially impacting phenomenon from such a computational perspective means creating a precise conceptual and linguistic model, thereby achieving a greater understanding of its characteristics, its dynamics, and, not least, how people perceive it. On the other hand, it is a matter of creating tools that can help policymakers and citizens define strategies to address the problems associated with the phenomenon, bearing in mind that the impact of an intervention depends meaningfully on how it is proposed by governments and political parties and accepted by citizens.

Among the issues that have a unique social importance today are certainly those related to the environment in which we live. As far as the emergency related to the environment, at first sight, one cannot but notice that the environmental issues underlie a great complexity. This is due to the mixing of natural and human entities and related interests, such as individuals, public and private organisations on the one side, and climate, animals and plants on the other one. The language used to describe and discuss environmental topics also mirrors this complexity and is featured by a certain degree of specialization.

Modelling this reality can be therefore especially complex but also particularly useful because it ultimately allows us to better understand the relationship between humans and the environment and to be more aware of the sensitivity towards the environment which is hidden in us.

The characteristics of the discourse about the environment can make especially challenging the classification of opinions expressed about it. We may hypothesize that an accurate annotation of data about environmental topics can be helpful in order to achieve reliable results, e.g., in the detection of the polarity or stance in these texts. According to this hypothesis, we are following two major directions: a) to preliminary analyze the linguistic features of the discourse about the environment carried on in different text genres and b) to design specific annotation schemes that take into account the specific features of these texts and to apply them on selected corpora.

The first direction allowed us to better understand the meaning of the wide-spreading discussion about the language used in *green* communication. This was also useful in preparing the ground for the second direction of research, in which we want to model a specific form of communication about *green* issues, namely that realized in social media. Notwithstanding the relevance of the topics we are addressing, in agreement with the results of the systematic survey of the studies about SA applied to the environment [1], it can be observed that currently in this research area there is a gap and we want to fill it out. Only a few projects indeed exist, also for English, in

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
cristina.bosco@unito.it (C. Bosco); muhammadokky.ibrohim@unito.it (M. O. Ibrohim); valerio.basile@unito.it (V. Basile); indra@cs.ui.ac.id (I. Budi)
0000-0002-8857-4484 (C. Bosco); 0000-0002-6943-6553 (M. O. Ibrohim); 0000-0001-8110-6832 (V. Basile); 0000-0002-2107-6552 (I. Budi)

which environmental topics are addressed by applying SA and in which only fairly rough techniques were used.

In this paper, we describe a variety of experiences carried on at the Department of Computer Science of the University of Turin in the development of corpora and tools for SA applied to environmental topics during the last few years.

The paper is organized as follows. The next section briefly surveys previous work related to the application of SA to environmental topics. Section three focuses on the collection of data, while the fourth is about the annotation schemes we adopted. Finally, the last section provides some conclusions and hints about our future works.

2. Background

There is a huge amount of divulgation and communication about environmental issues related in particular to products and services. A 2020 EU Commission study found that more than half of the environmental claims examined in the EU were vague, misleading or unfounded, while 40% were completely unfounded[1]. In section 3.3, we moreover show that it can be difficult for citizens to understand the exact meaning of texts discussing issues related to the environment, making easier to mislead their content.

To explore SA applied to environment topics, researchers have conducted reviews and surveys providing different perspectives. In particular, in [2], a review is conducted to explore the application of SA in the climate change debate. [3] explore the use of SA for analyzing opinions on several smart city issues like climate change, urban policy, energy, and traffic. While [2] explore papers that used various types of data sources (i.e. news articles, social media, etc.), [3] explore only papers that analyze sentiment in social media. However, both [2] and [3] do not provide an in-depth exploration of the NLP techniques (from the creation of dataset to the evaluation of SA models) that researchers used applying SA on natural environment topics, since they only cover a few among the large variety of topics closely related to nature and environment, like food or carbon issues.

3. Exploring *Green* Language

The first step in our investigation consisted of a linguistic analysis of the discourse about the environment and we applied it to documents from public institutions or online journals to inform citizens about these topics. Applying a multilingual perspective we collected texts from an institutional website in Italian and English, and from some Italian journals in which are discussed environmental topics. The first sample of data, described in section 3.1, is the result of a random collection while the second one, described in section 3.2, is collected using keywords about a specific topic related to the environment, i.e. livestock.

3.1. European Environment Agency

The *European Environment Agency*[2] (EEA) is an agency of the European Union that delivers knowledge and data to support Europe's environment and climate goals. Since 1994, EEA and the *European Information Network Environmental training and observation* [3] (Eionet) provides data and information on Europe's climate and environment to citizens and decision-makers European politicians, publishing articles and more extensive reports which address the state of air quality, or a set of inter-connected or systemic issues, such as the mobility system.

We collected Italian and English data from the EEA website and we built two comparable corpora composed of 10 reports each. The Italian corpus (henceforth EEA-Ita) includes 14,612 tokens corresponding to 556 sentences, while the English corpus (henceforth EEA-Eng) is composed of 11,778 tokens corresponding to 562 sentences.

A qualitative analysis based on the lists of frequency, obtained with SketchEngine, shows that the most used terms in both corpora, Italian and English, refer to the theme of sustainable-environmental quality, but with a slight nuance that differentiates the Italian with respect to English. The most frequent terms in the Italian corpus concern especially the sphere of the fight against the conservation of oceans and seas, the sustaining of the Earth's ecosystem and conservation. In the English corpus, instead, we find a higher frequency of terms related to climate change. In both cases, these are not terms of high specialisation, that is, terms that are difficult to understand by the great majority of citizens, but technical terms relating to the field of reference, and therefore not easily traceable in other contexts. For example, in the Italian corpus, we can highlight words such as "siccità" (drought), "effetto serra" (greenhouse effect), "ecosistema" (ecosystem), "inquinamento" (pollution), "suolo" (soil), "microplastiche e nano plastiche" (microplastics and nano plastics), while in the English one "pollution", "climate change", "adaptation", "mitigation", "habitat".

3.2. Livestock Issues

The livestock sector is currently at the center of a heated debate that has focused mainly on intensive farming. Among the several publications in which these issues are

[1] https://quifinanza.it/green/stop-al-greenwashing-in-etichetta-cosa-vuole-fare-lue/699054/

[2] https://www.eea.europa.eu/en

[3] https://www.eionet.europa.eu/

presented and discussed, we selected a sample of texts from online journals, namely mostly from *CREA Futuro* but also from *L'informatore agrario* and *agricultura.it*. Our corpus is composed of 20,854 words (4,386 different lemmas) corresponding to 24,383 tokens, organized into 725 sentences and 21 documents.

CREA Futuro is an initiative of CREA (*Consiglio per la Ricerca in Agricoltura e l'analisi dell'Economia agraria*)[4], the leading Italian research organization dedicated to the agri-food supply chains, supervised by the Ministry of Agriculture, Food Sovereignty and Forests, and organized in 12 research centres. This online publication[5] is aimed at citizens to combine authoritative information, based on scientific evidence. From the CREAfuturo website, we selected a sample composed of 11 documents. The other texts are from the freely accessible web version of two journals, namely *L'informatore agrario*[6] (8 documents) and *agricultura.it*[7] (2 documents).

As expected the frequency lists collected using SkecthEngine show that the words occurring more than 40 times are "produzione" (production), "animali" (animals), "carne" (meat), "acqua" (water), "latte" (milk), "allevamento" (farming), "zootecnia" (livestock), "benessere" (welfare) and "stress.

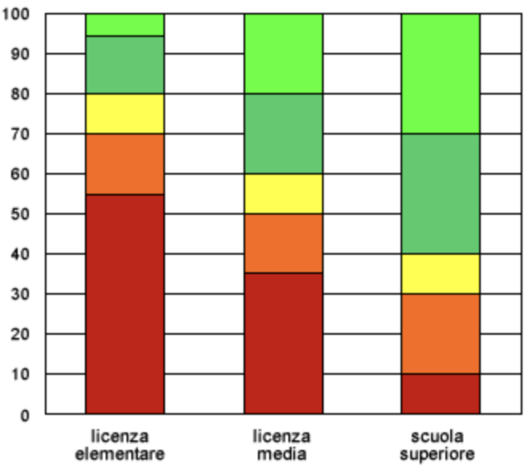


Figure 1: The scales for readability according to the Gulpease index for the three main levels of schooling (primary, secondary and high school): indexes in red for almost unreadable, in orange for very hardly readable, in yellow hardly readable, in dark green easy readable and in light green very easy readable.

3.3. How difficult is to read *green* texts?

All the texts we collected about *green* topics are intended for a general audience, but we want to understand how specialized they are, and thus less or more readable for a citizen. We calculated the readability scores for each of them. Different metrics are used for expressing the readability of different languages and we selected two of the most used ones for the two observed languages.

For Italian texts, we used the Gulpease index[8] whose scales are reported in Figure 1. The Gulpease index has been separately calculated for the 10 reports of the EEA-Ita corpus, showing values that vary from 45 to 53, for the less and the more readable text respectively (see Table 1). This means that the reports are unreadable for readers having primary school diplomas, but hard readable for readers having secondary school diplomas and easily readable for the other ones. According to this index, our texts are on average readable and not particularly specialized with the exception of some terms.

The Gulpease index was calculated also for the 21 documents of the Livestock-Ita corpus showing that are also less readable than the EEA's reports. Considering that the index of the harder-to-read document has a Gulpease index of 28 and the easier an index of 45, they are also featured in a larger variation.

Finally, we used the Flesch–Kincaid index[9] for evaluating the readability of English texts. The values of this index broadly correspond to those of the Gulpease index: values from 100 to 90 are associated with very easy readable texts, from 89 to 80 with easy readable, from 79 to 70 with fairly easy readable, and from 69 to 60 with standard readable. Values below 59 are instead associated with difficult-to-read texts: from 59 to 50 fairly difficult, from 49 to 30 difficult and from 29 to 0 very difficult or almost unreadable without a higher level of schooling.

Corpus	Max G	Min G	Var G
EEA-Italian	53	45	8
lives-Italian	45	28	17
	Max F	**Min F**	**Var F**
EEA-English	46.25	20.24	26.01

Table 1
Indexes of readability: Gulpease index for Italian data (EEA and livestock issues) and Flesh–Kincaid index for English data (EEA).

For English EEA's reports, the Flesch–Kincaid index score varies from 20.24 to 46.25, calculated for the less and the more readable text respectively. This means that

[4] https://www.crea.gov.it/en/home
[5] https://creafuturo.crea.gov.it/
[6] https://www.informatoreagrario.it/
[7] https://www.agricultura.it/
[8] The index can be calculated using the formula provided in [4] and implemented in online calculators, such as https://www.webandmultimedia.it/site/index.php?area=5&subarea=1&formato=scheda&id=36.
[9] This index is described in [5].

the same typology of texts observed for Italian is featured by a higher specialization and meaningfully lower readability. The harder-to-read reports are suitable only for post-graduated people, but also the less difficult ones can be hard to read for undergraduate people.

4. Developing corpora from social media about environmental topics

The observations based on texts published by EEA and in online journals helped us in having a more clear idea of how the language is used for communicating with the citizens and discussing environmental topics. Similar topics are discussed also in social media and we collected data from Twitter in order to build some datasets useful for advancing the application of classification tasks and SA on environmental topics.

Italian data: We collected from Twitter, in a time slot spanning from February 2nd 2022 to March 4th 2022, a total of 8,756 (including some duplicated messages in which more than one of the keywords occurs). They were filtered using the following set of keywords: "Transizione energetica" (energy turnaround), "Agenda 2030", "Crisi climatica" (climate crisis), "Combustibili fossili" (fossil fuel), "Deforestazione" (deforestation), "Greenwashing", "Riscaldamento globale" (global warming), "Impatto ambientale" (environmental impact), "Climate Change", "Green Deal", "Sviluppo sostenibile" (sustainability), "COP26", "Energie rinnovabili" (renewable energy).

English data: we collected from Twitter, within the date range 12 September 2022 until 30 September 2022, a larger amount of data. In collecting this dataset, we used 120 queries from 10 environmental topics including "Environment", "Green", "Sustainability", "Food", "Organism", "Climate Change", "Carbon", "Energy", "Waste", and "Pollution". These 10 environmental topics are obtained from the systematic review conducted by [1], while the queries are obtained from the surveyed papers. We obtained a total of 495,970 tweets, including several duplicated messages, since we use many keywords to collect the data.

4.1. Annotation Schemes for Environmental Topics

We applied three different forms of annotation to our data: one is based on the stance of the user against or in favour of the environmental topics and related politics, one is a fine-grained structured sentiment analysis annotation, while the last one is a sentiment term extraction annotation. The first and second schemes have been applied to the Italian data only, while the last scheme has been applied to the English corpus.

As far as **stance** is concerned, we used the basic scheme based on 3 labels, i.e. Against, Favour, Neutral, also considering Off-topic for the annotation of unclear messages.

In the **fine-grained structured SA** scheme, there are instead two label types that need to be annotated i.e. Spans and Relations. While Span labeling means to identify a set of adjacent or closely connected words, Relation labeling means to identify a relation between two entities annotated as Spans.
Each Span may represent a Holder, an Expression, a Target, or a Topic. A Holder can be a Citizen (an ordinary person/group not affiliated with any official community/ organization), a Government (a central or sub-unit government or its stakeholders), a Political Party (a political party or its stakeholders), a Media (a mass media or its stakeholders), a Company (a company or its stakeholders), a Private Foundation (a private foundation or its stakeholders), or an NGO (Non-Governmental Organization). An Expression can be Positive or Negative. The same entities that can be annotated as Holders can be annotated also as Targets. Topics include the general label Environment, but also more specific labels, i.e., the 10 environmental topics we used to collect the English dataset obtained from [1].

Relations are used for labeling the relationship between the Expression and its Holder, Target, or Topic. This allow us to group the Expression and its proper Holder, Target, or Topic, also considering that one tweet can include more Expressions and each Expression may be to be linked to a different Holder, Target and Topic. We also annotate the Coreference as the additional relation label. For the annotation of this fine-grained structured SA annotation, we used the annotation tool provided by Langing Annotate[10]. The example of annotation for this fine-grained scheme can be seen in Figure 2: the text con-

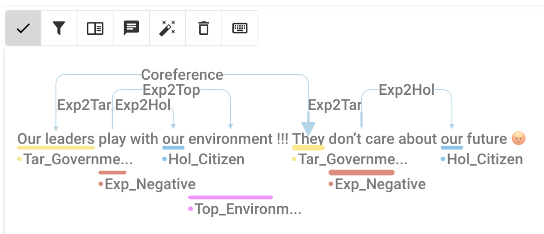


Figure 2: Example of fine-grained structured sentiment analysis

tains two Expressions of negative sentiment. If we wrap each Expression and its Holder, Target, and Topic using a quintuple format (similar to quadruple format used in

[10]https://annotate.langing.ai/

Text	Label
18 gradi a febbraio e rompete i coglioni col riscaldamento globale.. Ne vorrei 30 fissi (18 degrees in February and bust your balls with global warming.. I'd like 30 fixed)	Against
Bottigliette di plastica e collaborazione per ridurre l'impatto ambientale (Plastic bottles and collaboration to reduce environmental impact)	Favour
"#ClimateChange Nel 2021 la crisi climatica è costata 343 miliardi di dollari a livello globale (#ClimateChange In 2021, the climate crisis cost $343 billion globally)	Neutral
Interisti state rosicando così tanto che contribuite alla deforestazione della foresta Amazzonica. #InterMilan (Interisti are so gnawed that you contribute to the deforestation of the Amazon rainforest. #InterMilan)	Off-topic

Table 2
Example of stance annotation.

[6]), i.e. *(Holder, Target, Topic, Expression, Polarity)* we will get two quintuple as follows:

1. *("our", "Our leaders", "environment", "play", negative)*
2. *("our", "They", "", "don't care", negative)*

Notice that in this fine-grained scheme annotation, a Holder, Target, or Topic span should be connected to an Expression span. However, an Expression span can also occur without a Holder, Target, or Topic[11].

Lastly, for **sentiment term extraction** annotation, this scheme is a subset of our fine-grained scheme annotation. Instead of annotating Expression span with its Holder, Target, and Topic, we only annotate the Expression span. Following the guidelines for crowdsourcing datasets conducted by [7], we limit the annotation of English data to Expressions only as a first step, in order to avoid overloading crowdsourcing contributors with a too complex task.

4.2. Annotation of the Italian data

A portion of the Italian data from Twitter, namely 3,254 tweets without duplicates (corresponding to 58,893 words and 1,990 sentences), have been manually annotated for stance, while its annotation with the fine-grained SA scheme is currently ongoing.

4.2.1. Stance annotation

The annotation for this scheme was done using Google Sheets, and some examples of annotation are provided in Table 2.

The agreement occurs in around one-third of the data (2,233 over 3,254), while the disagreement in the other ones (1,021). The higher percentage of disagreement is referred to as the label against, as reported in Table 3. The disagreement has been considered as strong when Annotator-1 has annotated the message as Against and Annotator-2 as Favour, or vice versa, weak in the other cases. The strong disagreement, occurring in 201 annotated tweets, has been annotated also by a third skilled annotator that solved 168 cases by selecting the label used by the first or that chosen by the second annotator.

	Annotator-1 tweets (%)	Annotator-2 tweets (%)
Against	121 (3.7%)	710 (21.8%)
Favour	1032 (31.7%)	733 (22.5%)
Neutral	1789 (54%)	1691 (52%)
Off-topic	312 (9.6%)	119 (3.7%)

Table 3
Number of labels annotated for each label of the category Stance in the Italian corpus.

4.2.2. Fine-grained structured sentiment analysis annotation

For the annotation of the fine-grained structured SA, we used the same Italian dataset described in Section 4, from which we drew the corpus annotated for stance. In this case, we only selected a portion of the corpus composed of the tweets that contain the keyword "green" (whether a word or subword as in "greenwashing"). Using this filter term, we obtained 1,396 tweets and after dropping the duplicate tweets, we randomly chose 500 tweets to be annotated by two other master's degree students.
For span-level analysis, we analyze the annotation agreement level by calculating the pairwise weighted $F_1 - Score$[12] between annotators using SeqEval library[13]. In this case, $F_1 - Score$ is used to evaluate the span-level agreement because it not only evaluates the entity span agreement but also evaluates the *Beginning, Inside, Outside* (BIO) tagging structure. In this annotation, we obtain

[11]For more examples and details about this fine-grained structured SA annotation see the guidelines: https://github.com/okkyibrohim/environmental-topics-in-corpora/tree/main/annotator_guidelines

[12]We calculate a weighted average of $F_1 - Score$ instead of the macro one since we only annotate 500 tweets for this scheme, making many entities have no enough tweets to be calculated the $F_1 - Score$.

[13]https://github.com/chakki-works/seqeval

a 63.67% of weighted $F_1 - Score$, indicating the annotators have a moderate agreement and can be used for experiments in future works.

To see the sentiment distribution for each annotator, we convert the span-level label to the document-level label into a Negative, Positive, or Neutral, polarity label via majority voting between the Expression label. The distribution of document-level labels between annotators can be seen in Table 4. From Table 4, we see that the sentiment polarity in document-level distribution is quite balanced for Annotator-1. However, in Annotator-2, the Positive polarity has a significant amount more than the other two polarity labels. or this document-level label, we evaluated the agreement score using Cohen's Kappa score and got a score of 0.5718, indicating the document-level label has a moderate agreement and can be used for experiments in future works.

	Annotator-1 tweets (%)	Annotator-2 tweets (%)
Negative	164 (32.8%)	131 (26.2%)
Positive	178 (35.6%)	220 (44.0%)
Neutral	158 (31.6%)	149 (29.8%)

Table 4
Number of labels annotated for each label of the sentiment polarity for document-level in the Italian corpus.

4.3. Annotation of the English data

From the total of 495,970 collected tweets, we randomly select 700 tweets for English sentiment term annotation For this English annotation, we use crowdsourced annotators from Prolific[14] who must have English as their first language and a 100% of approval rate for their previous works in the Prolific platform. Annotators were paid £9/h to perform tasks up to one hour of duration. In this annotation scheme, each data chunk will be annotated by 3 anonymous Prolific workers, which means we have 27 workers in total.

The Fleiss' Kappa score for this annotation, computed at the document level as for Italian, can be seen in Table 5.[15]

Data Chunk	Fleiss' Kappa Score	Kappa Interpretation
1	0.4617	moderate
2	0.5374	moderate
3	0.1673	slight
4	0.4510	moderate
5	0.2778	fair
6	0.4048	moderate
7	0.2538	fair

Table 5
Fleiss' Kappa score for each data chunk for English annotation.

5. Conclusion and future work

This paper presents a report on the activities we are carrying on at the University of Turin in the application of SA to environmental topics. Starting with a linguistic analysis of texts extracted from different genres, we are developing data sets for stance detection, fine-grained structured SA, and sentiment term extraction[16]. Notwithstanding the relevance of these topics, very few applications of textual classification techniques and SA has been developed until now. With our activities, we want to start filling out this gap for Italian and English. Nevertheless this is only a starting point and in future work we will address a more extended domain of texts, for example news and interviews, so as to provide a more reliable barometer of sentiments towards climate topics as found in a general audience.

Acknowledgments

The work of English annotation is funded by PUTI Q1 research grant from Universitas Indonesia with number NKB-394/UN2.RST/HKP.05.00/2022.

Muhammad Okky Ibrohim thanks to FSE REACT-EU for PhD Research Projects funding dedicated to GREEN topics on Ministerial Decree 1061/21.

We thank for their contribution the master's degree students Fabiola Summa, Marco Stella, Martina Gagliardi, Gaia Miele and Maria Comandè.

References

[1] M. O. Ibrohim, C. Bosco, V. Basile, Sentiment analysis for the natural environment: A systematic review, ACM Comput. Surv. (2023). URL: https://doi.org/10.1145/3604605. doi:10.1145/3604605, just Accepted.

[2] M. Stede, R. Patz, The climate change debate and natural language processing, in: Proceedings of the 1st Workshop on NLP for Positive Impact, Association for Computational Linguistics, Online, 2021, pp. 8–18. URL: https://aclanthology.org/2021.nlp4posimpact-1.2. doi:10.18653/v1/2021.nlp4posimpact-1.2.

[3] X. Du, M. Kowalski, A. S. Varde, G. de Melo, R. W. Taylor, Public opinion matters: Mining social media

[14] https://www.prolific.co/

[15] All agreement score interpretation used in this research is obtained from [8]

[16] The dataset and code for agreement evaluation can be seen on this GitHub page: https://github.com/okkyibrohim/environmental-topics-in-corpora

text for environmental management, SIGWEB Newsl. (2019). URL: https://doi.org/10.1145/3352683.3352688. doi:10.1145/3352683.3352688.
[4] P. Lucisano, M. Piemontese, Gulpease: Una formula per la predizione della difficoltà dei testi in lingua italiana, Scuola e città 3 (1988) 110–124.
[5] J. Kincaid, R. Fishburne, R. Rogers, C. B.S., Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel", Research Branch Report (1975) 8–75.
[6] J. Barnes, L. Oberlaender, E. Troiano, A. Kutuzov, J. Buchmann, R. Agerri, L. Øvrelid, E. Velldal, SemEval 2022 task 10: Structured sentiment analysis, in: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), Association for Computational Linguistics, Seattle, United States, 2022, pp. 1280–1295. URL: https://aclanthology.org/2022.semeval-1.180. doi:10.18653/v1/2022.semeval-1.180.
[7] M. Sabou, K. Bontcheva, L. Derczynski, A. Scharl, Corpus annotation through crowdsourcing: Towards best practice guidelines, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), European Language Resources Association (ELRA), Reykjavik, Iceland, 2014, pp. 859–866. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/497_Paper.pdf.
[8] J. R. Landis, G. G. Koch, The measurement of observer agreement for categorical data, Biometrics 33 (1977) 159–174. URL: http://www.jstor.org/stable/2529310.

“Ti blocco perché sei un trollazzo”. Lexical Innovation in Contemporary Italian in a Large Twitter Corpus

Paolo Brasolin[1,*], Greta H. Franzini[1] and Stefania Spina[1,2]

[1]Eurac Research (Institute for Applied Linguistics), Viale Druso 1, 39100 Bolzano BZ, Italy
[2]University for Foreigners of Perugia, Piazza Fortebraccio 4, 06123 Perugia PG, Italy

Abstract
This study investigates emerging vocabulary in contemporary Italian in a corpus of 5.32 M timestamped and geotagged tweets extracted from the Italian timeline throughout 2022. We automatically identify and manually distill 8 133 candidate neologisms down to 346 unattested word forms, shedding light on their spatio-temporal circulation patterns.

Keywords
twitter, social media, corpora, italian, lexical innovation, language change

1. Introduction

Lexical innovation is one of the driving mechanisms of language change [1, 2]: through the creation of new words[1] and their integration into existing lexical systems [3], languages evolve and adapt to new social and technological contexts, which are constantly and rapidly changing. The process of creating new words can be approached from different standpoints. Firstly, the choice of sources necessary to trace the process of lexical innovation has great methodological relevance. One of the main traditional sources have been newspaper texts, which have the double benefit of being easily available and quantitatively relevant [4]. Secondly, lexical innovation follows different steps and usually develops from the initial emergence of new words in specific contexts to their proliferation to wider contexts and domains. This process may end with the institutionalisation of new word forms [5, 6] through their inclusion in dictionaries and consolidation in standard use. Thirdly, the linguistic processes leading to the creation of new words can be different and can include phenomena of derivation, composition, transcategorisation, creation of portmanteau forms, semantic shifts, and borrowing from other languages.

The aim of this study is twofold. On the one hand, we present an analysis of emerging vocabulary in contemporary Italian stemming from Twitter interactions using the 2022 Italian timeline as a source; social media represents an opportunity to analyse new word forms surfacing in everyday conversation, and provide vast amounts of data produced in real time by a large, heterogeneous and representative sample of speakers. Furthermore, the availability of geotagged texts enables the investigation of possible patterns of lexical innovation related to specific geographical areas [7]. This possibility is particularly promising in languages, like Italian, characterised by deep and articulated geographical variation. On the other hand, we propose a novel methodology to process and filter word forms acquired from a sizeable Twitter corpus, with the aim of detecting those that represent the best candidates to become new words.

The result of the study is a list of 346 word forms, classified into 15 categories based on the linguistic process of lexical creation and yet unattested in two of the most up-to-date Italian lexicographic resources.

2. Related Work

Studies on lexical innovation in Italian have a long tradition [8], and have produced extensive lexicographic works dedicated to neologisms (e.g., [9], to mention one of the most recent), as well as a vast body of research (e.g., [10], [11] and [12]). One of the most widely discussed topics is the classification of the linguistic processes leading to the creation and spread of new words.

Traditionally, it is acknowledged that the means by which languages enrich their vocabulary are essentially four: the acquisition of new elements from other languages, the formation of new words from pre-existing lexical elements, the change of grammatical category and the shift in the meaning of words already in use [13]. In the last few decades, the *Osservatorio neologico della lin-*

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

*Corresponding author.
paolo.brasolin@eurac.edu (P. Brasolin); greta.franzini@eurac.edu (G. H. Franzini); stefania.spina@unistrapg.it (S. Spina)
https://paolobrasolin.github.io/ (P. Brasolin)
0000-0003-2471-7797 (P. Brasolin); 0000-0003-1159-5575 (G. H. Franzini); 0000-0002-9957-3903 (S. Spina)

[1]In this paper, “word” and “form” are used interchangeably.

gua italiana[2] (ONLI) [4] has been tracking new words emerging in Italian newspapers, producing a database which, to date, includes 2 986 forms with definition, date of attestation and first retrieved occurrence in the press.

More recently, several studies have highlighted the benefits of using social media to track new word forms cropping up in informal contexts, such as everyday conversation, as opposed to newspaper texts, which are more formal and draw from different registers [14, 15, 16]. Additionally, as a populous repository of conversations held in real time by a large number of speakers, social media can capture lexical creativity originating in communities of people rather than inventive journalism [17]. This use of social media has produced a number of studies [18, 7, 19] focussed on the initial and less documented phase of the lexical innovation process, right after the words' creation and first use, and well before their final institutionalisation and inclusion in dictionaries [5, 6].

It is well-known that only a small portion of the words coined in everyday language use become new entries in dictionaries and thus part of the vocabulary: many remain ephemeral but are nevertheless compelling, as they provide evidence of the linguistic mechanisms driving the lexical innovation process. Generally, social media allow researchers to extract and use an unprecedented amount of conversational data [20, 21], which can provide reliable computations of lexical innovation and thus give a significant boost to the study of language variation and change [22, 23].

3. Corpus

In order to investigate emerging vocabulary in contemporary Italian, we used a corpus of timestamped and geotagged tweets extracted from the Italian Twitter timeline throughout 2022. The corpus comprises 5.32 M tweets written by 153 k unique users, amounting to 71.5 M tokens (or 564 M characters).

To the best of our knowledge, this is the first and largest study yet to address lexical innovation in Italian Twitter. Regrettably, this could also be the last. The recent takeover of Twitter collapsed its value for academia: as of summer 2023, publicly accessible data has been severely restricted, API prices have sharply risen, and academic access has been cancelled outright.

4. Methodology

Manual annotation aside, all our procedures are implemented as code and organised into a series of modular stages. To facilitate operation, they are accompanied and coordinated by an executable dependency tree specifying the relations between them, their inputs and their outputs. Together, they constitute a cohesive and reproducible data pipeline.

We exclusively used Open Source Software, mostly in the form of well-known PYTHON packages and GNU[3] tools. An exhaustive list including version numbers can be found in Appendix A.

In the following, we only discuss the general implementation design. The full source code is documented and available in [24].

Condition	Explanation
`lang:it`	written in Italian
`near:italy`	geotagged near Italy
`since:2022-01-01`	on or after 2022/01/01
`until:2023-01-01`	before 2023/01/01

Table 1
List of Twitter's search query language conditions defining the *Italian Twitter timeline of 2022*.

4.1. Acquisition

Our corpus samples the *Italian Twitter timeline of 2022*. We define this notion as the conjunction of the conditions listed in Table 1, expressed using Twitter's advanced search query language[4].

Thus, our corpus is a subset of the results given by the search combining the aforementioned conditions at the time of sampling.

4.2. Preparation

4.2.1. Geographic Data

Tweets can bear geolocation data in two independent forms: a latitude/longitude pair and an association with a place. A place is an administrative division or a point of interest and it is characterised by an id, a country code, a geographical bounding box and other metadata. In our corpus, 99.43 % of tweets bear a place, 0.04 % only bear a lat./long. pair, and 0.53 % bear neither[5]. Consequently, despite lat./long. pairs being more precise, we chose to deal with places only, as they cover the vast majority of tweets and already include the country code necessary to restrict the data exactly to Italy.

We extracted 34.8 k unique places, keeping their id and country code (47.0 % are IT), and computed the

[2]https://www.iliesi.cnr.it/ONLI/intro.php

[3]https://www.gnu.org/

[4]Extensive unofficial documentation for the query language is available at https://github.com/igorbrigadir/twitter-advanced-search/. The user interface is found at https://www.twitter.com/search-advanced.

[5]This is possible because Twitter data can be redacted.

`"Hi #twitter!"` $\mapsto$ `"Hi □#twitter□!"`

range of hashtag entity — U+E000 — U+E001

Figure 1: Schematic representation of how we inline entity range metadata as custom delimiters. This example shows how a hashtag entity is handled.

centroid of their bounding box as a reference point for geographical calculations.

4.2.2. Textual Data

Tweets are rich structures. They include an id, a user id, a timestamp, the full text, the geolocation data discussed above, a list of entities and other metadata. An entity is a character range in the full text labelled by a type (either *url*, *user mention*, *hashtag*, *symbol* or *media*) and other metadata.

First, we extracted all full texts into a flat data file to be loaded into AntConc [25] as an aid to the downstream manual annotation process.

Then, realising the entity metadata could greatly support the tokeniser at a later stage, we inlined them into the full text as delimiter markers, picking a different pair for every entity type from a set of reserved Unicode code points[6]. Figure 1 illustrates an example of how the procedure is carried out for hashtag entities.

Finally, we extracted 5.32 M tweets, keeping their id, user id, timestamp, full text with inlined entities, and place id.

91.77 % of tweets refer to places with the `IT` country code; we assigned these to Italian regions by matching their centroid with governmental data[7] on administrative boundaries in order to plot choropleth maps of Italy. Of the remaining tweets, 8.16 % refer to places with other country codes and 0.07 % refer to a generic place representing the entirety of Italy: the number of occurrences of candidate forms from these two categories are included in the choropleth maps under a legend titled “Not shown”.

4.3. Cleanup and Tokenisation

We used the spaCy v3.6.1 Italian tokeniser. However, tweets are challenging for a stock tokeniser and some issues need to be addressed.

The first problem is the extensive use of Unicode (especially emojis), along with liberal usage of casing and whitespace. This can be easily addressed: we replaced all emojis with spaces, lowercased the whole text, and replaced any streaks of whitespace with a single space.

The second, trickier, problem is the liberal usage of punctuation marks. Solving this required extending the tokeniser's default infix matcher to also match any sequence of these commonly abused punctuation marks: `?!;:,."()[]{}`.

The third and last problem is the presence of entities (urls, hashtags, etc.). This is where our previously inlined entity annotations came into play, quickly enabling us to make the tokeniser aware of them as follows:

- wrap all delimited regions in the text with spaces to nudge the tokeniser into correctly detecting their beginning,
- define a custom token matcher detecting any sequence whose extrema are our delimiter character pairs, and
- disable the tokeniser's default url matcher to avoid conflicts with our custom matcher.

The stratagems above allowed us to execute the tokeniser producing a negligible amount of spurious tokens. We then filtered its output, discarding tokens that were pure space, pure punctuation, pure numbers, broken and/or non-existent handles (i.e., tokens beginning with `@` but not marked as entities), and all entities except hashtags.

Processing all tweets as described, we extracted 71.5 M tokens, with 926 k types.

	$\mathcal{A}$	$\mathcal{B}$	$\mathcal{A} \cap \mathcal{B}$	$\mathcal{A} \cup \mathcal{B}$
Size	6 737	21 132	979	26 890
Fraction	0.73 %	2.28 %	0.11 %	2.90 %

Table 2
Sizes of the candidate subsets as a count and as a fraction of the extracted forms.

4.4. Candidate Selection

To select the candidates for annotation we applied two separate strategies, producing two subsets $\mathcal{A}$ and $\mathcal{B}$ with a slight overlap as detailed in Table 2.

$\mathcal{A}$ derives from an established method in literature, and $\mathcal{B}$ from our attempt to reach for a more interpretable and computationally lighter alternative. We now describe them both in detail.

4.4.1. Subset $\mathcal{A}$: Spearman's ρ

The first strategy follows in the steps of previous studies [18, 7] and amounts to calculating a measure of how monotonically the usage of a token increases in time in order to reject tokens below a fixed threshold. The

[6] We picked from the Private Use Area in the Basic Multilingual Plane, which is a set of code points left undefined by The Unicode Consortium [26, chapter 23.5] and reserved for special custom usage.

[7] Official ISTAT data is archived at https://www.istat.it/it/archivio/222527. We used the GeoJSON version maintained by the community, available at https://github.com/openpolis/geojson-italy/tree/2023.1.

chosen measure of monotonicity is the Spearman rank correlation coefficient between the daily occurrences of a token (normalised by daily total token count) and the day number; we denote it with ρ_O. The choice of threshold is arbitrary: while the cited studies operated on multi-billion tweet corpora picking very restrictive thresholds at 0.7 and 0.8, our corpus is much smaller so we can afford to lower the threshold until the size of the produced subset is still comfortable to annotate. We picked $\rho_O > 0.2$ selecting a subset of $4\,090$ candidates.

However, setting a positive lower bound to ρ_O penalises usage patterns we consider plausible for an emerging form (e.g., a sharp rise before midyear followed by a slow descent to a stable non-zero plateau). Therefore, we chose to extend the criteria to $|\rho_O| > 0.2$ selecting $2\,336$ additional candidates. In other words, we are discarding the central values of ρ_O, where it is less predictive. Furthermore, we decided to perform the same calculation on the daily unique users of a token; we denote the result with ρ_U. We allowed tokens with $|\rho_U| > 0.2$, selecting 311 additional candidates.

Our decision to be so permissive, at the cost of extra annotation effort, was dictated by the intention to experimentally evaluate the effectiveness of the bounds over a wide range of threshold choices.

Subset $\mathcal{A}$ is thus defined by the combined condition $\max(|\rho_O|, |\rho_U|) > 0.2$, selecting $6\,737$ candidates ($0.73\,\%$ of the total).

4.4.2. Subset $\mathcal{B}$: An Alternative Approach

ρ_O quantifies how much a form's usage increases monotonically during the year. As previously mentioned, while this complex measure correlates with the behaviour of some emerging forms, it also excludes plausible usage patterns.

We take the complementary approach and try instead to formulate *simple* criteria to *exclude* usage patterns that we would *not* expect from emerging forms:

- to reject accidental and sporadic phenomena (e.g., typos, inside jokes, etc.), we set a lower bound to the count of unique users U and occurrences O;
- to reject forms already in use from the past, we set a lower bound to the day of first occurrence A;
- to reject forms disappearing early, we set a large lower bound to the day of last occurrence Z;
- to reject ephemeral forms, we set a lower bound to the length of the usage lapse $\mathrm{Z} - \mathrm{A}$.

We chose the following thresholds: $U > 9$, $O > 9$, $\mathrm{A} > 7$, $\mathrm{Z} > 351$ and $\mathrm{Z} - \mathrm{A} > 28$. They read out as: we want forms that are used at least ten times by at least ten people, appear from the second week of January, do not disappear before mid December and last more than four weeks.

The specific values were tuned to cut off the markedly heavier tails from the distributions of the respective variables. This furthers the intention underlying our criteria to exclude the most common behaviours expected from non-emerging forms.

Appendix D contains charts showing how $\mathcal{A}$ and $\mathcal{B}$ partition the dataset and comparing the effect of their defining criteria over the parameter space.

Subset $\mathcal{B}$ defined by the conditions above includes $21\,132$ candidates ($2.28\,\%$ of the total).

4.5. Annotation

The subset for annotation $\mathcal{A} \cup \mathcal{B}$ amounts to $26\,890$ candidates ($2.90\,\%$ of the total extracted forms). To reduce the amount of handiwork, we used a lexicon of $514\,$k Italian forms specifically built for part-of-speech tagging tasks [27] to automatically tag already attested forms as uninteresting (including hashtags, to be analysed separately at a later stage) and thus excluding $18\,757$ candidates. This left us with $8\,133$ candidate forms for manual annotation, which was performed in two stages by the second and third author of the present paper, trained as a classicist and a corpus linguist respectively. Firstly, we loaded the corpus into AntConc [25] to look up each form's context (*KWIC - KeyWord in Context* format), while concurrently cross-checking two freely available online dictionaries and the ONLI neologisms database for attestation[8]. As a result of this search, the annotators rated forms as either *innovative* or *non-innovative*. Inter-annotator disagreement was settled with a negotiating phase until agreement could be reached for all forms. Examples of discarded entries include forms attested in at least one of the consulted dictionaries; mistypes caused by key proximity; popular terms, e.g., *bimbominchia*; foreign words well attested in the media but not in dictionaries (yet), e.g., *foliage*, *spending review*, *sponsorship*; adapted loanwords, e.g., *followo*, *crashare*; infrequently used foreign words, e.g., *smoothie*, *veggie*, *waffle*; infrequently used foreign acronyms, e.g., *PTSD*; regionalisms and regional variants, e.g., *annassero*, *ciolla*, *giargiana*; gender-inclusive graphic variants, e.g., *cittadinə*; nicknames, e.g., *pupone* for footballer Francesco Totti, and the unfriendly portmanteau *Cessica* (*cesso* + *Jessica*).

Next, and as shown in Table 3, we grouped innovative forms into one or more categories according to the ONLI typology scheme with minor adaptations and integrations. Specifically, we only relied on categories referring to formal properties, and thus ignored the *expressive*

[8] Garzanti at https://www.garzantilinguistica.it/ and Treccani at https://www.treccani.it/vocabolario/. The Slengo https://slengo.it/ urban dictionary was also used for the occasional look-up of slang forms.

Category	Forms	Examples
orthographic variation	109	*minkiate*, *rix*, *scienzah*
univerbation	48	*lho*, *miraccomando*
suffixation	45	*cinesata*, *sfanculamento*
loanword	40	*fancam*, *scammer*
portmanteau	33	*gintoxic*, *nazipass*
loanword adaptation	24	*flexo*, *droppare*
alteration	17	*fattoni*
prefixation	8	*bidosati*, *pregirata*
acronym	6	*lmv*, *sgp*
transcategorisation	6	*cuora*
compounding	3	*contapalle*
deonymic derivation	3	*drum*
redefinition	2	*maranza*
acronymic derivation	1	*effeci*
tmesis	1	*facenza*
Total form count	**346**	

Table 3
Categories used with respective candidate form counts and examples.

emphasis category used in the ONLI: emphasis is very common in Twitter interactions [21] and falls under all other categories. In addition, we merged multiple ONLI categories into one: e.g., *suffissazione*, *suffissoide*, *deverbale* and *denominale* were merged into *suffixation*, while *prefissazione* and *prefissoide* were merged into *prefixation*. Finally, a new *tmesis* category was added to account for forms deriving from the splitting of compounds (e.g., *facenza* from *nullafacenza*). Appendix C provides the complete list, and a machine-readable dataset of annotated candidates is available in Franzini et al. [28].

5. Results and Discussion

5.1. Emerging Forms

The most productive categories of lexical innovation in our corpus are:

- orthographic variation, often used either for emphasis (e.g., *minkiate*), to shorten existing words (e.g., *rix* for *risposta*), to conceal online conversation (also known as "leetspeak", e.g., *f4scist4*), for fun (e.g., *gomblotto*) or for sarcasm (e.g., *scienzah* with a final *-h* expressing scepticism towards scientific advances);
- univerbation, with forms such as *miraccomando*, *lho* or *senzapalle*;
- suffixation, featuring many forms ending in *-ato/a* (e.g., *cinesata*, *quarantenato*), *-mento* (e.g., *sfanculamento*) or with the intensifying *-issimo/a* applied to verbs (e.g., *riderissimo*) and to inherently intensified adjectives (e.g., *incantevolissimissima* from *incantevole*);
- (adapted) loanword, chiefly borrowed from English, with forms like *flexo*, *loser* and *trollazzo*;
- portmanteau, mostly relating to politics, with words such as *cessodestra*, *sinistronzi* and the amusing *lettamaio* (the combination of politicians Enrico Letta's and Luigi Di Maio's surnames reading as "pigsty"), but also *gintoxic* and *maxipass*.

Overall, the 346 forms give insights into the most common means by which potential new words are created by Italian speakers. Some of these are those traditionally detected in neologism studies: the *-ata* (*poverata*), *-ismo* (*cialtronismo*) and *-mento* (*sfanculamento*) suffixes, for example, are among the most common morphological resources used to derive new words from existing ones [12]. However, other forms seem particularly productive as potential sources of lexical innovation. Adapted loanwords, for instance, draw on the broad mechanism of inclusion of foreign verbs in the first conjugation in *-are* (*droppare*, *followo*, *switchare*), but also on less common phenomena, such as alteration through the suffixes *-ino* (*trollini*) or *-azzo* (*trollazzo*). Moreover, the widespread attitude towards evaluative language in social media interactions is witnessed by the presence of several emphatic and intensifying forms relying on different expressive means: in addition to the superlative suffix *-issimo/a* applied to verbs (*adorissimo*, *riderissimo*) or even employed as an autonomous word, particularly noteworthy is the use of augmentative suffixes like *-one* (*personaggione*, *garone*), univerbated forms (*opperbacco*, *eddaiii*, *masticazzi*), or portmanteaus such as *nazipass* and *sinistronzi* where emphasis blends with wordplay. Indeed, ironic and catchy wordplay frequently leads to lexical innovation and is typical of social media conversations.

Overall, a non-negligible part of the detected innovative forms are tied to the online sphere, and, in specific cases, are not expected to be used in different contexts or to establish themselves as new Italian words (e.g., *f4scista* or *mer*a*, which are mainly used to conceal content). Nevertheless, their emerging use in Twitter interactions evidences the linguistic mechanisms underlying lexical innovation in Italian. For each form we produce a choropleth map showing its usage. Appendix E presents the maps of all emerging forms mentioned in the article, while Figure 2 illustrates four notable examples from different categories. The map of *gomblotto* shows that orthographic variation, when used for emphasis or ludic purposes, is widespread in almost all regions, though predominantly in Lombardy. Conversely, when orthographic variation is not primarily intended as a joke (e.g., *poki* or *qndo*), the spread of new forms is not as far-reaching. Similar considerations can be made for univerbated forms, which appear to be evenly –albeit thinly– spread out with the

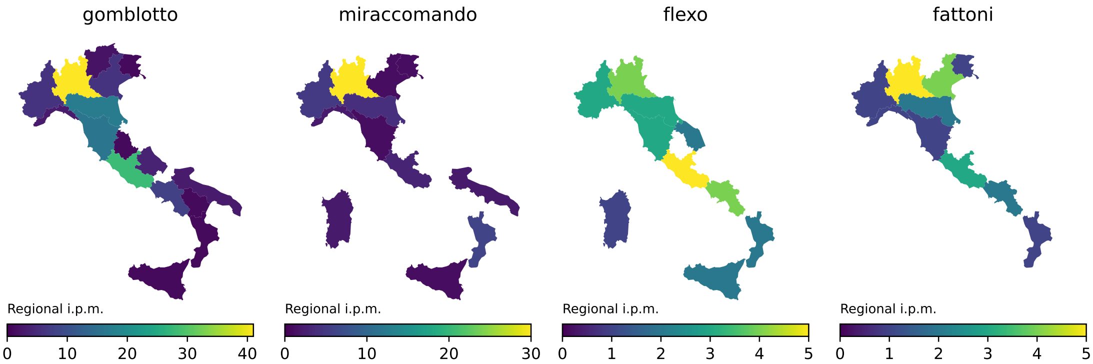


Figure 2: Choropleth maps showing the number of instances per million tokens at a regional level for the following forms: *gomblotto* (139 total instances), *miraccomando* (58), *flexo* (29) and *fattoni* (21). As previously mentioned, instances of forms found in tweets without an IT place association are not mapped: *gomblotto* (10), *miraccomando* (4) and *flexo* (1).

	$\mathcal{A}_O^+$	$\mathcal{A}_O^+ \cap \mathcal{B}$	$\mathcal{B}$
Innovative forms	70	14	281
Adjusted yield	**5.19 %**	4.11 %	4.41 %
Projected yield	3.79 %	3.13 %	**4.20 %**

Table 4
Comparison of innovative form counts and yields between $\mathcal{A}_O^+$, $\mathcal{B}$ and their intersection.

occasional regional peak: *miraccomando*, for instance, is popular in Lombardy but less so in other regions. Other words reveal different patterns: the loanword *flexo*, for instance, meaning "to flaunt", is mostly used in the western part of the country with little to no attestation in the lower eastern regions; *fattoni*, an alteration of "fatto" to denote unreliable individuals and junkies, appears to be in use in the northern regions of Lombardy and Veneto but not so in either the eastern part of the country or the islands. Although, intuitively, spatial variation in social media has different characteristics from traditional geographical variation in relation to language use, previous research has detected a broad alignment between regional lexical variation in Twitter corpora and traditional survey data [29]. The geographical patterns revealed by the data, therefore, provide curious insight into the analysis of lexical innovation in Italian.

5.2. Yields Comparison

To evaluate our $\mathcal{B}$ strategy, we compare subset $\mathcal{B}$'s yield with $\mathcal{A}_O^+$, which is defined as the partition of $\mathcal{A}$ with $\rho_O > 0.2$, in order to fairly represent the approach of previous studies [18, 7]. Table 4 shows the results.

The adjusted yield, computed excluding attested forms and hashtags, favours $\mathcal{A}_O^+$. However, the projected yield, computed including hashtags and assuming the previous yield on them, favours $\mathcal{B}$.

Even without hashtags, $\mathcal{B}$ is noteworthy: its intersection with $\mathcal{A}_O^+$ yields less than the other two, indicating non-redundancy and hence the success of $\mathcal{B}$ in isolating behaviours excluded by $\mathcal{A}_O^+$.

Despite requiring five thresholds, $\mathcal{B}$'s are intuitively meaningful, unlike Spearman's more abstract ρ. Additionally, ρ is computationally expensive[9], making our approach more suitable for data exploration on weaker machines or larger datasets.

5.3. Limitations

Although the one-year time frame considered is both effective in the context of Twitter, where linguistic phenomena appear and spread in a short span of time, and coherent with our objective to investigate the initial emergence of new words, it could well fail to detect new forms that spread more slowly albeit at a constant rate.

Annotation with AntConc revealed the sporadic presence of tweets in French and Spanish. These had no impact on the identified forms but on the selection of the subsets. However, we expect this impact to be negligible and refrain from quantifying the effect at this time. Conversely, the `lang:it` filter most likely excluded some tweets in Italian, but no further assessment is possible with our dataset; there is also no public information about Twitter's proprietary language identification algorithm. Some instances of local Italian varieties were also noticed, confirming previous work [30], but they had no bearing on our analysis as we discarded regionalisms.

[9] A full-fledged time/space analysis is beyond the scope of this work, but we estimate our approach to be upwards of 50 times faster. More details are provided in Appendix B.

6. Conclusions and Future Work

Lexical innovation in Twitter seems to stem mostly from creativity, amusement and attention-seeking behaviour rather than a need for specific new words to indicate new objects, events or situations. The sense of belonging to a large and cohesive community such as Twitter plays a key role in the creation and dissemination of new words. The possibility of being adopted and reused in traditional oral conversation, in large (online) communication streams or, in a trans-medial perspective, by the press, makes at least some of these forms reliable candidates to become institutionalised neologisms.

Next steps in this ongoing study, to appear in Spina et al. [31], will focus on refining the list of candidate neologisms with additional dictionary look-ups (e.g., Zingarelli [32]) and on extending the analysis to the hashtags we put aside by virtue of their multi-functional and natively univerbated nature. Furthermore, we intend on leveraging our annotation data to examine how the yields of the two methods vary in restricting the threshold choices, in the hope of locating *sweet spots* to use as a rule of thumb in future studies. Finally, we will experiment with an estimator for the convexity of the cumulative usage, which, while computationally comparable to ρ, has better interpretability.

Should Twitter die out, planned efforts to scale-up our analysis to multiple Italian timelines will be redirected to other text-based microblogging and social networking platforms, namely Mastodon[10], Bluesky[11] and Threads[12].

Acknowledgments

The authors wish to thank the reviewers for their thorough and constructive feedback.

References

[1] W. Croft, Explaining Language Change: An Evolutionary Approach, Pearson Education, 2000.
[2] W. Labov, Principles of Linguistic Change, volume 2, Wiley-Blackwell, Oxford, 2001.
[3] E. Jezek, Lessico. Classi di parole, strutture, combinazioni, Itinerari, 2 ed., Il Mulino, 2011.
[4] G. Adamo, V. Della Valle, Osservatorio Neologico della Lingua Italiana. Lessico e parole nuove dell'italiano, volume 1 of *Temi e Strumenti*, ILIESI Digitale, 2019.
[5] R. Fischer, Lexical change in present-day English: a corpus-based study of the motivation, institutionalization, and productivity of creative neologisms, number 17 in Language in performance, G. Narr, Tübingen, 1998.
[6] D. Kerremans, A Web of New Words, Peter Lang, Frankfurt am Main, 2015.
[7] J. Grieve, A. Nini, D. Guo, Mapping Lexical Innovation on American Social Media, Journal of English Linguistics 46 (2018) 293–319. URL: https://doi.org/10.1177/0075424218793191. doi:10.1177/0075424218793191.
[8] G. Adamo, V. Della Valle, Che cos'è un neologismo, Carocci, Roma, 2017.
[9] AA.VV., Neologismi (parole nuove dai giornali 2008-2018), Istituto dell'Enciplopedia Treccani, Roma, 2018.
[10] G. Adamo, V. Della Valle (Eds.), Innovazione lessicale e terminologie specialistiche, Olschki, Firenze, 2003.
[11] G. Adamo, V. Della Valle, Neologismi quotidiani. Un dizionario a cavallo del millennio, Olschki, 2003.
[12] F. Marri, I neologismi dentro e fuori dei repertori recenti, Quaderns d'Italià 23 (2018) 11–26. URL: https://doi.org/10.5565/rev/qdi.238.
[13] P. Zolli, Come nascono le parole italiane, Rizzoli, 1989.
[14] B. Rodríguez Arrizabalaga, Social Networks: A Source of Lexical Innovation and Creativity in Contemporary Peninsular Spanish, Languages 6 (2021). URL: https://www.mdpi.com/2226-471X/6/3/138. doi:10.3390/languages6030138.
[15] L. Tarrade, J.-P. Magué, J.-P. Chevrot, Detecting and categorising lexical innovations in a corpus of tweets, Psychology of Language and Communication 26 (2022) 313–329. URL: https://www.sciendo.com/article/10.2478/plc-2022-15. doi:10.2478/plc-2022-15.
[16] Q. Würschinger, Social Networks of Lexical Innovation. Investigating the Social Dynamics of Diffusion of Neologisms on Twitter, Frontiers in Artificial Intelligence 4 (2021). URL: https://www.frontiersin.org/articles/10.3389/frai.2021.648583/full. doi:10.3389/frai.2021.648583.
[17] J. Eisenstein, B. O'Connor, N. A. Smith, E. P. Xing, Diffusion of Lexical Change in Social Media, PLoS ONE 9 (2014). URL: https://dx.plos.org/10.1371/journal.pone.0113114. doi:10.1371/journal.pone.0113114.
[18] J. Grieve, A. Nini, D. Guo, Analyzing lexical emergence in Modern American English online, English Language & Linguistics 21 (2016) 99–127. doi:10.1017/S1360674316000113.
[19] D. Kershaw, M. Rowe, P. Stacey, Towards Modelling Language Innovation Acceptance in Online Social Networks, in: Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, WSDM '16, Association for Comput-

[10] https://joinmastodon.org/
[11] https://blueskyweb.xyz/
[12] https://www.threads.net/

ing Machinery, New York, NY, USA, 2016, pp. 553–562. URL: https://doi.org/10.1145/2835776.2835784. doi:10.1145/2835776.2835784.

[20] M. Laitinen, M. Fatemi, J. Lundberg, Size Matters: Digital Social Networks and Language Change, Frontiers in Artificial Intelligence 3 (2020). URL: https://www.frontiersin.org/article/10.3389/frai.2020.00046/full. doi:10.3389/frai.2020.00046.

[21] S. Spina, Fiumi di parole. Discorso e grammatica delle conversazioni scritte in Twitter, Aracne, 2019.

[22] D. Nguyen, A. Seza Doğruöz, C. P. Rosé, F. De Jong, Computational Sociolinguistics: A Survey, Computational Linguistics 42 (2016) 537–593. URL: https://direct.mit.edu/coli/article/42/3/537-593/1536. doi:10.1162/COLI_a_00258.

[23] D. Hovy, A. Rahimi, T. Baldwin, J. Brooke, Visualizing Regional Language Variation Across Europe on Twitter, in: S. D. Brunn, R. Kehrein (Eds.), Handbook of the Changing World Language Map, Springer International Publishing, Cham, 2019, pp. 3719–3742. URL: http://link.springer.com/10.1007/978-3-030-02438-3_175. doi:10.1007/978-3-030-02438-3_175.

[24] P. Brasolin, Breviloquia italica: data pipeline, 2023. URL: https://doi.org/10.5281/zenodo.10010427. doi:10.5281/zenodo.10010427.

[25] L. Anthony, AntConc (Version 4.2.0) [Computer Software], https://www.laurenceanthony.net/software, 2022. Tokyo, Japan: Waseda University.

[26] The Unicode Consortium, The Unicode Standard, Technical Report Version 15.0.0, Unicode Consortium, Mountain View, CA, 2022. URL: https://www.unicode.org/versions/Unicode15.0.0/.

[27] S. Spina, Il Perugia Corpus: una risorsa di riferimento per l'italiano. Composizione, annotazione e valutazione, in: R. Basili, A. Lenci, B. Magnini (Eds.), Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014, volume 1, Pisa University Press, Pisa, 2014, pp. 354–359.

[28] G. H. Franzini, S. Spina, P. Brasolin, Breviloquia italica: annotations, 2023. URL: https://doi.org/10.5281/zenodo.10010528. doi:10.5281/zenodo.10010528.

[29] J. Grieve, C. Montgomery, A. Nini, A. Murakami, D. Guo, Mapping Lexical Dialect Variation in British English using Twitter, Frontiers in Artificial Intelligence 2 (2019). URL: https://www.frontiersin.org/articles/10.3389/frai.2019.00011/full. doi:10.3389/frai.2019.00011.

[30] A. Ramponi, C. Casula, DiatopIt: A corpus of social media posts for the study of diatopic language variation in Italy, in: Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023), Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 187–199. URL: https://aclanthology.org/2023.vardial-1.19. doi:10.18653/v1/2023.vardial-1.19.

[31] S. Spina, P. Brasolin, G. H. Franzini, Mapping emerging vocabulary in a large corpus of italian tweets, Research in Corpus Linguistics (in preparation).

[32] N. Zingarelli, lo Zingarelli 2022, I grandi dizionari, 2022.

Name	Version	Webpage
JQ	1.6	jqlang.github.io/jq
GNU Parallel	20230622	gnu.org/software/make
GNU Bash	5.1.16	gnu.org/software/bash
GNU Make	4.3	gnu.org/software/parallel
Python	3.10.8	python.org
NumPy	1.25.2	numpy.org
SciPy	1.11.1	scipy.org
Pandas	2.0.3	pandas.pydata.org
Modin	0.23.0	modin.readthedocs.io
JupyterLab	4.0.4	jupyterlab.readthedocs.io
topojson	1.5	mattijn.github.io/topojson
Shapely	2.0.1	shapely.readthedocs.io
GeoPandas	0.13.2	geopandas.org
emoji	2.7.0	github.com/carpedm20/emoji
spaCy	3.6.1	spacy.io
Matplotlib	3.7.2	matplotlib.org
seaborn	0.12.2	seaborn.pydata.org

Table 5
Software and Python packages used in our data pipeline.

A. Data Pipeline Software Stack

The broad strokes of how we used Open Source Software to build our data pipeline are as follows: jq for bulk JSONL data manipulation parallelised with GNU Parallel; NumPy, SciPy and Pandas for general data manipulation and analysis parallelised with Modin; JupyterLab for data exploration; topojson, Shapely and GeoPandas for geographical data manipulation; emoji and spaCy for textual data cleanup and tokenisation; Matplotlib and seaborn for visualisation. All logic and glue code is written using Python and GNU Bash. GNU Make is used to codify an executable dependency tree between the pipeline stages, inputs and outputs. The versions of all stand-alone software and Python packages we used are listed in Table 5. Indirect Python dependencies are listed in the `requirements.txt` file of [24].

B. Computational Complexity

A full-fledged time/space complexity analysis is beyond the scope of this work, as it would require delving into

Figure 3: Code benchmarking the two methods we used and returning the speedup at various dataset sizes.

the implementation details of NumPy, SciPy, Pandas and Modin. However, we can still provide some general considerations and empirical measures on the behaviour of the two proposed methods on a dataset with c columns and r rows. In our case, $c = 365$ (days of the year) and $r \simeq 926k$ (token types).

Calculating Spearman's ρ for a row involves ranking two time series and calculating their Pearson correlation coefficient, so it is safe to assume its *best-case run-time is linear in* c (and probably log-linear on average depending on implementation details). Applying our method to a row involves (cumulative) sums and finding minima/maxima, so its *worst-case run-time is linear in* c. Naïve implementations using either method would simply iterate on the rows of the dataset, so they have *linear run-time in* r.

Given this rough time complexity analysis, we can expect our method to have *some* advantage regardless of implementation details. To quantify it, we ran a benchmark abstracting the core computations of the two methods and comparing their run-times for $c = 365$ and values of r up to the scale of our dataset. The code is presented in Figure 3 and the results are charted in Figure 4: we observe that our method is more than 50 times faster on bigger datasets.

The benchmark was run on a single core and expressed only as a speedup ratio to give a sense of what to *generally* expect. The implementation in Brasolin [24] is parallelised using Modin because we could run it on a hefty Intel Xeon E5-2690 v4 CPU with 128 GB RAM: we traded heavy memory usage for a further speedup, essentially making data exploration in a Jupyter notebook not only viable but pleasant. As a result, performing a detailed space complexity analysis is a particularly delicate matter and one that we do not address here. However, we should stress that our alternative method was initially developed because our means at the outset were much more limited (memory in particular proved to be a bottleneck at 16 GB), and that the initial, sequential, memory-aware implementation is still present in a comment alongside the parallelised one for use on smaller machines.

C. Full List of Innovative Forms

See Figure 5.

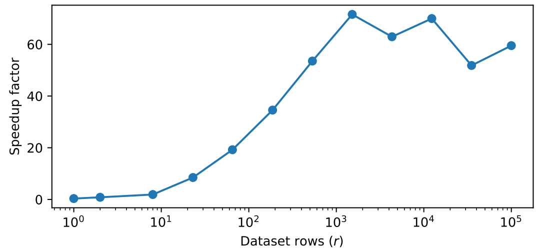


Figure 4: Chart showing the speedup of our method compared to calculating Spearman's ρ.

D. Comparison Charts for $\mathcal{A}$ and $\mathcal{B}$

See Figure 6.

E. Choropleth Maps of Examples

See Figures 7 and 8.

Orthographic variation
accaunt
adovo
affan
amerika
amiketti
amio
amïo
ancielo
anzia
assaj
azzzzz
babbà
benza
biutiful
c4zz0
caiser
cazxi
cazza
cme
collab
comple
coolo
csx
cuxo
dll
duddi
eu4ia
f4scist4
f4scista
fassisti
feffettissimo
gaz
gomblotto
graduidamende
graduidamente
graduido
gretina
grin
incaxxano
incaz
incazz
kaffè
kaimano
kazzate
kompagni
kultura
laik
leccac
lvi
madreh
mbeh
*mer*a*
merd@
merxa
minkiate
minkione
neanke
nerah
norde
nsomma
okk
okok
ovvove
pazzeska
pienah
pikkolo
pk
plis
poki
qlcosa
qlcuno
qlk
qndo
qnt
qt
qulo
qusto
reposta
rimba
rix
rubba
scienzah
sexi
sexo
singol
sinix
sll
snx
stronxate
stronz
tks
troya
trq
tuitt
ubri
urka
vafancul
vaff
vaffan
vaffanc
vairus
vaucher
vergonya
xazzo
xe
xhe
xsino
yessa
zola

Univerbation
ammiocuggino
anchio
buonagiornata
buonamattina
buontutto
cho
ciaobuogiorno
daltronde
demmè
diobono
dioca
diocan
dioporco
eddaiii
eropd
essu
estigrancazzi
evvaiiiiii
flattax
fuoriluogo
gintonic
graziealcazzo
ierisera
instagramstory
lho
lowcost
massí
masticazzi
mavalà
mavattelapijàn-d'erculo
miocuggino
miraccomando
ncazzo
nculo
noeuro
nowar
opperbacco
porcaputtana
porcodd
senzapalle
serietv
sottocasa
stemmerde
stica
streetart
terzopolo
tuttappost
ziocane

Suffixation
abilista
accannate
accannato
adorissimo
amorina
baguettari
benissimamente
cazzarone
ciacchera
cialtronismo
cinesata
cinesate
coglionazzo
ducessa
estaters
fisicati
godicchio
gretini
impiattamento
incantevolissimis-sima
inverners
legaiolo
mandrakata
memiamo
paccare
panchinato
pddizzato
piddini
pisellate
posturologo
poverata
presidenta
prosciutteria
quarantenati
riderissimo
rosiconi
senzadubbiamente
sfanculamento
sierare
sierata
tuitteri
twettini
twitteri
zanzarologi
zanzarologo

Loanword
admin
af
baller
banger
bollox
burp
champ
cishet
dilf
djset
drip
fail
fallout
fanbase
fancam
flu
horny
locals
loser
mentor
misunderstanding
reel
reminder
rimming
scammer
selca
shoutout
showrunner
slim
solution
soundbar
soundcheck
stats
terf
throwback
tier
topping
twitstar
venue
recap

Portmanteau
5stalle
assurdistan
cessodestra
deltacron
docuserie
fasciocomunista
fascioleghista
fascioleghisti
flurona
gintoxic
giornalanza
grillioti
grillopiddini
grillopitechi
intertristi
lettamaio
nazipass
naziucraini
pdiota
pdioti
piddiota
piddioti
pidiota
pidioti
presiniente
putler
renziota
renzioti
scansuolo
sinistronzi
tecnopolo
tridosato
triplodosati

Loanword adaptation
blastata
blessata
boyz
broder
condizionalità
cringiata
droppare
eppi
flex
flexo
followo
ghosta
matcha
pullato
schip
squirtare
stalkero
switchare
trollata
trollazzo
trolling
trollini
twerka
twitterino

Alteration
busoni
cazzaroni
eurini
falsona
fattoni
garone
paccotto
patati
patatino
personaggione
piagnina
pigiamone
pigiamoni
pirlotto
prezzemolina
ridolini
soggettone

Prefixation
appecorato
appecoronati
autoregalo
bidosati
biolaboratori
intrasezioni
iposcolarizzati
pregirata

Acronym
afc
lms
lmv
rdc
sgp
vfc

Transcategorisation
cuora
cuorare
cuoro
issima
issimo
vaffanculi

Compounding
contapalle
fotocazzo
fregacazzi

Deonymic derivation
cippalippa
drum
lippa

Redefinition
giornalaia
maranza

Acronymic derivation
effeci

Tmesis
facenza

Figure 5: Exhaustive list of the innovative forms we found, grouped by category.

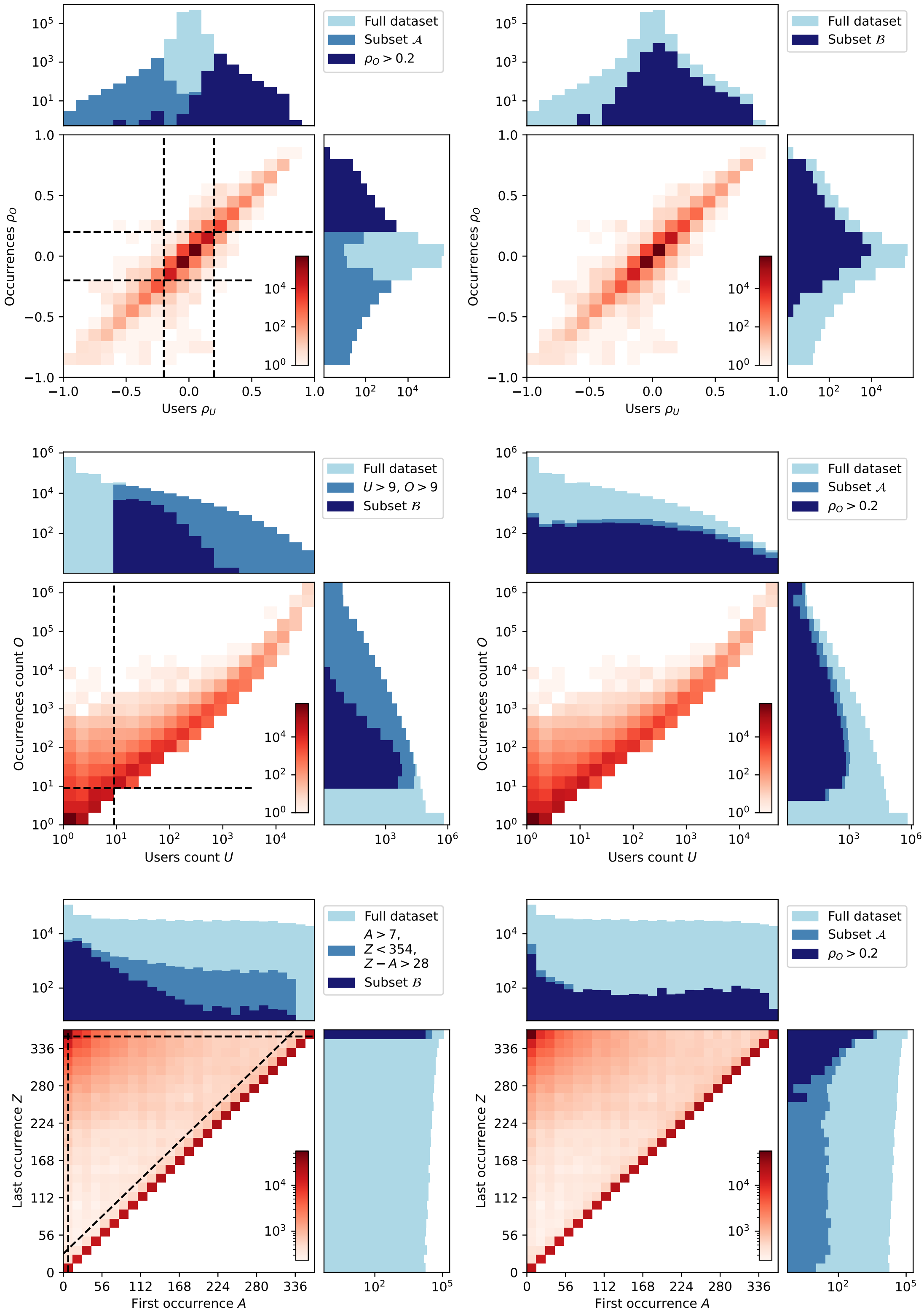


Figure 6: Charts comparing how $\mathcal{A}$ and $\mathcal{B}$ partition the dataset. Unlabeled axes are token counts. The dashed lines highlight how the thresholds act effectively discarding the densest areas. The last two charts reveal an intriguing pattern: a dense diagonal with tokens that appear and disappear quickly, and an opposite-facing dense corner with tokens that occur throughout the year.

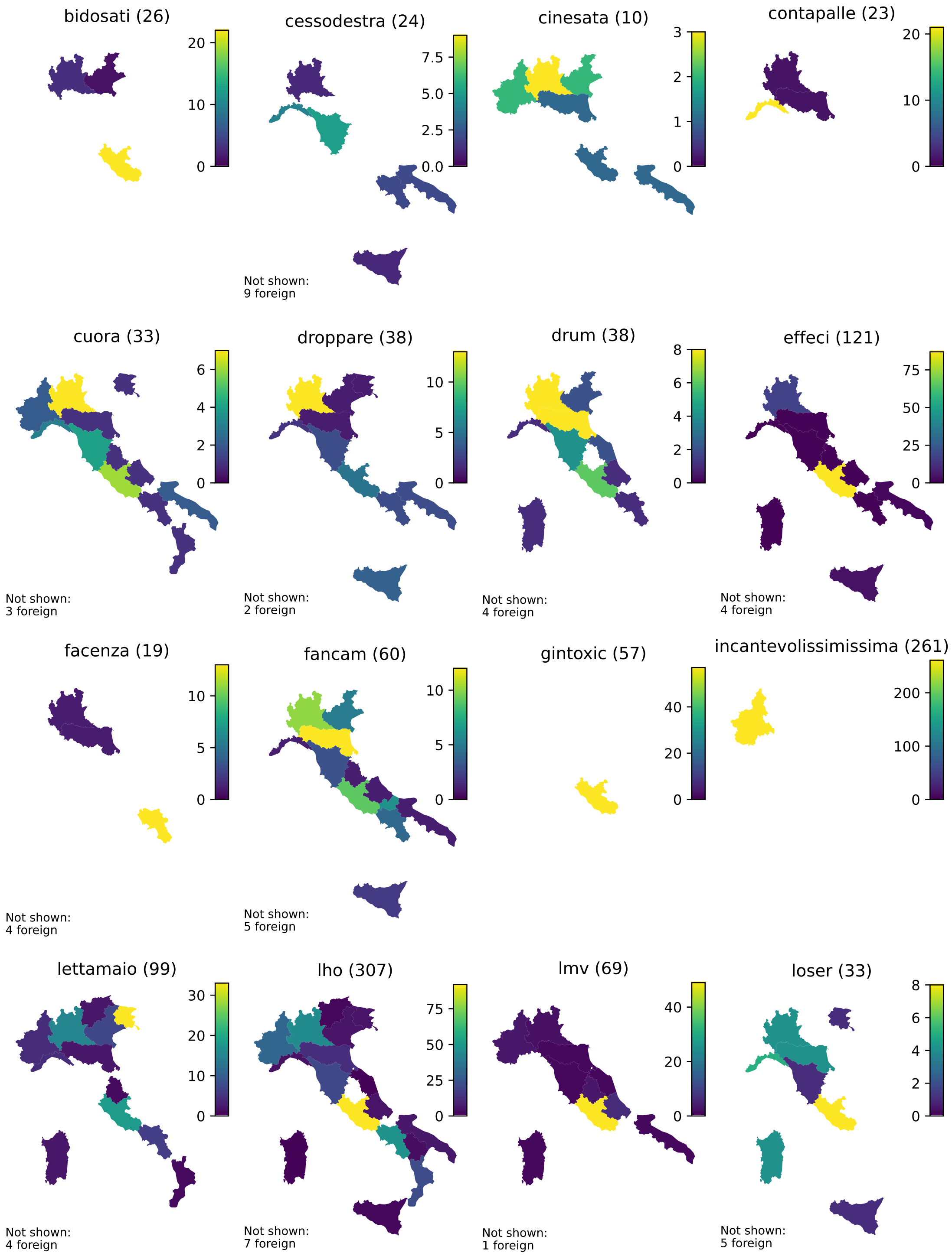


Figure 7: Choropleth maps of innovative forms mentioned as examples, from A to L. The colour scale represents instances per million tokens at the regional level. Total occurrences are provided with the titles, foreign ones in the legends. We omit *f4scist4* as it occurs outside of Italy only.

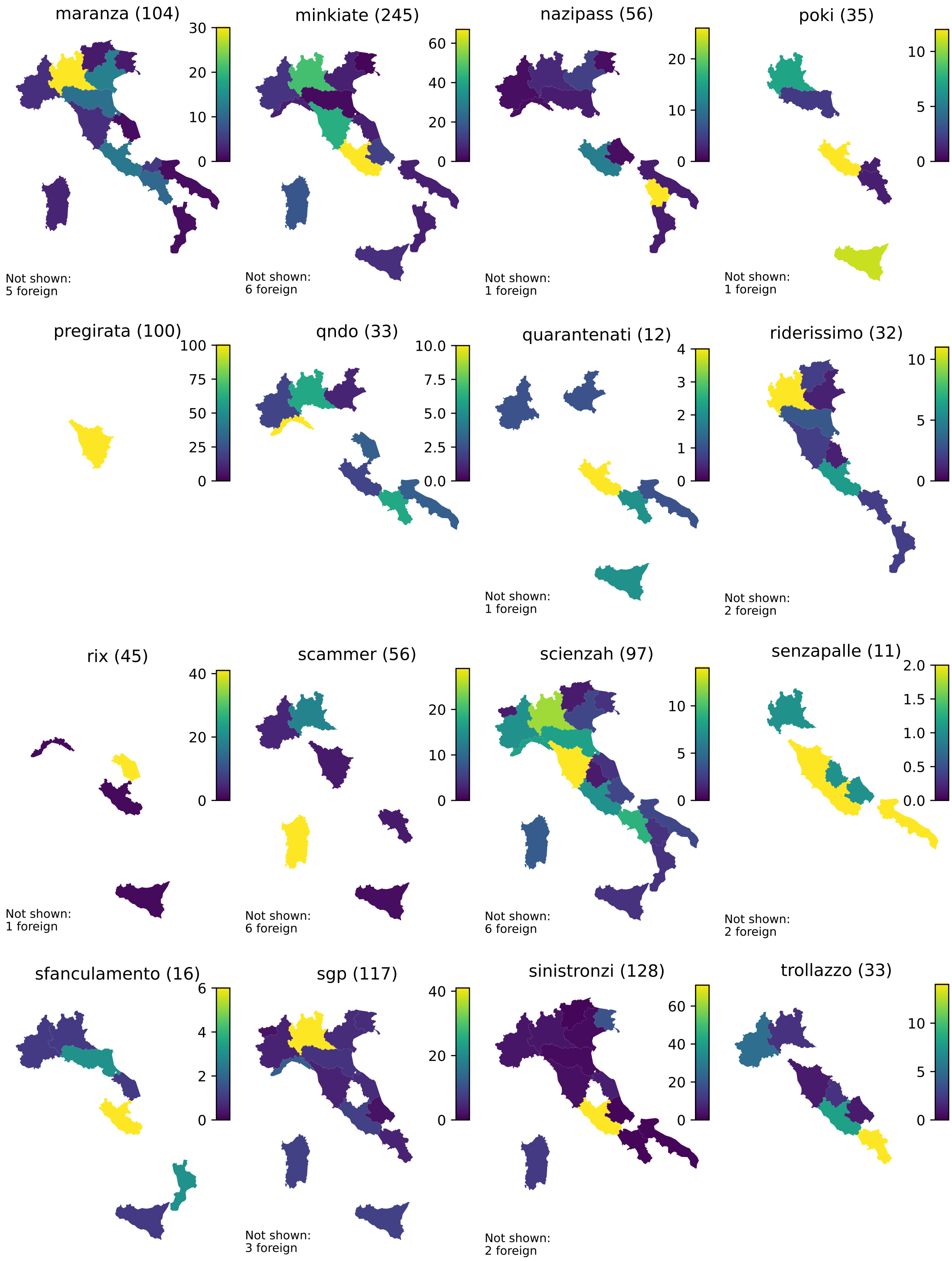


Figure 8: Choropleth maps of innovative forms mentioned as examples, from M to Z. The colour scale represents instances per million tokens at the regional level. Total occurrences are provided with the titles, foreign ones in the legends.

Testing ChatGPT for Stability and Reasoning: A Case Study Using Italian Medical Specialty Tests

Silvia Casola[1], Tiziano Labruna[2,3], Alberto Lavelli[3] and Bernardo Magnini[3]

[1]*University of Turin*
[2]*Free University of Bozen-Bolzano*
[3]*Fondazione Bruno Kessler*

Abstract

Although large language models (LLMs) are achieving impressive performance under zero- and few-shot learning configurations, their reasoning capacities are still poorly understood. As a step in this direction, we present several experiments on multiple-choice question answering, a setting that allows us to evaluate the stability of the model under different prompting, the capacity to understand when none of the provided answers is correct, and to reason on specific answering strategies (e.g., recursively eliminate the worst answer). We use the Italian medical specialty tests yearly administered to admit medical doctors to specialties. Results show that a gpt-3.5-turbo model achieves excellent performance in the absolute score (an average of 108 out of 140) while still suffering in certain reasoning capacities, particularly in failing to understand when none of the provided answers is correct.

Keywords

Large Language Models, ChatGPT, Stability

1. Introduction

Instruction-tuned Large Language Models (LLMs) have recently shown unprecedented results in various tasks in different languages [1]. Beyond the impressive performance, their popularity derives from the possibility of using them with no or little training data for multiple tasks and languages. In fact, instruction-tuned LLMs go beyond the previously established learning paradigm based on transfer learning — where a model, first pre-trained with no supervision, must be fine-tuned on downstream task-specific data — and are typically used in a zero- or few-shot manner.

LLMs performance and ease of use have attracted interest from Natural Language Processing researchers and practitioners. However, most previous work has focused on the models' performance and practical applications. Less relevance has been given to the models' stability and reliability, e.g., in the variability of their outputs or reasoning capacities in controlled settings. This is even more problematic since many of the most popular and performative instruction-tuned LLMs are proprietary, and the details of the exploited data, architecture, and training procedures are at best superficially discussed in technical reports [2] rather than proper research papers, conferences, and other scientific venues.

In this work, we choose a more holistic approach to analyzing results, stability, and consistency in Italian. We do so by considering a case study: the Italian medical specialty tests. The test consists of 140 multiple-choice questions in various medical areas, based on which Italian medical doctors are evaluated and ranked if they want to enroll in a medical specialty school. We chose this test for several reasons. First of all, we believe the task is intrinsically difficult. It requires domain-specific knowledge that doctors are expected to acquire after a six-year-long university career; moreover, the test contains both fact-based questions (for example, the criteria for a diagnosis) and clinical cases, which also require reasoning capabilities (for example, to decide on the most appropriate intervention given some symptoms). On the other hand, the structured nature of the test makes it more robust to the specific prompts used and allows us to measure performance easier and more reliably.

We perform experiments by using ChatGPT. This choice is due to several reasons: firstly, the model is undoubtedly very popular at the time of writing; secondly, according to our preliminary experiments, its performance is superior to those of other open-source LLMs available, e.g., Alpaca [3]. While we are aware of the limitations linked to the proprietary nature of the model, we believe its black-box nature, combined with its popularity and practical importance in NLP-related applications, make an analysis of its capabilities, limitations, and stability even more urgent.

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 – Dec 02, 2023, Venice, Italy
silvia.casola@unito.it (S. Casola); tlabruna@fbk.eu (T. Labruna); lavelli@fbk.eu (A. Lavelli); magnini@fbk.eu (B. Magnini)
0000-0002-0017-2975 (S. Casola); 0000-0001-7713-7679 (T. Labruna); 0000-0002-7175-6804 (A. Lavelli); 0000-0002-0740-5778 (B. Magnini)

2. Background and Related Work

In our work, we benefit from recent approaches to instruction-tuning of LLMs, out of which several prompt-based techniques have been developed.

Instruction-tuned LLMs. In recent years, LLMs have been the focus of extensive research due to their ability to learn from large amounts of data in a self-supervised fashion and to achieve impressive results in various tasks [4, 5]. A recent trend in utilizing LLMs is the development of prompt-based techniques, where a textual prompt is given to the model as input to generate the desired output. Such techniques have shown to be highly effective, especially for tasks that require specific outputs and have the advantage of (i) not requiring any parameter updates in the LLM; (ii) being human readable, and (iii) not requiring in-domain data, unlike fine-tuning techniques. An example of such a model is GPT-3.5, a pre-trained language model that uses the Transformer architecture and an attention mechanism to generate natural language text. For an extensive survey on prompt-based techniques, refer to [6]. Prompting has led to a shift from *objective engineering* to *prompt engineering*: this includes both the manual design of templates [7] and automatic prompt learning [8], as well as various options to ensemble [9] and compose [10] multiple prompts.

Reinforcement Learning from Human Feedback and ChatGPT. We leverage the "gpt 3.5 turbo" model, which is the basis of the interactive interface of ChatGPT [11], and part of the InstructGPT family [12] based on the GPT-3 language model [13]. Unlike standard GPT-3 models, however, InstructGPT models are optimized for interactive use, are particularly suited to take instructions as input prompts, and can modify their outputs when asked in a dialogue, making them more aligned with users' requests. This is accomplished by a reward mechanism, Reinforcement Learning from Human Feedback (RLHF) [14] used to optimize the model. After unsupervised pretraining, conversation data – generated by human trainers who act as both the user and the AI assistant – were collected; the model was then fine-tuned through supervised learning. Given several possible model responses to each prompt, human annotators ranked the desirability and alignment of each response; a reward model was thus trained to mimic their preference. Finally, the reward model was used to further fine-tune the LLM, making it more aligned with human preferences.

Taking advantage of the multilingual pretraining at the base of the GPT-3.5 models, ChatGPT is also available for Italian.

3. Experimental setting

We collect questions from the 2022 Italian medical specialty test. The test contains 140 short questions in Italian, each with five possible answers. Only one answer is correct. A small fraction of the original questions require considering a picture (e.g., an ECG or a medical image). We remove those questions. This leaves us with 136 questions. Since we have collected questions and corresponding correct answers from a published solution (where the correct answer was always the first), we randomize the order of the answers. Unless otherwise specified, the order of the answers is consistent for all experiments.

After constructing a prompt, we input it to a *gpt-3.5-turbo* model with 4K tokens of context. We set the temperature to 0 to avoid hallucinations and leave all other parameters at their default value. Unless otherwise specified, the prompt is inputted through a user role, and no system role is used[1].

We measure the model's performance using accuracy; we also compute the associated test score (normalized to 140 answers to be comparable with human performance), which assigns one point to correct answers, -0.25 to incorrect answers, and 0 to unanswered questions.

4. Experiments and Results

4.1. Baseline performance

To measure the model's baseline performance on our task, we construct a simple prompt (see Example 1[2]).

Since doctors are allowed not to answer questions for which they do not feel confident enough, we also experimented with adding an option (5-choice + IDK) to allow the model not to choose any of the options (*F: I do not know or there is not enough information to answer the question*).

Finally, we also experimented with allowing the model to select an answer according to which none of the provided answers were correct (*F: None of the previous answers is correct*).

Table 1 reports the results. For the cases in which the model was allowed, we also report the number of questions for which it chose not to answer or to answer that none of the options were correct.

[1]Three different roles can be specified through APIs: 'assistant' i.e. the model (used to show expected responses in a chain of interactions); 'system' (used to give "developer-like" instructions and modify the overall behavior of the model), and 'user' (the user that is interacting with the model).

[2]We always prompted the model in Italian. For the sake of simplicity, we will only report the English translation in the continuation of this paper.

Rispondi alla seguente domanda a scelta multipla in formato json. Per esempio {"lettera": <la tua scelta>}.
Domanda: 'Quali dei seguenti Score è utilizzato per valutare la gravità di un paziente affetto da cirrosi epatica?'
Possibili risposte (una sola risposta è corretta):
{ "lettera":"A", "contenuto": "GCS"}
{ "lettera":"B", "contenuto": "Chads-VASC"}
{ "lettera":"C", "contenuto": "ABCD"}
{ "lettera":"D", "contenuto": "Child-Pugh"}
{ "lettera":"E", "contenuto": "Curb-65"}

Answer the following multiple-choice question in json format. For example {"letter": <your choice>}.
Questions: 'Which of the following Scores is used to assess the severity of a patient affected by liver cirrhosis?'
Possible answers (only one answer is correct):
{ "letter":"A", "content": "GCS"}
{ "letter":"B", "content": "Chads-VASC"}
{ "letter":"C", "content": "ABCD"}
{ "letter:"D", "content": "Child-Pugh"}
{ "lettera":"E", "content": "Curb-65"}

Example 1: Basic prompt for a test question.

	Acc.	Score	?	None
5-choice	81.62	107.83	–	–
5-choice + IDK	77.94	101.4	0	–
5-choice + None	78.68	103.20	–	2

Table 1
Model performance on the test.

Notice the accuracy is very high, with a score comparable to that of the best-performing doctors.

Considering the minimal score needed to be admitted in different specialties in 2022, this performance would be sufficient to be admitted in all but one medical specialty school (Dermatology) in at least one of the Universities offering such specialty and to be able to choose among *all* University for 38 specialties (among the 51 available).

4.2. Stability

In this section, we consider the model's stability, with a focus on the consistency of the results.

Repeated questions. Despite setting the model temperature to 0, asking the model to repeatedly answer to the same exact prompt (free of modifications of any sort) can result in different outputs. To measure this effect, we ask the model to answer the test given the same inputs 5 times.

Outputs were not consistent between runs. In most cases, the differences were cosmetic (e.g., some answers reported only the key "letter" with the corresponding letter answer in the output, while others also reported the key "content" with the corresponding answers). In several cases, however, the different runs correspond to different answers to the same questions.

	P1	P2	P3	P4	P5
P1	-	(9, 0)	(8, 1)	(9, 2)	(9, 1)
P2	-	-	(11, 1)	(10, 2)	(12, 1)
P3	-	-	-	(11, 4)	(3, 0)
P4	-	-	-	-	(12, 3)
P5	-	-	-	-	-

Table 2
Results when predicting the same prompt. For each pair of predictions (P_i, P_j), we report a tuple (diff_all, diff_ans), where diff_all is the number of total cases in which the answers have some differences, while diff_ans is the number of cases in which the two runs gave different answers to the same question.

	Accuracy	Score
Order 1	81.62	107.5
Order 2	80.88	106.54
Order 3	83.09	110.40
Order 4	83.82	111.95
Order 5	79.41	104.23
Mean (std)	81.76 (1.57)	108.19 (2.74)

Table 3
Results when predicting the same prompt, changing the order of the given options.

Table 2 reports the experiment results. For each pair of predictions (P_i, P_j), we report the number of total cases in which the answers have some differences and the number of cases in which these differences correspond to different answers. In most cases, this difference has no or negligible effect on accuracy, as mistakes tend to be compensated between runs.

Stability to the order of the answers. We want to test whether changing the order of the given option affects the result and to measure the magnitude of such an effect. To do so, we show the same prompt to the model (see Example 1), but we randomly change the order of the answers for each run. Results are in Table 3. While not dramatic, we notice that the difference in the order corresponds to a visible difference in accuracy. We also noticed that, in just 2 of the 5 runs, one question[3] is blocked by the model due to the prompt triggering Azure OpenAI's content management policy.

Stability to the prompt. Finally, we want to test the effect of using different prompts on the results. To this

[3]The question regards the correct action a family doctor has to take when a person dies at home.

Read the following question from a medicine test and return the option you consider correct among the following. Return the answer in this format: {"letter": <your choice>}.
Question: 'Which of the following Scores is used to assess the severity of a patient affected by liver cirrhosis?'
Possible answers (only one answer is correct):
{ "letter":"A", "content": "GCS"}
...

Example 2: Prompt 2

This is a question with 5 possible answers.
Which of the following Scores is used to assess the severity of a patient affected by liver cirrhosis?
A. GCS
[...]
Select the correct answer. Do not provide any text beside the answer.

Example 3: Prompt 3

end, the authors of this paper constructed prompts independently. Table 4 reports the accuracy and related scores obtained by the different prompts.

Notice that while the prompts (see Example 1, 2 and 3) are not particularly different from each other — which can be expected, given the structured nature of the task —, there is a difference of more than 5 points in accuracy between the prompts that obtain the best and the worst performance.

To process the answer in an easier and more reliable way, all prompts try to condition the outputs to be structured or semi-structured. Using the first prompt, the output is always a valid JSON file; in 21 cases, however, the JSON does not only contain the letter (as required in the prompt) but also the "content" field (mimicking the way the possible answers are presented). For prompt 2, the output is not a valid JSON in 8 cases and presents other text (often corresponding to the answer text) outside brackets. In all cases, the JSON contains the field "letter" only. While prompt 3 requires the model to output the letter corresponding to the right answer only, in the vast majority of cases the output also included the answer text, e.g., in the format "*D. Child-Pugh*" (129 cases) rather than "*D*" (1 case) or "*D.*" (10 cases). For all experiments in this paper, we take into account the correctness even of those outputs that are not perfectly formatted.

4.3. No Correct Answer

We want to understand whether the model is able to understand when none of the provided answers is correct. Thus, we remove the correct answer and add the option *E: None of the answers is correct*, which is expected to be the correct answer. Not all questions, however, can

	Accuracy	Score
Prompt 1	81.62	107.5
Prompt 2	85.29	114.26
Prompt 3	80.15	105.26
Mean (std)	82.4 (2.65)	109.01 (4.69)

Table 4
Results when using different human-generated prompts.

	Accuracy	Score
Prompt 1 (substituted)	23.08	5.38

Table 5
Results obtained when none of the provided answers is correct. We removed the correct answer and added a "No answer is correct" option.

be adapted to this setting. Some questions, for example, require a relative judgment[4]; thus, we first manually selected adequate questions only. This leaves us with 130 questions, for which we build a counterfactual version. We experimented with a slight variation of our default prompt, where we specify that if none of the options seems correct, it must choose *option E*.
Table 5 reports the results of the experiment. We notice that the model performance drastically decreases in this setting: the model tends to very rarely pick the "No answer is correct" options, resulting in an accuracy that is only slightly above random.

4.4. Recursive Reasoning

Instead of choosing the best answer strategy (implemented with the baseline prompt), an alternative solution strategy is to recursively remove the worst answer, choosing the last that is not filtered out.
Previous research has demonstrated that instructing models to perform intermediate steps [15] or explicitly encouraging them to do so in the prompt [16] leads to improved performance. This methodology is commonly referred to as Chain of Thoughts (CoT).
While the direct application of this approach to the multiple-choice context is not straightforward, we sought to explore how a multi-step approach influences performance. We experimented with two different methods: (1) in a single prompt, we asked the model to remove one wrong answer at each step recursively and to give us the correct answer at the end of the process; the chain of thoughts and the resulting correct answers needed to be provided in the same output; (2) we asked the model to identify the answer most likely to be incorrect; we then construct an identical prompt where the model choice

[4]For example: *For a 60-year-old patient affected by metastatic gastric carcinoma at the liver level, HER-2 positive (stage IV), which of the following treatments is the most recommended?*

This is a multiple choice question in the medical domain. Only one answer is correct.
Question:
...
Possible answers (only one answer is correct):
{"letter":"A", "content": "GCS"}
...
Recursively remove one wrong answer at a time until only one answer is left. You will need to provide 4 wrong questions.
At each step, provide the output in the following format:
{"wrong_letter": <your choice>, "reason": <the reason for the esclusion>}
Finally, provide the only correct answer in the format:
"Correct answer: <the letter corresponding to the correct answer >"

Example 4: Prompt for recursive approach 1.

Choose the option that is most likely WRONG among the following. Return the wrong option in the following format: "letter": <choice>
Question:
...

Example 5: Prompt for recursive approach 2.

was removed and repeated the process until only two options were left. In this scenario, we prompted the model 4 times in 4 different conversations.
Examples 4 and 5 show our resulting prompts. Note that, in the first case, the prompt needed to be overengineered and pleonastic as the model was not able to follow instructions with simpler versions consistently — in some cases, for example, it would remove one option only, or output one answer only with no clear indication of whether it considered it as wrong or correct.
The recursive strategy does not seem complementary to the baseline one, as only in one case a question that is answered incorrectly by the baseline prompt is answered correctly by using elimination.

The results in Table 6 indicate a significant decrease in accuracy compared to the baseline experiment. We observed that the model particularly struggled to handle the high logical complexity required by understanding the question and intentionally avoiding the correct answer by selecting a different one. This challenge was particularly evident when our request was performed on questions that themselves asked to identify the wrong option among the given ones; in fact, the model failed to recognize the need for a double negation.

	Accuracy	Score
Approach 1	56.62	63.82
Approach 2	55.88	62.79

Table 6
Results when recursively removing wrong answers.

	Accuracy	Score
Prompt 1	80.88	106.54
Prompt 2	82.35	109.12

Table 7
Results for prompt correction.

Carefully analyze the following multiple-choice medical question. Consider all the available options and provide the choice that you believe is the most accurate. Please indicate your response in the following format: "letter": <your choice >. Remember that your answer should be based on your ability to analyze and comprehend the information available up to September 2021.
Question:
...

Example 6: Prompt for prompt-correction, approach 1

4.5. Prompt correction

In all the experiments conducted thus far, we utilized human-generated prompts to obtain the results from the model. However, using such prompts introduces biases and may not necessarily yield the most optimal results. To explore the potential for improvement, we decided to leverage ChatGPT itself to enhance the prompts. We experimented with two different approaches: (i) we provided ChatGPT with all the human-generated prompts and requested it to improve upon them, and (ii) we granted ChatGPT the freedom to choose the best prompt independently, without any specific examples, but by merely describing the required task.

Both prompt versions are considerably long and elaborated if compared to the human-generated ones. The first version is shown in Example 6. The outcomes of both approaches are summarized in Table 7. Interestingly, the results obtained from ChatGPT-generated prompts closely aligned with those from human-written prompts. Therefore, this particular approach yielded no significant benefits, as the performance remained consistent with the original prompts.

5. Conclusions

We presented several experiments to test the stability and reasoning capacities of an LLM on a multiple-choice question-answering task in the medical domain and for Italian. We evaluated several aspects of the model behavior: the stability of the model (e.g., repeated questions, stability under different prompts and under different orders of answers), the capacity to understand counterfactual reasoning (e.g., when all answer choices are incorrect), the capacity to manage specific answering strategies (e.g., recursively eliminating wrong answers). Results show that a gpt-3.5-turbo model achieves excellent performance in terms of absolute score (an average of 108, out of 140), which is surprising given the technical nature of the test. The model is also relatively stable under different prompts. The model was also able to interpret and manage prompts asking to perform recursive reasoning, even though the resulting performance is considerably worse than the baseline. The major weakness that was found is related to understanding when none of the provided answers is correct, as the model performed only slightly better than random.

Acknowledgments

This work has been partially supported by the eCREAM (enabling Clinical Research in Emergency and Acute care Medicine) Horizon Europe project.
Project name: enabling Clinical Research in Emergency and Acute care Medicine through automated data extraction
Project acronym: eCREAM
Project number: 101057726
Call: HORIZON-HLTH-2021-TOOL-06
Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

We also acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU.

References

[1] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, J.-R. Wen, A survey of large language models, 2023. arXiv:2303.18223.

[2] OpenAI, GPT-4 technical report, 2023. arXiv:2303.08774.

[3] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, T. B. Hashimoto, Stanford Alpaca: An instruction-following LLaMA model, https://github.com/tatsu-lab/stanford_alpaca, 2023.

[4] J. Howard, S. Ruder, Universal language model fine-tuning for text classification, arXiv preprint arXiv:1801.06146 (2018).

[5] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.

[6] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, ACM Computing Surveys (CSUR) (2021).

[7] F. Petroni, T. Rocktäschel, P. Lewis, A. Bakhtin, Y. Wu, A. H. Miller, S. Riedel, Language models as knowledge bases?, arXiv preprint arXiv:1909.01066 (2019).

[8] Z. Jiang, F. F. Xu, J. Araki, G. Neubig, How can we know what language models know?, Transactions of the Association for Computational Linguistics 8 (2020) 423–438.

[9] T. Schick, H. Schütze, Few-shot text generation with natural language instructions, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 390–402.

[10] X. Han, W. Zhao, N. Ding, Z. Liu, M. Sun, Ptr: Prompt tuning with rules for text classification, AI Open 3 (2022) 182–192.

[11] OpenAI, Introducing ChatGPT, OpenAI Blog (2022). URL: https://openai.com/blog/chatgpt.

[12] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., Training language models to follow instructions with human feedback, arXiv preprint arXiv:2203.02155 (2022).

[13] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[14] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, D. Amodei, Deep reinforcement learning from human preferences, Advances in neural information processing systems 30 (2017).

[15] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems, volume 35, Curran Associates, Inc., 2022, pp. 24824–24837. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

[16] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems,

volume 35, Curran Associates, Inc., 2022, pp. 22199–22213. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/8bb0d291acd4acf06ef112099c16f326-Paper-Conference.pdf.

Assessing Language and Vision-Language Models on Event Plausibility

Maria Cassese[1], Alessandro Bondielli[1,2] and Alessandro Lenci[1]

[1]CoLing Lab, Department of Philology, Literature, and Linguistics, University of Pisa, 36 S. Maria St, Pisa, I-56126, Italy
[2]Department of Computer Science, University of Pisa, 3 Largo Bruno Pontecorvo, Pisa, 56127, Italy

Abstract

Transformer-based *Language Models (LMs)* excel in many tasks, but they appear to lack robustness in capturing crucial aspects of event knowledge due to their reliance on surface-level linguistic features and the mismatch between language descriptions and real-world occurrences. In this paper, we investigate the potential of Transformer-based *Vision-Language Models (VLMs)* in comprehending *Generalized Event Knowledge (GEK)*, aiming to determine whether the inclusion of a visual component affects the mastery of GEK. To do so, we compare multimodal Transformer models with unimodal ones on a task evaluating the plausibility of curated minimal sentence pairs. We show that current VLMs generally perform worse than their unimodal counterparts, suggesting that VL pre-training strategies are not yet as effective to model semantic understanding and resulting models are more akin to bag-of-words in this context.

Keywords
multimodal semantics, vision language models, language models, generalized event knowledge,

1. Introduction

Humans have rich knowledge about events and their typical participants. This is known as **Generalized Event Knowledge** (GEK) [1]. GEK is a fundamental part of commonsense knowledge, and plays a key role in language processing as well as in reasoning. For instance, GEK supports our intuitions about likely events (e.g., *A cop arrested a thief*), possible but implausible events (e.g., *A thief arrested a cop*), and impossible events (e.g., *A stone arrested a thief*). Event knowledge is intuitive for humans because we perceive the world by simultaneously processing information from different modalities such as textual, visual, and auditory [2]. In fact, GEK is acquired through linguistic (e.g., reading and talking about events) and sensorimotor experiences based on observing and participating in real-world events.

Several works have investigated to what extent Language models (LMs) possess GEK [3, 4]. These analyses reveal that LMs have remarkable aspects of GEK, though with important differences with respect to humans. This prompts the question whether such differences might stem from the way LMs acquire their knowledge. In fact, even the most recent Transformer-based ones [5], do not possess the same level of multimodal integration of human learners, since they are trained solely on textual data, lacking key visual information like an object's shape and color. In this context, it is natural to ask whether the recently introduced Vision-Language Models (VLMs) possess capabilities that surpass those of text-only LMs in modelling GEK due to their multimodal knowledge of the world. Recent literature has shown that language interpretation appear to not be improved using multimodal architectures [6], and that in some cases VLMs behave as bag-of-words models when it comes to interpreting texts [7, 8].

We contribute to this line of research by carrying out a comparative study of the performance of LMs and VLMs in recognizing event plausibility. The dataset is formed by sentences that differ for the degree of plausibility of the event they express and the argument animacy. Furthermore, we explore the effect of event concreteness on the performances of the models. Finally, we evaluate the impact of actually including images describing test events sentences as inputs for multimodal models. Our analyses reveal that VLMs do not exhibit better performances than LMs on semantic plausibility recognition, with or without images as inputs. Further, we show how more challenging sentences impact the performances of VLMs, suggesting that they are less capable than LMs in recognizing semantic differences that are affected by word orders (e.g., with subject-patient inversion).

This paper is organized as follows. In Section 2 we describe related work. Then, Section 3 details the datasets (Sec. 3.1), the tested models (Sec. 3.2), and the evaluation procedure (Sec. 3.3). We show and discuss the obtained results in Section 4. Finally, Section 5 draws some conclusions and highlights possible future works.

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
m.cassese4@studenti.unipi.it (M. Cassese); alessandro.bondielli@unipi.it (A. Bondielli); alessandro.lenci@unipi.it (A. Lenci)
0009-0007-4765-1221 (M. Cassese); 0000-0003-3426-6643 (A. Bondielli); 0000-0001-5790-4308 (A. Lenci)

2. Related Work

The introduction of multimodal models in NLP stems from the intrinsic limitations of computational models that are trained exclusively on distributional statistics extracted from textual data [9, 3, 10]. In fact, they lack referential competence [11], which prevents them from grounding linguistic structures onto real-world experiences [12, 13].

The earliest multimodal distributional models already showed the ability to improve the semantic representation of concrete concepts and properties [14], as well as abstract verbs that lack direct perceptual information, but benefit from integrating linguistic inputs and perceptual information [15]. However, they proved to be less effective in representing verbs, adjectives, and abstract concepts [16].

The introduction of Transformer-based VLMs such as Visual-BERT [17] and FLAVA [18], and effective techniques for Vision-Language Pre-training [19] paved the way for new research in a multimodal setting. While numerous studies has shown VLMs success on different multimodal tasks, less effort has been put in analyzing their differences with unimodal counterparts on natural language understanding (NLU). [20] show that both dual-stream and single-stream VLMs are equally capable of preserving NLU capabilities. The analysis conducted by [6] shows that multimodal models do not significantly outperform the text-only variants in a language-only setting. This was attributed to the use of narrow domain data and direct extensions of NLP architectures. Our work support these findings by focusing on understanding the plausibility of events.

3. Experiments

Our goal is to evaluate the ability of LMs and VLMs to predict the semantic plausibility of sentences with respect to human judgements. In the following, we describe the data used in the experiments (Sec. 3.1), the models we considered in the evaluation (Sec. 3.2), and detail the evaluation procedure itself (Sec. 3.3).

3.1. Data

We sourced our data from a number of existing datasets containing pairs of sentences describing transitive event distinguished by patient plausibility within the context of the sentence. Plausibility is rated by humans and expressed on a 1-7 Likert scale. Formally, each data point consists of a plausible sentence S_p and its corresponding implausible one S_i obtained through a modification of S_p.

We considered the following datasets:

DTFit [21]. It includes past tense sentences distinguished by patient prototipicality. For each plausible sentence, the implausible (i.e., atypical) one is obtained by replacing the patient with an atypical filler for that role (e.g., *The actor won the award* vs *The actor won the battle).*

EventsAdapt [22]. It includes pairs of plausible-implausible sentences where the implausible one is obtained by reversing the noun phrases (e.g., *The cop arrested the criminal* vs. *The criminal arrested the cop*). The dataset is divided into two sub-datasets. In the former, henceforth referred to as $EventsAdapt_{AN-AN}$, both the agent and the patient are animate. In the latter, denoted as $EventsAdapt_{AN-IN}$, the agent of the original sentence is animate, while the patient is not. Thus the implausible sentence is also semantically impossible.

EventsRev [23]. It includes concrete sentences describing events in the present progressive tense. Like EventsAdapt, implausible sentences are obtained by reversing the noun phrases, which in this case always depict animate entities (e.g., *The cat is chasing the mouse* vs. *The mouse is chasing the cat*). Each sentence, both plausible and implausible, is accompanied by an image depicting the interaction between the two animated participants described in the sentence. The images are simple black and white drawings.

As we are interested in considering also the effect of concreteness in VLMs' ability to recognize plausibility, we further grouped sentences of EventsAdapt (and its subgroups) and DTFit into concrete and abstract ones. We categorized the sentences based on the level of concreteness of the verb, subject, and object in each sentence. We chose to consider sentences that refer to abstract concepts with high imageability as concrete (e.g., *The priest celebrated the marriage*).

To the best of our knowledge, none of the data used in this study was included in the training set of the evaluated models.

3.2. Models

We test various popular multimodal VLMs and compare them with baseline unimodal LMs: BERT [24] and RoBERTa [25]. As for the VLMs, our analysis includes:

VisualBERT [17]. A single-stream early fusion encoder model initialized from pre-trained BERT-base weights and further trained on multimodal datasets. Visual features are extracted from a pre-trained Faster R-CNN network [26] and fed into the transformer model alongside the text.

LXMERT [27]. A dual-stream early fusion encoder model including some modality-specific layers and allowing cross-attention in specific co-attention layers. Visual features are extracted with a Faster R-CNN network.

ViLT [28]. A single-stream model employing a BERT model for textual feature extraction and a ViT model for visual feature extraction, respectively. Resulting representations are then concatenated and fed into the final model.

FLAVA [18]. A foundation VLM including an image encoder, a textual encoder, and a multimodal encoder. It is jointly pre-trained on both unimodal and multimodal data, thus learning high-quality visual and textual representations. It is capable of achieving both crossmodal alignment and multimodal fusion objectives.

To adapt multimodal models to the text-only task, we simply modified the inputs, e.g. by feeding them empty image tensors. FLAVA does not require to be adapted to text-only inputs, as it can directly be evaluated by using only the textual encoder.

All models and their pre-trained weights are available on Huggingface Transformers [29].[1]

3.3. Evaluation procedure

To evaluate the ability of a LM (or VLM) to distinguish between plausible and implausible sentences, we first have to compute a plausibility score for each sentence. Since we are dealing with bi-directional masked language models, we can approximate this plausibility score via pseudo-log-likelihood (PLL), defined as the sum of logarithmic probabilities of each token based on the remaining tokens in the sentence [30]. To avoid bias favoring multi-token words, we apply an additional mask that covers tokens to the right of the target, as proposed in [4]. To compare PLL scores with human judgements expressed on a Likert scale, we normalized both using a min-max scaler function.

First, we evaluated the models using an accuracy metric. Specifically, considering all (S_p, S_i) sentence pairs for a dataset, we computed accuracy as the percentage of cases for which $PLL(S_p) > PLL(S_i)$.

To provide a more detailed analysis of the performances, we further evaluate the models via distribution analyses. We used the Pearson correlation coefficient between each model's score for the plausible and implausible sentences. More in detail, for each pair of (S_p, S_i), we plot the correlation between normalized $PLL(S_p)$ and $PLL(S_i)$. High correlation implies similar scores for plausible and implausible sentences, indicating that the model is less able to distinguish between them. Thus, negative correlation values indicate good performances. We also analyzed the density of the distributions for PLLs. This is essential to comprehend how humans and models differentiate between plausible and implausible classes, aiding in evaluating sentence complexity and comparing model behavior to humans'.

[1] https://huggingface.co/

4. Results and Discussion

We first verify the performances of VLMs in plausibility recognition via accuracy. Results of all models on the datasets are reported in Table 1.

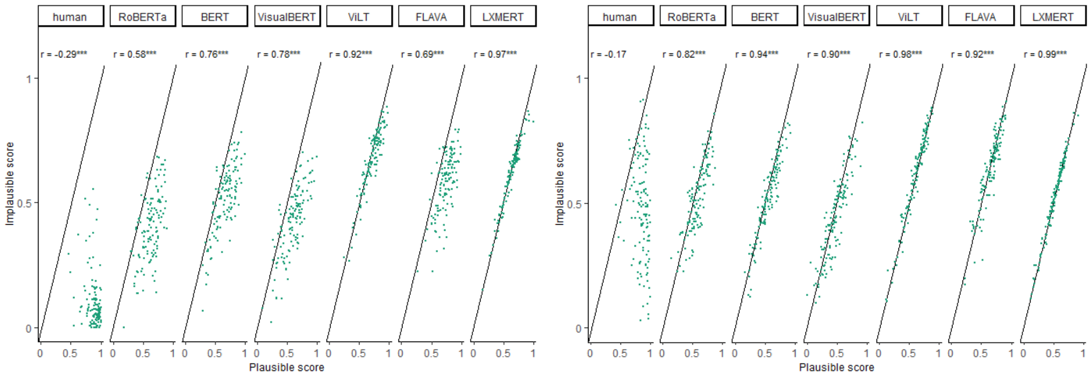
Both LMs and VLMs show significantly higher performances on $EventsAdapt_{AN-IN}$, where implausible sentences describe impossible events, than on $EventsAdapt_{AN-AN}$ where implausible sentences depict unlikely but not impossible events. On AN-AN sentences, BERT, RoBERTa, VisualBERT, and FLAVA performed above chance levels, while ViLT and LXMERT performed at chance. This indicates that extracting information about AN-AN sentence plausibility is generally challenging, and more so for VLMs. Among VLMs, FLAVA performs best, with results generally close to RoBERTa.

Going further, we consider EventsAdapt and we provide the density plot of PLLs divided by plausibility for each model and human raters in Figure 2, and plot the correlation between PLLs of plausible and implausible sentences in Figure 1.

Both LMs and VLMs do not clearly distinguish between the two classes and exhibit very similar distributions for plausible and implausible sentences. The complexity of the task affect the results as well: for tasks where humans have no difficulty in distinguishing between the two classes, as the implausible sentence violates the verb selection preferences (AN-IN), the models can better identify patterns that differentiate the two sentences (Fig 1a); for tasks where even humans are more uncertain (AN-AN), the models tend to assign very similar scores to the two sentences (Fig. 1b). This is also clearly shown by the density distribution plot in Figure 2. for the AN-IN case, the density distribution for humans show a clear separation, while models show more modest but still evident signs of separation. The density distribution for AN-AN sentences shows a less separated distribution for human scores and almost entirely overlapped distributions for models. One possible reason for this is that the grammaticality of a sentence depends on syntactic rules that can be more easily detected through statistical inference. In contrast, linguistic acceptability may depend on extralinguistic information requiring multiple inference levels.

Table 1
Model accuracy on the different datasets

Dataset	Size	Human	BERT	RoBERTa	VisualBERT	LXMERT	ViLT	FLAVA
$DTFit$	395	0.99	0.85	0.89	**0.90**	0.70	0.80	0.86
$EventsAdapt_{AN-IN}$	128	1.00	0.93	**0.95**	0.93	0.72	0.84	**0.95**
$EventsAdapt_{AN-AN}$	129	0.95	**0.78**	**0.78**	0.64	0.53	0.50	0.66
$EventsRev$	38	1.00	0.76	**0.79**	0.76	0.66	0.76	**0.79**



(a) AN-IN (128 tuples)

(b) AN-AN (129 tuples)

Figure 1: Correlation plot on the EventsAdapt dataset for plausible and implausible sentences tuples. Significant differences are marked with asterisks ($*p < 0.05, **p < 0.01, ***p < 0.001$).

The role of concreteness. The experiments show that VLMs do not show improved abilities to deal with GEK and event plausibility with respect to textual LMs. However, we could expect that this might also depend on the event concreteness, as concrete concepts are more directly grounded on visual information than abstract ones. The concreteness of an event depends on the the predicate itself, as well on its arguments. For instance, the verb *to fight* has a concrete use in the sentence *The wrestler fought the opponent* and an abstract use in *The patient fought the cancer*. Cognitive research has shown that abstract concepts require more linguistic experiences to be understood [31]. Thus they are generally more difficult to acquire and process. This is influenced by two main factors, namely imageability and familiarity [32]. For instance, the abstract verb *to celebrate* becomes more concrete in the context of *The priest celebrated the wedding* because it is easier to form mental images of the event and it is very frequent in language use.

To evaluate how concreteness affects the models' abilities, we first compute the accuracy on concrete and abstract subsets of the dataset. Results are reported in Table 2. Multimodal models seem to perform worse on abstract sentences with a higher degree of complexity: on the $EventsAdapt_{AN-AN}$ dataset, the average performance gap between abstract and concrete sentences is higher for VLMs than for LMs (0.06 for LMs, 0.09 for VLMs); when considering the simpler sentences of $EventsAdapt_{AN-IN}$, the differences are less marked. On the other hand, multimodal models demonstrate excellent recognition of abstract events in the DTFit dataset. Note however that abstract sentences are an order of magnitude less than concrete ones in the dataset.

We also show a comparison of Pearson correlation scores of results between $EventsAdapt^{concr}_{AN-AN}$ and $DTFit^{concr}$, shown respectively in Figures 3a and 3b. While VLMs exhibit high correlation values, i.e. less prowess on the task, values for DTFit are generally lower, suggesting a better ability to assess plausibility. VLMs' performance difference in the two datasets may be due to how implausible sentences are generated. EventsAdapt uses noun phrase order reversal, while DTFit only replaces the typical patient with an incompatible one. If VLMs behave more like bag-of-words models, they may struggle to recognize semantic differences between sentences with the same words but different order. This would explain their worse performances on $EventsAdapt_{AN-AN}$.

The impact of images Finally, we analyze whether including images of the (im)plausible test events in the input is beneficial for VLMs. We provide accuracy scores

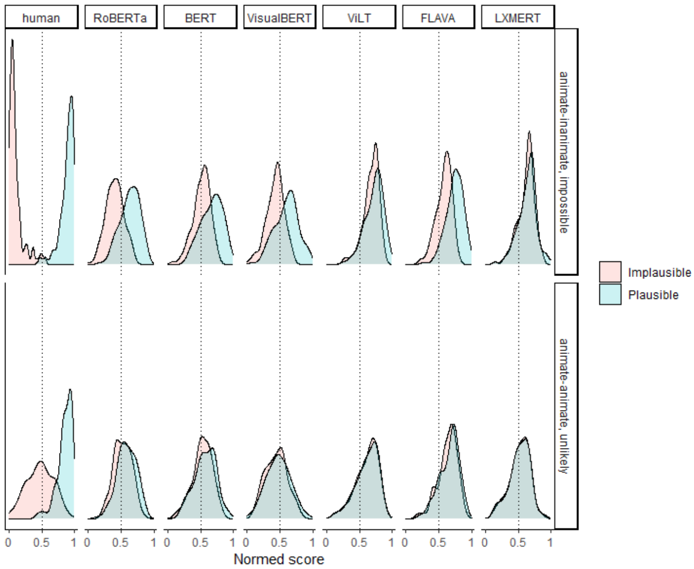


Figure 2: Density plots (EventsAdapt AN-IN (128 pairs) and AN-AN (129 pairs))

Table 2
Accuracy on DTFit and EventsAdapt sentences distinguished by concreteness.

Dataset	Size	Human	BERT	RoBERTa	VisualBERT	LXMERT	ViLT	FLAVA
$DTFit^{abstr}$	45	0.99	0.89	0.86	**0.93**	0.55	0.80	**0.93**
$DTFit^{concr}$	350	0.99	0.85	**0.90**	0.89	0.72	0.80	0.85
$EventsAdapt^{abstr}_{AN-IN}$	31	1.00	0.87	0.94	0.90	0.71	0.71	**0.97**
$EventsAdapt^{concr}_{AN-IN}$	97	1.00	**0.95**	**0.95**	0.94	0.72	0.88	**0.95**
$EventsAdapt^{abstr}_{AN-AN}$	64	0.96	0.75	**0.80**	0.56	0.47	0.47	0.62
$EventsAdapt^{concr}_{AN-AN}$	65	0.96	**0.82**	0.75	0.70	0.57	0.55	0.67

for VLMs on the EventsRev dataset in Table 3. Including event images does not lead to any improvement: performances either remain the same or slightly degrade.

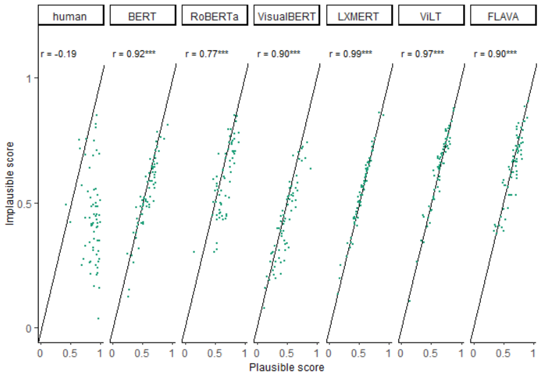
Dataset	VisualBERT	LXMERT	ViLT	FLAVA
$EventsRev_t$	0.76	0.66	0.76	**0.79**
$EventsRev_{t+i}$	0.61	0.66	0.71	**0.79**

Table 3
Accuracy of VLMs on EventsRev with $(t + i)$ and without (t) images in the input.

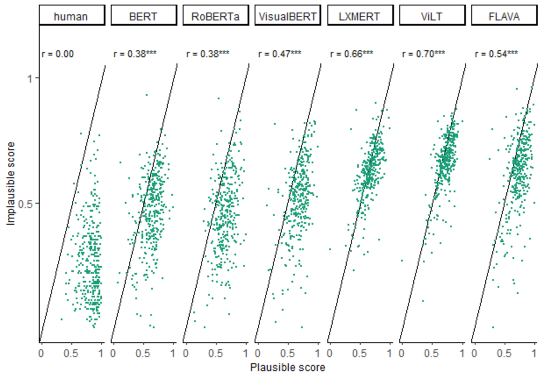
4.1. Discussion

Several interesting findings have emerged from our analysis. First, VLMs do not achieve significantly higher accuracy values than unimodal ones in a semantic plausibility recognition task. Second, we saw that performances of VLMs is worse when dealing with more challenging sentences represented by $EventsAdapt_{AN-AN}$, exhibiting lower accuracy and a high correlation between plausible and implausible sentences. Third, we saw that including images of events in the input does not lead to improved model performances.

We discuss a possible interpretation of these findings in the following. First, the generally high correlation between PLL scores for pairs of (S_p, S_i) for VLMs suggest that these models struggle to recognize semantic differences, especially between sentences with different word orders (e.g., with subject-patient inversion), and relationships between sentence components, like semantic roles. This may be further indication that VLMs model language in a bag-of-words fashion [7, 8]. The pre-training method used in masked language modelling for VLMs, adding visual features to language models already specialized on linguistic tasks, may also compromise learning as suggested by [33]. The high-dimensional space learnt by these models could make it difficult to identify se-



(a) $EventsAdapt^{concr}_{AN-AN}$.



(b) $DTFit^{concr}$.

Figure 3: Correlation plots for sentences in $EventsAdapt^{concr}_{AN-AN}$ and $DTFit^{concr}$.

mantic errors. Moreover, models using pre-trained LM weights for text processing may face limitations in the type of visual information they can capture during training. Some models rely on object categories trained on bounding boxes. This is computationally expensive, and the learned representations may not adequately capture shapes and relationships. Other models, such as ViLT, that leverage ViT representations and use a linear function to extract embeddings for image patches, are less costly but may result in lower-quality representations. These results are in line with [33].

A possible explanation of why VLMs do not benefit from including test images is that in this specific case (minimal sentence pairs with subject-object inversion) the images for both sentences are very similar, and differ only for the relationship between the entities. The visual encoders of the models might be too weak to differentiate substantially similar images, leading the models to rely on their LM priors and make random choices. Finally, we saw that even the foundation Large VLM we considered – FLAVA – does not show significantly improved accuracy compared to other VLMs.

5. Conclusions and Future Works

In this paper, we presented a set of experiments aimed at evaluating the ability of VLMs to model event plausibility in both language-only and vision-language tasks against LMs. We find that VL pre-training does not lead to a significant improvement compared to unimodal LMs in this task aiming at testing their GEK. Specifically, we observed that VLMs tend to perform worse when the implausible sentence has a higher semantic complexity, because it contains two animate nouns. Our analysis also brings further support to argument that VLMs models still behave similarly to Bag-of-Words models, regardless of the degree of concreteness of the events.

In the future, we plan to focus on the analysis of models with visual grounding as their training objective, such as PaLM-E [34], a large embodied multimodal language model that directly incorporates real-world continuous sensor modalities into language processing. This may shed more light into the abilities of large multimodal models to achieve more human-level grounded language understanding.

Acknowledgments

Research partially supported by the Italian Ministry of University and Research (MUR) in the framework of the PON 2014-2021 “Research and Innovation" resources – Innovation Action - DM MUR 1062/2021 - Title of the Research: “Modelli semantici multimodali per l’industria 4.0 e le digital humanities.”, and by PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - “FAIR - Future Artificial Intelligence Research" - Spoke 1 “Human-centered AI", funded by the European Commission under the NextGeneration EU programme.

References

[1] K. McRae, K. Matsuki, People use their knowledge of common events to understand language, and do so as quickly as possible., Lang Linguist Compass. 3(6) (2009) 1417–1429. doi:10.1111/j.1749-818X.2009.00174.x.

[2] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, IEEE transactions on pattern analysis and machine intelligence 41 (2018) 423–443.

[3] P. Pedinotti, G. Rambelli, E. Chersoni, E. Santus, A. Lenci, P. Blache, Did the cat drink the coffee? challenging transformers with generalized event knowledge, arXiv preprint arXiv:2107.10922 (2021).
[4] C. Kauf, A. A. Ivanova, G. Rambelli, E. Chersoni, J. S. She, Z. Chowdhury, E. Fedorenko, A. Lenci, Event knowledge in large language models: the gap between the impossible and the unlikely, 2022. doi:10.48550/ARXIV.2212.01488.
[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, p. 6000–6010.
[6] T. Yun, C. Sun, E. Pavlick, Does vision-and-language pretraining improve lexical grounding?, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 4357–4366. doi:10.18653/v1/2021.findings-emnlp.370.
[7] S. Castro, O. Ignat, R. Mihalcea, Scalable performance analysis for vision-language models, in: Proceedings of the The 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 284–294.
[8] T. Thrush, R. Jiang, M. Bartolo, A. Singh, A. Williams, D. Kiela, C. Ross, Winoground: Probing vision and language models for visio-linguistic compositionality, 2022. arXiv:2204.03162.
[9] J. Gordon, B. Van Durme, Reporting bias and knowledge acquisition, in: Proceedings of the 2013 Workshop on Automated Knowledge Base Construction, AKBC '13, Association for Computing Machinery, New York, NY, USA, 2013, p. 25–30. doi:10.1145/2509558.2509563.
[10] A. Lenci, M. Sahlgren, Distributional Semantics, Cambridge University Press, Cambridge, 2023.
[11] D. Marconi, Lexical Competence, The MIT Press, Cambridge, MA, 1997.
[12] S. Harnad, The symbol grounding problem, Physica D: Nonlinear Phenomena 42 (1990) 335–346. doi:https://doi.org/10.1016/0167-2789(90)90087-6.
[13] E. M. Bender, A. Koller, Climbing towards nlu: On meaning, form, and understanding in the age of data, in: Proc. ACL, Seattle, WA, 2020, pp. 5185–5198.
[14] E. Bruni, G. Boleda, M. Baroni, N.-K. Tran, Distributional semantics in technicolor, Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (2012) 136–145.
[15] F. Hill, A. Korhonen, Learning abstract concept embeddings from multi-modal data: Since you probably can't see what I mean, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 255–265. doi:10.3115/v1/D14-1032.
[16] R. Shekhar, S. Pezzelle, Y. Klimovich, A. Herbelot, M. Nabi, E. Sangineto, R. Bernardi, FOIL it! find one mismatch between image and language caption, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 255–265. doi:10.18653/v1/P17-1024.
[17] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, K.-W. Chang, Visualbert: A simple and performant baseline for vision and language, arXiv preprint arXiv:1908.03557 (2019).
[18] A. Singh, R. Hu, V. Goswami, G. Couairon, W. Galuba, M. Rohrbach, D. Kiela, Flava: A foundational language and vision alignment model, 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 15617–15629.
[19] Z. Gan, L. Li, C. Li, L. Wang, Z. Liu, J. Gao, et al., Vision-language pre-training: Basics, recent advances, and future trends, Foundations and Trends® in Computer Graphics and Vision 14 (2022) 163–352.
[20] T. Iki, A. Aizawa, Effect of visual extensions on natural language understanding in vision-and-language models, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 2189–2196. doi:10.18653/v1/2021.emnlp-main.167.
[21] P. Vassallo, E. Chersoni, E. Santus, A. Lenci, P. Blache, Event Knowledge in Sentence Processing: A New Dataset for the Evaluation of Argument Typicality, in: LREC 2018 Workshop on Linguistic and Neurocognitive Resources (LiNCR), Miyazaki, Japan, 2018.
[22] E. Fedorenko, I. A. Blank, M. Siegelman, Z. Mineroff, Lack of selectivity for syntax relative to word meanings throughout the language network, Cognition 203 (2020) 104348. doi:https://doi.org/10.1016/j.cognition.2020.104348.
[23] A. A. Ivanova, Z. Mineroff, V. Zimmerer, N. Kanwisher, R. Varley, E. Fedorenko, The Language Network Is Recruited but Not Required for Nonverbal Event Semantics, Neurobiology of Language 2 (2021) 176–201. doi:10.1162/nol_a_00030. arXiv:https://direct.mit.edu/nol/article-pdf/2/2/176/189
[24] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for

language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.
[25] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. Cite arxiv:1907.11692.
[26] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, Advances in neural information processing systems 28 (2015).
[27] H. Tan, M. Bansal, Lxmert: Learning cross-modality encoder representations from transformers., in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), EMNLP/IJCNLP (1), Association for Computational Linguistics, 2019, pp. 5099–5110.
[28] W. Kim, B. Son, I. Kim, Vilt: Vision-and-language transformer without convolution or region supervision, in: International Conference on Machine Learning, 2021.
[29] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. doi:10.18653/v1/2020.emnlp-demos.6.
[30] J. Salazar, D. Liang, T. Q. Nguyen, K. Kirchhoff, Masked language model scoring, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 2699–2712. URL: https://aclanthology.org/2020.acl-main.240. doi:10.18653/v1/2020.acl-main.240.
[31] G. Vigliocco, L. Meteyard, M. Andrews, S. T. Kousta, Toward a theory of semantic representation, Language and Cognition 1 (2009) 219–247.
[32] G. Löhr, Does the mind care about whether a word is abstract or concrete? why concreteness is probably not a natural kind, Mind & Language n/a (2023). doi:https://doi.org/10.1111/mila.12473.
[33] C. Fields, C. Kennington, Vision language transformers: A survey, 2023. arXiv:2307.03254.
[34] D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang, Y. Chebotar, P. Sermanet, D. Duckworth, S. Levine, V. Vanhoucke, K. Hausman, M. Toussaint, K. Greff, A. Zeng, I. Mordatch, P. Florence, Palm-e: An embodied multimodal language model, 2023. arXiv:2303.03378.

GPT-based Language Models meet Emojitaliano: A Preliminary Assessment Test between Automation and Creativity

Francesca Chiusaroli[1,†], Tiberio Uricchio[1,*,†], Johanna Monti[2], Maria Laura Pierucci[1] and Federico Sangati[3]

[1]*Università degli Studi di Macerata*
[2]*Università di Napoli "L'Orientale"*
[3]*OIST Graduate University, Japan*

Abstract

English. Starting from the crowdsourcing experience of Pinocchio in Emojitaliano [1], the present paper intends to test Chat-GPT's ability to take on the Emojitaliano grammar and dedicated glossary to verify and reapply the Emojitaliano rules in order to produce translations on its own. A test of re-translation of Pinocchio is presented here.

Italiano. A partire dall'esperienza in crowdsourcing di Pinocchio in Emojitaliano [1], il presente contributo intende testare la capacità di Chat-GPT di assumere la relativa grammatica e il glossario dedicato per verificare e riapplicare le regole della emojilingua allo scopo di svolgere traduzioni in proprio. Si presenta qui un test di ritraduzione di Pinocchio.

Keywords

Emojitaliano, LLM, Assessment, Evaluation

1. Introduction

Consisting today in over three thousand pictograms and symbols, and regularly updated by Unicode Consortium, the emoji international catalog contains signs for facial expressions (smileys) and for human gestures, portraits of people, plants and the animals, reproductions of food and objects for everyday activities and sports, symbols of travel and places. Whereas the visual content seems to provide an encyclopaedic catalog with a universal status, ideally able to signify language-independent meanings, the interpretation of emojis is, on the contrary, heavily arbitrary, subject to ambiguities and differences due to linguistic and cultural specificities [2].

Some efforts were made to develop an emoji based language that could be shared among different cultural peoples. The first notable project that made an effort of translating a classical novel ("Moby Dick" of Herman Melville), was the Emoji Dick project [1] by Fred Benenson (2010). Starting from the English version of the novel, each sentence was translated into an emoji version via crowdsourcing. Each of Moby Dick's 6,438 sentences has been translated 3 times by different Amazon Mechanical Turk (MTurk) workers. The resulting emoji sentences were then chosen by voting by another set of workers, and the most popular version of each sentence was selected for inclusion in the book. The outcome is a wonderful but inconsistent translation of the same terms according to the wisdom of the crowd in good sense, but without any shared rules, structure or grammar, leading to the impossibility of recovering the original text or meaning. Another project was the translation of Lewis Carroll's "Alice's Adventures in Wonderland" by Joe Hale[2] (2014). In this case, each word was directly translated into a corresponding emoji. Consistency was thus guaranteed as the same word was translated with the same emoji, introducing a de-facto lexicon. Nonetheless, no grammar structure is developed as the translation follows verbatim the original text and its English-based word order.

In order to counteract the natural polysemy of emojis [3], Emojitaliano[3] was created through a social community on Twitter (#scritturebrevi #emojitaliano), devoted to the experimental crowdsourcing construction of an international emoji code 'emojilingua' [4, 5]. The aim of the project includes linguistic simplification and the possibility of reproducing a text in emoji that will be com-

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
*Corresponding author.
†These authors contributed equally.
f.chiusaroli@unimc.it (F. Chiusaroli); tiberio.uricchio@unimc.it (T. Uricchio); jmonti@unimc.it (J. Monti); marialaura.pierucci@unimc.it (M. L. Pierucci); federico.sangati@oist.jp (F. Sangati)
https://docenti.unimc.it/f.chiusaroli (F. Chiusaroli); https://docenti.unimc.it/tiberio.uricchio (T. Uricchio)
0000-0003-1923-3974 (F. Chiusaroli); 0000-0003-1025-4541 (T. Uricchio); 0000-0002-4563-5988 (J. Monti); 0000-0003-3637-2757 (M. L. Pierucci); 0000-0001-6088-415X (F. Sangati)

[1]https://www.emojidick.com
[2]https://www.joehale.info/visual-poetry/wonderland.html
[3]https://www.treccani.it/vocabolario/emojitaliano

prehensible and readable in all languages of the world. For this reason, Emojitaliano consists in a unique project that provides a grammatical structure and a shared vocabulary.

Emojitaliano is thus based on the assessment of conventional meanings, capable of guaranteeing the sharing of sense by means of intersemiotic translation [6]. The iconic property of emoji, together with the coded glossary and grammar, makes Emojitaliano a unique tool for communicative accessibility and for multilingual and language (L1 and L2) teaching. Born with the translation of Collodi's *Pinocchio, The Story of a Puppet*[4] [1], the structure and glossary of Emojitaliano have been later usefully reapplied for the translation of texts of different genres such as the technical declaratory prose of the Italian *Constitution*, the narrative prose of moral tales (i.e. *The Wolf and the Lamb*), Dante's allegorical poetry of *The Comedy*, Giacomo Leopardi's lyrical poem *The infinite*[5].

The process of such translation relied heavily on manual labor and human expertise, often time-consuming and subject to human limitations.

With the rapid advancements in artificial intelligence, a new era has dawned upon the world of translation. Large Language Models (LLMs), such as BLOOM [7], GPT-3 [8], LLaMa-1 [9], have revolutionized the way we approach language translation. Recent chat interfaces enable an easy human interaction with them, that has led to a rapid and widespread adoption among the public, also due to the recent high performance closed models Chat-GPT and GPT-4, which remarkable are reported to be able to pass several human admission exams[6]. This preliminary assessment, thus, is about evaluating and establishing the utility of such models, even in the challenging application that is Emojitaliano, where figurative and idiomatic writing is needed, beyond the basic rules of the language.

2. Preliminaries on Emojitaliano: lexicon and grammar

Emojitaliano consists of an iconic-based shared conventional code, first of all a simplified grammar of an isolating and analytical type, constructed on the model of 'interlanguage'; it is anyway not a truly formalized algebraic language, since it is adopted as a conventional code in a social media environment, as a 'living human language'[7]. As for the vocabulary, Emojitaliano is a semantic-based code. In fact, word-emoji pairings are fixed by exploiting the semiotic value of the icons, such as 'scarpa' (shoe). When lacking direct matches, compound emojis are created ('bottega', *atelier* = [casa + attrezzi, *house + tools*]), also adopting the techniques of symbolic and figurative transposition, such as metaphors and similes ('volare', *to fly* = ["aereo", *airplane*]). Linguistic simplification is achieved through synonymy and semantic generalization, based on the fact that the very same emoji may have more than one meaning (i.e. the 'monkey' translates 'birba', 'monello', 'capriccio'; the compound emoji 'man+heart' stays for 'Geppetto' and for 'babbo'). Each new translation experiment subsequent to Pinocchio has been conducted by the community and also by new groups, university and high school students in particular, and sometimes by single translators: everyone was required to use the fixed grammar and, if already present, the fixed vocabulary, and to reapply the simple rules for the creation of new vocabulary, starting from the semiotic value of the emoji. In this way, Emojitaliano has been able to benefit from an ever-growing community, capable of using creativity within a codified scheme of rules. Hosted in a specific bot on Telegram (@emojitalianobot) [10][8], the Emojitaliano language consists today in 3.522 recorded matches[9].

Figure 1: A few sentences from the first chapter of *Pinocchio*, given to the model to derive rules. Best seen zoomed in the PDF version.

[4]https://it.wikisource.org/wiki/Le_avventure_di_Pinocchio
[5]www.scritturebrevi.org
[6]https://openai.com/research/gpt-4
[7]https://www.treccani.it/magazine/lingua_italiana/speciali/ludolinguistica/Chiusaroli.html

3. GPT-4 meets Emojitaliano

Given the project's goal of establishing an international emoji code, we assumed that LLMs can be a useful tool to speed up translation, as well as to spread the language beyond the currently active community. For this purpose, we chose GPT-4 as our reference model, reported as currently the best general performing model in language tasks [11] also for translation [12], even compared by several users in benchmarks with other open source models[10]. Since we do not have access to the training data of such model, we preliminarily verified that it does not know about Emojitaliano making multiple empirical questions (e.g. "What is Emojitaliano?")[11]. Then, we proceeded to gradually instruct the model using the prompt method [13], by providing pairs of Italian and Emojitaliano translated text or a description of the grammar and lexicon of Emojitaliano.

3.1. Automatic grammar derivation

Our first experiment aimed at observing the capability of the model in acting as a human translator of Emojitaliano, its capability to apply grammar rules, and its ability to use the already known tags by Unicode Consortium to creatively associate italian words, to memorize and enlarge the lexicon. We developed a prompt where we provided a passage from the first chapter of *Pinocchio in Emojitaliano*, initially without providing grammar and the lexicon, asking it to derive them, instead. Then we asked the model:

1. to derive the syntactic and semantic rules;
2. to translate on its own some sentences by applying the derived rules;
3. to correct and refine the rules by comparing the personal version and the official one.

[8]Developed and updated by Federico Sangati: https://fede.sangati.me/project/emojitaliano/

[9]On July 30, 2023

[10]https://chat.lmsys.org/?arena

[11]The first answer to "What is Emojitaliano?" was generic and incorrect, that is: "Emojitaliano is a term used to describe the use of Italian words and phrases in combination with emojis to create a unique form of communication. It is a type of digital language that blends written words with visual symbols, allowing people to express emotions, thoughts, and ideas in a more dynamic and creative way. Emojitaliano is particularly popular among young people in Italy and has become a common feature of social media and messaging platforms. It has also led to the development of new emojis that reflect Italian culture and traditions, such as the pizza, the espresso, and the gondola. The use of Emojitaliano is not limited to Italy, however, and it has spread to other countries where Italian is spoken or appreciated. It is an interesting example of how digital communication is transforming language and culture, and how people are adapting to new forms of expression in the digital age."

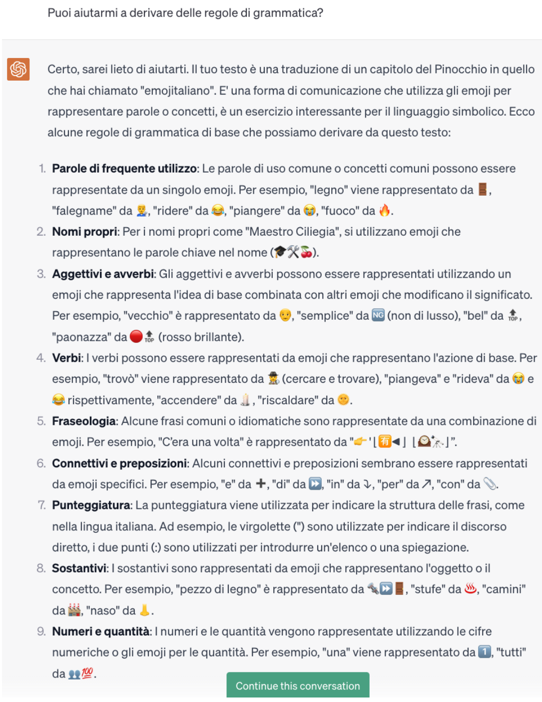


Figure 2: Grammar rules automatically derived by the model. Best seen zoomed in the PDF version.

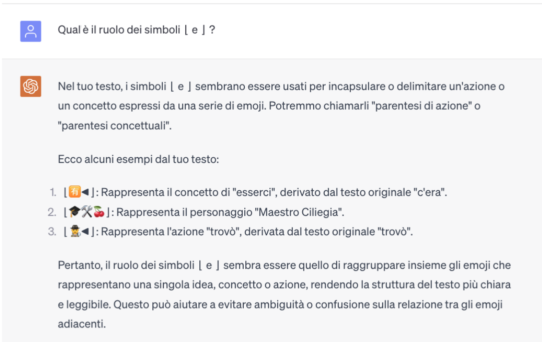


Figure 3: The role of the delimiters, automatically derived by GPT-4. Best seen zoomed in the PDF version.

In Fig. 1, we report the sentences from the first chapter that were given to the model, with the subsequent request to derive the language rules. The output is reported in Fig. 2. The interpretation of conceptual bracket signs in Emojitaliano was then requested, as considered a relevant information missing in the first answer. The result is particularly interesting since a metalinguistic label is assigned by the model, as can be seen in Fig. 3

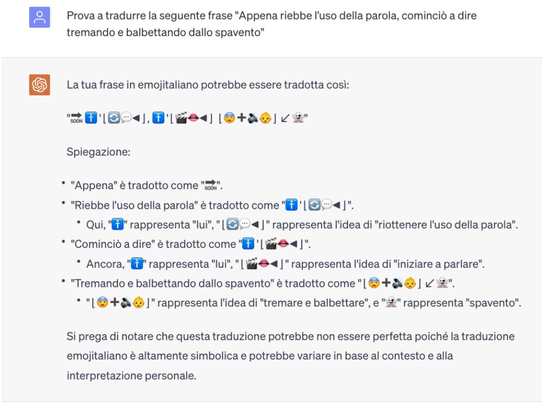


Figure 4: The translation of the first sentence of *Pinocchio*. Best seen zoomed in the PDF version.

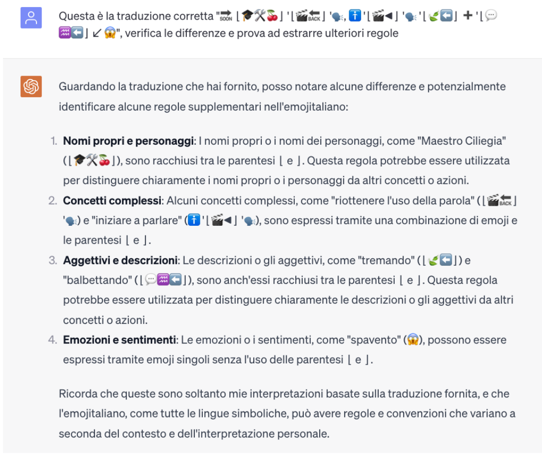


Figure 5: Providing the official version as a correction to the translation and subsequently asking the model to derive new rules. Best seen zoomed in the PDF version.

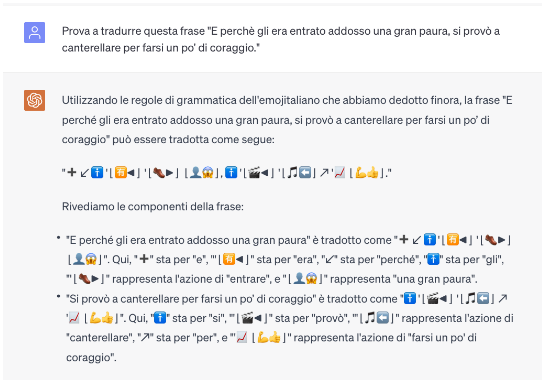


Figure 6: Further sentences and subsequent derivations of new rules. Best seen zoomed in the PDF version.

3.2. Re-translating *Pinocchio*

We then proceeded to ask the model to translate other sentences and lastly the whole first chapter of Pinocchio. The first sentence translation is shown in Fig. 4. The most relevant considerations are that the model correctly applies the grammatical rules to translate a sentence for the first time. Notably, we can observe that it applies the principle of semantic reduction to the lexicon, e.g. in assigning the emoji 'soon' the meaning of 'appena' ('as soon as'), based on the the already available pairings 'subito' ('now') and 'quando' ('when').

Having no access to the registered Emojitaliano glossary, the model succeeds in the translation by applying the common principles of semantic and rhetorical relations in lexicon for the choice of word vs. emoji pairings. Interesting choices include the use of the 'wheel' symbol for the iteration verb 'riottenere' ('to get back') and the use of the baby for 'balbettare' ('stuttering'), together with the 'speaker' emoji for the speaking action.

The experiment continues by providing the official version as a correction, and the request to derive the rules, which has the output shown in Fig. 5. The test involves two further sentences and the subsequent derivation of the rules, as reported in Fig. 6. The figurative expression 'gli era entrato addosso una gran paura' ('a great fear had come upon him') is not simplified in 'to get scared', as it should be, but the literal meaning of 'entrare' as 'to come in' is rendered through the 'shoe' emoji. As a relevant fact, in the derivation of the rule, the model does not catch the meaning of the first point (the three emojis with the clip in Fig. 7) as 'to be with', since in Emojitaliano the 'clip' emoji is 'with' and 'to have' is 'to be + with'.

At the final stage of the test the whole set of the Emojitaliano grammatical rules is provided, with the request to realize the translation of chapters 1, 2, 3 of the text. The 1st chapter is reported in Fig. 8. One notable mistake is that, at the end of the learning process, the model appears to ignore the rule of the explicit subject in basic sentences. An important rule of the Emojitaliano is that the subject is always to be expressed, as the verb does not have inflection. The personal pronoun appears in imperative sentences, instead. With regard to vocabulary, the present model makes use of the emoji repertoire updated to 2022: this implies the availability of sometimes more appropriate choices compared to the past. The emoji catalog includes, for example, the 'tree trunks' for 'legno, tronchi', the 'machine workshop' for 'bottega', which appears particularly suitable compared to the official pairings for 'falegname' as 'mechanic' and the 'robot' for 'puppet'. Among the right choices is the 'volleyball

Figure 7: Further sentences and subsequent derivations of new rules. Best seen zoomed in the PDF version.

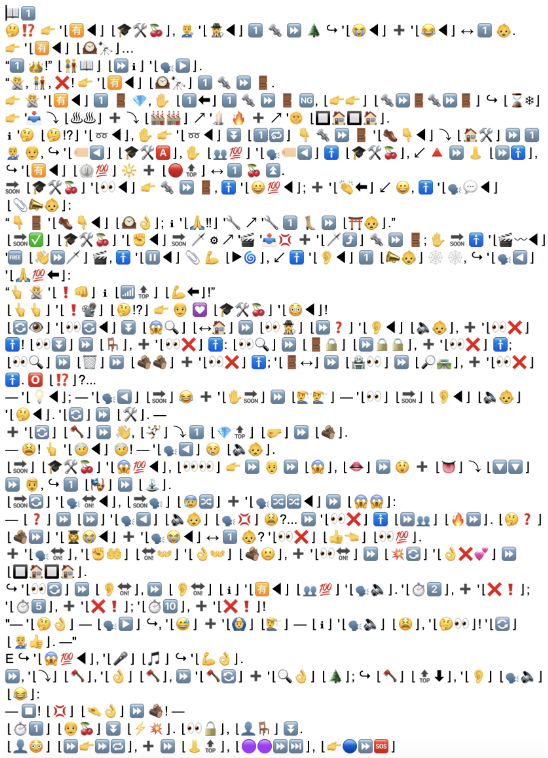


Figure 8: The 1st chapter of *Pinocchio* translated by the model. Best seen zoomed in the PDF version.

player' emoji for 'tirare un colpo'. Although the model is familiar with the rule for the semantic plain translation of figurative language, it does not succeed in applying it, as in 'occhi fuori dal capo', which is translated literally 'eyes going out of the head' instead of rendering the meaning of 'to be shocked' or 'to be upset'.

A relevant issue emerged when Chat-GPT was asked to translate some specific words: for example, 'legno' ('pezzo di legno', 'piece of wood') is translated with the 'fir' ('abete') emoji and, somewhere later, with the 'wooden door' emoji. This is against one of the main Emojitaliano lexical rules which aims at reducing semantic ambiguity. In fact, each word within the same text, should always be translated the same way. Chat-GPT is to be trained accordingly.

We also noticed that grammar and rules mistakes can be corrected by the model upon casually reminding rules in long interactions. The model leaned to progressively forget the rules and, thus, a restart of the session was required after a few sentences. We believe that this is due to the limited window of attention of LLMs and the encoding of emoji that require several tokens for each of them.

4. Performance evaluation

According to our preliminary exploration, we established that GPT-4 is able to derive the semantic rules and translate text to Emojitaliano. To evaluate the latter, we performed a more extensive evaluation by building a dataset of text pairs Italian-Emojitaliano and asking the model for the translation.

We constructed the dataset by considering the first 3 chapters of Pinocchio [14], previously translated in Emojitaliano [1]. The chapters are split respectively in 27, 50, and 45 sentences for a total of 122, ranging from 5 to ∼80 Italian words and from 2 to ∼70 emojis each. For each sentence, we constructed a pair made of the original Italian text and the relative human translation. Each sentence is given to the model for translation independently from the others.

To perform the evaluation, we constructed a textual prompt where the grammar and the basic rules are carefully explained in Italian, where we include as the training set, the first chapter as given examples of translation. The remaining two chapters are used as the test set. Measuring quantitatively the quality of the translation is more challenging than the typical translation tasks metrics, given the creative use of emojis and their combinations in expressing a meaning. Given the low number of samples, we resorted to human evaluation and the use of GPT-4 as a judge following [15]. For both human and GPT-4 evaluators, we provided the rules of grammar,

	GPT-4 Evaluation			Human Evaluation		
Translation by	Human	GPT-4	Equal	Human	GPT-4	Equal
Preferences	27	51	17	46	36	13
Average Score	7.23	7.80	/	7.34	7.21	/

Table 1
Results of the performance evaluation.

the original text and blindly the translated Emojitaliano from the ground truth and the output of the model. The evaluators were asked to vote for the best translation (i.e. choose the preferred translation) according to relevance, accuracy, creativity, correct use of grammar. In addition to choosing the preferred translation, we also asked the evaluators to provide a quality score from 1 to 10 for each sentence.

Results are reported in the Table 1.

GPT-4 and Human evaluators disagree on their preference of translations. The Human evaluators, generally, tend to prefer the Human translations while GPT-4 the opposite. From the evaluators and GPT-4 feedbacks, we noted that the Human evaluators put more emphasis on the correct structure of the sentences (e.g. the subject verb object rule), while GPT-4 generally reported better scores for creativity and direct matching of the emojis (e.g. emojis that match the words). This is consistent, since the translation in the ground truth was realized in 2017, when most of the modern emojis where still not defined at the time. Due to the absence of a proper matching, many emojis where chosen even if they were distant from the corresponding words. Moreover, GPT-4 has consistently not fully caught the rules of Emojitaliano, leading to less awareness of errors in the sentences structure.

5. Conclusions and work in progress

Emojitaliano was born thanks to the free dedication and commitment of an enthusiast devoted Twitter social community, then also of student groups, willing to share the goal of building an emoji-based artificial language model, to be used as a communicative code across language barriers [5, 1]. The effort to adapt to the rules and to join the common glossary, as well as to expand it according to the common rules, was challenging as well as a hard task, but it was the only way to ensure an essential linguistic basis, by giving rise to a language, validated and practiced by a community of ‘speakers’. The intensive crowdsourcing experience made Emojitaliano a unique case among the (actually not many) examples of integral translations in emoji, which are mostly represented by intentionally non-systematic or solipsistic works. The regular expansion of the international emoji set by the Unicode Consortium constantly extends the range of choice by enriching the emoji-language with ‘emoji-neologisms’, as happens in every living natural language, but the core of the Emojitaliano glossary and grammar provides a settled authoritative translation method. Translating Pinocchio into Emojitaliano today would certainly involve new, and sometimes more relevant, pairings, synonymic pairs that do not exclude the previous ones; but the method remains fixed, because the syntax alone guarantees, through the instrument of translation, mutual understanding. Teaching Emojitaliano to GPT-4 (and the like) does not mean replacing a human translator with a machine, but rather is like having a tool to enhance human work to the maximum: automation ensures the speed, the iconic base of the emoji embeds and guides creativity, therefore setting limits against the arbitrary drift of individual subjective interpretation. Following our design, the year of work spent in the ‘human’ translation of the original 15 chapters of Pinocchio will be matched by a few minutes’ work in the translation of the entire work (35 chapters) by Chat-GPT, and in the translation of other works from any world’s language. Extreme speed is comfortable and convenient, but the results cannot be achieved without training: that is, by learning a “language” and its rules.

Acknowledgments

This work has been funded by the European Union - NextGenerationEU under the Italian Ministry of University and Research (MUR) National Innovation Ecosystem grant ECS00000041 - VITALITY - CUP D83C22000710005.

References

[1] F. Chiusaroli, J. Monti, F. Sangati, Pinocchio in Emojitaliano, Apice libri, Sesto Fiorentino, 2017.
[2] V. Evans, The emoji code: How smiley faces, love hearts and thumbs up are changing the way we communicate, Michael O’Mara Books, 2017.
[3] M. Danesi, The semiotics of emoji: The rise of visual language in the age of the internet, Bloomsbury Publishing, 2017.
[4] F. Chiusaroli, Da emojipedia a pinocchio in emojitaliano: l’“emojilingua” tra scritture e riscritture, in: Homo Scribens 2.0. Scritture ibride della modernità, Franco Cesati, 2019, pp. 45–87.
[5] F. Chiusaroli, Emoji e semplificazione linguistica, in: Comunicare il patrimonio culturale. Accessibilità comunicativa, tecnologie e sostenibilità, FrancoAngeli, 2021, pp. 164–193.
[6] F. Chiusaroli, La scrittura in emoji tra dizionario e traduzione, in: Proceedings of the Second Italian Conference on Computational Lin-

guistics CLiC-it 2015, Accademia University Press, Torino, 2015. URL: http://books.openedition.org/aaccademia/1437. doi:https://doi.org/10.4000/books.aaccademia.1437.
[7] T. Le Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, et al., Bloom: A 176b-parameter open-access multilingual language model, arXiv preprint arXiv:2211.05100 (2022).
[8] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.
[9] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).
[10] J. Monti, F. Sangati, F. Chiusaroli, B. Martin, M. Sina, et al., Emojitalianobot and emojiworldbot-new online tools and digital environments for translation into emoji, in: Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016), 2016.
[11] J. A. Baktash, M. Dawodi, Gpt-4: A review on advancements and opportunities in natural language processing, arXiv preprint arXiv:2305.03195 (2023).
[12] W. Jiao, W. Wang, J. Huang, X. Wang, Z. Tu, Is chatgpt a good translator? yes with gpt-4 as the engine, arXiv preprint arXiv:2301.08745 (2023).
[13] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, ACM Computing Surveys 55 (2023) 1–35.
[14] C. Collodi, Le avventure di Pinocchio. Storia di un burattino, illustrata da Carlo Chiostri., [etc.] Bemporad & figlio„ 1907.
[15] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, et al., Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023.

Linking Stance and Stereotypes About Migrants in Italian Fake News

Alessandra Teresa Cignarella[1,2], Simona Frenda[1,2], Tom Bourgeade[1], Cristina Bosco[1] and Francesca D'Errico[3]

[1]Dipartimento di Informatica, Università di Torino, Turin, Italy
[2]aequa-tech, Turin, Italy
[3]Dipartimento di Formazione, Psicologia, Comunicazione, Università di Bari "Aldo Moro", Italy

Abstract
This paper investigates stance and stereotypes within a dataset of Twitter conversational threads in Italian. The starting point of these conversations are tweets containing misinformation, in the form of racial hoaxes targeted at migrants, identified as untrustworthy by fake news debunking websites. The conversational structure of the dataset gives us the opportunity to observe and collect evidence about some linguistic and social phenomena at play in the propagation of stereotypes and the interactions between users which stem from them. We propose a theoretical background, as well as quantitative and qualitative analyses of our annotated data, at different levels of granularity, which can provide insights into the dynamics of Italian online discourses on the topic of migration.

Keywords
Stance Detection, Stereotypes, Rumors, Fake News, Misinformation, Italian

Warning: *This work contains words and expressions that could be considered vulgar or offensive to varying degrees. We emphasize that all authors of this paper are deeply involved in activities to counter the spread of online hatred and do not condone the use of such expressions in any way.*

1. Introduction and Motivation

In the era of information overload and widespread digital communication, the terms "disinformation", "fake news", "hoaxes", "misinformation", and "rumors" have become buzzwords that dominate public discourse [1]. While these terms are often used interchangeably, it is important to recognize the nuanced differences between them and establish some order in the terminology. "Fake news" refers to fabricated or misleading information presented as legitimate news, often with the intention to deceive or manipulate public opinion. "Disinformation" encompasses a broader range of intentionally false or misleading information disseminated with the aim of influencing beliefs or actions. On the other hand, "rumor" refers to unverified or unsubstantiated pieces of information that circulate widely within communities [2]. Finally, a particular type of rumor that disseminates information including threat claims to the health or safety of a person or group based on their race, ethnicity or religion, has been recently defined as "racial hoax" by Cerase and Santoro [3], reshaping a previous definition from Russell [4]. In this paper, we aim at stressing a connection with the area of research that investigates the above-mentioned phenomena and the research conducted so far in the field of Stance Detection (SD). Indeed, as it can also be seen in Figure 1, in a recent survey paper, Küçük and Can [5] illustrate the relationships occurring among the different tasks and subtasks in the field of Sentiment Analysis, putting at the center of attention SD. In the boxes hilighted in blue in Figure 1 we see the tasks of *Rumor Stance Classification* and *Fake News Stance Classification* being strictly related to SD. In this work, we connect the dimensions of stance and stereotypes, based on a re-annotation of our previous corpus, in particular studying the conversational structure of the data.

The paper is organized as follows. In Section 2 we briefly survey the related work on Fake News and Stance Detection on one side, and Racial Hoaxes and Stereotypes on the other side. In Section 3 we describe the corpus collection, the annotation process concerning especially the dimension of stance. In Section 4 we provide a corpus-based analysis taking into consideration also the dimension of stereotype, and we show some examples from the corpus. Finally, in Section 5 we discuss some insights gained by the observation of the annotated phenomena, and we conclude the paper with final remarks and possible ideas for future research.

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
alessandrateresa.cignarella@unito.it (A. T. Cignarella); simona.frenda@unito.it (S. Frenda); tom.bourgeade@unito.it (T. Bourgeade); cristina.bosco@unito.it (C. Bosco); francesca.derrico@uniba.it (F. D'Errico)

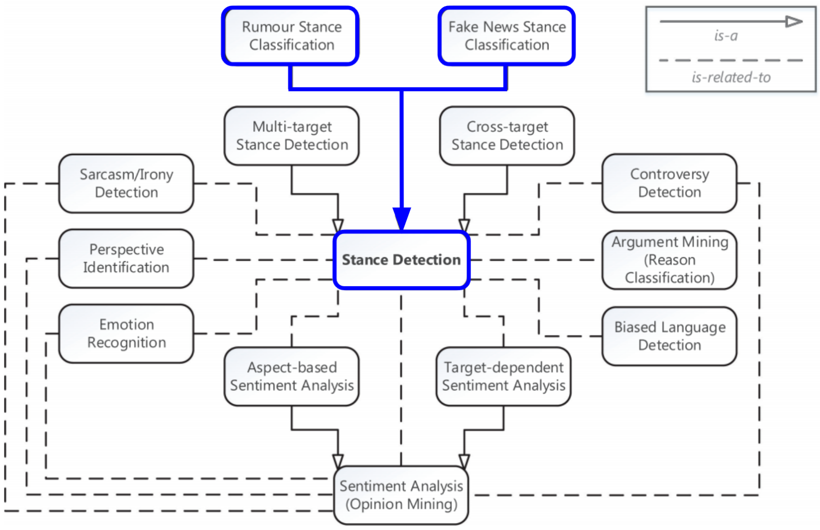


Figure 1: NLP detection tasks and subtasks related to Stance Detection and Sentiment Analysis. Adapted from Küçük and Can [5].

2. Related Work

2.1. Stance Detection, Fake News and Rumors

In SemEval 2016 [6] introduced the first shared task in the domain of stance detection establishing for the first time a formal framework for target-specific stance classification, with admissible labels: *Against*, *Neutral* and *Favor*. The task of stance detection has also proven valuable in distinguishing misinformation from genuine stories. Within the Fake News Challenge participants had to classify the stance towards a claim made in a news headline [7]. By categorizing headlines and news bodies as *Agrees*, *Disagrees*, *Discusses* (a given topic), or *Unrelated*, researchers aimed to identify and combat the spread of fake news more effectively.

In a more general context, researchers started to study the development of stance in online conversational contexts and have employed a slightly different annotation scheme, to classify attitudes toward rumors or broader topics: *Support*, *Deny*, *Query*, *Comment*, often represented as SDQC. This categorization has provided a versatile approach especially classifying tweets belonging to the same conversational thread.

Aker et al. [8] proposed the four labels described above, for the first time at a SemEval shared task: *RumorEval 2017*, which provided a standardized framework for evaluating rumor detection techniques and assessing their effectiveness. The same setting was also proposed in a second edition [9] by introducing additional exercises, such as stance prediction and veracity prediction. This allowed for a more comprehensive evaluation of rumor detection systems, focusing on not only identifying rumors, but also understanding their stance.

Finally, in the context of the Italian language, the *SardiStance* shared task was introduced in EVALITA 2020[1], offering a pioneering challenge for Italian stance detection [10]. As far as the conversational dimension is concerned, Stranisci et al. [11] presented the MoralConvITA corpus, in which moral values and conversational relations linking the components of pairs of messages are annotated with similar categories: *Attack*, *Support* or *Same topic*.

2.2. Racial Hoaxes and Stereotypes

In Bosco et al. [12] we conducted an insightful investigation into the presence of Italian stereotypes on Facebook using a combination of psychology and natural language processing frameworks. We delved into the dynamics of racial stereotyping by extracting replies and comments written below a controversial post written by the famous Italian singer Gianni Morandi, where he compared nowadays migrants in the Mediterranean Sea to Italians immigrating to the USA in the 1920s. We explored how these stereotypes manifest and spread, providing valuable insights into the prevalence and impact of Italian stereotypes in online spaces.

Similarly, D'Errico et al. [13] examines stereotypes and prejudices that arise from racial hoaxes using a psycholinguistic analysis approach. The study investigates in-

[1]http://www.di.unito.it/~tutreeb/sardistance-evalita2020/index.html

stances where false information or hoaxes related to immigrants are spread, leading to the reinforcement or creation of negative stereotypes and prejudices.

Building upon this line of research, our team presented our latest paper at the EACL 2023 conference. In the paper titled *"A Multilingual Dataset of Racial Stereotypes in Social Media Conversational Threads"* we introduced a novel multilingual dataset called Multi-StereoHoax [14]. Our study aimed at studying racial hoaxes and stereotypes in three different languages: Italian, Spanish and French. The dataset is labeled with a complex annotation scheme, based on the Stereotype Content Model (SCM) proposed by Fiske et al. [15]. It is a theoretical framework that provides a psychological understanding of how stereotypes are formed, maintained, and applied in social contexts.

These studies provide crucial insights into the origins, dissemination, and potential consequences of stereotypes, paving the way for future efforts to mitigate their harmful effects and promote a more inclusive online environment. In this study, we attempt to bridge the gap between these two areas of research. Specifically we extracted the Italian portion of the dataset (StereoHoax-It), which was created for the study of racial hoaxes and stereotypes, and we further annotated it with stance information, enabling a more comprehensive analysis of the propagation and impact of racial stereotypes in Italian online discourse.

3. Describing the Corpus

StereoHoax-It [14] is the Italian subset of a corpus of conversations collected on Twitter originated from hoaxes targeting migrants. We started from an initial list of hoaxes deemed *racial* as they tend to explicitly or implicitly attack immigrants, inciting to adopt a contestant stance to the phenomenon of immigration. This initial list was created by consulting debunking websites (`bufale.net`[2] and BUTAC[3]).

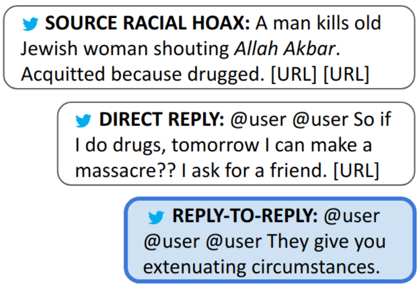


Figure 2: Example of conversational thread.

We were able to collect 273 conversations that discuss these racial hoaxes. The dataset is composed of a total of 2,850 tweets of which 597 are direct replies to the tweets that mention the racial hoax, and 2,253 are replies-to-replies, that is, replies to direct replies. Therefore, the corpus preserves the conversational structure of the Twitter threads, allowing a better analysis of the relations between these conversations' participants. An example of a conversational thread is reported in Figure 2.

In this work, we are interested in studying the stance expressed in the messages of the conversations towards the veracity of the hoax. Considering the purpose of our analysis, we chose to adopt the SDQC schema of annotation adding a label called "*Head*" to identify the texts that spread the hoax or start the conversational thread (identified in Figure 2 as "Source Racial Hoax"). Inspired by Aker et al. [8], we conceived the schema as follows:

- **H** (*Head*): the tweet contains the racial hoax at the root of the conversation;
- **S** (*Support*): the author of the message supports the veracity of the hoax;
- **D** (*Deny*): the author of the message denies the veracity of the hoax;
- **Q** (*Query*): the author of the message asks for additional evidence in relation to the veracity of the hoax;
- **C** (*Comment*): the author of the message makes their own comment without a clear contribution to assess the veracity of the hoax.

As an example, in Figure 3 a source racial hoax, i.e., the *Head* of a Twitter conversation, and four replies (one per SDQC label):

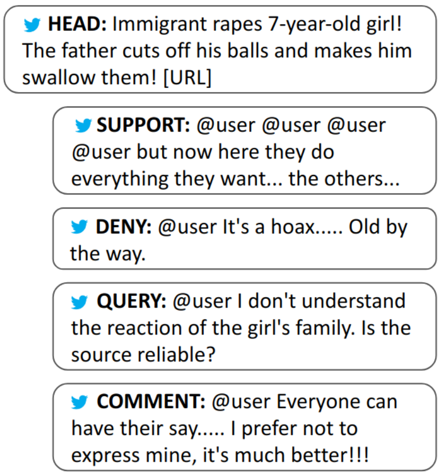


Figure 3: Example of Head with four replies showing different labels.

[2]https://www.bufale.net/
[3]https://www.butac.it/

Differently from the standard schema used by Mohammad et al. [16] where annotators determine if the author of the message is in *favor*/*against*/*neutral* towards a specific phenomenon, the SDQC+H gives us the possibility to identify, more precisely, the attitude of the author with respect to the hoax that targets immigrants. This annotation was applied to direct replies and replies-to-replies toward the hoax declared in the *Head* (or "Source Racial Hoax" in Figure 2 as defined in Bourgeade et al. [14]).

3.1. Enriching the Corpus with Stance Labels

Two different annotators, a male and a female Italian native between 25 and 35 years old (one master student in Linguistics and a PhD student in Digital Humanities) have participated in the annotation campaign. They both annotated all the tweets contained in StereoHoax-It and later additionally annotated it for the dimension of stance as described above. The annotation was performed using *Label Studio*[4] – an open-source annotation platform. Annotator 1 (A1) and Annotator 2 (A2), were both assigned 5,255 tweets in total, and they were asked to label them accordingly to the scheme presented in the previous section (i.e., SDQC+H).

Due to the complexity of the task, and to the fact that annotators could *skip* annotating a tweet in case of uncertainty, in this phase, we were able to collect only 3,123 complete annotations. Once the first round of labeling was completed, we performed an inter-annotator agreement test by calculating Cohen's kappa coefficient, which resulted in $\kappa = 0.3318$ (*fair agreement*). The cases in which A1 and A2 provided two different labels were solved by a third experienced female annotator (A3), an Italian native, 25-35 years old post-doc researcher in NLP.

Thanks to this, some tweets with disagreements were adjudicated, thus increasing the size of the gold-labeled data. However, despite this effort, some disagreements remained for some instances, and we refer to them as "*complex cases*". In Table 1 we report the numbers that are the outcome of the annotations and some more details regarding their nature.

n# tweets	**details**
2,132	skipped tweets / off-topic
449	incomplete annotation from either A1 or A2
202	*complex cases*
2,472	agreement between A1, A2 + A3 (*gold*)
5,255	**total**

Table 1
Number of tweets annotated for stance.

[4] http://labelstud.io/

In the remainder of the paper we provide analyses only focusing on the 2,472 tweets that present agreement between annotators, and leave the study of "complex cases" and incomplete annotations for future versions of the corpus.

4. Analyzing the Corpus

4.1. Annotation Analysis

In this section, we provide quantitative and qualitative analyses regarding annotation of stance. In Figure 4, the bar chart shows the distribution of labels annotated by A1 and A2. We can observe how both annotators had similar judgements when handling users' stance towards migrants. From the same figure, it can also be seen that for both annotators, *Comment* is the predominant label (blue), followed by *Support* (green), *Query* (yellow) and *Deny* (red). The same percentages are respected in the final label distribution of stance calculated over the gold-labeled portion of the dataset, i.e. 2,472 tweets (see Figure 5).

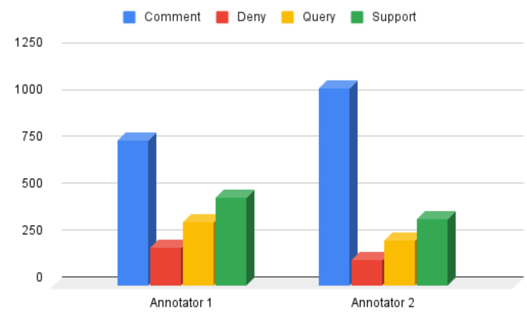


Figure 4: Annotations of A1 and A2.

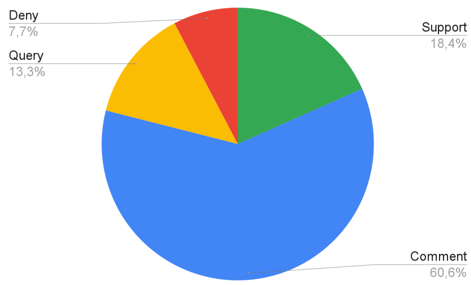


Figure 5: Stance distribution over 2,472 tweets.

Finally, in Table 2 we show a confusion matrix which intersects the newly annotated dimension of stance with the pre-existing annotation of stereotypes. The numbers reported in this table do not add up to 100% because we removed the percentages relative to the tweets annotated with the label *Head*, since they are not relevant for the analyses.

	stance			
stereo	Comment	Deny	Query	Support
no	63.60%	8.16%	14.42%	12.07%
yes	12.77%	1.28%	1.06%	34.68%
Total	**53.91%**	**6.85%**	**11.88%**	**16.38%**

Table 2
Confusion matrix showing the percentages (%) of Support, Deny, Query and Comment labels with respect to the dimension of Stereotype.

The results do not seem to show a particularly significant co-occurrence of one phenomenon with the other. Although, as expected, the majority of tweets annotated as *Support*, also contain racial stereotypes (34.68%), and the majority of tweets annotated as *Comment* do not contain forms of stereotyping towards migrants (63.60%).

We could have expected a significant portion of the *Deny* label to co-occur with the absence of stereotypes, but the tweets annotated with that label are very sparse (they are only 6.58% in total), therefore it is not sufficient for drawing meaningful conclusions.

4.2. Analysis on the Conversational Structure

In order to evaluate the influence of conversational structure on the distribution of stance labels within our dataset, we measured the "conversation depth" of each individual tweet. Specifically, each *Head* tweet was assigned a depth of 0 (however, these *Head* tweets were not considered for the rest of this analysis). The conversation depth of each subsequent tweet was then determined by calculating the length of the reply-chain leading back to the original *Head* of its conversation. Unfortunately, due to the nature of the phenomenon we are investigating here, numerous tweets ($1,947$) presented gaps in their respective reply-chains, due to the deletion of content (either by their authors, or by moderation of the microblogging platform). In these cases, we assigned the minimum potential depth value of 2, given that all *Heads* are accounted for in our dataset. Table 3 thus depicts the distribution of varying stance labels according to depth within the dataset.

Although the label distribution across depths largely mirrors the overall dataset distribution, we can observe a higher proportion of *Support* messages in direct replies (depth 1). This might suggest that users who aim to challenge the veracity of a racial hoax might be more inclined to express themselves as replies to replies, rather than directly under the initial posts. To estimate the correlation

	stance			
depth	Comment	Deny	Query	Support
1	88	16	26	103
***2**	1158	134	247	269
3	74	15	18	27
4	9	3	2	3
5	1	1	0	1
Total	**1330**	**169**	**293**	**403**

(a) In number of labels

	stance			
depth	Comment	Deny	Query	Support
1	6.62%	9.47%	8.87%	25.56%
2*	87.07%	79.29%	84.30%	66.75%
3	5.56%	8.88%	6.14%	6.70%
4	0.68%	1.78%	0.68%	0.74%
5	0.08%	0.59%	0.00%	0.25%
Total	**100%**	**100%**	**100%**	**100%**

(b) In percentages of labels

Table 3
Confusion matrices showing the distribution of *Support*, *Deny*, *Query* and *Comment* labels with respect to the depth of each tweet in its conversation. *the minimum depth for conversations with missing links is 2.

of stance labels with depth, we perform a Chi-squared test and compute Cramér's V: we find a Chi^2 value of 130.762, with a p-value of 4.34×10^-22, as well as a V of 0.134, which thus only indicates a small association [17]. To investigate differences among specific conversations, we compute two measures of "controversiality":

1. *Support-Deny Balance* (SD-B) is simply derived from the proportion of *Support* minus *Deny* messages, as a positive or negative percentage of their sum:

$$\frac{\text{count}(Support) - \text{count}(Deny)}{\text{count}(Support) + \text{count}(Deny)}$$

2. `max P-index` implements the measure proposed by Akhtar et al. [18], the *Polarization Index*, where we consider each conversation as an instance with its *Support* and *Deny* replies as annotations. We then iterate over all possible $k = 2$ partitions of these annotations (*proponents* and *opponents*) for each conversation, and find the maximum *P-index* which we report here.

Table 4 presents these measures for the 10 largest conversations (in number of tweets) in the dataset, as well as their percentage of messages containing stereotypes. We only display the top-10 both for space reasons and because further conversations are too small to compute meaningful metrics (starting from the 27th largest conversation the number of messages is 4 or less, and many

Conv #	SD-B	max P-idx	Stereo %	Size
1	14.29%	97.96%	3.45%	898
2	50.48%	74.52%	5.18%	657
3	100.00%	0.00%	39.06%	64
4	64.44%	58.47%	20.00%	60
5	-71.43%	48.98%	7.69%	52
6	55.56%	69.14%	5.00%	40
7	100.00%	0.00%	11.43%	35
8	-33.33%	88.89%	20.59%	34
9	100.00%	0.00%	21.21%	33
10	100.00%	0.00%	20.00%	30

Table 4
Balance (SD-B) of *Support* (100%) vs. *Deny* (-100%), maximum *P-index* [18], and percentage of stereotypes in messages, for the 10 largest conversations in the dataset (in number of tweets).

have no *Support* and *Deny* replies). Conversation `#1` can be considered the most "controversial" in this dataset (`SD-Balance` closest to 0%, largest `max P-index`), and it also happens to be the largest. The *Head* of this conversation is the following tweet (adapted into English): *"Now Matteo Salvini is in court in Catania for defending the borders, please also tweet #IstandWithSalvini, let's make him feel our affection!"*. As this is a call for support for a controversial figure in Italian politics, this explains the relative balance of *Support* and *Deny* replies, though the number of messages presenting stereotypes remains relatively low, possibly due to supporters' intent not to have their messages moderated by the platform. Examples of more polarized conversations are Conversations `#3` and `#5`: the former does not have a single *Deny* response, with the *Head* being a tweet criticizing the verdict for the 2017 Kobili Traoré murder trial in France, which attracted a significant number of replies containing stereotypes against immigrants and Muslims; whereas the latter provoked a larger proportion of *Deny* responses compared to *Support*, with its *Head* propagating a racial hoax about the Italian government supposedly secretly bringing in illegal immigrants by plane during the COVID-19 pandemic. Interestingly, Conversation `#8` (also displayed in Figure 2) concerns the same subject as Conversation `#3`, but displays a greater proportion of *Deny* responses than the former, indicating that the same subject may be received wildly differently, depending on the context it is introduced in.

4.3. Lexical analysis

To investigate the vocabulary employed by users supporting and denying the hoaxes expressed in the heads of conversational threads, we present: the most relevant n-grams (unigrams, bigrams, and trigrams) of the messages annotated with the presence of stereotypes. The n-grams are weighted using the TF-IDF measure on normalized texts after a specific phase of preprocessing that involves: the deletion of all user mentions, stop-words, punctuation and URLs, leaving only words that were lexically significant. For the tokenization and lemmatization, we employed the small model for the Italian language available in the *SpaCy*[5] library.

By looking at the resulting lists of words and expressions, we noticed that texts labeled as *Support* are explicitly offensive towards immigrants. On the contrary, the ones that are labeled as *Deny* tend to stress the condition of need and poverty of immigrants, and are more empathetic. In this second list, we also noticed some offensive words but towards political parties leaning far-right. Some of the most relevant n-grams of both lists with their TF-IDF values are reported in Table 5.

Support	TF-IDF	Deny	TF-IDF
casa	5.357	nave	0.348
governo	3.889	disperato	0.346
entrare	3.750	pelle povero	0.346
clandestino	3.323	propaganda	0.346
bastardo	2.901	difeso	0.281
cinese	2.677	minacciare	0.281
potere	2.599	povero cristo	0.266
dare	2.521	poveraccio	0.242
merde	1.822	andare cagare leghista	0.175
schifoso	1.605	affamato malato	0.167
invasione	1.574	indifese	0.167
risorsa	1.338	raccogliere pomodoro	0.157
peso	1.314	approdo coraggio	0.121

Table 5
The most relevant n-grams extracted from messages that support and deny the hoax.

In Table 5 we can also notice how the TF-IDF scores greatly vary between *Support* and *Deny*. This is due to the different number of tweets labeled with one or the other category (see Table 4). The top ranking term for the *Support* category is the word *"casa"* (lit. house), probably coming from derogatory expressions like *"rimandiamoli a casa loro"* (lit. let's send them back to their house).

5. Conclusions and Future Work

In this article, we explored the expression of stance and stereotypes as occurring in a dataset of Twitter conversational threads in Italian, focused on the topic of migration. The dataset consists of dialogues originating from tweets containing misinformation marked as untrustworthy by experts.

The analysis of the dataset shed light on the distribution of stance labels and their relationship with stereotypes. The majority of tweets were annotated as *Comment*, followed by *Support*, *Query*, and finally *Deny*. While

[5]https://spacy.io/

there was no significant co-occurrence between stance and stereotypes, tweets annotated as *Support* were more likely to contain racial stereotypes. On the other hand, tweets annotated as *Comment* were less likely to exhibit forms of stereotyping.

The corpus analysis provided insights into how the structure and nature of conversations, and lexical choices in messages, affect the perceived stance of users towards racial hoaxes.

In conclusion, this work paves the way for further investigations about topics closely related to the social phenomenon of misinformation that should be countered to stimulate accurate information dissemination and create a more inclusive online environment. In future research, we may increase the size of the dataset, improve the annotation guidelines and consider the feedback provided by the annotators. We may moreover further investigate the relationship between stance and stereotypes, as well as explore interventions to mitigate the harmful effects of stereotypes in online conversations.

Limitations

In line with the recent trend of the main NLP conferences, we add a brief section addressing the limitations of our work. In this work, we enrich our corpus previously introduced in Bourgeade et al. [14], using a similar annotation framework, and therefore the same limitations brought forward in this work still apply here: more specifically, regarding the practical reliability of the theoretical social-psychological framework used to derive the annotation guidelines. In addition, the Italian subset of the multilingual StereoHoax corpus has a very limited size and presents many unbalanced dimensions and high data sparsity. If in the future it will be used for computational tasks, as it is intended, it should be made more balanced and more inclusive in terms of data sources.

Acknowledgments

This work was partially funded by the International project *STERHEOTYPES - Studying European Racial Hoaxes and sterEOTYPES*, funded by the Compagnia di San Paolo and VolksWagen Stiftung under the ‘Challenges for Europe’ Call for Projects (CUP: B99C20000640007). The work of T. Bourgeade is funded by the project StereotypHate, funded by the Compagnia di San Paolo for the call ‘Progetti di Ateneo - Compagnia di San Paolo 2019/2021 - Mission 1.1 - Finanziamento ex-post’.

References

[1] G. Ruffo, A. Semeraro, A. Giachanou, P. Rosso, Studying fake news spreading, polarisation dynamics, and manipulation by bots: A tale of networks and language, Computer Science Review 47 (2023) 100531.

[2] E. Aïmeur, S. Amri, G. Brassard, Fake news, disinformation and misinformation in social media: A review, Social Network Analysis and Mining 13 (2023) 30.

[3] A. Cerase, C. Santoro, From racial hoaxes to media hypes: Fake news’ real consequences., Amsterdam University Press, 2018, pp. 333–354. URL: http://www.jstor.org/stable/j.ctt21215m0.20.

[4] K. K. Russell, The color of crime: Racial hoaxes, white fear, black protectionism, police harassment, and other macroaggressions, New York University Press New York, 1998.

[5] D. Küçük, F. Can, Stance detection: A survey, ACM Computing Surveys (CSUR) 53 (2020) 1–37.

[6] S. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, C. Cherry, SemEval-2016 Task 6: Detecting Stance in Tweets, in: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016), ACL, 2016.

[7] FNC, Fake News Challenge Stage-1 (FNC-1): Stance Detection, http://www.fakenewschallenge.org/, 2017.

[8] A. Aker, L. Derczynski, K. Bontcheva, Simple Open Stance Classification for Rumour Analysis, in: Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2017), INCOMA Ltd., 2017.

[9] G. Gorrell, E. Kochkina, M. Liakata, A. Aker, A. Zubiaga, K. Bontcheva, L. Derczynski, SemEval-2019 Task 7: RumourEval 2019: Determining Rumour Veracity and Support for Rumours, in: Proceedings of the 13th International Workshop on Semantic Evaluation: NAACL HLT 2019, ACL, 2019, pp. 845–854.

[10] A. T. Cignarella, M. Lai, C. Bosco, V. Patti, P. Rosso, SardiStance@EVALITA2020: Overview of the Stance Detection in Italian Tweets, in: Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020), 2020.

[11] M. Stranisci, M. De Leonardis, C. Bosco, V. Patti, The expression of moral values in the twitter debate: a corpus of conversations, Italian Journal of Computational Linguistics 7 Special Issue: Computational Dialogue Modelling: The Role of Pragmatics and Common Ground in Interaction (2021).

[12] C. Bosco, V. Patti, S. Frenda, A. T. Cignarella, M. Paciello, F. D’Errico, Detecting racial stereotypes: An

Italian social media corpus where psychology meets NLP, Information Processing & Management 60 (2023) 103118.
[13] F. D'Errico, C. Papapicco, M. T. Delor, 'Immigrants, hell on board': Stereotypes and prejudice emerging from racial hoaxes through a psycho-linguistic analysis., Journal of Language & Discrimination 6 (2022).
[14] T. Bourgeade, A. T. Cignarella, S. Frenda, M. Laurent, W. Schmeisser-Nieto, F. Benamara, C. Bosco, V. Moriceau, V. Patti, M. Taulé, A Multilingual Dataset of Racial Stereotypes in Social Media Conversational Threads, in: Findings of the Association for Computational Linguistics: EACL 2023, 2023, pp. 674–684.
[15] S. T. Fiske, A. J. Cuddy, P. Glick, Universal dimensions of social cognition: Warmth and competence, Trends in cognitive sciences 11 (2007) 77–83.
[16] S. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, C. Cherry, A dataset for detecting stance in tweets, in: N. C. C. Chair), K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), ELRA, Paris, France, 2016.
[17] C. M. B. McLaughlin, The association between two categorical variables: A review of cramér's v and alternative statistics, The Journal of the Royal Statistical Society. Series D (The Statistician) 26 (1977) 187–195.
[18] S. Akhtar, V. Basile, V. Patti, A new measure of polarization in the annotation of hate speech, in: AI*IA 2019 – Advances in Artificial Intelligence, volume 11946, Springer International Publishing, Cham, 2019, pp. 588–603. URL: http://link.springer.com/10.1007/978-3-030-35166-3_41. doi:10.1007/978-3-030-35166-3_41.

Interpretation of Generalization in Masked Language Models: An Investigation Straddling Quantifiers and Generics

Claudia Collacciani[1,*,†], Giulia Rambelli[1,†]

[1]University of Bologna, Italy

Abstract

Generics are statements that express generalizations and are used to communicate generalizable knowledge. While generics convey general truths (e.g., *Birds can fly*), they often allow for exceptions (e.g., penguins do not fly). Nonetheless, generics form the basis of how we communicate our commonsense about the world [1, 2]. We explored the interpretation of generics in Masked Language Models (MLMs), building on psycholinguistic experimental designs. As this interpretation requires a comparison with overtly quantified sentences, we investigated i) the probability of quantifiers, ii) the internal representation of nouns in generic vs. quantified sentences, and iii) whether the presence of a generic sentence as context influences quantifiers' probabilities. The outcomes confirm that MLMs are insensitive to quantification; nevertheless, they appear to encode a meaning associated with the generic form, which leads them to reshape the probability associated with various quantifiers when the generic sentence is provided as context.

Keywords

Generics, Quantifiers, Masked Language Models, Commonsense Knowledge, Pragmatics

1. Introduction

Generic generalizations, or *generics*, are sentences such as *Birds fly* and *Cars have four wheels*, which allow us to convey information about categories, or *kinds*, of individuals. They are used to communicate information that extends beyond the present context and express our knowledge about the world, including beliefs, stereotypes, or prejudices (e.g., *Women are more sensitive than men*, as well as the less harmful *Italians eat spaghetti*). Generics can be considered one of the cornerstones of human cognition since they allow us to conceptualize the properties we attribute to categories and thus organize our experience of the world [3].

The most distinctive feature of generics is that they allow for exceptions [4]. For example, *Birds fly* is judged true even if there are birds that cannot fly (e.g., penguins): in this case, therefore, the corresponding universal statement (*All birds fly*) is false. Different generalizations tolerate exceptions to varying degrees. Thus, some generic statements might be better paraphrased with *all*, others with *most*, and others with *some*, but —unlike quantified statements —they do not explicitly contain information about the prevalence of the property in the category (i.e., how many members of the category possess the property). Similarly, there is no unambiguous relationship between the prevalence of a property among category members and the acceptability of the corresponding generic as true. For example, the generalization *Lions have manes* is accepted even if only male adult lions have manes, but the generalization *Lions are males* is rejected.

Given these properties, the meaning of generalizations can be considered "vague", and their interpretation can be assumed to be derived by people through world knowledge and pragmatic skills [5]. Most of the experimental studies conducted on generics are cognitively driven and based mainly on contrasting generics with overtly quantified sentences [3]; in other words, quantifiers are used to approximate the vague meaning of generics.

In this paper, we investigate the interpretation of generalizations in Large Language Models (LLMs) of the Transformer family, building on psycholinguistic experimental designs (in particular, Leslie et al. [6] and Cimpian et al. [7]). Since comparison with quantification seems to be necessary to decode the meaning of generics, we also used quantifiers.

We present three tasks related to different but complementary research questions:

1. *Are LLMs biased towards some quantifiers more than others?* We computed the probability of several quantifiers appended to a generic statement. This analysis serves as a baseline to understand the probability distribution of quantifiers and if

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 – Dec 02, 2023, Venice, Italy

*Corresponding author.
†These authors contributed equally.
claudia.collacciani2@unibo.it (C. Collacciani); giulia.rambelli4@unibo.it (G. Rambelli)

CEUR Workshop Proceedings (CEUR-WS.org)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

there is any bias towards some of them (for example, a generic overgeneralization effect such as that found in humans by Leslie et al. [6]).

2. *Are the hidden representations of generics similar to those of quantified phrases?* We extracted the hidden representation of words in generic and quantified sentences and compared their representation pairwise to understand which quantified nominal phrase approximates the meaning of generics better.
3. *Are LLMs showing the same prevalence effect as humans?* We reproduced the experimental design of Cimpian et al. [7] Implied Prevalence task to test whether the presence of the generic as a premise impacts the probability of selected quantifiers.

The data and code that we used for the experiments are publicly available[1].

2. Related Works

2.1. Generics in Human Cognition

Experimental evidence has revealed a generic bias for which people tend to overgeneralize from the truth of a generic to the truth of the corresponding universal statement [6, 8, 9]. For example, people tend to accept the statement *All lions have manes* as true, even though it is not, because they rely on the truth of the corresponding generic *Lions have manes.* This effect is known as the generic overgeneralization (GOG) effect [6]. It is detected only on certain categories of generics, namely those of minority characteristic and majority characteristic generics, i.e., generics that predicate properties that are true for a minority or the majority of category members.

Cimpian et al. [7] conducted a series of studies investigating the relationship between genericity and prevalence and found an inferential asymmetry in the meaning of generics. People tend to judge a generic sentence about a novel category as true even if they have been informed that only a certain percentage of the kind (on average, up to less than 70 percent) possess the property in question (Truth Condition task). However, when asked to estimate how many members of the kind possess the property, given the generic (Implied Prevalence task), they tend to assign very high percentages (on average, very close to 100 percent). This study indicates that generic sentences require little evidence to be judged true but have substantial implications, since the properties they predicate tend to be interpreted as applying to virtually all members of the category.

[1] https://github.com/claudiacollacciani/Interpretation-of-Generalization-in-Masked-Language-Models

2.2. Genericity in NLP

Most of the NLP literature on genericity has focused on the creation and annotation of resources for identifying generic expressions as opposed to non-generic ones and, based on these resources, on the development of automatic annotation systems [10, 11, 12, among others]

To the best of our knowledge, there are no studies investigating the interpretation of generalizations by LLMs, except for the recent work by Ralethe and Buys [13], which addresses the generic overgeneralization effect in BERT and RoBERTa. The authors argue that these models suffer from overgeneralization by assessing how many times one or more of *all, every, most, some, few* and *many* are predicted in a masked sentence like *[MASK] lions have manes*: the higher the rank of the quantifiers, the stronger the LM exhibits the GOG effect. However, the GOG effect refers to the acceptance of *universally quantified* sentences, not just quantified ones; therefore, we can speak of overgeneralization only when the preferred quantifier is the universal one (*all* or *every*). For this reason, we first propose a similar task to evaluate the probability distribution of various quantifiers, distinguishing between them qualitatively.

3. Materials and Methods

Data For this study, we selected the generic sentences from the dataset of Allaway et al. [14]. The authors extracted 653 generics about objects, animals, and plants from Bhagavatula et al. [15] and annotated them into three categories obtained unifying theories from linguistics and philosophy, by condensing the five types of generics proposed by Leslie [16, 17] and Khemlani et al. [18]. In **quasi-definitional** sentences, the property is essential to a concept, thus is considered a defining characteristic of the concept (e.g., *triangles have three sides*). In this type of sentences, the generic is *de facto* equivalent to the corresponding universal quantified statement (e.g., *all triangles have three sides*). In **principled** sentences, the property has a strong association with the concept. This category includes both properties that are viewed as inherent, or connected in a principled way with a concept (e.g., *birds can fly.*), and properties that are uncommon and often dangerous (e.g., *sharks attack swimmers*); this last case is the one that Leslie [16, 17] defines *striking*. Finally, **characterizing** sentences express a non-accidental relationship between property and concept, based only on absolute or relative prevalence among category members. These generics concern properties that are neither deeply connected to the concept nor striking, but occur in the majority (*majority characteristic* generics for [16, 17], e.g., *Cars have radios*) or in the minority of members of the category (*minority characteristic* generics for [16, 17], e.g., *Lions have manes*).

From the original batch, we restricted our choice to 207 generic sentences, picking only the ones in the bare plural form (e.g., *Tigers are striped*), excluding indefinite and definite singulars (e.g., *A/The tiger is striped*). All these syntactic forms can express generic meanings, but the bare plural is the only surface form in English that gives rise to a generic interpretation unambiguously [19]. For this and other reasons, this is considered as the paradigmatic case, and it is the one that has been used in the psycholinguistic experiments from which we draw inspiration.

Models We experimented with BERT and RoBERTa, two bidirectional Masked Language Models (MLMs) based on the Transformer architecture. **BERT** [20] is trained both on a masked language modeling task and on a next sentence prediction task, as the model receives sentence pairs in input and has to predict whether the second sentence is after the first one in the training data. BERT has been trained on the BookCorpus and the English Wikipedia for around 3300M tokens. We employed the `bert-base-uncased` and `bert-large-uncased` pre-trained versions, which differ in terms of parameters (110M and 340M parameters, respectively). On the other hand, **RoBERTa** [21] has the same architecture as BERT; however, it introduces several parameter optimization choices, such as dynamic masking, a larger batch and vocabulary size, and the removal of the next sentence prediction objective. Another key difference is the larger training corpus: RoBERTa was trained on 160GB of texts. We relied on the Huggingface's Transformers[2] Library to load the models and carry on our experiments.

4. Experiments

4.1. Experiment 1: Probability distribution of Quantifiers in MLMs

In the first place, we needed to assess what was the most expected quantifier for the sentences in our dataset. Therefore, we modified the original generic sentences by placing the special token [MASK] at the beginning of each sentence, as in '[MASK] *strawberries have a sweet flavor.*' Then, we computed the conditional log probability of quantifiers *few, some, many, most*, and *all* in the masked position, following previous works in quantification [22, 13, 23]. The conditional log probability is defined as

$$p(w_i) = logP_{MLM}(w_i|c) \qquad (1)$$

where c are the words preceding and following the critical word in the sentence.

[2]https://huggingface.co/docs/transformers/index

This analysis serves as a baseline to understand the probability distribution of quantifiers, if there is any bias towards some of them (i.e., overgeneralization effect), and possibly whether the belonging of sentences to different categories impacts it (as observed for humans by Leslie et al. [6]).

Results Figure 1 reports the quantifiers distributions for the base models, as the larger counterparts show a similar trend (all boxplots are in Appendix B). Overall, all models consider *few* the least likely. This outcome reflects our expectations: as the selected sentences are all generalizations, they are, in most cases, referable to a substantial number of category members, rarely to 'few' members. Apart from that, BERT and RoBERTa show different probability distributions of quantifiers, regarding *some* and *all* in particular. BERT models assign a higher probability to the existential and proportional quantifiers (*some*, *many*, and *most*) than the universal quantifier *all*, and *some* is overall the most expected. The differences among the quantifier scores are statistically significant[3], with few exceptions (see Appendix B). It is

[3]By relying on Wilcoxon Signed-Rank Test statistical test.

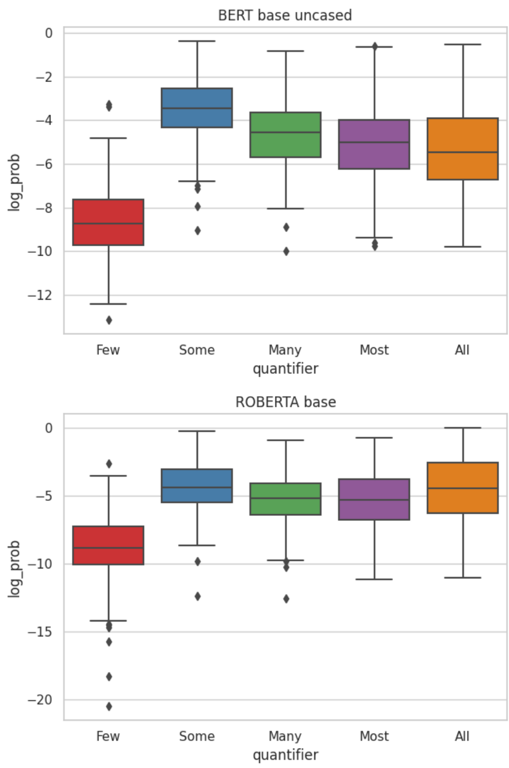


Figure 1: Probability distributions per quantifier for MLM-base variants in Experiment 1.

worth noticing that the reported distributions remain the same even when we separate the analysis by sentence categories. In other words, BERT models are not sensitive to the pragmatic differences of the selected sentences.

Conversely, both *all* and *some* are the most expected quantifiers for RoBERTa. Accordingly, the universal quantifier *all* seems to be more expected by RoBERTa than by BERT, but this distribution is not constant for all three sentence conditions. In quasi-definitional and characterizing sentences, *some* and *all* have the same probability; alternatively, *some* is more expected in principled sentences. We could draw that, for principled sentences, the model prefers not to overgeneralize the property to all members of the category. This behavior seems to approximate Leslie et al. [6] results: people tend to overgeneralize (i.e., to accept as true the universal sentence corresponding to the generic) in the case of characterizing sentences, while they do not overgeneralize (correctly) in the case of striking sentences (included in the principled category).

Regardless, the observed trends could be determined by the overall frequency of the quantifiers. As a sanity check, we extracted their frequency from a large corpus of English, enTenTen21 [24, 25]. We found that $freq(some) > freq(all) > freq(many) > freq(few) > freq(most)$ (frequencies are reported in Appendix A). This pattern confirms that *few* is not the less probable because of a frequency effect but for the properties of the sentence. Conversely, *most* is the less frequent but has a probability score similar to the more frequent *many* and *all*. Finally, *some* is overall the most frequent quantifier. This observation could partially reflect the probability outputs of BERT; however, it is not the case for RoBERTa scores.

4.2. Experiment 2: Representation of words in Generics and Quantified Sentences

The architecture of MLMs allows us to follow the transformations of each token throughout the neural network. Previous works in BERTology have reported that internal representations, also known as contextualized embeddings, encode syntactic and semantic properties in different hidden layers [26]. However, it is complex to localize semantic phenomena, as they spread across the entire model [27]. For our purposes, we decided to compare the contextualized embedding of a target token (*strawberries*) in the generic sentence (<u>*strawberries*</u> *have a sweet flavor*) with the embedding of the target token in each of the corresponding quantified sentences (e.g., *all* <u>*strawberries*</u> *have..*). Following Timkey and van Schijndel [28], for each layer, we computed the similarity of the two contextualized embeddings by relying on Spearman's ρ correlation[4].

This study has a twofold aim: i) Identify the quantifier that shifts the noun representation closer to that of the generic statement (if possible), and ii) Localize the layers where quantification emerges. To the best of our knowledge, no previous work has explored the internal representations of quantified expressions in relation to genericity.

Results Figure 2 illustrates how the Spearman's ρ between the noun and the quantified version changes with respect to each hidden layer in BERT and RoBERTa base (see Appendix C for the plots of larger models and the ones reporting correlations by sentence categories). The first layers do not show a difference among correlations, meaning that representations characterized by the different quantifiers are practically identical; this is expected, as the context is limitedly attended by the model in these layers. The following layers show a gradual change in correlation values, but BERT and RoBERTa show different patterns. For the first, we observe a slight decrease in scores from layers 3 to 9 (but the correlation values are still above 0.9). Conversely, the peak of the curve is at layer 5 for RoBERTa (ρ 0.76), while the other internal

[4] The authors reported that Spearman's ρ is more robust to rogue dimensions contextual language models than cosine or Euclidean similarity measures.

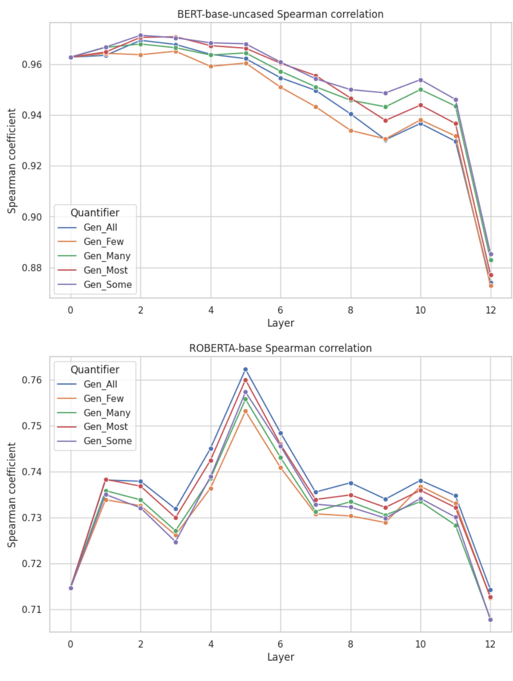


Figure 2: Spearman's ρ against model layers.

layers have a constant value of around 0.73. In the last intermediate layers (8-10), we observe some drifting away between the values for the different quantifiers, indicating that the contribution of the quantifiers differs from one to another. For instance, in BERT-base and large the noun quantified by *some* is the most similar to the corresponding non-quantified. At the same time, RoBERTa has generics similar to *all*, which could be interpreted as an overgeneralization effect.

Intriguingly, the most similar representations are the most probable quantifiers of the previous experiment, thus confirming that the choice of the quantifier is not a frequency effect but is related to providing a representation closer to that of the generic statement. However, the values for the various quantifiers always remain close to each other and follow the same trend, so it is hard to disclose how the meaning of the quantifier affects the noun representation. Finally, all models worsened their performance at the last layer - the one producing the most context-specific representations [29], indicating that contextual information weakens the quantification signal.

4.3. Experiment 3: Implied Prevalence effects in MLMs

As observed above, MLMs are not particularly sensitive to quantifiers, and the probability choices are independent of the sentence's meaning. This outcome is mainly due to the fact that these models are agnostic to world knowledge. Therefore, we decided to test the relation between quantification and generalization from a more formal point of view: we examined how models interpret generalizations aside from their content, that is, whether they contain any linguistic information associated with the form of generics.

We reproduce the experimental design of Cimpian et al. [7] Implied Prevalence task, in which people were presented with a generic sentence about a novel animal category and then asked to estimate how many members of the category possess the characteristic predicated by the generic (e.g., Information: *Morseths have silver fur.* Question: *What percentage of morseths do you think have silver fur?*). In this case, world knowledge is not called into play, unlike in Leslie et al. [6]'s experiments: the categories employed are made up, and thus lack associations to properties in the speakers' mind. Since models do not seem to encode the world knowledge necessary to interpret generics on account of their content (with the partial exception of RoBERTa), this experimental design may be suitable for investigating instead the default interpretation they associate with a generic form.

We build the stimuli using the generic sentence as the premise in the following way: *Strawberries have a sweet flavor means that* [MASK] *strawberries have a sweet flavor.* As in Experiment 1, we compute the log probability of quantifiers (*few, some, many, most,* and *all*) in the masked position. This last study should answer the following question: Does the presence of a generic sentence impact the quantifier preference, compared to the a-contextualized version of the first analysis?

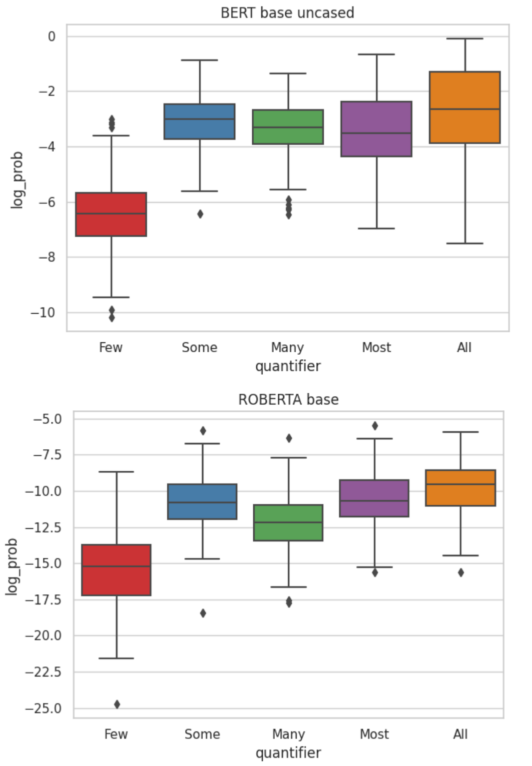


Figure 3: Probability distributions per quantifier for MLM-base variants in Experiment 3.

Results Figure 3 reports the quantifiers distributions for the base models, as the larger counterparts show the same trend (all boxplots are in Appendix D). Surprisingly, the most expected quantifier is now *all* for all models[5]. BERT shows an inversion in the ratio between the probability associated with the universal quantifier *all* and that associated with the other quantifiers (apart from *few*), which in the first experiment were all more expected than *all*, whereas now are all less expected. The pattern exhibited by RoBERTa shows a less striking change than BERT; however, the probability of *all* with respect to the other quantifiers is still higher than in Experiment 1. As an additional check, we performed the same test with a very small dataset of 24 generic sentences about novel

[5]The differences among the quantifier scores are statistically significant; a few exceptions are listed in Appendix D.

categories (non-words) from Cella et al. [30],[6] to be sure that the content of the sentences would not affect the results. As we expected, we obtained the same pattern as with our dataset.

The results of this experiment, when compared with the baseline for the choice of quantifiers in Experiment 1, suggest that the presence of a generic sentence as a premise does indeed have an impact on the preference of the quantifier by the models. When the generalization is provided as context, the preferred quantifier becomes *all*. This behavior mirrors that of people, observed in Cimpian et al. [7] experiments and later replications: people tend to estimate very high percentages (on average very close to 100 percent) in the Implicit Prevalence task.

5. Discussion and Conclusions

In this paper, we analyzed the interpretation of generics in MLMs through psycholinguistic experimental designs that exploited quantified expressions to investigate the understanding of generic ones.

The first two experiments raise questions about the codification of quantifiers, as it seems that the models do not substantially exhibit a strong sensitivity to quantifiers and do not encode a semantic difference in the representation of quantification. Altogether, our results suggest that the models do not appear to contain the commonsense knowledge required to interpret generics that differ in content through quantifiers. However, they seem to have encoded a meaning associated with the generic form, which leads them to reshape the probability associated with various quantifiers when the generic sentence is provided as context. In the last experiment, we observed that the models prefer the universal quantifier unanimously if preceded by a generic utterance. People behave similarly when tested on novel categories, that is, non-word categories for which subjects have no prior understanding. However, people can modulate their interpretations of generalizations in a real language setting through their world knowledge of real categories. Regardless, MLMs tend to treat real and invented categories equally, being agnostic to world knowledge. For this reason, this could be a potentially harmful bias.

The presented analysis has theoretical and methodological implications. First, we observed that the language of generalization is a complex phenomenon that is hard to investigate in human processing and even more in LLMs, mostly because the investigation of generics' interpretation makes use of quantifiers, and language models often fail in tasks related to quantification. Another problem lies in the fact that it is difficult to test autoregressive models (e.g., GPT family) on tasks such as the one used in Experiment 1 because, as they do not have access to the right context, they do not have sufficient information to modulate the probabilities associated with the various quantifiers accordingly. Finding ways to test autoregressive models in addition to MLMs would be desirable.

In this paper, we have not directly investigated the aspect of the attention in MLMs. Future research could address this aspect. Furthermore, future work could involve the definition of alternative tasks for investigating generalizations to make comparing models and human interpretations easier. Psycholinguistic tests on this phenomenon often rely on truth judgments, and we should be cautious about comparing human truth judgments with model outputs since they lack commonsense knowledge comparable to that of humans. Overall, further investigations are needed to clarify the interpretation of generics in language models.

[6] The authors reproduced the experiment of Cimpian et al. [7], obtaining the same results.

Acknowledgments

This research is funded by the European Union (GRANT AGREEMENT: ERC-2021-STG-101039777). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. We thank the anonymous reviewers for their useful suggestions.

References

[1] J. A. Hampton, Generics as reflecting conceptual knowledge, Recherches linguistiques de Vincennes (2012) 9–24.
[2] S.-J. Leslie, Carving up the social world with generics, Oxford studies in experimental philosophy 1 (2014).
[3] D. L. Chatzigoga, Genericity, in: The Oxford Handbook of Experimental Semantics and Pragmatics, Oxford University Press, 2019, pp. 156–177.
[4] M. Krifka, F. J. Pelletier, G. Carlson, A. ter Meulen, G. Chierchia, G. Link, Genericity: An introduction, in: G. N. Carlson, F. J. Pelletier (Eds.), The Generic Book, University of Chicago Press, 1995, pp. 1–124.
[5] M. H. Tessler, N. D. Goodman, The language of generalization., Psychological review 126 (2019) 395.
[6] S.-J. Leslie, S. Khemlani, S. Glucksberg, Do all ducks lay eggs? the generic overgeneralization ef-

fect, Journal of Memory and Language 65 (2011) 15–31.
[7] A. Cimpian, A. C. Brandone, S. A. Gelman, Generic statements require little evidence for acceptance but have powerful implications, Cognitive science 34 (2010) 1452–1482.
[8] S. Khemlani, S.-J. Leslie, S. Glucksberg, Inferences about members of kinds: The generics hypothesis, Language and Cognitive Processes 27 (2012) 887–900.
[9] S.-J. Leslie, S. A. Gelman, Quantified statements are recalled as generics: Evidence from preschool children and adults, Cognitive psychology 64 (2012) 186–214.
[10] N. Reiter, A. Frank, Identifying generic noun phrases, in: Proceedings of the 48th annual meeting of the association for computational linguistics, 2010, pp. 40–49.
[11] A. Friedrich, A. Palmer, M. P. Sørensen, M. Pinkal, Annotating genericity: a survey, a scheme, and a corpus, in: Proceedings of the 9th Linguistic Annotation Workshop, 2015, pp. 21–30.
[12] V. Govindarajan, B. V. Durme, A. S. White, Decomposing generalization: Models of generic, habitual, and episodic statements, Transactions of the Association for Computational Linguistics 7 (2019) 501–517.
[13] S. Ralethe, J. Buys, Generic overgeneralization in pre-trained language models, in: Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 3187–3196. URL: https://aclanthology.org/2022.coling-1.282.
[14] E. Allaway, J. D. Hwang, C. Bhagavatula, K. McKeown, D. Downey, Y. Choi, Penguins don't fly: Reasoning about generics through instantiations and exceptions, arXiv preprint arXiv:2205.11658 (2022).
[15] C. Bhagavatula, J. D. Hwang, D. Downey, R. Le Bras, X. Lu, L. Qin, K. Sakaguchi, S. Swayamdipta, P. West, Y. Choi, I2D2: Inductive knowledge distillation with NeuroLogic and self-imitation, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 9614–9630. URL: https://aclanthology.org/2023.acl-long.535.
[16] S.-J. Leslie, Generics and the structure of the mind, Philosophical perspectives 21 (2007) 375–403.
[17] S.-J. Leslie, Generics: Cognition and acquisition, Philosophical Review 117 (2008) 1–47.
[18] S. Khemlani, S.-J. Leslie, S. Glucksberg, Generics, prevalence, and default inferences, in: Proceedings of the 31st annual conference of the cognitive science society. Austin, TX: Cognitive Science Society, 2009.
[19] G. N. Carlson, Reference to kinds in English., University of Massachusetts Amherst, 1977.
[20] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of NAACL, 2019.
[21] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).
[22] M. Apidianaki, A. Garí Soler, ALL dolphins are intelligent and SOME are friendly: Probing BERT for nouns' semantic properties and their prototypicality, in: Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 79–94. URL: https://aclanthology.org/2021.blackboxnlp-1.7. doi:10.18653/v1/2021.blackboxnlp-1.7.
[23] A. Gupta, Probing quantifier comprehension in large language models, arXiv preprint arXiv:2306.07384 (2023).
[24] A. Kilgarriff, Kovář; v.; rychlý, p.; suchomel, v. the tenten corpus family, in: 7th International Corpus Linguistics Conference CL, 2013.
[25] V. Suchomel, Better web corpora for corpus linguistics and nlp, Doctoral Theses. Brno: Masaryk University, Faculty of Informatics (2020).
[26] A. Rogers, O. Kovaleva, A. Rumshisky, A primer in BERTology: What we know about how BERT works, Transactions of the Association for Computational Linguistics 8 (2020) 842–866. URL: https://aclanthology.org/2020.tacl-1.54. doi:10.1162/tacl_a_00349.
[27] I. Tenney, D. Das, E. Pavlick, BERT rediscovers the classical NLP pipeline, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 4593–4601. URL: https://aclanthology.org/P19-1452. doi:10.18653/v1/P19-1452.
[28] W. Timkey, M. van Schijndel, All bark and no bite: Rogue dimensions in transformer language models obscure representational quality, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 4527–4546. URL: https://aclanthology.org/2021.emnlp-main.372. doi:10.18653/v1/2021.emnlp-main.372.
[29] K. Ethayarajh, How contextual are contextualized word representations? Comparing the geometry

of BERT, ELMo, and GPT-2 embeddings, 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), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 55–65. URL: https://aclanthology.org/D19-1006. doi:10.18653/v1/D19-1006.

[30] F. Cella, K. A. Marchak, C. Bianchi, S. A. Gelman, Generic language for social and animal kinds: An examination of the asymmetry between acceptance and inferences, Cognitive Science 46 (2022) e13209.

A. Quantifiers Frequencies in enTenTen21

We extracted the frequencies from enTenTen21. The corpus is made up of texts collected from the Internet consisting of more than 60 billion tokens. The texts were downloaded in October–December 2021 and January 2022. We relied on the concordance tool provided by SketchEngine to extract the frequencies in the form '[quantifier][noun]'.

	N. hits	N. hits per million tokens	% of whole corpus
some	43,436,065	705.29	0.07053%
all	39,201,717	636.54	0.06365%
many	30,529,532	495.72	0.04957%
few	19,067,122	309.6	0.03096%
most	11,187,850	181.66	0.01817%

Table 1
Distribution of quantifiers in enTenTen21 corpus.

B. Experiment 1: Boxplots and Wilcoxon statistical analysis

We report the boxplots for the base and large versions of BERT and RoBERTa for Experiment 1.

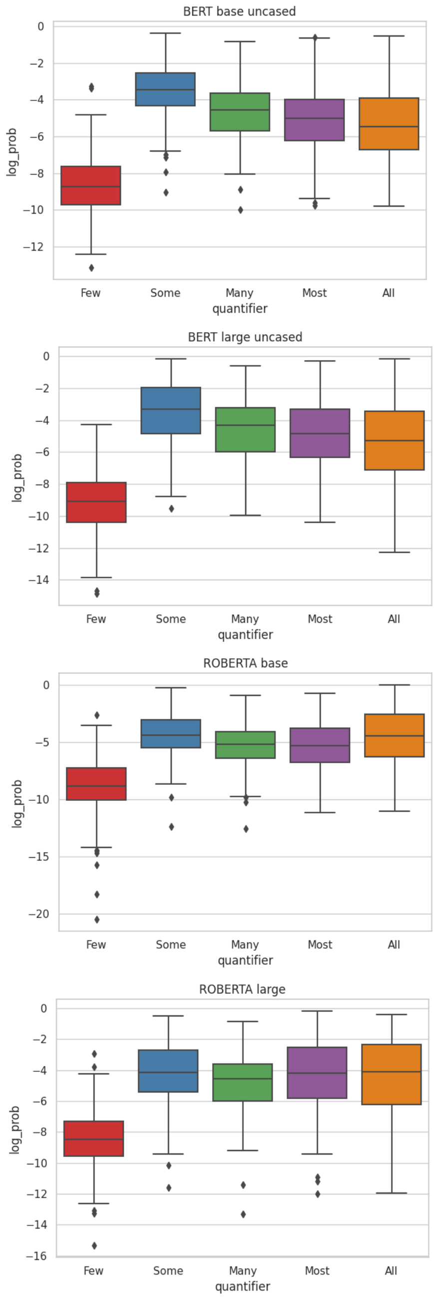


Figure 4: Probability distributions per quantifier for BERT and RoBERTa variants in Experiment 1.

In the tables below, we also report the results of the statistical test performed to verify if the difference in probabilities of quantifiers is statistically significant or not.

Model	Group1	Group2	p-value	Significance
bert-base	few	some	1.023e-35	significant
bert-base	few	many	1.023e-35	significant
bert-base	few	most	7.204e-35	significant
bert-base	few	all	2.538e-34	significant
bert-base	some	many	4.439e-30	significant
bert-base	some	most	2.390e-26	significant
bert-base	some	all	6.709e-21	significant
bert-base	many	most	8.841e-04	significant
bert-base	many	all	8.854e-06	significant
bert-base	most	all	3.054e-03	significant
bert-large	few	some	1.023e-35	significant
bert-large	few	many	1.084e-35	significant
bert-large	few	most	1.536e-35	significant
bert-large	few	all	7.630e-35	significant
bert-large	some	many	4.051e-24	significant
bert-large	some	most	1.296e-19	significant
bert-large	some	all	1.958e-17	significant
bert-large	many	most	5.408e-02	not significant
bert-large	many	all	3.407e-05	significant
bert-large	most	all	6.600e-05	significant

Table 2
Wilcoxon Signed-Rank Test on BERT variants for Experiment 1.

Model	Group1	Group2	p-value	Significance
RoBERTa-base	few	some	1.450e-35	significant
RoBERTa-base	few	many	1.291e-35	significant
RoBERTa-base	few	most	6.319e-33	significant
RoBERTa-base	few	all	9.364e-34	significant
RoBERTa-base	some	many	2.023e-19	significant
RoBERTa-base	some	most	2.724e-08	significant
RoBERTa-base	some	all	4.231e-02	significant
RoBERTa-base	many	most	8.949e-01	not significant
RoBERTa-base	many	all	8.148e-03	significant
RoBERTa-base	most	all	6.013e-05	significant
RoBERTa-large	few	some	1.023e-35	significant
RoBERTa-large	few	many	1.272e-35	significant
RoBERTa-large	few	most	7.414e-35	significant
RoBERTa-large	few	all	3.472e-34	significant
RoBERTa-large	some	many	4.334e-14	significant
RoBERTa-large	some	most	1.612e-01	not significant
RoBERTa-large	some	all	5.842e-02	not significant
RoBERTa-large	many	most	6.901e-06	significant
RoBERTa-large	many	all	1.639e-02	significant
RoBERTa-large	most	all	5.972e-01	not significant

Table 3
Wilcoxon Signed-Rank Test on RoBERTa variants for Experiment 1.

C. Experiment 2: A layer-wise analysis of MLMs representations

We report the plots for the base and large variants of BERT and RoBERTa, with respect to each hidden layer.

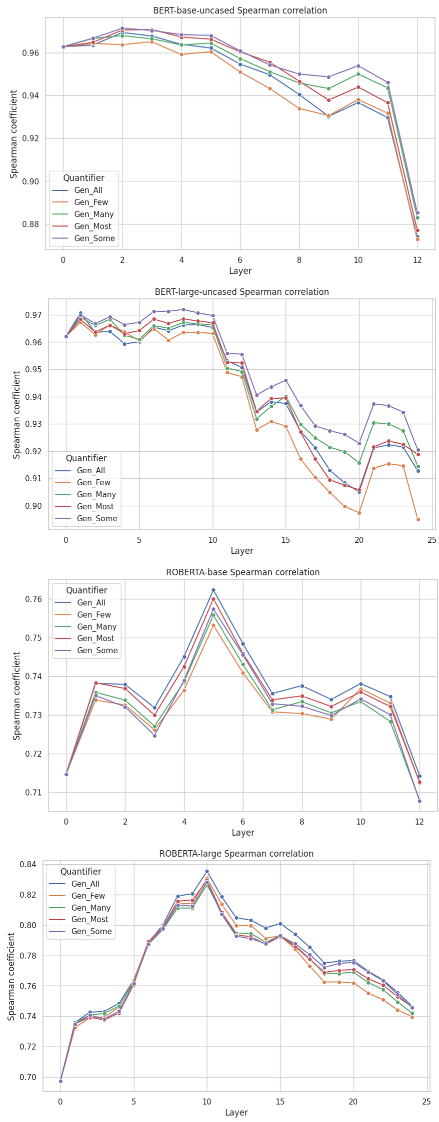


Figure 5: Spearman's ρ between the noun in generic sentence and its quantified variant across models layers.

We also report the same plots with the correlation scores for each sentence category. While the trends are the same for the three conditions, the values have a slight difference in the means, with quasi-definitional sentences having a higher correlation than the other two types.

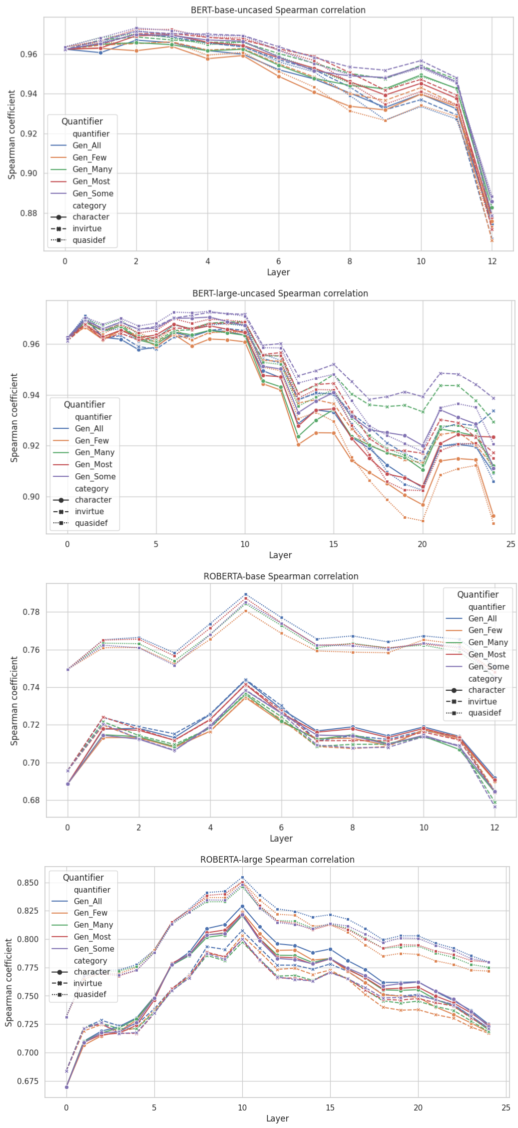


Figure 6: Spearman's ρ between the noun in generic sentence and its quantified variant across models layers, by sentence categories.

D. Experiment 3: Boxplots and Wilcoxon statistical analysis

We report the boxplots for the base and large versions of BERT and RoBERTa for experiment 3.

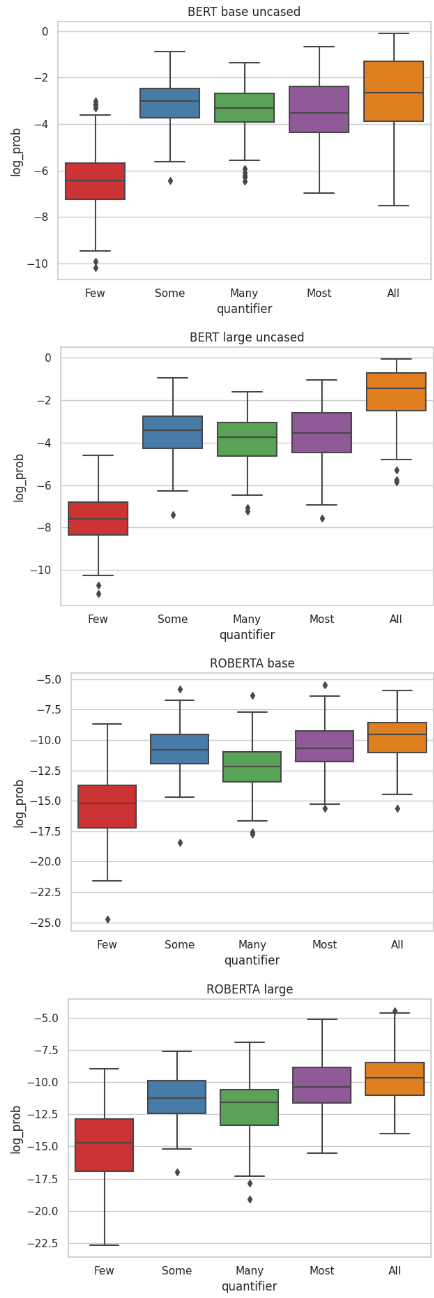


Figure 7: Probability distributions per quantifier for BERT and RoBERTa variants in Experiment 3.

In the tables below, we also report the results of the statistical test performed to verify if the difference in probabilities of quantifiers is statistically significant or not.

Model	Group1	Group2	p-value	Significance
bert-base	few	some	1.701e-35	significant
bert-base	few	many	1.116e-35	significant
bert-base	few	most	1.140e-34	significant
bert-base	few	all	3.357e-35	significant
bert-base	some	many	2.399e-07	significant
bert-base	some	most	1.335e-04	significant
bert-base	some	all	2.599e-02	significant
bert-base	many	most	7.970e-01	not significant
bert-base	many	all	7.696e-07	significant
bert-base	most	all	3.920e-10	significant
bert-large	few	some	1.023e-35	significant
bert-large	few	many	1.084e-35	significant
bert-large	few	most	1.536e-35	significant
bert-large	few	all	7.630e-35	significant
bert-large	some	many	4.051e-24	significant
bert-large	some	most	1.296e-19	significant
bert-large	some	all	1.958e-17	significant
bert-large	many	most	5.408e-02	not significant
bert-large	many	all	3.407e-05	significant
bert-large	most	all	6.600e-05	significant

Table 4
Wilcoxon Signed-Rank Test on BERT variants for Experiment 3.

Model	Group1	Group2	p-value	Significance
RoBERTa-base	few	some	1.084e-35	significant
RoBERTa-base	few	many	5.788e-34	significant
RoBERTa-base	few	most	3.992e-35	significant
RoBERTa-base	few	all	1.652e-35	significant
RoBERTa-base	some	many	4.769e-21	significant
RoBERTa-base	some	most	2.381e-01	not significant
RoBERTa-base	some	all	5.573e-08	significant
RoBERTa-base	many	most	1.141e-19	significant
RoBERTa-base	many	all	2.409e-27	significant
RoBERTa-base	most	all	4.718e-10	significant
RoBERTa-large	few	some	3.169e-35	significant
RoBERTa-large	few	many	3.326e-34	significant
RoBERTa-large	few	most	7.521e-35	significant
RoBERTa-large	few	all	2.515e-35	significant
RoBERTa-large	some	many	3.979e-10	significant
RoBERTa-large	some	most	5.351e-11	significant
RoBERTa-large	some	all	3.152e-20	significant
RoBERTa-large	many	most	1.723e-23	significant
RoBERTa-large	many	all	7.571e-28	significant
RoBERTa-large	most	all	3.265e-07	significant

Table 5
Wilcoxon Signed-Rank Test on RoBERTa variants for Experiment 3.

Highway to Hell.
Towards a Universal Dependencies Treebank for Dante Alighieri's Comedy

Claudia Corbetta[1,2,*], Marco Passarotti[3], Flavio Massimiliano Cecchini[3] and Giovanni Moretti[3]

[1] *Università degli studi di Bergamo, via Salvecchio 19, 24129 Bergamo, Italy*
[2] *Università di Pavia, corso Strada Nuova 65, 27100 Pavia, Italy*
[3] *Università Cattolica del Sacro Cuore, largo A. Gemelli 1, 20123 Milan, Italy*

Abstract

In this paper, we describe the creation in Universal Dependencies of a treebank for Dante's *Comedy*, the first syntactically annotated text for Old Italian following a dependency-based schema. We detail the phase of treebanking the first part of the *Comedy*, the *Inferno*, and we describe some annotation issues. Then, we perform an evaluation of automated dependency parsing with models trained on the currently available annotated portion of the text.

Keywords

Dante Alighieri, Old Italian, treebank, Universal Dependencies

1. Introduction

Over the past two decades, there has been a growing convergence between the world of corpora for ancient languages and the scholarly community working in the area of technologies for Natural Language Processing (NLP). Because of the absence of native speakers and newly written texts, dealing with ancient languages means lacking the possibility of introspective analysis or field inquiries. The only empirical evidence historical linguists can engage with is confined to old texts, many of which are fortunately digitally available today. Enhancing these data sources with meta-linguistic annotation provides scholars with enriched data to support their investigations. Moreover, building annotated sets of textual data for an ancient language following *de facto* standards is a way to make these old texts compatible with several ready-made NLP tools, as well as to make them comparable with annotated corpora for other (modern) languages.

Universal Dependencies[1] (UD) [1] is an annotation framework started in 2015 which aims to provide a universal formalism for dependency-based syntactic annotation, with the goal of facilitating cross-linguistic comparison. Currently, the project boasts 245 treebanks for 141 languages,[2] including historical languages such as Ancient Greek, Latin, Old French, Akkadian and Classical Chinese. With regard to the Italian language, there are 9 UD treebanks, covering a diverse range of genres,[3] amounting to 879 657 tokens and 37 871 sentences.

This paper details the process of developing a UD treebank out of Dante's *Comedy*, starting from the annotation of the *Inferno*, the first out of the three parts (*cantiche*) of the work. The motivation for this is the current absence of any dependency-based treebank for Old Italian.[4] Besides providing the scholarly community of historical linguistics with a valuable resource, we create gold data that can be used for the supervised training and testing of stochastic NLP tools.

This paper is organized as follows: in Section 2, we introduce Old Italian and the resources available for this language, with a specific focus on the DanteSearch corpus. In Section 3, we describe the creation of the treebank, starting from the *Inferno*. In Section 4, we describe training and evaluation of a number of models for parsing. Section 5 concludes the paper by summarizing our findings and sketching future work.

2. Old Italian

Although in earlier stages of linguistic research there were claims of similarity between Old Italian, particu-

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
*Corresponding author.
claudia.corbetta@unibg.it (C. Corbetta);
marco.passarotti@unicatt.it (M. Passarotti);
flavio.cecchini@unicatt.it (F. M. Cecchini);
giovanni.moretti@unicatt.it (G. Moretti)
0009-0000-7425-196X (C. Corbetta); 0000-0002-9806-7187 (M. Passarotti); 0000-0001-9029-1822 (F. M. Cecchini); 0000-0001-7188-8172 (G. Moretti)

[1] https://universaldependencies.org.

[2] UD version 2.12, May 2023 [2].

[3] Including "legal, news, wiki, nonfiction, government legal, social, learner-essays and grammar-examples". No literary texts have been included thus far.

[4] Whereas, with regard to Dante Alighieri, his works in Latin are already part of UD, see [3, 4].

larly Dante Alighieri's vernacular, and Modern Italian,[5] especially when compared to the evolution of other Romance languages like French, where differences between old and modern varieties are more pronounced [7], numerous studies have now recognized and emphasized the distinction between Old and Modern Italian [8], particularly from a syntactic perspective [9].

The *Grammatica dell'italiano antico* (GIA; 'Grammar of Old Italian') [10], defines Old Italian as the language spoken in Florence during the 13th century and the early 14th century. The authors of the GIA justify their choice of selecting Florentine texts (later expanded to texts from all the Tuscan region) on the basis of the abundant documentation of vernacular *scripta* in Florence, driven also by the diligence and productivity of the Florentine scribes. However, it should be noted that there are numerous written varieties that characterize Medieval Italy, albeit in a minority when it comes to documentation and written evidence.

Regardless of whether Old Italian should be strictly limited to the Tuscan area or can also encompass non-Tuscan varieties, the significance and influence of Tuscan on the evolution of the Italian language is undeniable. Therefore, while choosing an Old Italian text for a UD treebank, it seems obvious to select a Tuscan text, specifically a Florentine one, namely the *Comedy* of Dante Alighieri.

Dante Alighieri was born in Florence in 1265 and he is legitimately considered one of the greatest poets and writers of the Middle Ages. His most important work is the *Comedy*, which was written between 1308 and 1320, and is crucial to Italian literature, due to its historical (and still continuing) success among readers, and relevance among scholars. The decision of Dante to write the *Comedy* in the Florentine vernacular represents a pivotal moment in the history of Italian literature and language, as it contributed to spreading and elevating the vernacular to a literary language [11].

Together with the undeniable significance of the text, the availability of a digital resource, DanteSearch [12], containing all of Dante's works enhanced with a number of fundamental layers of annotation, further supports our decision to choose the *Comedy* as the text for the first UD treebank of Old Italian.

2.1. Resources for Old Italian

There is quite a substantial amount of texts and lexical resources in digital format available for Old Italian. Among them, the *Opera del Vocabolario Italiano* corpus[6] (OVI) contains Old Italian texts dating before the 15th century and is one of the major corpora, containing 3 443 texts of Old Italian for a total of 30 176 628 word occurrences. Strictly related to the the historical dictionary of Old Italian built by OVI is the *Tesoro della Lingua Italiana delle Origini* corpus (TLIO) [13], which collects 3 173 texts for a total of 23 685 634 occurrences. Additionally, there are corpora that cover a wider temporal span, such as the MIDIA corpus [14], a lemmatized and morphologically annotated collection of Italian texts from the 13th century to the first half of the 20th century, and the CODIT corpus [15], a diachronic corpus of Italian that covers the period from the 13th century until 1947.

Although a preliminary effort has been made towards the creation of a digital corpus of Old Italian with respect to the quotations reported in the *Grande dizionario della lingua italiana* [16],[7] no dependency-based syntactic annotation of Old Italian texts is currently available.

2.2. DanteSearch

Among the resources available for Old Italian, DanteSearch (DS) [12] is an annotated corpus containing all of Dante Alighieri's works, including both the Latin and the vernacular texts. The resource has been developed by the University of Pisa and consists of a set of (downloadable)[8] XML files providing both textual data and linguistic annotation.

Concerning the *Comedy*, the text included in DS is based on Petrocchi's edition [17] and is recorded in two separate XML files: one file provides the grammatical layer of annotation (featuring tokens, lemmas, and tags representing both parts of speech and morphosyntactic features), while the other contains a clause-based layer of syntactic annotation [18].

The clause-based annotation of syntax distinguishes main and subordinate clauses, the latter being assigned a label for their function, such as "declarative", "temporal", and "relative" [19].

3. Treebanking Dante's *Comedy*: the *Inferno*

Dante's *Comedy* is composed of three parts, called *cantiche*, which are *Inferno* 'Hell', *Purgatorio* 'Purgatory' and *Paradiso* 'Heaven'. These *cantiche* are divided respectively into 34, 33 and 33 subsections called *canti*. This Section details the process of annotating the *Inferno* according to UD's formalism.

[5] As exemplified by a statement by [5, p. 124], cf. [6, ch. VI].

[6] http://www.ovi.cnr.it/Il-Corpus-Testuale.html

[7] The work by Favaro consists of a conversion from an XML source file to the CoNLL-U format adopted in UD, for tokenization, lemmatization, and morphological annotation.

[8] https://dantesearch.dantenetwork.it.

3.1. From DanteSearch to UD

In DS, the *Inferno* consists of 33 416 tokens out of a total of 99 390 (without punctuation marks).

We perform a conversion from the grammatical XML file of the *Inferno* provided by DS to the CoNLL-U format adopted by UD's treebanks.[9] The conversion focuses on tokens (i. e. forms), lemmas, parts of speech (PoS), and morphological features. However, in the CoNLL-U file we do not report the syntactic annotation contained in the XML syntactic file of DS, due to its incompatibility with the word-based UD syntactic analysis [1, §2.2].

The conversion of tags happens on a 1:1 basis (DS:UD) whenever possible. Different criteria for the assignment of PoS and morphological tags between the two annotation styles are managed case by case. For instance, DS alternately assigns the tag for "pronouns" (p) or "adjectives" (a) to possessives such as *mio* 'my', while in UD we always tag them as "determiners" (DET).

With regard to tokenization and lemmatization, in a few cases we modify the criteria followed by DS to fit the ones of UD. Specifically, this applies to the tokenization and lemmatization of what are referred to as *locuzioni* 'locutions' in DS, i. e. sets of two or more words arranged in a fixed sequence [20], such as *mentre che* 'while' and *davanti a* 'in front of'. In DS, such multiword expressions are analyzed as single tokens, while the UD annotation schema requires that the words they are composed of be analyzed individually and considered as separate tokens. As a consequence, for locutions we employ a distinct tokenization, lemmatization, and PoS tagging in contrast to DS, as shown in Table 1 with regard to the following example:[10]

> *Inferno*, V, vv. 95–96
>
> noi udiremo e parleremo a voi, / *mentre che* 'l vento, come fa, ci tace.
>
> 'will please us, too, to hear and speak with you, / now *while* the wind is silent, in this place.'[11]

	DS	UD	
no. tokens	1	2	
lemma(s)	*mentre che*	*mentre*	*che*
tag(s)	`clst`	ADV	SCONJ

Table 1
Example of locution *mentre che*

Modifications of lemmatization and PoS tagging are required also for multiword proper nouns, which are lemmatized under a unique lemma in DS in contrast to UD. Table 2 shows the example of the multiword proper name *Filippo Argenti*:[12]

	DS	UD	
no. tokens	1	2	
lemma(s)	*Filippo Argenti*	*Filippo*	*Argenti*
tag(s)	n	PROPN	PROPN

Table 2
Example of the proper noun *Filippo Argenti*

Further, we also want to adjust the lemmatization of articles. In DS, there are separate lemmas *la*/*una* and *il*/*uno* for the definite/indefinite feminine and masculine articles respectively, whereas, following the convention of most UD Italian treebanks, we lemmatize both under the respective masculine forms.

3.2. Syntactic annotation

We perform the syntactic annotation of the *Inferno* manually[13] using *ConlluEditor* [22] and with the support of a few critical commentaries on the work, namely those by Chiavacci Leonardi [23] and Inglese [24]. Following the UD guidelines, annotation is made at sentence level; we base sentence splitting on full stops and question or exclamation marks followed by an uppercase letter, according to Petrocchi's edition of the *Comedy* recorded in DS [17].

A sentence corresponds to a syntactic tree, i. e. an acyclic, oriented, rooted graph [25], whose nodes correspond to tokens in the text.[14] Nodes are related to each other through dependencies, i. e. hierarchical binary relations, which are labeled with a syntactic function, such as

---

[9] CoNLL-U is a format with tab-separated values where lines contain the annotation of tokens into 10 fields; see https://universaldependencies.org/format.html.

[10] In this example, the DS tag `clst` stands for a subordinating conjunction (`cs`) used in a locution (`l`) within a temporal clause (`t`), while the UD PoS tags ADV and SCONJ stand respectively for "adverb" and "subordinating conjunction".

[11] The English translations of the examples from the *Comedy* are by Allen Mandelbaum, available at: https://digitaldante.columbia.edu/dante/divine-comedy/.

[12] The DS tag n stands for "onomastics" and the UD tag PROPN stands for "proper noun".

[13] The syntactic annotation is performed by a single annotator with expertise in Italian studies. Annotating pre-parsed data has been ruled out after evaluating the accuracy of the UDPipe model [21] trained on the largest UD treebank of Italian (ISDT) and tested on the first three *canti* of *Inferno*: its LAS score is 63,52% (see Section 4).

[14] In UD, a distinction between "token" and "syntactic word" is made: while "token" refers to an orthographic unit of segmentation, "syntactic word" refers to the actual level of analysis in the syntactic tree. These two levels often, but not always, coincide, e. g. the token *nel* 'in the' would be analyzed into the syntactic words *in* 'in' and *il* 'the', each bearing its own annotation. Refer to https://universaldependencies.org/u/overview/tokenization.html. In this paper, the term "token" will be used throughout as an equivalent to UD's "syntactic word".

nsubj for “nominal subject”. Dependency-based annotation schemes are predicate-centered, with the sentence’s main predicate serving as the tree’s root. In UD’s formalism, function words depend on the content words they modify.[15]

While annotating the *Inferno* according to the UD formalism, we encounter several issues that require taking specific decisions. In the following, we discuss the annotation of ellipses and comparative clauses.

The total number of sentences in the *Inferno* is 1 228, for a total of 41 367 tokens.

3.2.1. Ellipsis

“Ellipsis” refers to the omission of words or phrases that can be inferred from the context of a sentence or utterance.[16] While annotating the *Inferno*, we encounter several cases of ellipses, including nominal ellipses, i. e. [27, p. 526]:

> different types of anaphoric phenomena involving a gap within the internal structure of the nominal phrase.

and predicate ellipsis [28, p. 504]:

> a type of ellipsis that leaves the main predicate of the clause unpronounced, most often together with one or more of its internal arguments or (low) adjuncts.

In the matter of nominal ellipsis, we follow the solution of promotion, as outlined in the UD guidelines.[17] We present here an example of nominal ellipsis (Figure 1):

> *Inferno*, IX, vv. 28–29:
>
> Quell’è ’l più basso loco e ’l più *oscuro* / e ’l più *lontan* dal ciel che tutto gira
>
> ‘That is the deepest place and the darkest place, / the farthest from the heaven that girds all’

where *oscuro* ‘dark’ and *lontan* ‘far’ depend on the omitted noun (NOUN) *loco* ‘place’, as shown by the repetition of the definite article (DET) *’l* ‘the’, which modifies the noun. In this case, we promote the adjectives (ADJ) *oscuro* and *lontan* to heads of their respective coordinate clauses using the dependency relation conj “conjunct”.

Following to the UD guidelines, we handled predicate ellipsis by using the dependency relation orphan (orphan relation), like in the following example:

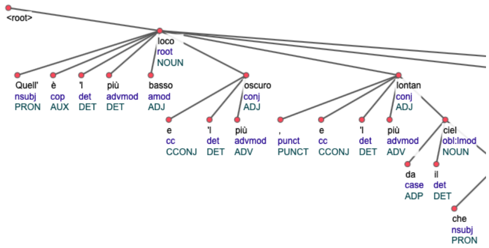


Figure 1: Nominal ellipsis (*Inf.*, IX, vv. 28–29)

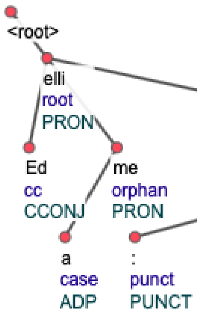


Figure 2: Predicate ellipsis (*Inf.*, III, v. 76)

> *Inferno*, III, v. 76:
>
> Ed elli a me:
>
> [And he (said) to me:]

where the predicate of the sentence, namely the *verbum dicendi*, is omitted. This structure is extremely common to introduce a reported speech. As shown in Figure 2, the omission of the predicate requires promoting the subject of the sentence, *elli* “he”, to the root of the tree (root) and annotating the underlying oblique relation (obl) of the phrase *a lui* “to him” with an orphan relation (orphan).

Currently, the syntactic annotation in UD handles these cases of ellipsis with the promotion mechanism, which involves promoting an element to function as the omitted element in the sentence and replacing it in its dependency relation without explicitly signaling this omission, and the use of the orphan dependency relation, whose function is to indicate that the element subject to

---

[15] This is not the case for all dependency-based schemes, like for instance for the analytical layer of annotation of the Prague Dependency Treebank for Czech (PDT), where e. g. conjunctions govern conjuncts and adpositions are the heads of adpositional phrases. Refer to https://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/a-layer/html/index.html.

[16] See [26] for an introduction to the topic.

[17] Promotion involves selecting an element to take the place of the omitted element in the syntactic tree, following a specific hierarchy. Promotion is used without explicitly signaling the ellipsis. See UD guidelines: https://universaldependencies.org/u/overview/specific-syntax.html#ellipsis.

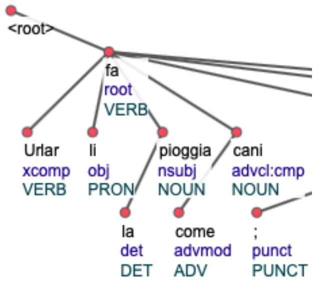


Figure 3: Comparative clause (*Inf.*, VI, v. 19)

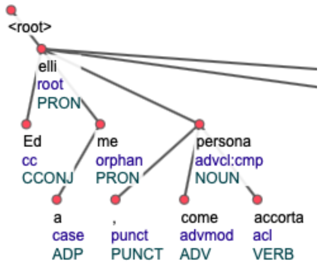


Figure 4: Secondary predication or comparative clause (*Inf.*, III, v. 13)

the `orphan` relation does not have an overt dependent element in the syntactic structure.

3.2.2. Comparative clauses

In the *Inferno*, we find a diverse usage of comparative clauses, ranging from sentences where the comparative clause is longer than the main clause it depends on, to others where comparatives consist of just a few tokens. In light of the long-lasting discussion on the treatment of comparative clauses in UD,[18] we annotate such clauses by labeling their head tokens with the dependency relation `advcl` "adverbial clause modifier" specified for the subtype `cmp` for comparative clauses.[19]

A number of issues concerns cases of clauses where our annotation, following the UD framework, parts from the interpretation provided by DS, like in the following example (Figure 3):

> *Inferno*, VI, v. 19:
>
> Urlar li fa la pioggia *come cani*
>
> 'That downpour makes the sinners howl *like dogs*'

In the annotation of DS, the portion *come cani* 'like dogs' is considered a phrase that is part of a declarative clause. Instead, we consider *come cani* as a comparative clause with an elliptical predicate, namely *Urlar li fa la pioggia come [fa urlare] i cani* 'That downpour makes the sinners howl like [it makes] dogs [howl]'.

We observe a few cases where a *come*-clause can be considered either a comparative clause or a secondary predication. In such cases, we rely on the interpretation provided by commentaries, like in the following sentence (Figure 4):

> *Inferno*, III, v. 13:
>
> Ed elli a me, *come persona accorta*:
>
> 'And he to me, *as one who comprehends*:'

In this sentence, the *come*-clause can be interpreted either as a secondary predication (therefore, annotated using the subtyped relation `advcl:pred`[20]), 'He, being a comprehensive person, answered to me', or as a comparative clause (with subtyped relation `advcl:cmp`), 'He answered to me like a comprehensive person'. In this case, we follow the interpretation of Chiavacci Leonardi [23] in considering the *come*-clause as a comparative construction.

4. Evaluation

We use the manually annotated *Inferno* to train models with UDPipe 1[21] [29] and to assess their performances in view of employing them for parsing *Purgatorio* and *Paradiso*, so as to facilitate their subsequent manual annotation.[22] In our evaluation framework, we employ a cross-validation based on 10%/90% splits of the data: each test set will then consist of approximately 4 137 out of 41 367 tokens and 123 out of 1 228 sentences, while train sets of approximately 37 230 tokens and 1 105 sentences. The evaluation of the models' accuracies is performed by measuring Labeled (LAS) and Unlabeled Attachment Score (UAS) [30].

The training and evaluation process is based on one eleven- and one tenfold partition of the data, for a total of 11+10 iterations: the first partition patterns upon the original division into *canti*, with batches of 3 consecutive

[18] Cf. the discussion group on comparatives in UD: https://universaldependencies.org/workgroups/comparatives.html.

[19] Cf. documentation at https://universaldependencies.org/la/dep/advcl-cmp.html (for Latin).

[20] Cf. documentation at https://universaldependencies.org/la/dep/advcl-pred.html (for Latin).

[21] https://github.com/ufal/udpipe.

[22] We acknowledge that doing tests within a single *cantica* may not guarantee the same performances when compared to other *cantiche*.

Partition	Scenario	Avg. UAS	Avg. LAS
random	+Morph	81,95±0,94%	77,07±1%
consecutive	+Morph	81,79±1,38%	77,09±1,34%
random	-Morph	75,32±0,91%	67,97±0,80%
consecutive	-Morph	74,90±1,37%	67,71±1,17%

Table 3
Averages and standard deviations of accuracy metrics

canti[23] assigned to the test set and the remaining 31[24] forming the training set; the second partition is obtained by fully random selection of sentences.[25] Moreover, evaluation is carried out according to two scenarios: one (+Morph) in which lemmas, parts of speech and morphological features are given, and one (-Morph) in which every annotation level has to be tagged from scratch.[26]

The accuracy of each model is calculated using `eval.py`,[27] an evaluation script provided by the UD project. As shown in Table 3, evaluations conducted on the random partition result into slightly higher average accuracy scores than those based on triplets[28] of consecutive *canti*: in the +Morph scenario, a difference of 0,16% is observed for UAS, whereas in the opposite -Morph scenario the improvement is more marked, but still minor, at 0,42% for UAS and 0,26% for LAS. The only exception regards LAS in the +Morph scenario, though the difference of 0,02% encountered there is negligible.

Consistently with our expectations, we also observe that parsing performed with prior assignment of the other annotation levels produces better results compared to the case where the parser has to handle all annotation levels simultaneously. Specifically, in the +Morph scenario the average of models trained on the random partition exhibits an improvement of 6,63% for UAS and 9,10% for LAS , and similarly models trained on consecutive *canti* show an improvement of 6,89% for UAS and 9,38% for LAS.

We can conclude that, on the one hand, sampling the dataset randomly or by selecting consecutive parts of the text does not seem to significantly affect performances, and this could point to the fact that, at least in this *cantica*, morphosyntactic phenomena are uniformly distributed across the text, as also standard deviation is very low. On the other hand, LAS and UAS metrics improve significantly when the text is already enriched with linguistic annotation. This allows us to have positive expectations with regard to the parsing of *Purgatorio* and *Paradiso*, *cantiche* for which lemmatization and morphosyntactic taggings are inherited from the conversion from DS.

5. Conclusions and future perspectives

Building a UD treebank for Dante's *Comedy* is the first step towards incorporating Old Italian among the languages of UD. This paper describes the development of the first part of this treebank, which consists of the first *cantica* of the *Comedy*, the *Inferno*.

We also present the results of an experiment of supervised automated dependency parsing using both as training and test sets data from the *Inferno*. We run this experiment to understand to what extent the process of syntactic annotation of the *Comedy*, which has been performed so far fully manually, can benefit from the results of the application of an NLP tool. Although the accuracy rates reported in the paper are fairly good (≈77% LAS), in the near future we will have to evaluate how and to what extent they will drop once a model trained and evaluated on the *Inferno* is applied to a different *cantica*. Should the accuracy rates drop heavily, even such a negative result might prove helpful in pointing out syntactic differences between the three *cantiche*. Moreover, the use of other parsers, based on different algorithms and resources (like embeddings), might lead to better and, most importantly, diverging results and errors.

As for annotation issues, we will suggest to introduce a specific subtype, e. g. `ellp`, in UD's documentation, so as to properly identify cases of ellipses, as they are not explicitly captured by the current annotation strategies mentioned in the paper, namely promotion and the use of the relation `orphan`: the former does not signal the presence of ellipsis, while the latter obscures the real dependency relations which are replaced by it. While adopting a subtype like `ellp` would make it possible to collect cases of ellipses, their resolution is up to the annotation of so-called enhanced dependencies, which are a kind of advanced annotation that augments dependency

[23] We actually note that, since the number of *canti*, 34, is not divisible by 3, one *canto* would be left out, and is instead aggregated to the last batch, which then consists of 4 consecutive *canti* (31, 32, 33, 34).
[24] Or 30; see fn. 23.
[25] Please refer to the GitHub page https://github.com/ClaudiaCorbe/Inferno_treebank for the data and detailed statistics on the partitions.
[26] Corresponding respectively to `–parse` and `–tag –parse` options for UDPipe; see https://ufal.mff.cuni.cz/udpipe/1/users-manual, §3.6.
[27] https://github.com/UniversalDependencies/tools/blob/master/eval.py.
[28] Or a quadruplet; see fn. 23.

labels to facilitate disambiguation.[29]

We plan to engage additional annotators with expertise in Old Italian to expedite the process of annotation of *Purgatorio* and *Paradiso*. Additionally, we intend to apply error detection processes (like, for instance, those described in [31]) to retrieve possible mistakes or inconsistencies in syntactic annotation.

Another task we intend to address is the extension of the UD documentation for Italian in order to make the validator[30] correctly deal with some peculiarities of Old Italian, like for instance enclitic adpositions (e.g., *meco* 'with me'), which require the introduction of the feature `Clitic=Yes` combined with the PoS tag ADP, currently permitted only with the PoS tag PRON.

Finally, we plan to include enhanced dependencies [31] in the UD treebank of Dante's *Comedy*, once the basic syntactic annotation of the entire work will be completed.

[29] https://universaldependencies.org/u/overview/enhanced-syntax.html.
[30] https://github.com/UniversalDependencies/tools/blob/master/validate.py
[31] https://universaldependencies.org/u/overview/enhanced-syntax.html

References

[1] M.-C. de Marneffe, C. D. Manning, J. Nivre, D. Zeman, Universal Dependencies, Computational Linguistics 47 (2021) 255–308. URL: https://direct.mit.edu/coli/article/47/2/255/98516/Universal-Dependencies. doi:10.1162/coli_a_00402.

[2] D. Zeman, et alii, Universal dependencies 2.12, 2023. URL: http://hdl.handle.net/11234/1-5150, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. Available at http://hdl.handle.net/11234/1-5150.

[3] F. M. Cecchini, R. Sprugnoli, G. Moretti, M. Passarotti, UDante: First Steps Towards the Universal Dependencies Treebank of Dante's Latin Works, in: J. Monti, F. Dell'Orletta, F. Tamburini (Eds.), Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020, Bologna, Italy, March 1–3 2021), Associazione italiana di linguistica computazionale (AILC), Accademia University Press, Turin, Italy, 2020, pp. 99–105. URL: http://ceur-ws.org/Vol-2769/paper_14.pdf.

[4] M. Passarotti, F. M. Cecchini, R. Sprugnoli, G. Moretti, *UDante*, Studi Danteschi LXXXVI (2022) 309–338.

[5] G. I. Ascoli, L'Italia dialettale, Archivio glottologico italiano VIII (1882–1885) 98–128. Available at https://archive.org/details/archivioglottolo08fireuoft/page/n5/mode/2up.

[6] L. Tomasin, Il caos e l'ordine, Piccola Biblioteca Einaudi, Giulio Einaudi, Turin, Italy, 2019.

[7] M. Dardano (Ed.), Sintassi dell'italiano antico, Lingue e Letterature Carocci, Carocci, Rome, Italy, 2013. URL: https://www.carocci.it/prodotto/sintassi-dellitaliano-antico.

[8] M. Dardano, G. Frenguelli (Eds.), SintAnt, Aracne, Rome, Italy, 2004.

[9] R. Tesi, Parametri sintattici per la definizione di "Italiano antico", in: M. Dardano, G. Frenguelli (Eds.), SintAnt. La sintassi dell'italiano antico, Aracne, Rome, Italy, 2004, pp. 425–444.

[10] G. Salvi, L. Renzi (Eds.), Grammatica dell'italiano antico, il Mulino, Bologna, Italy, 2010. URL: https://www.mulino.it/isbn/9788815134585.

[11] P. Manni, La lingua di Dante, Le vie della civiltà, il Mulino, Bologna, Italy, 2013.

[12] M. Tavoni, DanteSearch: il corpus delle opere volgari e latine di Dante lemmatizzate con marcatura grammaticale e sintattica, in: A. Cerbo, R. Mondola, A. Žabjek, C. D. Fiore (Eds.), Lectura Dantis 2002-2009. Omaggio a Vincenzo Placella per i suoi settanta anni, volume 2 (2004–2005), Il Torcoliere – Officine Grafico-Editoriali di Ateneo, Naples, Italy, 2011, pp. 583–608.

[13] P. G. Beltrami, Il Tesoro della Lingua Italiana delle Origini (TLIO), in: N. Maraschio, T. Poggi Salani, M. Bongi, M. Palmerini (Eds.), Italia linguistica anno mille. Italia linguistica anno duemila. Atti del XXXIV Congresso Internazionale di Studi della Società di Linguistica Italiana, Firenze 19–21 ottobre 2000, number 45 in Società di linguistica italiana, Bulzoni, Rome, Italy, 2003, pp. 695–698. URL: https://www.torrossa.com/it/catalog/preview/2280318. doi:10.1400/28371.

[14] P. D'Achille, M. Grossmann (Eds.), Per la storia della formazione delle parole in italiano, Quaderni della Rassegna, eighth ed., Franco Cesati, Florence, Italy, 2017. URL: https://www.francocesatieditore.com/catalogo/per-la-storia-della-formazione-delle-parole-in-italiano/.

[15] M. S. Micheli, CODIT. A new resource for the study of Italian from a diachronic perspective: Design and applications in the morphological field, Corpus 23 (2022). URL: https://journals.openedition.org/corpus/7306. doi:10.4000/corpus.7306.

[16] M. Favaro, E. Guadagnini, E. Sassolini, M. Biffi, S. Montemagni, Towards the Creation of a Diachronic Corpus for Italian: A Case Study on the GDLI Quotations, in: R. Sprugnoli, M. Passarotti (Eds.), Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA), European Language Resources

Association (ELRA), Marseille, France, 2022, pp. 94–100. URL: https://aclanthology.org/2022.lt4hala-1.13/.

[17] D. Alighieri, La Commedia secondo l'antica vulgata voll. I–IV, number 7 in Edizione nazionale delle Opere di Dante Alighieri a cura della Società Dantesca Italiana, Le Lettere, Florence, Italy, 1994. URL: https://www.lelettere.it/libro/9788871661483, editor: Giorgio Petrocchi.

[18] M. Tavoni, Allestimento, fruizione e prospettive di DanteSearch, in: E. Cresti, M. Moneglia (Eds.), Corpora e Studi Linguistici. Atti del LIV Congresso della Società di Linguistica Italiana (Online, 8-10 settembre 2021), number 6 in nuova serie, Officinaventuno, Milan, Italy, 2022, pp. 255–273. URL: https://www.societadilinguisticaitaliana.net/wp-content/uploads/2022/11/017_Tavoni_Atti_LIV_Congresso_SLI.pdf. doi:10.17469/O2106SLI000017.

[19] S. Gigli, La codifica sintattica della Commedia di Dante, in: M. D'Amico (Ed.), Sintassi dell'italiano antico e sintassi di Dante. Atti del seminario di studi (Pisa 15/16 ottobre 2011), Felici, Ghezzano (PI), Italy, 2015, pp. 81–95.

[20] L. Serianni, A. Castelvecchi, Grammatica italiana, Universitaria, second ed., UTET Università, Turin, Italy, 2006.

[21] M. Straka, J. Hajič, J. Straková, UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing, in: N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. O. Odijk, S. Piperidis (Eds.), Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), European Language Resources Association (ELRA), Portorož, Slovenia, 2016, pp. 4290–4297. URL: https://aclanthology.org/L16-1680.

[22] J. Heinecke, ConlluEditor: a fully graphical editor for Universal Dependencies treebank files, in: A. Rademaker, F. Tyers (Eds.), Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019), Association for Computational Linguistics (ACL), Paris, France, 2019, pp. 87–93. URL: https://aclanthology.org/W19-8010. doi:10.18653/v1/W19-8010.

[23] D. Alighieri, Inferno, number 613 in Oscar classici, Arnoldo Mondadori, Milan, Italy, 2005. Editor: Anna Maria Chiavacci Leonardi.

[24] D. Alighieri, Commedia. Inferno, number 1 in Opere, Carocci, Rome, Italy, 2007. Editor: Guglielmo Inglese.

[25] J. Havelka, Mathematical Properties of Dependency Trees and their Application to Natural Language Syntax, Ph.D. thesis, Univerzita Karlova – Matematicko-fyzikální fakulta, Prague, Czech Republic, 2007. URL: https://dspace.cuni.cz/handle/20.500.11956/12614?locale-attribute=en.

[26] J. Merchant, Ellipsis: A survey of analytical approaches, in: J. van Craenenbroek, T. Temmerman (Eds.), The Oxford Handbook of Ellipsis, Oxford Handbooks, Oxford University Press, Oxford, UK, 2019. URL: https://academic.oup.com/edited-volume/41718/chapter/353990361.

[27] A. Saab, Nominal Ellipsis, in: J. van Craenenbroek, T. Temmerman (Eds.), The Oxford Handbook of Ellipsis, Oxford Handbooks, Oxford University Press, Oxford, UK, 2019, pp. 526–561. URL: https://academic.oup.com/edited-volume/41718/chapter-abstract/353995808. doi:10.1093/oxfordhb/9780198712398.013.26.

[28] A. Lobke, W. Harwood, Predicate Ellipsis, in: J. van Craenenbroek, T. Temmerman (Eds.), The Oxford Handbook of Ellipsis, Oxford Handbooks, Oxford University Press, Oxford, UK, 2019, pp. 504–525. URL: https://academic.oup.com/edited-volume/41718/chapter-abstract/353995176.

[29] M. Straka, J. Straková, Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe, in: J. Hajič, D. Zeman (Eds.), Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Association for Computational Linguistics (ACL), Vancouver, BC, Canada, 2017, pp. 88–99. URL: https://aclanthology.org/K17-3009. doi:10.18653/v1/K17-3009.

[30] S. Buchholz, E. Marsi, CoNLL-X Shared Task on Multilingual Dependency Parsing, in: L. Màrquez, D. Klein (Eds.), Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X), Association for Computational Linguistics (ACL), New York City, NJ, USA, 2006, pp. 149–164. URL: https://aclanthology.org/W06-2920.

[31] M. Dickinson, Error Detection and Correction in Annotated Corpora, Ph.D. thesis, The Ohio State University, 2005. URL: https://sifnos.sfs.uni-tuebingen.de/decca/publications/dickinson-dissertation.html.

Towards an Italian Corpus for Implicit Object Completion

Agnese Daffara[1] and Elisabetta Jezek[1]

[1] University of Pavia, Corso Strada Nuova, 65, Pavia, Italy

Abstract

This study centers on the creation of an Italian corpus designed for the task of Implicit Object Completion. In this corpus, every sentence contains a token [MASK] denoting the position of the Object's head, along with the annotation of a Gold Standard filler word. The completion of the Object is conceived as a masked word task, theoretically executable by a BERT-based transformer model. In the next phase of the project, this task will be applied to a range of Italian language models, and their performance will be assessed. Overall, this project seeks to offer insights into the capabilities and constraints of such models in successfully completing Implicit Objects within various contexts.

Keywords

BERT, Implicit Object, masked word

1. Introduction

When coming across the verb-argument structure of a sentence, individuals have the cognitive ability to comprehend its meaning by forming a semantic representation of the situation in their minds. Even in cases where one argument is implicit, they are still capable of understanding the overall sense, thanks to the verb's inherent lexical meaning and the neighbouring words. The Distributional Hypothesis, as proposed by Harris (1954) and Firth (1951), suggests that it is possible to infer the meaning of a word purely on the basis of the context.

In the field of Natural Language Understanding, Artificial Intelligence must replicate this ability in order to reconstruct the scenario of the event, specifically identifying its semantic participants. Given the requirement for a computational model to fill in the missing information, we propose that this task can be conceived and construed as a masked word completion task, for which transformer-based technologies such as BERT (Devlin et al., 2019) have proven to be the most suitable.

This paper focuses on building an Italian corpus for this specific purpose while hinting at the same time at the forthcoming work of evaluation.

The corpus centers on verbs that exhibit an Optional Object, i.e. an Object that can be Implicit or Explicit. The ontological set of verbs on which the corpus is constructed is presented in section 4. Following these verbs' ambivalent possibility of expressing or implying the Argument, the corpus is divided into two datasets: on one side, an IMPLICIT dataset of sentences with Implicit Objects; on the other side, with a contrastive role, an EXPLICIT dataset of sentences containing Explicit Objects.

Our decision to create two different datasets is motivated by the idea of observing the differences in the performance of the models: do they perform better when the original Object is Implicit? This issue is grounded in the findings of a prior guide experiment conducted by Ye et al. (2020); according to their results, the model's performance would be notably improved when fine-tuned on an IMPLICIT dataset, because of the greater richness of contextual information available. We want to investigate if such observations can be generalized to our experiment.

Regarding the annotation of the masked word, the two datasets are treated differently. In the IMPLICIT dataset, we inserted two [MASK] tokens right after the verb or the adverb, and allowed the model to generate either a single Noun, as in Table 1, sentence 1., or, if not found, a Noun Phrase (NP) consisting of a Determiner plus a Noun. Further explanation of this possibility can be found in section 5. Furthermore, a Gold Standard (GS) Noun representing the optimal completion of the Object's head position was annotated aside each sentence, together with the type of omission (see section 2 for theoretical references).

In the EXPLICIT dataset, on the other hand, we removed the Explicit Object's nominal head, consisting of one word, and we annotated it as a GS. Two examples of annotation for sentences 1. and 2., belonging respectively to the IMPLICIT and the EXPLICIT dataset, are provided in Table 1.

1. Da quel 26 dicembre non vuole più bere [MASK][MASK] né lavarsi. 'Since that December 26, they no longer want to drink [MASK][MASK], nor wash themselves.'
2. Infilo la fetta velocemente nel sacchetto delle fragole e tiro un sospiro, bevo un [MASK] di caffè. 'I slide the slice quickly into the strawberry bag and let out a sigh, I drink a [MASK] of coffee.'

Table 1
Example of annotation on sentences 1. and 2.

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 – Dec 02, 2023, Venice, Italy.

agnese.daffara01@universitadipavia.it (A. Daffara); jezek@unipv.it (E. Jezek)

dataset	sent	GS_obj_head	omission_type
IMP.	1.	acqua 'water'	LD
EXP.	2.	sorso 'sip'	-

As referenced above, the corpus will allow us to undertake a selection of BERT Italian models and systematically evaluate their performance on the task of Implicit Object Completion, which we define in section 6.

We firmly believe that an annotated Italian dataset containing masked Optional Objects, their categorisation and their corresponding Gold Standard completions, as well as the subsequent experiment and evaluation, will greatly contribute to research endeavors in the field of NLP.

2. Related Work

Previous computational works have primarily focused on the task of Implicit Argument Detection rather than on the mere completion of a masked Object. SemEval 2010 task 10 (Ruppenhofer et al., 2009) introduced a variety of approaches aimed at detecting the semantic participants to the event, specifically identifying the Null Instantiations of the Arguments. The term Null Instantiation was introduced by Fillmore (1982) within the theory of Frame Semantics. In fact, most of the proposals relied on this theoretical background, adopting as a starting point the Framenet dataset and annotation.

While the rise of transformer-based models has brought significant improvement for this task, as shown, for example, by Zhang et al. (2020), it still remains an interesting and challenging issue.

For what concerns the Italian language, we identified a potential gap in the literature on the computational detection and processing of Implicit Arguments, probably due to the lack of annotated corpora designed for this task. It is, therefore, of utmost importance to investigate this topic and create new computational suitable resources.

Significant progress has been made in the training of BERT-based Italian models, including AlBERTo (Polignano et al., 2019), UmBERTo (Parisi et al., 2020), GilBERTo (Ravasio and Di Perna, 2020), and the distilled Italian version of DistilBERT called BERTino (Muffo and Bertino, 2020). Thanks to the availability of this generation of open-source BERT models, the masked word task has been applied to a variety of different linguistic and cognitive topics, such as the study of Agentivity and Telicity (Lombardi and Lenci, 2021) or connectives (Albertin et al., 2021). In particular, our study consistently builds upon the prior application of the masked word Task to the semantic topic of Logical Metonymy by Ye et al. (2020).

The existing linguistic literature has extensively explored the concept of Implicit Argument and the phenomenon of Argument omission in Italian. Notably, Cennamo (2017) proposed a meticulous comparative analysis of the parameters involved in this process. In our study, we adopt the notion of "Defaulting", first introduced by Pustejovsky (1995) and further refined by Jezek (2018). Following Fillmore's distinction between Definite and Indefinite Null Instantiation, we delineate Pragmatic Defaulting (PD) as the omission of the Object based on contextual cues and Lexical Defaulting (LD) as the omission of the Object licenced by the core meaning of the verb. Overall, it is undeniable that both the contextual cues and the semantics encoded into the verb contribute to the possibility of implying and reconstructing an Argument, and we believe it is necessary to consider this difference when studying Implicit Objects.

3. Data Preparation

As a first step towards the corpus preparation, we established a set of 30 verbs that allow for their Object to remain implied. We refer to this set with the term 'ontology' since it contains the basic verbal structures of reference for the building of the corpus. Our selection of such verbal structures draws upon the resource T-PAS (Jezek et al., 2014), a repository of Typed Predicate-Argument Structures (T-PAS) which was developed at the University of Pavia in collaboration with the Bruno Kessler Foundation in Trento (I) and the Masaryk University in Brno (CZ) by adopting a corpus-driven methodology.

Each pattern in T-PAS corresponds to a distinct contextual meaning of the verb (Predicate), plus the list of all the possible semantic participants to the event associated with that specific meaning (Arguments). Notably, T-PAS not only captures information concerning the syntactic structure but also provides insights into the semantic types of the Arguments. This resource is a valuable foundation for our data collection, as it also annotates (in round brackets) the potential for exhibiting an Implicit Argument for each structure. An example of three patterns displayed on the online T-PAS website for the verb 'bere' ('drink') (including a metonymic use) is given in Figure 1.

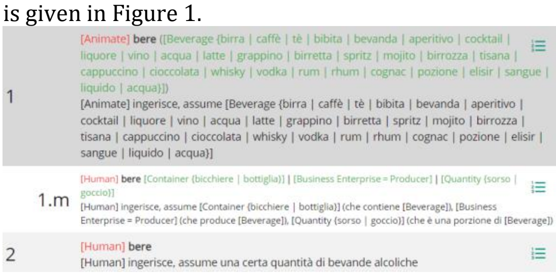

1 [Animate] bere ([Beverage {birra | caffè | tè | bibita | bevanda | aperitivo | cocktail | liquore | vino | acqua | latte | grappino | birretta | spritz | mojito | birrozza | tisana | cappuccino | cioccolata | whisky | vodka | rum | rhum | cognac | pozione | elisir | sangue | liquido | acqua}])
[Animate] ingerisce, assume [Beverage {birra | caffè | tè | bibita | bevanda | aperitivo | cocktail | liquore | vino | acqua | latte | grappino | birretta | spritz | mojito | birrozza | tisana | cappuccino | cioccolata | whisky | vodka | rum | rhum | cognac | pozione | elisir | sangue | liquido | acqua}]
1.m [Human] bere [Container {bicchiere | bottiglia}] | [Business Enterprise = Producer] | [Quantity {sorso | goccio}]
[Human] ingerisce, assume [Container {bicchiere | bottiglia}] (che contiene [Beverage]), [Business Enterprise = Producer] (che produce [Beverage]), [Quantity {sorso | goccio}] (che è una porzione di [Beverage])
2 [Human] bere
[Human] ingerisce, assume una certa quantità di bevande alcoliche

Figure 1. Example of three patterns for the verb 'bere' ('drink') in T-PAS.

From the comprehensive dataset of patterns available, we first identified the ones containing one or more Optional Arguments, using a simple RegEx match search to detect round brackets. Afterwards, we isolated 'fundamental' verbs (verbi fondamentali) according to the Nuovo Vocabolario di Base della Lingua Italiana (NVdB) (De Mauro, 2016). These particular verbs were chosen due to their presence in 90% of Italian texts, making them a suitable representative set for constructing an ontology.

We then conducted a cleansing process that excluded causatives, passives, and idiomatic expressions, as well as other multiword expressions or subpatterns with relatively infrequent occurrences and less common meanings. We constantly consulted the online version of the NVdB to decide whether a structure was fundamental or not. This cleansing process finally yielded a comprehensive list of 324 patterns with an Optional Argument, spanning across 213 distinct verb types.

We finally proceeded to further narrow down our focus, isolating the structures with an Optional Object. The final ontology comprehends 30 different verbs, corresponding to 60 T-PAS patterns and over 50 possible Object's Semantic Types, which represents a consistent variety. The detailed list is provided in Appendix A, whereas a summary of the quantities of verbs and patterns contained in T-PAS can be found in Table 2.

Table 2
Verb types and patterns in T-PAS

List	Verb types	Patterns
Tot.	1000	5392
Fundamental	492	
Fundamental with Optional Arg. (clean)	213	324
Fundamental with Optional Obj. (clean)	30	60

4. Data Collection

After the assessment of the ontology, our attention turned to collecting the sentences. The resource of reference is the T-PAS dataset, comprising 252,943 Manually Annotated Corpus Instances. All the sentences were selected from the It-Wac reduced corpus (Baroni and Kilgarriff, 2006) and annotated with the corresponding T-PAS number, denoting the specific semantic pattern being used in that sentence. For example, as shown in Figure 1, when the verb 'bere' ('drink') was used with a metonymic Object, like 'sorso' ('sip'), it was tagged with T-PAS number 1m, while, when it was found without an Object and implying an alcoholic drink, it was tagged with T-PAS number 2.

After isolating the T-PAS structures contained in our ontology, we proceeded to manually select the sentences for the corpus. We removed those with a Noun as an Object and preferred those with a linear order (the Object following the Noun). In our pursuit of a more extensive and diverse dataset, especially concerning the variety of Objects, we also conducted searches in the whole It-Wac reduced corpus through the Sketch Engine online platform (Kilgarriff et al., 2014). Eventually, 40 sentences were selected for each verb. The resulting 1200 sentences are divided into the two datasets, each containing 600 sentences, as illustrated in Table 3.

Table 3
Structure of the corpus

	IMP. dataset	EXP. dataset	Whole corpus
Verbs	30	30	30
Sentences for verb	20	20	40
Tot. sentences	600	600	1200

5. Annotation

The annotation process was handled differently for the IMPLICIT and the EXPLICIT dataset.

In the IMPLICIT dataset, the token [MASK] was manually inserted after the verb or the adverbial modifier, in order to signal the position to be filled by the model. However, we've observed that when the model encounters only one masked word, it tends to generate either mass Nouns or plural Nouns due to the lack of a Determiner. This presents a significant limitation in the evaluation process. The chosen approach involves the following steps: 1. Initially, annotate two [MASK] tokens to indicate the positions of the Determiner and the Noun. 2. Instruct the model to first look for a Noun for the second position. (3) If a Noun is not generated, proceed to search for a Determiner. (4) Generate a new sentence containing that Determiner, which will later be filled with a Noun.

As a following step in the annotation of the IMPLICIT dataset, the GS word constituting the optimal Object's head was manually inserted on the basis of the pragmatic context and the strength of the possible collocations. This value, achieved using the LogDice metric, can be obtained by querying Sketch Engine on the ItTenTen20 Italian corpus[1], as shown in Figure 2.

The last step of the IMPLICIT dataset's annotation regards the type of omission. As already mentioned, we adopted the classification proposed by Jezek (2018) following Pustejovsky (1995) between Pragmatic and Lexical Defaulting. This categorization serves as a valuable tool during the final evaluation of the model, enabling an assessment of its performance across different kinds of omission.

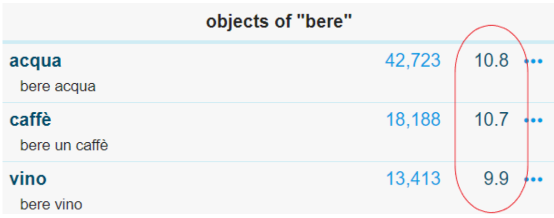


Figure 2: Collocation scores of verb 'bere' ('drink') plus Object on ItTenTen20.

For what concerns the EXPLICIT dataset, the Object's head was manually detected, removed and replaced by the token [MASK]. Subsequently, it was annotated aside the sentence. Note that by removing just a single word, these sentences retain their rich syntactic context, displaying the modifiers of the removed word. Such cues may improve the models' ability to detect the original filler. As the EXPLICIT dataset primarily has a contrastive function, we anticipate that comparing results from both datasets will help determine whether the model's output is closer to the original when it receives a significant amount of

[1] https://www.sketchengine.eu/ittenten-italian-corpus/

syntactic information or, conversely, when the context is semantically richer, as seen in Implicit Object sentences.

6. Task Definition

We define Implicit Object Completion as the task of substituting the masked Object in a sentence, previously marked with the token [MASK], with the most appropriate word or filler. When tested on each sentence, the transformer model is expected to produce the word that best fits the context of the sentence.

However, alternative outputs are possible, potentially encompassing other Parts of Speech. As an example, we employed the online demo of bert-base-italian-cased, made accessible by the MDZ Digital Library team (dbmdz) at the Bavarian State Library on Hugging Face[2]. The model generated the most probable candidates for sentence 1.. Predictably, the first output was the punctuation sign "," and the expected Nouns were found in lower positions, as depicted in Figure 3. In order to mitigate this issue and ensure more accurate results, during the model's interrogation, we implemented a two-step filter that isolates Nouns. In particular, we exclusively considered the Noun with the highest probability score.

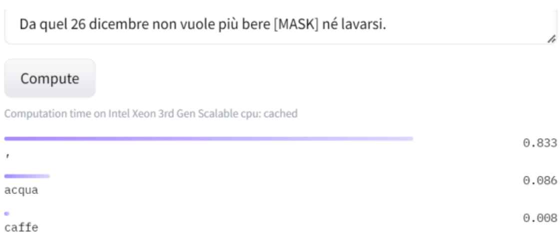


Figure 3: Outputs of the model bert-base-italian-cased for sentence 1.

7. Evaluation

An issue in the design of the task is the possibility of getting a synonym or a word that is only partially correct or doesn't perfectly align with the Gold Standard.

The Theory of Prototypes, as proposed firstly by Rosch (1973), posits that within a semantic category, certain members are more representative of the category's core meaning. In contrast, less central members demonstrate greater variability and may deviate further from the core concept. By taking into consideration both the Theory of Prototypes and the Distributional Hypothesis (cited in section 1), during the evaluation phase, we will systematically calculate the similarity score (sim) between the output word and the Gold Standard completion, corresponding to the cosine between the two word vectors. This value will be obtained by running the Python library SpaCy (Honnibal and Montani, 2017) on the Italian model it_core_news_lg, a large language model with a size of 541 MB[3].

An example of output of the model bert-base-italian-xxl-cased (bert-base-xxl), the bigger version of bert-base-italian-cased from dbmdz, and its relative annotation for sentence 1. and 2. is shown in Table 4.

Table 4
Example of sentence 1. and 2. bert-base-xxl outputs

data set	sent	GS_obj_head	bert-base-xxl	sim
IMP.	1.	acqua 'water'	acqua 'water'	1.0
EXP.	2.	sorso 'sip'	bicchiere 'cup'	0.62

With these annotation parameters, we aim to extend our linguistic analysis beyond the model's ability to complete the cloze test by providing the right word. Instead, we will also investigate the model's capability to effectively cluster words within the same domain.

8. Discussion and Results

This paper discusses the ongoing construction of a corpus specifically tailored for the task of Implicit Object Completion. This resource contains sentences exhibiting both Implicit and Explicit Objects, thus enabling the assessment of two distinct datasets that will be treated separately.

In the IMPLICIT dataset, the position for the Noun/NP is signaled by manually inserting two tokens [MASK] right after the verb or the adverb. The GS Object's head is manually added, considering both the context of the sentence and the general strength of the verb-Object collocation, which can be quantified through the typicality score on the ItTenTen20 corpus (Jakubíček et al., 2013) using Sketch Engine. Additionally, in the case of Implicit Objects, we provide information about the type of omission, which may either depend on the contextual cues (Pragmatic Defaulting) or the lexical verbal root (Lexical Defaulting). On the other hand, the annotations within the EXPLICIT dataset include the manual identification of the Object's nominal head, which is substituted with the token [MASK] and annotated aside the sentence.

The forthcoming second phase of our project involves an in-depth analysis of the outputs generated by a selection of the primary BERT Italian models. As a metric for the evaluation, we will adopt cosine similarity. This value measures the similarity between the output word provided by the model and the GS word, thus measuring the ability of the model to generate a filler which is semantically close to the original. As an example of a comparison between two models, consider the results of bert-base-xxl and umberto-commoncrawl-cased-v1 (UmBERTo), on sentence 2., which are reported in Table 5.

---

[2] https://huggingface.co/dbmdz/bert-base-italian-cased

[3] https://github.com/explosion/spacy-models/releases/tag/it_core_news_lg-3.7.0

Table 5
Comparison of two models' outputs on sentences 2. from the EXPLICIT dataset

sent	GS_obj_head	bert-base-xxl	sim	UmBERTo	sim
2.	sorso 'sip'	bicchiere 'cup'	0.62	cucchiaino 'teaspoon'	0.4

Bert-base-xxl returns a slightly higher score, as the vectors of 'bicchiere' ('cup') and 'sorso' ('sip') have an higher cosine similarity than those of 'cucchiaino' ('teaspoon') and 'sorso' ('sip'). Although both the models fail to understand the exact word and categorize the filler as a [CONTAINER] rather than a [QUANTITY], both the results are satisfactory and plausible. More results for the EXPLICIT dataset can be found in Appendix B.

In conclusion, we expect the results to raise a number of theoretical questions and possible investigations. By conducting this analyses, we will compare the models' performance on a novel topic and investigate their ability to identify the semantic category of the Objects, while effectively clustering words within the same domain. In addition, the annotation of the type of omission will allow further insights on the importance of the context in reconstructing Implicit Objects.

Acknowledgements

The authors gratefully acknowledge the contributions of the anonymous reviewer for CLiC-it 2023. Their insights and feedback have significantly enhanced the quality of this paper.

Upon completion of the corpus, the complete dataset will be hosted on a public GitHub repository in accordance with the FAIR principles.

References

G. Albertin, A. Miaschi, D. Brunato, On the Role of Textual Connectives in Sentence Comprehension: A New Dataset for Italian, in: Proceedings of the Eighth Italian Conference on Computational Linguistics CLiC-it 2021, Milano, 2021.

M. Baroni and A. Kilgarriff, Large Linguistically-Processed Web Corpora for Multiple Languages, Demonstrations (2006) 87-90.

M. Cennamo, Object omission and the semantics of predicates in Italian in a comparative perspective, In: L. Hellan, A.L. Malchukov, M. Cennamo (eds.), Contrastive Studies in Verbal Valency, Benjamins, Amsterdam, 2017, pp. 251-273.

T. De Mauro (ed.), Il Nuovo vocabolario di base della lingua italiana, Internazionale, 2016. URL: https://dizionario.internazionale.it/

J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171-4186.

C. J. Fillmore, Frame semantics. In Linguistics in the Morning Calm, Hanshin Publishing Co, Seoul, 1982, pp. 111-137.

J.R. Firth, Modes of Meaning, Papers in Linguistics 1934-1951 (1951) 190-215.

Z.S. Harris, Distributional Structure, word 10, 2-3 (1954) 146-162.

M. Honnibal, I. Montani, SpaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing, 2017. URL: https://spacy.io/

M. Jakubíček, A. Kilgarriff, V. Kovář, P. Rychlý, V. Suchomel, The TenTen Corpus Family, 2013. URL: https://www.sketchengine.eu/ittenten-italian-corpus/

E. Jezek, Partecipanti impliciti nella struttura argomentale dei verbi, In: S. Dallabrida, P. Cordin (eds.), La Grammatica delle Valenze, Franco Cesati, Firenze, 2018, pp. 55-71.

E. Jezek, B. Magnini, A. Feltracco, A. Bianchini, O. Popescu, T-PAS; A resource of Typed Predicate Argument Structures for linguistic analysis and semantic processing, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), European Language Resources Association (ELRA), Reykjavik, Iceland. 2014, pp. 890–895. URL: https://tpas.sketchengine.eu/

A. Kilgarriff, V. Baisa, J. Bušta, M. Jakubíček, V. Kovář, J. Michelfeit, P .Rychlý, V. Suchomel, The Sketch Engine: ten years on, Lexicography, 1 (2014) 7-36 URL: https://www.sketchengine.eu/

A. Lombardi, A. Lenci, Agentività e telicità in GilBERTo: implicazioni cognitive, in: Proceedings of the Eighth Italian Conference on Computational Linguistics CLiC-it 2021, Milano, 2021.

M. Muffo, E. Bertino, BERTino: an Italian DistilBERT model, in: Proceedings of the Seventh Italian Conference on Computational Linguistics CLiC-it 2020, Bologna, 2020.
L. Parisi, S. Francia, P. Magnani, Umberto: An Italian language model trained with whole word masking, 2020. URL: https://github.com/musixmatchresearch/umberto

M. Polignano, P. Basile, M. De Gemmis, G. Semeraro, V. Basile, AlBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets, in: Proceedings of the Sixth Italian

Conference on Computational Linguistics CLiC-it 2019, Bari, 2019.

J. Pustejovsky, The Generative Lexicon. Cambridge, MA. MIT Press, 1995.

G. Ravasio, L. Di Perna, GilBERTo: An Italian pretrained language model based on RoBERTa. URL: https://github.com/idb-ita/GilBERTo

E. Rosch, Cognitive representations of semantic categories, Journal of Experimental Psychology: General, 104,3 (1975) 192-233.

J. Ruppenhofer, C. Sporleder, R. Morante, C. Baker, M. Palmer, SemEval-2010 Task 10: Linking Events and Their Participants in Discourse, in: Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009), Association for Computational Linguistics, Boulder, Colorado, 2009, pp. 106-111.

B. Ye, J. Tu, E. Jezek, J. Pustejovsky, Interpreting Logical Metonymy through Dense Paraphrasing, in: Proceedings of the Annual Meeting of the Cognitive Science Society, 44, 2022.

Z. Zhang, X. Kong, Z. Liu, X. Ma, E. Hovy, A Two-Step Approach for Implicit Event Argument Detection, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 7479–74.

Appendix A: Ontology of selected verbs and patterns from T-PAS

<table>
<tr><th>verb</th><th>n.</th><th>pattern</th></tr>
<tr><td>ascoltare</td><td>1
1m</td><td>[Human] ascoltare ([Sound] | [Musical Performance {concerto}])
[Human1] ascoltare ([Sound Maker] | [Medium] | [Musical Composition] | [Tv Program] | [Event] | [Part of Language] | [Speech Act] | [Narrative] | [Human2 = Singer | Musician] | [Human Group = Band])</td></tr>
<tr><td>attendere</td><td>1
1m</td><td>[Human] | [Institution] attendere ([Event]) | (che [Event])
[Human] | [Institution] attendere ([Human2] | [Vehicle] | [Time Point {data}] | [Document {visto | passaporto}])</td></tr>
<tr><td>bere</td><td>1

1m

2</td><td>[Animate] bere ([Beverage {birra | caffè | tè | bibita | bevanda | aperitivo | cocktail | liquore | vino | acqua | latte | grappino | birretta | spritz | mojito | birrozza | tisana | cappuccino | cioccolata | whisky | vodka | rum | rhum | cognac | pozione | elisir | sangue | liquido | acqua}])
[Human] bere ([Container {bicchiere | bottiglia}] | [Business Enterprise = Producer] | [Quantity {sorso | goccio}])
[Human] bere ([Alcool])</td></tr>
<tr><td>cantare</td><td>1
2
3</td><td>[Human] cantare
[Human] cantare
[Human] cantare ([Musical Composition {canzone | canto | inno | brano | testo | salmo}])</td></tr>
<tr><td>chiamare</td><td>6
6m</td><td>[Human1] chiamare ([Human2] | [Institution {polizia}])
[Human] chiamare ([Number] | [Device {telefono}] | [Location {call center}] | [Vehicle {ambulanza}])</td></tr>
<tr><td>combattere</td><td>2
3</td><td>[Human] combattere ([War {guerra | battaglia}])
[Human1] | [Human Group1] combatte ([War]) (con|contro [Human2] | con|contro [Human Group2])</td></tr>
<tr><td>condurre</td><td>5</td><td>[Human] condurre ([TV Program])</td></tr>
<tr><td>consumare</td><td>2</td><td>[Human] | [Human Group] | [Machine] | [Device] consumare ([Energy] | [Gas] | [Inanimate])</td></tr>
<tr><td>correre</td><td>3</td><td>[Human = Runner | Pilot] correre ([Competition {maratona | palio | rally}])</td></tr>
<tr><td>cucinare</td><td>1
2</td><td>[Human] cucinare ([Food] | [Meal {pranzo | cena}])
[Human] cucinare</td></tr>
<tr><td>dirigere</td><td>2
3</td><td>[Human] dirigere ([Musical Performance {concerto}])
[Human = Director] dirigere ([Movie])</td></tr>
<tr><td>disegnare</td><td>1
2
3
4</td><td>[Human] disegnare ([Image] | [Physical Entity])
[Human] disegnare ([Inanimate])
[Human] disegnare ([Document {fumetto | comics | copertina}])
[Human] disegnare</td></tr>
<tr><td>fumare</td><td>1
2</td><td>[Human] fumare ([Drug {sigaretta | pipa | sigaro | marijuana}])
[Human] fumare</td></tr>
<tr><td>giocare</td><td>3
4</td><td>[Human] giocare
[Human] | [Human Group = Team] giocare ([Competition {partita}] | {mano | set | stagione | tempo})</td></tr>
<tr><td>guadagnare</td><td>1</td><td>[Human] guadagnare ([Money])</td></tr>
<tr><td>guidare</td><td>1</td><td>[Human] guidare ([Road Vehicle])</td></tr>
<tr><td>leggere</td><td>2
3
8</td><td>[Human] leggere
[Human] leggere ([Document])
[Human1] leggere ([Document]) (a [Human2])</td></tr>
</table>

mangiare	1 2 3	[Human] mangiare ([Food {cibo | carne | pane | uovo | pizza | panino | gelato | biscotto | torta | bistecca | hamburger | salsiccia | salame | polpetta | frutta | mela | verdura | banana | riso | patata | carota | formaggio | minestra | insalata | polenta | zuppa | antipasto | spaghetto | pasta | patatina | panettone | brioche | piadina | cornetto | focaccia | pasticcino | pappa | pasto | biada}]) [Human] mangiare ([Food] {cibo}) [Human] mangiare ([Meal])
ordinare	1 2	[Human] ordinare ([Artifact]) [Human] ordinare ([Food] | [Beverage] | [Meal])
pagare	4	[Human] pagare ([Abstract Entity {conseguenza | debito | errore}])
perdere	7	[Human] | [Human Group] perdere ([Competition])
pregare	1 2 4	[Human] pregare ([Deity]) [Human1] | [Institution] pregare ([Deity]) (per [Human2]) [Human1] pregare ([Human2]) di [Activity)
preoccupare	2	[Anything] preoccupare ([Human])
provare	5	[Human = Artist] | [Human Group = Artist] provare ([Artwork])
respirare	1	[Animate] respirare ([Vapor])
scrivere	1 2 4 6 7	[Human] scrivere [Human] scrivere ([Part of Language]) [Human] scrivere ([Document]) [Human] scrivere ([Document]) (a|per [Human2]) [Human = Writer] scrivere
servire	6 7	[Human] servire ([Food] | [Meal]) (Manner) [Human1 = Waiter] servire ([Food] | [Meal]) a [Human2 = Customer])
suonare	3 2 1 5	[Human] suonare ([Musical Instrument]) [Human] suonare [Human = Artist] suonare ([Musical Composition {canzone | brano | pezzo | concerto}] | {musica}) [Human] suonare ({il campanello} | {il citofono}) | (alla {porta})
tirare	3	[Human = Football Player] tirare ([Ball])
vincere	1	[Human] | [Human Group] vincere ([Activity {gara | competizione | festival | elezioni}] | [War])

Appendix B: Example of results for the EXPLICIT dataset

The following table reports an example of the outputs of two models, italian-BERT-xxl-cased and UmBERTo. The models were tested on the 20 sentences from the EXPLICIT dataset with the verb 'bere' ('drink'). The column 'sent' displays the sentence with the masked word, corresponding to the Object's NP's head. The removed word is shown in the column 'GS_obj_head'. The columns 'bert-base-xxl' and 'UmBERTo' report the outputs of the models. The similarity score between the output word and the GS word is shown in the columns 'sim'.

sent	GS_obj_head	bert-base-xxl	sim	UmBERTo	sim
Tutto meno che restare a guardare la televisione a **bere** [MASK] e divorare patatine.	[birra]	birra	1	birra	1
Giganti americani del valore di Theodore Dreiser, Ernest Hemingway e Thornton Wilder, quando erano stanchi della routine andavano nei locali a **bere** [MASK] di whisky e a sentire la grande musica per ricercare la giusta ispirazione.	[litri]	litri	1	fiumi	0,14
Grazie anche per avermi fatto **bere** l' [MASK] di barbabietola e per avermi fatto svegliare tutte le mattine alle 6 per prendere la pappa reale...	[estratto]	acqua	0,36	acqua	0,36
Bere due [MASK] di latte di soia o mangiare una tazza di tofu è responsabile di livelli ematici di Isoflavoni che possono essere 500 o 1000 volte più elevati dei normali livelli di Estrogeni nelle donne.	[tazze]	bicchieri	0,64	bicchieri	0,64
Da circa 3 settimane Federico ha cominciato a **bere** [MASK] parzialmente scremato alta qualità e sin dai primi giorni ha mostrato di gradire il nuovo alimento.	[latte]	latte	1	latte	1
La disposizione potrebbe essere utile anche nelle altre stagioni dell'anno, specie se si vieta ai minori di **bere** [MASK].	[alcool]	alcolici	0,69	alcolici	0,69
Bevi ogni giorno [MASK] in abbondanza è infatti la quinta regola della sana alimentazione, che invita a **bere**, ma rigorosamente acqua e non altre bevande, frequentemente e in piccole quantità.	[acqua]	acqua	1	acqua	1
Per il pubblico in generale e per i giovani studenti c'è come al solito il padiglione 9 aperto gratuitamente in cui si può fare shopping, **bere** del [MASK] e rilassarsi oppure informarsi su ciò che accade nella fiera.	[vino]	vino	1	caffè	0,54
Pruneddu, che forse aveva bevuto qualche [MASK] di troppo prima di affacciarsi sulla porta del bar, non si è accorto che il proprietario e le altre persone presenti avevano organizzato una castagnata per stare insieme a **bere** un po'di vino.	[bicchiere]	bicchiere	1	bicchiere	1
Ma io non bevo [MASK] e gioco a freccette mentre dico parolacce.	[alcolici]	alcolici	1	alcolici	1
Davanti al Castello c'è il Ritz, dove Mordecai e Florence spesso andavano a **bere** un [MASK].	[drink]	caffè	0,61	caffè	0,61
E poi pensa un po' che GiPo ormai non può più dir niente perchè **ha bevuto** la [MASK] ed è morto.	[cicuta]	birra	0,33	birra	0,33
Dopo **aver** recuperato un maglione e **bevuto** un buon [MASK] al cardamomo, rientriamo in chiesa per l'ora del silenzio.	[caffè]	caffè	1	tè	0,77
Continuai a **bere** in silenzio il [MASK], mentre il sole che tramontava tingeva di rosso il cielo.	[the]	vino	-0,16	vino	-0,16
Gli uruguayani, come gli argentini, **bevono** moltissimo [MASK], un the fatto con le foglie secche della pianta omonima, sorseggiato da una piccola zucca attraverso una cannuccia di metallo, la bombilla.	[mate]	tè	0,34	caffè	0,32
La vita umana andrebbe rispettata, ma non sentirti mai in colpa di	[sangue]	cibo	0,39	sangue	1

bere [MASK] umano: è una cosa naturale.					
Infilo la fetta velocemente nel sacchetto delle fragole e tiro un sospiro, **bevo** un [MASK] di caffè.	[sorso]	bicchiere	0,62	cucchiaino	0,4
Questi giovani, ci scommetterei, **han bevuto** [MASK] ed ascoltato musica rock come gli altri coetanei, cosa é scattato ad un certo punto nel loro animo per un totale stravolgimento e per abbracciare un'ideologia perversa?	[coca cola]	birra	0,6	birra	0,6
Un turista che chiede un caffè in tazza, molto lungo e con latte - dice il barman Umberto - dissimula la voglia di **bere** un [MASK] e pagarlo come caffè ristretto.	[cappuccino]	caffè	0,69	caffè	0,69
Il cubetto grande è molto richiesto soprattutto sul mercato spagnolo; nei locali il consumatore vuole **bere** un [MASK] in un bicchiere grande (generalmente un tumbler alto) e gradisce che lo stesso gli venga presentato colmo di distillato.	[whisky]	cocktail	0,54	cocktail	0,54

Linking the Corpus CLaSSES to the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin

Irene De Felice[1,*], Lucia Tamponi[2], Federica Iurescia[3] and Marco Passarotti[3]

[1]Università di Genova, Italy
[2]Università di Pisa, Italy
[3]Università Cattolica del Sacro Cuore, Milan, Italy

Abstract
In this paper, we describe the process of linking the corpus CLaSSES (which collects non-literary Latin texts of different periods and places) to the LiLa Knowledge Base of linguistic resources for Latin made interoperable through their publication as Linked Data. The paper details the RDF modeling of the (meta)data provided by CLaSSES and presents three queries on data from different resources that interact in LiLa.

Keywords
Latin, Textual resources, Linguistic Linked Open Data

1. Introduction

The Latin language shows a large diversity, in the light of its wide usage in terms both of diachrony (spanning across two millennia) and diatopy (all over Europe and beyond). Such diversity is mirrored in the set of linguistic resources currently available for Latin, ranging from collections of literary texts of the Classical era,[1] to corpora of documentary texts of the Medieval times,[2] dictionaries,[3] and glossaries.[4]

Like for many other languages, one limitation that affects the wealth of resources for Latin is their sparseness, which prevents the full exploitation of the data they provide. The LiLa Knowledge Base was built to overcome such limitation, making distributed resources for Latin interact through their publication as Linked Data, by using a set of commonly used vocabularies provided by ontologies for the representation of linguistic (meta)data.

Among the resources interlinked in LiLa is the CLaSSES corpus, which enhances the set of lexical and textual data made interoperable by the Knowledge Base with a peculiar kind of non-literary Latin texts (such as inscriptions, writing tablets, and letters) written in different periods and provinces of the Roman Empire, thus contributing to extend the coverage of LiLa with a typology of texts not present so far in the Knowledge Base.

This paper details the process of linking CLaSSES to LiLa, and is organized as follows. Section 2 presents the corpus CLaSSES. Section 3 describes the LiLa Knowledge Base. Section 4 discusses the modeling and the linking of CLaSSES into LiLa. Section 5 reports three examples of queries that exploit the interoperability of CLaSSES with other resources in LiLa. Finally, Section 6 provides some conclusions on the results of the linking, and outlines directions of future work.

2. CLaSSES

CLaSSES (*Corpus for Latin Sociolinguistic Studies on Epigraphic textS*) is a digital resource created by the Laboratory of Phonetics and Phonology at Pisa University. Freely accessible on the internet,[5] it consists of over 3,400 non-literary Latin texts such as inscriptions, private letters, ink tablets, ostraka and papyri from various periods (6th century BCE to 6th century CE) and regions of the Roman Empire. The goal of CLaSSES is to use non-literary texts that exhibit (ortho-)graphic variants as a source to study the sociolinguistic variation of Latin [3, 4]. The identification of these spelling variants is the most crucial aspect of the corpus: words like *dedet*

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
*Corresponding author.
irene.defelice@edu.unige.it (I. De Felice); lucia.tamponi@fileli.unipi.it (L. Tamponi); federica.iurescia@unicatt.it (F. Iurescia); marco.passarotti@unicatt.it (M. Passarotti)

Although this paper was conceived and discussed jointly by the authors, solely for academic purposes scientific responsibility is to be divided up as follows: I. De Felice wrote Sections 2 and 4.2; L. Tamponi Sections 5, 5.1, 5.2, 5.3; F. Iurescia wrote Section 4.1; M. Passarotti wrote Sections 1 and 3. Section 6 is to be attributed to all authors.

[1]Such as the LASLA corpus: https://www.lasla.uliege.be/cms/c_8570411/fr/lasla-textes-latins.
[2]Such as the corpus of Computational Historical Semantics: https://www.comphistsem.org/home.html.
[3]Such as the bilingual Latin-English dictionary curated by Ch. T. Lewis and Ch. Short [1].
[4]Such as the Medieval Latin *Glossarium Mediae et Infimae Latinitatis* by du Cange [2].
[5]http://classes-latin-linguistics.fileli.unipi.it.

(CIL-I^2-9-26) and *Vivia* (ILLRP-S-99-8) are categorized as "non-classical" forms in comparison to the standard spelling of Classical Latin, which would be *dedit* and *Vibia* respectively. CLaSSES is divided into four sections based on the place of provenance of the texts: Rome and Italy, Roman Britain, Sardinia, Egypt and Eastern Mediterranean. The database includes 3,415 texts, which were first automatically tokenized, resulting in 46,888 tokens. Then, expert annotators lemmatized the entire corpus manually, given the high number of incomplete and misspelt words that cannot be easily processed by automatic tools. They also provided a meta-linguistic and extra-linguistic annotation, including additional information about each document (place of provenance, dating, text type, author/addressee) and about each token of the corpus (graphic form, language). Finally, the linguistic annotation identifies non-classical variants and classifies them according to the variation phenomena [5, 6].

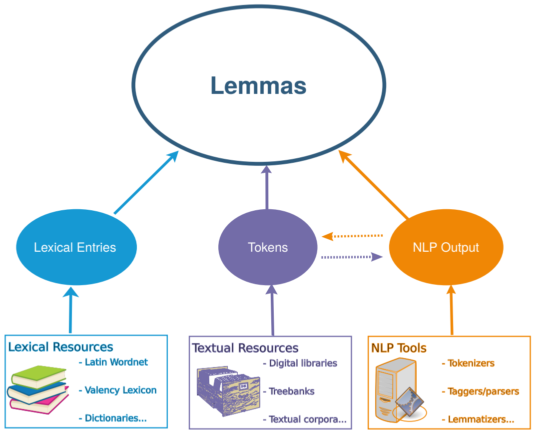


Figure 1: The architecture of LiLa.

3. The LiLa Knowledge Base

The aim of the "LiLa - Linking Latin" ERC project (2018-2023)[6] was to reach interoperability between the wealth of existing lexical and textual resources that have been developed in the last decades for Latin. One of the main problems that LiLa solved is the fact that such resources and tools are often characterized by different conceptual and structural models, which makes it difficult for them to interact with one another.

To this goal, LiLa undertook the creation of an open-ended Knowledge Base, following the principles of the Linked Data paradigm.[7] All content involved or referenced in the linguistic resources connected in LiLa is made unambiguously findable and accessible by assigning an HTTP Uniform Resource Identifier (URI) to each data point. Data reusability and interoperability between resources are achieved by establishing links between different URIs and by using web standards such as: [a] the RDF data model, which is based on triples: (i) a predicate-property connects (ii) a subject (a resource) with (iii) its object (another resource, or a literal) [7]; and [b] SPARQL, a query language specifically devised for RDF data.

Furthermore, the LiLa Knowledge Base makes reference to classes and properties of already existing ontologies to model the relevant information. The main ones are POWLA for corpus data [8], OLiA for linguistic annotation [9], and Ontolex-Lemon for lexical data [10, 11].

Within this framework, LiLa uses the lemma as the most productive interface between lexical resources, annotated corpora and NLP tools. Consequently, the architecture of the LiLa Knowledge Base is highly lexically based (Figure 1), grounding on a simple, but effective assumption that strikes a good balance between feasibility and granularity: textual resources are made of (occurrences of) words ("tokens"), lexical resources describe properties of words (in "lexical entries"), and NLP tools process words (producing "NLP outputs").[8]

The core of the Knowledge Base is the so-called Lemma Bank,[9] a collection of about 200,000 Latin lemmas taken from the database of the morphological analyzer LEMLAT [12]. Interoperability is achieved by linking all those entries in lexical resources and tokens in corpora that point to the same lemma.

4. CLaSSES into LiLa

4.1. Modeling (Meta)data

The Lemma Bank of the LiLa Knowledge Base is modeled as a collection of Lexical Forms of Ontolex-Lemon. Lexical Forms are the inflected forms of Lexical Entries and are assigned one, or more graphical variants (`ontolex:writtenRep`).[10] One of the Lexical Forms of a Lexical Entry is linked to the latter by the property `ontolex:canonicalForm`, to model that it is the form that is canonically chosen to represent the entire lexical entry, i.e., the lemma. As a consequence, the Lemma Bank is not a lexical resource (as it does not contain Lexical Entries), rather it is a collection of Ontolex-Lemon Lexical Forms that can be used as Canonical Forms in the re-

[6] https://lila-erc.eu/.
[7] https://www.w3.org/DesignIssues/LinkedData.html.

[8] In Figure 1 the arrows going from and to the node for "NLP Output" represent the fact that tokens that are the output of a specific NLP tool (a tokenizer) become the input of further tools (like, for instance, a syntactic parser).
[9] http://lila-erc.eu/lodview/data/id/lemma/LemmaBank.
[10] http://www.w3.org/ns/lemon/ontolex#writtenRep.

sources for Latin to be interlinked in the LiLa Knowledge Base.

In particular, textual resources are connected to the Lemma Bank through the property `lila:hasLemma`,[11] which links a token in a corpus with its lemma in the Lemma Bank. In LiLa, textual resources are modeled as objects of the type Corpus from the POWLA ontology .[12] Each Corpus includes one, or more `powla:Document`,[13] which are the parts in which the corpus is divided, like for instance the different texts that it contains, or its sections. In the case of the Corpus entitled CLaSSES, there are 10 documents, corresponding to as many sections of the resource.[14] Every document of CLaSSES is assigned two layers, namely (1) a Document Layer, which collects all the tokens of a section, and (2) a Citation Layer, which records the full citation path of each token of a section.

For instance, Figure 2 shows the modeling of one token from CLaSSES. The token (*sacra*) is linked to its lemma in the Lemma Bank (*sacer*) by the `lila:hasLemma` property, and to the Document Layer by the `POWLA:hasLayer` property.[15] The properties `lila:isLayer`,[16] `lila:hasCitSubUnit`[17] and `POWLA:hasChild`[18] link the Citation Layer to the token. In the example of Figure 2, the token *sacra* occurs in the inscription number 27 of volume S of the Document entitled *Inscriptiones latinae liberae rei publicae*, to which both its Document and Citation Layers are linked through the property `POWLA:hasDocument`.[19]

4.2. Linking Process

Out of the 46,888 tokens of CLaSSES, only those that are assigned a lemma are linked to the Lemma Bank of LiLa. Around 14k tokens of CLaSSES are not lemmatized due to the fragmentary nature of the texts contained therein. By exploiting the original lemmatization of the corpus, the automatic linking of the tokens of CLaSSES resulted in the following three output categories.

1. Perfect match (or one-to-one lemma; 25,279 items), i.e. whenever the lemma-PoS couple in CLaSSES was linked to one single lemma-PoS couple in the LiLa Lemma Bank. For such cases, we conducted an evaluation of the mapping on 10% of the couples. The data were randomly selected; to ensure that the sample was representative, the original PoS distribution was maintained.

   In CLaSSES, 3,490 different couples are recorded, thus the evaluation was conducted on 349 couples. Only 7 errors were found, all due to a wrong PoS tagging in the source data that caused a mapping error. Thus, the rate error is very low, i.e., 2%.
2. No match (or one-to-zero lemma; 5,366 items), i.e. when the lemma in CLaSSES was not associated with any lemma in LiLa. In this case, with the addition of the new lemma in LiLa we have enriched the Lemma Bank. Proper names are the category more affected, since inscriptions typically feature a wide range of anthroponyms which can identify the committee of the text (e.g., in public texts), the honorand (e.g., in sacred inscriptions) or the name of the dead on epitaphs [13]. In addition, given the wide geographical extension of our corpus, CLaSSES features local proper names typical of specific areas (e.g., Sardinia, or Roman Britain) that do not occur easily in Classical texts; an example from Sardinia [14, 15] is *Scribonissa* in ANRW-B61-6 [15, p. 45]. A few lemmas pertaining to other parts of speech were also added to the Lemma Bank, consisting mainly of hapax, like *ansata* in BTT-196-47 (lemma *ansatus* 'provided with handles'),[20] *infrascribo* in CEL-I-232-8 (lemma *infrascribo* 'to write lower down'),[21] *internumero* in BTT-645-48 (lemma *internumero* 'to reckon among other things').[22]
3. Ambiguous match (one-to-many lemmas; 1,503 items), i.e. when the lemma in CLaSSES was associated with several possible lemmas in LiLa. In most cases, the correct lemma between two or more possible ones was identified manually by a

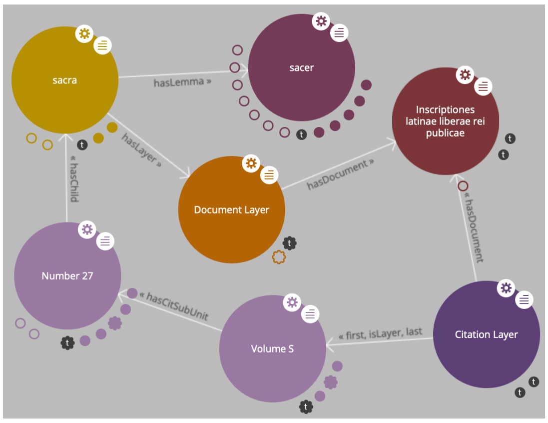


Figure 2: A token of CLaSSES in LiLa.

[11] https://lila-erc.eu/ontologies/lila/hasLemma.
[12] http://purl.org/powla/powla.owl#Corpus.
[13] http://purl.org/powla/powla.owl#Document.
[14] http://lila-erc.eu/data/corpora/CLaSSES/id/corpus.
[15] http://purl.org/powla/powla.owl#hasLayer.
[16] https://lila-erc.eu/lodview/ontologies/lila_corpora/isLayer.
[17] https://lila-erc.eu/lodview/ontologies/lila_corpora/hasCitSubUnit.
[18] http://purl.org/powla/powla.owl#hasChild.
[19] http://purl.org/powla/powla.owl#hasDocument.
[20] http://lila-erc.eu/data/id/lemma/89148.
[21] http://lila-erc.eu/data/id/lemma/142756.
[22] http://lila-erc.eu/data/id/lemma/142757.

disambiguation based on the linguistic context of the document; this happens, for instance, in the case of homographs, as for the word *dico*, linked both to *dĭco*, 'to proclaim' or 'to dedicate'[23] and to *dīco*, 'to name', 'to utter'.[24] On rare occasions (29 tokens), it was however not possible to disambiguate between the lemmas available in LiLa: as a consequence, we linked the ambiguous tokens to all their corresponding lemmas. This was due to the fragmentary nature of some texts, where an analyzable context for disambiguation was not available. This is the case, for example, of BTT-609-16 *mallus* (context: [...] *mallus alu*[...] [...]*us*), for which two senses are equally possible, that of 'pole'[25] and 'appletree'.[26]

5. Querying CLaSSES in LiLa

Thanks to the interoperability of CLaSSES with the other resources for Latin linked to the LiLa Knowledge Base, research questions related to non-literary texts can be empirically investigated on the several different textual resources interlinked in the Knowledge Base by running queries on the SPARQL endpoint of LiLa[27]. By focusing on the question of spelling variants attested in the inscriptions, in what follows we shortly consider two case studies, i.e., consonant doubling (see 5.1), and the writing of long /i:/ through the diphthong <EI> (see 5.2). Moreover, we report and briefly discuss a query that exploits the information on derivational morphology recorded in the Lemma Bank (see 5.3).

5.1. Consonant doubling

As is known, the spelling of Latin long consonants through *geminatio consonantium* was introduced at the end of the third century BCE [16, 17, 18]. Consonant doubling, however, generalized slowly, so it was seldom omitted in the second century BCE in inscriptions. For example, in the 2nd-century inscriptions included in CLaSSES, 28 tokens (20 lemmas) display single for double consonants over 72 spellings with *geminatio consonantium*. These tokens can be easily retrieved through the function "Search for linguistic phenomena" available in the CLaSSES online search interface, by selecting the label "single pro double consonant". Thanks to the interoperability between distributed resources provided by LiLa, it is possible to search the occurrences of the lemmas for these tokens in the corpora interlinked through the Lemma Bank. This is particularly useful for both quantitative and qualitative linguistic analysis. For example, among the forms found in CLaSSES, it is possible to find occurrences of the same lemma either in the form with a single consonant or with a double consonant – such as the name *Mummius* in the *tituli mummiani*, which is displayed either in the forms *Mumius*, in CIL-I^2-628, or *Mummius*, in CIL-I^2-627, 629 [13]. The presence of the alternation between <C> and <CC> in these inscriptions can be interpreted as a sign of an incomplete generalization of consonant doubling at this stage. However, it is fundamental to exclude the possibility that the form *Mumius* occurring in our corpus represents a commonly attested variant of the proper name *Mummius*. This information is not readily retrievable in the available sources, since such spelling variants of proper names are generally not recorded in the dictionaries. However, by collecting the occurrences of the lemma *Mummius* in the textual resources interlinked through LiLa, it is possible to ascertain that the variant without consonant doubling is never attested in any of the texts provided by such resources (e.g., Cicero's *De Lege Agraria*, *In Verrem* and Tacitus' *Annales*, included in the LASLA corpus).[28] Thus, we may assume that the form *Mumius* found in CIL-I^2-628 is a hint of the incomplete generalization of *geminatio consonantium*, in line with the chronology proposed in the literature.

5.2. <EI> for /i:/

The linking of the tokens of CLaSSES to the Lemma Bank of LiLa can also shed light on the writing of /i:/ through <EI> in Latin sources. It is known from the literature [19, 20, 21, 22] that, in the 'urban' Latin of the city of Rome, the monophthongization of the diphthong [ej] took place in two steps: (i) [ej] > [e:],[29] between the 3rd and mid-2nd century BCE; (ii) [e:] > [i:], between the 2nd and 1st century BCE. The data from CLaSSES, obtained through the function "Search for linguistic phenomena" (label "Diphthong - Classical <I> /ī/ = <EI>"), confirm the traditional picture, indicating that the spelling <EI> for /i:/ is either a conservative spelling retained in earlier documents, or an archaizing feature that characterizes the solemn register of later public and official inscriptions. More in detail, in CLaSSES the spelling <EI> for /i:/ is found in 225 occurrences (99 lemmas), mainly in older public inscriptions, before the 1st century BCE (212 occurrences over 225). A more comprehensive view of this phenomenon can be obtained thanks to the interoperability between different Latin corpora made possible in LiLa. By running a query on the corpora interlinked in the Knowledge Base, it is possible to collect all the tokens linked to the 99 lemmas concerned and select those

[23] http://lila-erc.eu/data/id/lemma/99301.
[24] http://lila-erc.eu/data/id/lemma/99302.
[25] http://lila-erc.eu/data/id/lemma/111421.
[26] http://lila-erc.eu/data/id/lemma/111423.
[27] https://lila-erc.eu/sparql/.
[28] https://lila-erc.eu/data/corpora/Lasla/id/corpus.
[29] Possibly a long lax [i:] [20, 23].

where the spelling <EI> for /i:/ takes place.

For instance, of particular interest is the form *sei* for *sī* 'if' that is found in Archaic Latin. By using LiLa, it is possible to find that out of the 22,161 occurrences of *si* in the corpora interlinked therein, 10 show the form *sei*. One relevant example is from Plautus' *Epidicus* (*Ep.* 567, twice). These 2 occurrences of *sei*, which in LiLa are recorded as 2 tokens from the LASLA corpus, testify to the above-mentioned first step of the monophthongization process ([ej] > [e:]), which takes place in the age of Plautus and which is attested elsewhere in his works.

5.3. Derivational Morphology

So far, we have discussed some very easy queries on specific lexical items that can be performed to compare information provided by CLaSSES to that provided by other resources. However, LiLa allows not only to explore and compare single corpora at the lexical level (via the Lemma Bank), but also to conduct in-depth linguistic analysis, concerning, for instance, morphology. For example, it is possible to compare the type and number of affixes found in CLaSSES, investigating how many (and which type of) derivational morphemes are represented in non-literary texts. The list of affixes that build up the lexicon of CLaSSES can be accessed with a SPARQL query that retrieves all the lemmas in the CLaSSES corpus that feature an affix (either prefix, or suffix) in their morphological form, and reports the number of their occurrences therein (see Listing 1[30]).

Morphological information was not annotated in CLaSSES. Thus, the link to LiLa allows to conduct more in-depth linguistic analyisis; most importantly, it also allows users to compare different corpora with relation to specific linguistic features. For instance, it is possible to investigate to what extent the derivational morphology found in non-literary texts deviates from that of Classical texts by performing the very same query on the LASLA corpus, by simply replacing the URI for CLaSSES in the SPARQL query (as subject of the `powla:hasSubDocument` property) with that for the LASLA corpus: http://lila-erc.eu/data/corpora/Lasla/id/corpus.

The affixes that most frequently occur in the CLaSSES corpus are three suffixes and a prefix:

- *-in*, 486 occurrences (7.3% of affixes extracted from the corpus);
- *-(t)or*, 456 occurrences (6.9%);
- *-t*, 442 occurrences (6.7%);
- *con-*, 440 occurrences (6.6%).

These affixes have a very different distribution in LASLA, in which only *con-* is among the most frequent affixes, with 32,763 occurrences (7,9%), whereas *-in* counts just for 1.2% of all affixes extracted from the corpus (5,137 occ.), *-(t)or* for 2.3% (9,593 occ.), and *-t* for 1% (4,024 occ.). Such differences are largely due to a number of lexemes that are highly frequent in epigraphic texts, in particular *dominus* 'master' (198 occ.) for the suffix *-in* and *imperator* 'general, emperor' (153 occ.) for the suffix *-(t)or*, which are most frequent in public inscriptions, or *libertus/liberta* 'freedman' (281 occ.) for the suffix *-t*, which is most frequent in funerary inscriptions, where the epitaph often refers to the civil status of freed slaves. Therefore, even if there is a major difference in dimension between the two corpora, a query such as the one here illustrated can bring to light specificities of the corpus CLaSSES that go beyond the lexical level and that could not be observed without comparison with other resources.

6. Conclusion and Future Work

The linking of CLaSSES into LiLa represents an added value for both the resources. As for CLaSSES, its (meta)data are now interoperable with the other resources interlinked in the Knowledge Base. As for LiLa, the non-literary texts of CLaSSES increased significantly its textual coverage, both in terms of size and in terms of register variation.

In the near future, we plan to model and interlink in LiLa other types of metadata provided by CLaSSES, such as information about the provenance and the dating of the texts. We plan to start from metadata on the time span of the texts, that we will model as Linked Data using data categories and properties from the CIDOC Conceptual Reference Model.[31]

Acknowledgments

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. The work of Lucia Tamponi is partly funded by the PRIN Project "Ancient languages and writing systems in contact: a touchstone for language change", prot. 2017JBFP9H.

References

[1] C. T. Lewis, C. Short, A Latin Dictionary. Founded on Andrews' edition of Freund's Latin dictionary, Clarendon Press, Oxford, 1879.

[30] The query outputs a table with four columns: the label of the lemma (`?lemmaLabel`), the type of affix, either prefix or suffix (`?affixType`), the label of the affix (`?affixLabel`) and the total number of tokens for the lemma in CLaSSES (`(count(?tokenClasses) as ?count)`).

[31] https://www.cidoc-crm.org/.

Listing 1: A SPARQL query on the LiLa Knowledge Base

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX lila: <http://lila-erc.eu/ontologies/lila/>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX powla: <http://purl.org/powla/powla.owl#>

SELECT ?lemmaLabel ?affixType ?affixLabel (count(?tokenClasses) as ?count)
WHERE {
      ?lemmaLiLa a lila:Lemma ;
            (lila:hasPrefix|lila:hasSuffix) ?affix ;
            rdfs:label ?lemmaLabel .
      ?affix rdfs:label ?affixLabel ;
            rdf:type ?affixType .
      ?tokenClasses lila:hasLemma ?lemmaLiLa ;
            powla:hasLayer ?docLayer ;
            rdfs:label ?tokenClassesLabel .
      ?docLayer powla:hasDocument ?subDocLayer .
      ?subDocLayer dc:title ?titlesubDocLayer .
      <http://lila-erc.eu/data/corpora/CLaSSES/id/corpus> powla:hasSubDocument ?subDocLayer .
}
GROUP BY ?lemmaLabel ?affixType ?affixLabel
ORDER BY DESC(?count)
```


[2] C. d. F. s. du Cange, bénédictins de la congrégation de Saint-Maur, d. P. Carpentier, J. C. Adelung, G. A. L. Henschel, L. Diefenbach, L. Favre, Glossarium mediae et infimae latinitatis, Favre, Niort, France, 1883–1887.

[3] G. Marotta, Talking stones. Phonology in Latin inscriptions?, Studi e Saggi Linguistici 53 (2015) 39–63.

[4] G. Marotta, Sociolinguistica storica ed epigrafia latina: il corpus CLaSSES I, Linguarum Varietas 5 (2016) 145–159.

[5] I. De Felice, G. Marotta, M. Donati, CLaSSES: A new digital resource for Latin epigraphy, Italian Journal of Computational Linguistics 1 (2015) 119–130.

[6] G. Marotta, F. Rovai, I. De Felice, L. Tamponi, CLaSSES: Orthographic variation in non-literary Latin, Studi e Saggi Linguistici 58 (2020) 39–65.

[7] O. Lassila, R. R. Swick, Resource Description Framework (RDF) Model and Syntax Specification, 1998.

[8] C. Chiarcos, POWLA: Modeling Linguistic Corpora in OWL/DL, in: E. Simperl, P. Cimiano, A. Polleres, O. Corcho, V. Presutti (Eds.), The Semantic Web: Research and Applications, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2012, pp. 225–239. doi:10.1007/978-3-642-30284-8_22.

[9] C. Chiarcos, M. Sukhareva, OLiA – Ontologies of Linguistic Annotation, Semantic Web 6 (2015) 379–386.

[10] P. Buitelaar, P. Cimiano, J. McCrae, E. Montiel-Ponsoda, T. Declerck, Ontology Lexicalization: The *lemon* Perspective, in: Proceedings of the Workshops-9th International Conference on Terminology and Artificial Intelligence (TIA 2011), 2011, pp. 33–36.

[11] J. McCrae, J. Bosque-Gil, J. Gracia, P. Buitelaar, P. Cimiano, The OntoLex-Lemon Model: Development and Applications, in: Proceedings of eLex, 2017, pp. 587–597.

[12] M. Passarotti, M. Budassi, E. Litta, P. Ruffolo, The Lemlat 3.0 Package for Morphological Analysis of Latin, in: Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language, 2017, pp. 24–31.

[13] L. Tamponi, Consonant gemination in Latin epigraphy between variation and standard, in: Latin Vulgaire-Latin Tardif XIV, Actes du XIVème colloque international sur le latin vulgaire et tardif, Brepols, Turnhout, Forthcoming.

[14] R. J. Rowland, Onomastic remarks on Roman Sardinia, Names 21 (1973) 82–102.

[15] G. Lupinu, Latino epigrafico della Sardegna: aspetti fonetici, Ilisso, Nuoro, 2000.

[16] M. Mancini, Lucilius and Nigidius Figulus on orthographic iconicity, Journal of Latin Linguistics 18 (2019) 1–34.

[17] M. Mancini, Repertori grafici e regole d'uso: il caso del latino <XS>, in: L. Agostiniani, M. P. Marchese (Eds.), Lingua, testi, storia. Atti della giornata di studi in ricordo di Aldo Luigi Prosdocimi, Giorgio

Bretschneider, Roma, 2019, pp. 13–53.
[18] L. Tamponi, La *geminatio consonantium*: studio su un corpus di epigrafi latine anteriori al I secolo d.C., Studi e Saggi Linguistici 60 (2022) 29–50.
[19] M. Niedermann, Phonétique historique du latin, Librairie C. Klincksieck, Paris, 1953.
[20] M. Leumann, Lateinische Laut- und Formenlehre, Beck, Munchen, 1977.
[21] W. S. Allen, Vox latina: A Guide to the Pronunciation of Classical Latin, Cambridge University Press, Cambridge, 1978.
[22] M. Mancini, *Dilatandis litteris*: uno studio su Cicerone e la pronunzia 'rustica', in: R. Bombi, G. Cifoletti, F. Fusco, L. Innocente, V. Orioles (Eds.), Studi linguistici in onore di Roberto Gusmani, Edizioni dell'Orso, Alessandria, 2006, pp. 1023–1046.
[23] M. Benedetti, G. Marotta, Monottongazione e geminazione in latino: nuovi elementi a favore dell'isocronismo sillabico, in: P. Molinelli, P. Cuzzolin, C. Fedriani (Eds.), Latin Vulgaire–Latin Tardif, Actes du Xe Colloque International sur le Latin Vulgaire et Tardif, Sestante, Bergamo, 2014, pp. 25–43.

Processing effort during reading texts in young adults: text simplification, readability assessment and preliminary eye-tracking data

Maria De Martino

University of Salerno, via Giovanni Paolo II, 132, 84084, Fisciano (SA), Italy

English. The paper reports empirical data about the impact of text simplification procedures supported by readability assessment measures on processing effort during reading. Sixty-six Italian native undergraduate students read original and simplified versions of TV news texts and answered a comprehension question. Accuracy data, single word-based and sentence-based measures collected by means of an eye-tracker show that reading simplified texts requires less cognitive demands than their original versions.

Italiano. Il lavoro riporta dati empirici relativi all'efficacia dell'applicazione di metriche di leggibilità nella produzione di testi semplificati. Sessantasei studenti italiani madrelingua hanno letto versioni originali e semplificate dei lanci di notizie tratte da telegiornali e hanno risposto ad una domanda di comprensione. I dati sull'accuratezza e diverse metriche ottenute con l'impiego di un eye-tracker e calcolate al livello di singole parole e singole frasi mostrano che le versioni semplificate dei testi riducono significativamente il carico cognitivo dei lettori rispetto alle relative versioni originali.

Keywords
Text simplification, Readability, Eye-tracking

1. Introduction

Reading is a recent achievement in human evolution, but reading proficiency is considered an important component of success and life outcome [1]. The ability to read and understand with minimum effort depends on both reader characteristics (e.g., literacy, multilingualism, presence/absence of cognitive disorders) and text properties (e.g., length, topic, lexical and syntactic complexity, cohesion, coherence) [2]. One approach to improve inclusion by coping with disadvantage in reading skills is to match readers with texts appropriate to their reading abilities. This goal can be accomplished by exploiting readability formulas to predict the reading and comprehension difficulty of a text for a given target audience and, then, to obtain simplified, i.e. easy-to-understand, texts [3].
Different readability formulas are available for several languages [4] [5]. Many of them rely on text linguistic features such as lexical and syntactic features. Lexical features include the values of frequency, familiarity, imageability and age of acquisition of words within a text. Syntactic features include the complexity of syntactic structures such as the incidence of types of clauses and phrases in the text. Both lexical and syntactic features have been shown to impact cognitive demands in reading processes [6]. Hence, more recent readability formulas take into account measures of natural text processing and try to express readability in terms of cognitive processing effort [7].
Actually, reading texts is a complex behavior subserved by automated interactions between different cognitive processes: visual perception, attention, lexical access, working memory, semantic processing. All these processes are involved in the two main aspects of reading: visual information decoding and meaning construction (comprehension). One of the techniques that has been extensively used to study the reading behavior is the recording of eye movements [8]. During reading, the reader's eyes move from one position to the next in order to process different levels of information that can be extracted from words' visual form. Psycholinguists assume that eye movements during reading reflect different stages of language processing. Some movement have a perceptual function: saccades are rapid movements that shift the eye's focus between two fixed points and are necessary to bring the visual information into the zone of the visual field where acuity is best. Other movements have more complex functions. Fixations are short periods of steadiness of the eye on a word and their duration is a marker of the ease of accessing the meaning of the word and integrating this into the current sentence. Regressions are backward-directed saccades and are related to the necessity of the reader

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
mdemartino@unisa.it (M. De Martino)

to re-analyze previously explored portions of the text because of processing difficulties.

Gaze behavior during reading is exploited in several contexts and for different aims both in cognitive psychology and in Natural Language Processing (NLP) literature. For instance, eye-tracking metrics are exploited to unfold mechanisms of reading in L2 learners, typical and atypical readers [9] [10] [11] [12] [13] [14]. Corpora of eye-tracking data are available in many languages and are profitably used to implement language models that can predict human reading behavior (e.g., [15] [16] [17]).

A less investigated issue is to what extent the use of specific text simplification strategies reduces the processing effort as measured through eye-movements [18] [19] .

The aim of the present study is to test the impact of texts' simplification and readability assessment on young adults reading behavior through the analysis of eye-movements.

2. Method

We conducted an eye-tracking reading study to obtain objective and reliable measures of processing effort. The advantage of monitoring readers' eye movements during reading is that it is considered to be the experimental situation that better resembles natural reading [11] [20] [21].

2.1. Stimuli

2.1.1. Selection

The same materials as used in [22] were employed. They consisted of 18 texts of news scripts as read by Italian TV news anchors. Such news texts are usually short but often linguistically and stylistically complex and can be difficult to comprehend for non-native speakers and/or for people with low literacy, reading disorders or cognitive and intellectual disabilities. Hence, they constitute suitable materials to be employed in an empirical study on simplification.

2.1.2 Readability Assessment and Text Simplification

Each selected text in its original version (OV) underwent a readability assessment through the READ-IT tool [3] and then to a double manual simplification process that generated 2 simplified versions: S1 and S2. The simplification strategies used to implement S1 involved sentence splitting, changing passive voice to active, lexical and syntactic ambiguity resolution, low frequency and long words replacement.

In addition to the above mentioned simplification strategies, specific interventions on the semantic content were used in order to achieve the S2 versions. They were mainly focused on the temporal ordering of events and on reporting each factual event into a separate sentence (See [22] for further details and for examples of the original and simplified materials).

The obtained simplified versions were matched for text length calculated in number of words (average number of words: OV=56; S1=58; S2=58).

The text complexity measures obtained through the application of the READ-IT tool [3] revealed that the OV texts scores were significantly lower than the S1 and S2 ones (see Table 1: values shown in bold are the readability scores, values reported in parentheses are the p values of the t-tests comparing OV vs. S1 and S2).

The GULPEASE score indicates the readability of the texts: a higher GULPEASE score indicates higher readability of a text.

Table 1
Readability assessment: comparisons between OV vs. simplified versions

READ-IT scores	OV	S1	S2
Base	**36**	**11** (p<.005)	**6** (p<.001)
Lexical	**96**	**80** (p<.05)	**83** (p=.05)
Syntactic	**51**	**13** (p<.01)	**7** (p<.001)
Global	**80**	**34** (p<.001)	**21** (p<.001)
GULPEASE	**51**	**58** (p<.001)	**65** (p<.001)

The comparison between the two simplified versions, S1 vs. S2, revealed that S2 obtained better readability scores than S1 scores only on the global score (t (17) = 215, p<.05) and on the GULPEASE score (t (17) = -5.08, p<.001)[2] .

2.1.3. Implementation of the experimental text lists

The whole set of 54 texts (18 OV, 18 S1, 18 S2) was split into 3 lists. Each list included 1 of the 3 versions of each text and comprised 6 OV, 6 S1 and 6 S2. Participants were randomly assigned to a given list. This strategy allowed to ensure that each participant was presented only once with a given text in order to avoid putative effects of the repetition of materials. Each participant was administered with texts presented in a shuffled order.

2.2. Participants

Sixty-six undergraduate students from University of

[2] Further user-based readability scores on the materials used in the study are reported in [22] and provide information about the speakers' perceived difficulty of the linguistic formulation and of the topic of the texts and the perceived naturalness/acceptability of the Italian language used to generate the texts. The judgments of speakers did not reveal significant differences between OV, S1 and S2.

Salerno (45 females) were recruited; they voluntarily took part in the experiment. Their age ranged from 19 to 30 years (average = 23 years). They were all native speakers of Italian, had normal or corrected-to-normal vision and did not report history of reading, language, learning or neurological/psychiatric disorder.

2.3. Apparatus

The reading experiment was implemented and administered via Tobii Pro Lab 1.194 software.
A screen-based Tobii Pro X2-30 eye-tracker was connected to the monitor of an HP computer available at the LaPSUS laboratory (University of Salerno). The range of head movement allowed was within a three-dimensional range of 50 cm W × 36 cm H × 70 cm and the allowed operating distance from the monitor was within 40 and 90 cm.

2.4. Procedure

The experimental procedure consisted of different steps.

2.4.1. Calibration

The participants sat in front of the screen; the distance to the screen was adjusted. For calibration in the Tobii Pro Lab software, participants were required to keep their heads as still as possible and to look at a fixation point moving on the screen.

2.4.2. Training and warm up trials

The participants were presented with a slide displaying written instructions about the reading task. Then, a training phase was administered. Participants were requested to perform 2 warm up trials: they were shown a text on the screen and were instructed to read it silently at their own pace of comprehension and to press any key of the PC keyboard to move to the subsequent slide. After reading the text, they were asked to use the mouse to select the correct response of a true–false question presented on the screen. Then, a second trial was administered.

2.4.3. Experimental Session

After the warm up trials, each participant was administered 1 of the 3 lists made up with 18 experimental trials.

2.5. Dependent variables

Different measures were analysed in the current study.

2.5.1. Global reading time

The global reading time was recorded.

2.5.2. Accuracy

Answers to the true-false question were recorded and analysed as indicators of the comprehension of the texts.

2.5.3. Eye-tracking metrics

The Tobii Pro X2-30 hardware and software equipment (Tobii Pro Lab 1.194) provides a large number of eye-tracking measures. However, the current study reports the most commonly measures found to be related to text difficulty and cognitive demand [23].
The following measures were analysed for word-based and sentence-based areas of interest (AOI):

- Number and Duration of Fixations: fewer and shorter fixations are supposed to be associated with lower reading effort.
- Number and Duration of Visits: visits can be defined as the number of times that the reader's eye move towards a given AOI with either progressive or regressive saccades. The entry and exit saccades are excluded. The number and duration of visits indicate that specific portions of the text receive specific amount of attentional and linguistic resources to be processed.
- Regression path duration: it describes the time that elapses between a first fixation on an AOI to the moment when gaze is directed away from that region to the right. Thus, it includes time spent re-reading earlier parts of the text before the reader is ready to proceed with the rest of the text.
- Re-reading duration: it corresponds to the regression-path duration minus first-pass duration and it is assumed to reflect strategic, controlled processes involved in reading comprehension.

3. Results

The results of reading times, accuracy rates and eye-tracking metrics were analyzed through a series of ANOVAs.

3.1.1. Whole text reading times and accuracy data

No significant effect was found on global reading time; this result replicates the findings of [22] obtained in a different experimental setting.
On the contrary, ANOVA on accuracy data (Figure 1) showed a significant effect of simplification ($F(2,1185) = 3{,}8094$, $p=.02$). LSD post hoc tests revealed that questions to S2 were responded significantly better than OV ($p=.008$). The difference between S1 and OV was only marginally significant ($p=.052$), while the difference between S1 and S2 was not significant ($p=.46$).

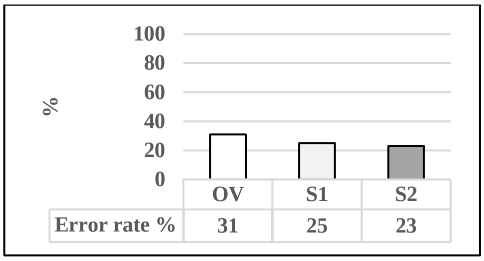


Figure 1: Error rate for OV, S1 and S2

3.1.2. Word-based eye-tracking data

The mean and standard deviation values obtained for word-based and sentence-based eye-tracking metrics are reported in Tables 2 and 3.

Table 2
Word-based eye-tracking metrics

WORD-BASED METRICS		OV	S1	S2
Number of Fixations	mean	2,11	2,0	2,0
	ds	1,59	1,5	1,47
Duration of Fixations (ms)	mean	413,94	393,5	389,46
	ds	370,97	342,49	340,43
Number of Visits	mean	1,75	1,71	1,69
	ds	1,13	1,08	1,07
Duration of Visits (ms)	mean	430,96	408,77	404,67
	ds	398,37	368,13	364,58
Regression-Path Duration (ms)	mean	413,78	391,46	389,28
	ds	1416,05	1195,97	1174,15
Re-Reading Duration (ms)	mean	285,36	259,77	254,21
	ds	1383,62	1167,14	1143,29

ANOVAs performed on word-based metrics showed a significant effect of simplification both on the number (F (2, 42566) = 21,098, $p<.001$) and the duration of Fixations (F (2, 42566) = 19,733, $p<.0001$). More specifically, the post hoc LSD tests revealed that the number and the duration of fixations is significantly higher in OV than in both S1 ($p<.001$) and S2 ($p<.001$). No significant difference was detected between S1 and S2.
An effect of simplification was observed both on the number (F (2, 42566) = 11,202, $p<.001$) and the duration of visits (F (2,42566) = 19,920, $p<.0001$): the OV texts received significantly higher scores than S1 and S2, while S1 and S2 resulted equivalent.
Data on the re-reading scores showed that simplification elicited a slight tendency to the statistical significance (F (2, 42560) = 2,5594, $p = .07$); however, additional planned comparisons showed that the OV texts required a re-reading time significantly higher than the S2 texts ($p<.05$), but did not differ from S1 ($p=.08$).

The regression-path duration did not show any significant effect of simplification at the word-level.

Table 3
Sentence-based eye-tracking metrics

SENTENCE-BASED METRICS		OV	S1	S2
Number of Fixations	mean	20,05	14,94	15,35
	sd	17,34	11,68	9,32
Duration of Fixations (ms)	mean	3928,22	2906,84	2967,14
	sd	3660,62	2471,19	1949,63
Number of Visits	mean	3,58	3,50	4,39
	sd	2,58	2,55	2,22
Duration of Visits (ms)	mean	4975,30	3615,57	3686,45
	sd	4660,07	3109,65	2430,72
Regression-Path Duration (ms)	mean	3970,04	2904,92	2967,20
	sd	4793,85	3588,49	3048,10
Re-Reading Duration (ms)	mean	2249,18	1806,07	2122,96

A slightly different picture emerged from the analyses performed on sentence-based metrics. The ANOVAs revealed that the effect of simplification reached the statistical significance for all the collected eye-tracking metrics:

- Number of fixations: F (2, 5692) =163,15, $p = .0000$;
- Duration of Fixations: F (2, 5692) = 144,61, $p = .0000$;
- Number of visits: F (2, 5692) = 5,2883, $p = .00507$;
- Duration of visits: F (2, 5692) = 160,30, $p=.0000$;
- Regression-path duration: F (2, 5651) = 75,393, $p = .0000$;
- Re-reading duration: F (2, 5651) = 20,504, $p =.00000$.

OV was found to be the version that required significantly higher processing effort when compared both to S1 and S2 ($p < .001$). In addition, S1 was found to be more demanding than S2 ($p < .05$). Only for the number of visits OV and S1 showed equivalent amount of processing effort ($p = .33$).

4. Conclusion

The paper investigates to what extent the application of text simplification strategies improves the readability of texts and reduces the reading processing effort as it emerges from cognitive indexes that are out of the awareness of the reader, i.e. eye movements patterns.
The preliminary data reported in the paper show that accuracy in comprehension questions increases significantly when texts undergo simplification procedures based on the reduction of the reader's

amount of processing inferences (i.e., event reordering or coreference chains explaining).

On the other hand, the physiological and cognitive measures related to the processing effort during reading are affected by simplification strategies that involve both the lexical-syntactic level and the content level.

Moreover, the metrics collected at the sentence-level and single-word level are found to be suitable and sensitive measures to detect respectively the efficacy of simplification procedures in modulating the strategic controlled processes involved in comprehension and the attentional and lexical processing effort during reading. Interestingly, the data were obtained by analyzing the performance of young adult skilled readers that are supposed to be less likely influenced by the readability of texts.

References

[1] Dehaene S. Reading in the brain: The science and evolution of a human invention. New York: Viking; 2009.

[2] Kuperman V, Matsuki K, Van Dyke JA. Contributions of reader-and text-level characteristics to eye-movement patterns during passage reading. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2018 Nov;44(11):1687. doi: 10.1037/xlm0000547

[3] Dell'Orletta F, Montemagni S, Venturi G. READ–IT: Assessing readability of Italian texts with a view to text simplification. In: Proceedings of the second workshop on speech and language processing for assistive technologies 2011 Jul (pp. 73-83).

[4] Crossley SA, Skalicky S, Dascalu M. Moving beyond classic readability formulas: New methods and new models. Journal of Research in Reading. 2019 Nov;42(3-4):541-61. doi:10.1111/1467-9817.12283

[5] Nahatame S. Text readability and processing effort in second language reading: A computational and eye-tracking investigation. Language learning. 2021 Dec;71(4):1004-43

[6] Keller TA, Carpenter PA, Just MA. The neural bases of sentence comprehension: a fMRI examination of syntactic and lexical processing. Cerebral cortex. 2001 Mar 1;11(3):223-37. doi.org/10.1093/cercor/11.3.223

[7] Huckin TN. A cognitive approach to readability. InNew essays in technical and scientific communication 2019 Mar 8 (pp. 90-108). Routledge.

[8] Hyönä J, Kaakinen JK. Eye movements during reading. Eye movement research: An introduction to its scientific foundations and applications. 2019:239-74.

[9] Cop U, Drieghe D, Duyck W. Eye movement patterns in natural reading: A comparison of monolingual and bilingual reading of a novel. PloS one. 2015 Aug 19;10(8):e0134008. doi.org/10.1371/journal.pone.0134008

[10] Franzen L, Stark Z, Johnson AP. Individuals with dyslexia use a different visual sampling strategy to read text. Scientific reports. 2021 Mar 19;11(1):6449. doi.org/10.1038/s41598-021-84945-9

[11] Godfroid A. Eye tracking in second language acquisition and bilingualism: A research synthesis and methodological guide. 2019.

[12] Prabha AJ, Bhargavi R. Predictive model for dyslexia from fixations and saccadic eye movement events. Computer Methods and Programs in Biomedicine. 2020 Oct 1;195:105538. doi.org/10.1016/j.cmpb.2020.105538

[13] Rayner K, Shen D, Bai X, Yan G, editors. Cognitive and cultural influences on eye movements. Taylor & Francis; 2023 May 31.

[14] Rayner K. Eye movements in reading and information processing: 20 years of research. Psychological bulletin. 1998 Nov;124(3):372.

[15] Hollenstein N, Rotsztejn J, Troendle M, Pedroni A, Zhang C, Langer N. ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. Scientific data. 2018 Dec 11;5(1):1-3. doi.org/10.1038/sdata.2018.291

[16] Luke SG, Christianson K. The Provo Corpus: A large eye-tracking corpus with predictability norms. Behavior research methods. 2018 Apr;50:826-33. doi.org/10.3758/s13428-017-0908-4

[17] Siegelman N, Schroeder S, Acartürk C, Ahn HD, Alexeeva S, Amenta S, Bertram R, Bonandrini R, Brysbaert M, Chernova D, Da Fonseca SM. Expanding horizons of cross-linguistic research on reading: The Multilingual Eye-movement Corpus (MECO). Behavior research methods. 2022 Dec;54(6):2843-63.

[18] Rivero-Contreras M, Engelhardt PE, Saldana D. Do easy-to-read adaptations really facilitate sentence processing for adults with a lower level of education? An experimental eye-tracking study. Learning and Instruction. 2023 Apr 1;84:101731. doi.org/10.1016/j.learninstruc.2022.101731

[19] Rets I, Rogaten J. To simplify or not? Facilitating English L2 users' comprehension and processing of open educational resources in English using text simplification. Journal of Computer Assisted Learning. 2021 Jun;37(3):705-17. doi: 10.1111/jcal.12517

[20] Raney GE, Campbell SJ, Bovee JC. Using eye movements to evaluate the cognitive processes involved in text comprehension. JoVE (Journal of Visualized Experiments). 2014 Jan 10(83):e50780. doi: 10.3791/50780

[21] Roberts L, Siyanova-Chanturia A. Using eye-tracking to investigate topics in L2 acquisition and L2 processing. Studies in Second Language Acquisition. 2013 Jun;35(2):213-35. doi:10.1017/S0272263112000861

[22] De Martino M, Colella A. La produzione di testi semplificati di notiziari televisivi italiani destinati a persone con disturbi cognitivi acquisiti: un'integrazione tra metodi psicolinguistici e analisi automatiche (Implementing Simplified TV News Texts in

Italian for People with Acquired Cognitive Disorders: Psycholinguistic Methods and Automatic Analyses). In Proceedings of the Eighth Italian Conference on Computational Linguistics (CLiC-IT).CLiC-it 2021.
[23] Rayner K, Chace KH, Slattery TJ, Ashby J. Eye movements as reflections of comprehension processes in reading. Scientific studies of reading. 2006 Jul 1;10(3):241-55. doi.org/10.1207/s1532799xssr1003_3

An experiment in error analysis of real-time speech machine translation using the example of the European Parliament's Innovation Partnership*

Elisa Di Nuovo[1,*]

[1]University of Turin, via Giuseppe Verdi, 8, 10124 Torino (TO) - Italy

Abstract

In recent years, technological progress has made Machine Translation (MT) a reality. Significant improvements have been obtained using deep learning models as opposed to rule-based and statistical MT models. Human evaluation still remains under-explored. In 2019 the European Parliament (EP) started an innovation partnership with commercial operators, with the purpose of developing a tool exploiting state-of-the-art, real-time Automatic Speech Recognition (ASR) and MT technologies to make parliamentary plenary sessions accessible to D/deaf and hard of hearing. In this paper, we present a quantitative and qualitative error analysis carried out on a test set consisting of 78 short speeches delivered by Members of the EP in 19 languages deployed in the EP prototype by November 2022. The taxonomy used for ASR and MT is adapted from the Multidimensional Quality Metrics framework. Results show that sentence segmentation is the biggest issue in the ASR output—not considered using automatic metrics—which often affects the MT output.

Keywords

Real-time speech machine translation, Error analysis, Human evaluation, Cascade system

1. Introduction

In recent years, the landscape of language translation has been fundamentally transformed by remarkable technological advancements. Machine Translation (MT), once an ambitious aspiration, has now become a tangible reality. This transformation has been primarily fueled by the advent of deep learning models, specifically neural machine translation and transformers. These cutting-edge models have ushered in a new era of translation, eclipsing the limitations of traditional rule-based and statistical MT methods. Deep learning models use the mechanism called attention to improve the performance [1] and have been usually evaluated on offline written translation tasks involving a few language pairs [2].

Very recently, research expanded its focus also on speech machine translation, tackled as a concatenation of Automatic Speech Recognition (ASR) and MT, or as an end-to-end task (i.e. direct translation of speech in language A into text in language B).[1] In the last evaluation campaign of the 19th International Conference on Spoken Language Translation (IWSLT 2022) [5], one of the eight shared tasks focused on real-time speech translation, addressed as translation of ASR output or directly from the audio source and involving English to German, English to Japanese and English to Mandarin Chinese. A novelty of this year campaign is the addition of manual evaluation of real-time outputs.

Like many natural language processing tasks, MT is difficult to evaluate. One of the reasons for this is the nondeterministic nature of translation, i.e. there is more than one correct way to translate from one language into another. Evaluation in shared tasks is usually carried out by means of automatic metrics, BLEU (Bilingual Evaluation Understudy) [6] being the standard for MT evaluation. This metric tries to overcome the nondeterministic nature of translation using multiple references. However, automatic metrics have several limitations [7].[2] On the other hand, human evaluation, if carried out using fined-grained guidelines to limit subjectivity, can give a clearer indication of the MT output quality. However, being resource expensive (i.e. it is hard to find skilled evaluators; skilled evaluators have a high cost), it has been used limitedly and in small studies. To avoid these limitations, crowdsourced annotators have been used. Unfortunately, crowdsourced annotators are frequently inexperienced. As [10] affirm, crowdsourced human evaluation can be used when MT quality is poor, because it can still provide

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

*This study and paper was written while the author was working for the European Parliament, in the Unit in charge of the administration and evaluation of the prototype. This is not the official evaluation methodology employed by the European Parliament for evaluation.

*The author, as of 1st October 2023 is employed by the Joint Research Centre, European Commission, Ispra (VA), Italy.

elisa.dinuovo@unito.it (E. D. Nuovo)

0000-0002-4814-982X (E. D. Nuovo)

[1]See for example the studies in [3, 4] for a comparison between cascade and end-to-end systems.

[2]See Moorkens et al. [8] on translation quality assessment and Chatzikoumi [9] for a comprehensive review of automatic and human MT evaluation.

a useful indication; but, as quality improves, it becomes unfit and leads to erroneous claims.[3]

In 2019 the European Parliament (EP) started an Innovation Partnership with commercial operators, with the purpose of developing a tool that can perform real-time ASR and MT from and into all the 24 official languages of the European Union (EU).[4] This partnership has the aim of making parliamentary plenary sessions accessible in near-real time to D/deaf and hard of hearing persons.[5] The challenges faced by this project are manifold: the high degree of multilingualism which is highly ambitious considering the technical limits of current MT particularly in a number of low-resource languages, the presence of non-native accents, the large variety of vocabulary/EU jargon required by the numerous specific domains tackled in the plenaries, the low latency constraints to have transcriptions and translations in near-real time, and the required high quality of the output. By November 2022, 19 language models have been developed and made available via a demo interface. These languages are English (EN), French (FR), German (DE), Spanish (ES), Italian (IT), Polish (PL), Greek (EL), Romanian (RO), Dutch (NL), Portuguese (PT), Bulgarian (BG), Czech (CS), Slovak (SK), Croatian (HR), Lithuanian (LT), Finnish (FI), Hungarian (HU), Swedish (SV) and Slovenian (SL).

In this paper, we present a quantitative and qualitative study—using Word Error Rate (WER) metric [14] and manual human evaluation—on a test set consisting of short speeches delivered by members of the EP in the 19 languages already deployed in the prototype.[6] The aim of this study is to evaluate the quality of both ASR and MT output and to reflect on the different insights of the same text given by different annotators. The manual human evaluation is based on an error taxonomy adapted from the Multidimensional Quality Metrics (MQM) framework[7] and applied to part of the test set covering 6 languages (EN, FR, ES, IT, RO, DE). The MQM framework, developed in the EU QTLaunchPad and QT21 projects, provides a hierarchy of translation errors that can be adapted according to the application. We devised our taxonomy consisting of different error categories for ASR and MT and 3 severity levels (i.e. neutral, minor and major). We decided to exclude critical errors as in [10]. The remainder of the paper is organised as follows: in Section 2 we describe the methodology applied for automatic and human evaluation; in Section 3 we report the results quantitatively and qualitatively analysed per language and per annotator; Section 4 concludes the paper.

2. Methodology description

We evaluated ASR using both automatic metrics, in particular WER and human evaluation, and MT only relying on human evaluation. Human evaluation for both ASR and MT is carried out under the MQM framework. We describe the procedures and the experimental setup in the subsequent sections.

2.1. Automatic evaluation

Automatic evaluation was used only for ASR. The metric used is WER.[8] The test set consists of 92 short speeches (minimum = 01:01; maximum = 05:10; average = 01:39; standard deviation = 00:39) delivered in March and May 2022 plenaries by members of the EP. The speeches are in the 19 languages deployed in the tool by November 2022. See Table 1 for more details. Languages are ordered according to deployment in the tool.[9]

The gold standard of this test set is made of the verbatim transcription of the speeches (often referred to by its French abbreviation, CRE, *Compte Rendu d'Evènement*), manually corrected from the published report available on the website of the EP.[10] The corrections are performed by two native speakers per language and a third annotator is involved to solve disagreement.

2.2. Human evaluation

The error taxonomy is germane to the MQM framework and includes different categories for ASR and MT, sharing the same severity scale—i.e. neutral, minor, major (see Figure 1 in Appendix A for the decision tree). The

[3] One (in)famous claim is that MT has achieved human parity [11, 12, 13].

[4] Specifications of the Innovation Partnership are available here: https://etendering.ted.europa.eu/cft/cft-document.html?docId=58722. All links were last access on 13/05/2023.

[5] Deaf with a capital D denotes individuals who are culturally and linguistically Deaf, often due to congenital deafness or early-life hearing loss. They identify with the Deaf community, characterized by its unique culture, sign languages, and traditions. In contrast, deaf (with a lowercase d) is a general term referring to individuals with a hearing impairment, irrespective of their cultural identification or community affiliation. It describes the audiological condition of partial or complete hearing loss, without specifying sign language usage or cultural ties.

[6] This study and paper was written while the author was working for the European Parliament Unit in charge of the prototype management and evaluation. This is not the official evaluation methodology employed by the Parliament to evaluate the prototype.

[7] MQM website available here: https://themqm.org/.

[8] Script written by Dr. Claudio Fantinuoli available here: https://github.com/fantinuoli/WERvisual/blob/main/wer.py.

[9] During Stage 1 of the project 10 language models were deployed; during Stage 2, 9 other language models were added. This order is maintained in Table 1. Stage 1 models were trained during 2020-2021, Stage 2 languages during 2021-2022. Stage 1 ASR models have been updated in August 2022. Both ASR and MT models are developed by Cedat85 consortium.

[10] Each parliamentary sitting is publicly available and the CRE and videos in the original language are available in the EP website: https://www.europarl.europa.eu/plenary/en/debates-video.html.

Language	# speeches	Time (hh:mm:ss)
EN	9	00:14:56
FR	3	00:04:52
DE	4	00:06:28
ES	4	00:05:03
IT	7	00:15:31
RO	4	00:04:37
PL	10	00:19:10
EL	1	00:01:05
NL	1	00:01:21
PT	1	00:01:10
BG	4	00:07:19
CS	4	00:07:58
SK	4	00:05:42
SL	4	00:06:10
HR	4	00:05:59
LT	4	00:06:13
FI	8	00:13:20
HU	9	00:13:03
SV	7	00:12:00
Total	92	02:31:57

Table 1
ASR automatic evaluation test set.

error categories used for ASR error annotation are *over-segmentation*, *under-segmentation*, *lexical substitution*, *lexical deletion*, *lexical addition*, *morpho-syntactic errors* (e.g. number agreement, part of speech substitution), *terminology* (e.g. named entities and terms). The categories used for MT error annotation are *accuracy* (e.g. meaning is not rendered in its entirety), *punctuation*, *grammar*, *register* (formality, gender-marked pronouns), *terminology* (including the presence of non-words, spelling errors or incorrect terms), *other* and *unintelligible*. *Unintelligible* is used to mark segments containing more than 5 major errors [10]. *Other* should be used in rare cases in which none of the existent error categories apply. Neutral errors weight 0 points, minor errors 1, major 5. Except for unintelligible which weights 5, if minor, and 25, if major. These weights are similar to those used in [10].

We involved four annotators. All received the annotation guidelines and a training. After a few annotations a further meeting was scheduled to clear doubts. We involved four annotators with different backgrounds and knowledge of the languages. For reference, we call them annotator A, B, C and D (henceforth, Ann for annotator). Ann A has a background in Translation studies and is an experienced translator at the EP. Ann B was a trainee at the EP with a master's degree in Translation and previous experience on the MQM framework. Ann C was a trainee at the EP with a master's degree in Translation and no previous experience on ASR and MT evaluation. Ann D is a communications assistant at the EP with a background in interpretation, with no experience on the MQM framework, but with experience on ASR and MT evaluation. In Table 2 we report their self-reported knowledge of the languages according to the CEFR levels.[11]

Annotator	Language	CEFR level
Ann A	RO	Native language
	EN	C2
	IT	C2
Ann B	IT	Native language
	RO	Native language
	EN	C2
	ES	C2
	DE	C2
Ann C	IT	Native language
	EN	C1
	FR	C1
	ES	B2
Ann D	IT	Native language
	EN	C2
	FR	C1

Table 2
Annotators and language knowledge.

The annotated test set consists of 48 documents in 6 languages (EN, IT, FR, ES, RO, and DE): 18 automatic transcriptions (3 speeches per language, with an identification number from 1 to 6) and 30 translations (from and into the 6 above mentioned languages). In Table 3 we report the evaluated task and the involved languages, the number of speeches (with an identification number between brackets to be able to identify them when used as source and target and also in the automatic evaluation results reported in Table 4), and the annotators providing the annotations.

3. Results

3.1. Automatic evaluation

ASR was evaluated in two different scenarios: first, in sessions with more than one speech but all in the same language; second, in sessions with more than one speech, each in a different language. This is possible because the tool has a feature called Language Identification (LID), which is used to identify the language spoken and subsequently transcribe the audio in the identified language. WER results (computed per session) are reported in Table 4. In the table body, from row 2 to 7, we report the WER obtained in the speeches also undergoing human evaluation. This is the reason why we have multiple rows for the same language (e.g. EN LID on with id code

[11] For more details about the CEFR levels see the website: https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions.

Evaluated task	# speeches (source id)	Annotations
ASR RO	3 speeches (1)	Ann A – B
ASR IT	3 speeches (2)	Ann B – C
ASR EN	3 speeches (3)	Ann B – C
ASR ES	3 speeches (4)	Ann B – C
ASR FR	3 speeches (5)	Ann C – D
ASR DE	3 speeches (6)	Ann B
MT EN-IT	3 speeches (3)	Ann B – C
MT EN-RO	3 speeches (3)	Ann A – B
MT EN-FR	3 speeches (3)	Ann C – D
MT EN-ES	3 speeches (3)	Ann B
MT EN-DE	3 speeches (3)	Ann B
MT RO-IT	3 speeches (1)	Ann A – B
MT IT-EN	3 speeches (2)	Ann B – C
MT ES-IT	3 speeches (4)	Ann B – C
MT FR-IT	3 speeches (5)	Ann C – D
MT DE-IT	3 speeches (6)	Ann B

Table 3
Human evaluation test set.

3, indicating the speeches subjected to human evaluation, and then again EN LID on, non subjected to human evaluation).

Language (source id)	LID	WER
All 19, 1 speech each	On	6.45
RO (1)	On	2.77
IT (2)	On	3.22
EN (3)	On	8.98
ES (4)	On	4.94
FR (5)	On	8.91
DE (6)	On	7.81
EN	Off	5.25
EN	On	5.48
IT	Off	5.58
BG	Off	5.83
PL	Off	5.05
PL	On	7.80
HU	Off	9.18
HU	On	9.57
CS	Off	4.03
SK	Off	2.52
SL	Off	5.02
HR	Off	5.63
LT	Off	11.14
FI	Off	5.48
SV	Off	10.78

Table 4
ASR: Averaged WER results. Source id in brackets links the speeches with those in Table 3.

The results show that LID does not have a big impact on WER (e.g. EN, HU), except for PL (almost 3% WER difference), but the main difference in WER is due to different speeches (e.g. IT, in which the 3 speeches with LID on have a lower WER than the 3 with LID off, or EN with LID on in different interventions with more than 4% WER difference).[12]

3.2. Human evaluation

We investigated manual annotation quantitatively and qualitatively. Quantitative evaluation is based on an average score per document and annotator. Qualitative evaluation takes error categories and severities into account.

3.2.1. Quantitative evaluation

Same speeches, different annotators. For each annotator, we calculated a score per document by averaging the segment-level scores. Results are shown in Figures 2–6 in Appendix A. In general, ASR output received higher scores than expected, especially in languages in which WER is lower than 5% (i.e. RO, IT, ES). This can be due to the fact that WER metric does not take punctuation into account, thus over- and under-segmentation issues are not counted. Also, in WER calculation all errors have the same weight (e.g. a missing negation changing completely the meaning of the sentence has the same weight of any other missing token). MT output received lower scores if compared to ASR output. This might mean that some errors in ASR are well handled in translation. When both annotators are native speakers of the target language, their scores are more similar. This is the case of Ann A and B in EN-RO (Figure 2), Ann B and C in EN-IT (Figure 4), Ann B and C in ES-IT (Figure 5). The same applies to Ann C and D (Figure 6) in the annotation of FR-IT MT, although Ann D displays a different annotation behaviour than Ann B and C. In fact, Ann D tends to annotate fewer errors. This could be influenced by their different backgrounds (interpreter vs. translator).

The annotation scores are the most similar when the annotators are native speaker of the target language, as in the annotation of IT-EN MT (Figure 3) and EN-FR MT (Figure 6). However, monolingual annotators (Ann A and C) show more severity in MT judgement into their native language (Figure 2 RO-IT MT and Figure 4 EN-IT MT) when compared with our IT-RO bilingual annotator (Ann B). This seems in line with research on bilingualism and acceptability, where results show that "bilinguals do not reject ungrammatical items with the same certainty as monolinguals" [15].

Averaging the scores attributed by the two annotators (ASR and translation into IT using the ASR output as source, except for IT, that is translated into EN), we obtain the following order (from the presumably best output to the worse): ES (average = 17.3), IT (average = 24.5), FR

[12]Please note that this could be due to pure chance and since the test set is small, we do not report statistical tests.

(average = 24.7), EN (average = 25.3), RO (average = 40.8). The order considering WER would be RO, IT, ES, FR, EN.

Same annotator, different speeches. Here we compare the annotations carried out by Ann B and C. We selected these two annotators because they performed the majority of the annotation task, so it is possible to compare their results in different languages. We report their scores in Figure 7–8, respectively (Appendix A).

According to Ann B (Figure 7), we can order the languages from the best output to the worse: ES (average = 14.33), IT (average = 27.00), DE (average = 29.44), EN (average = 36.67) and RO (average = 46.33). According to Ann C (Figure 8), the order would be: ES (average = 13.67), IT (average = 18.50), EN (average = 26.33) and FR (average = 27.50).

3.2.2. Qualitative evaluation

Each annotator draws a different picture of each text, being that the product of ASR or MT. As far as ASR output is concerned, we report the results in Figures 9–13 in Appendix A.

Despite we did not put a major emphasis on over- and under-segmentation errors during training, as they were considered to be straightforward (at least in their identification), the disagreement in annotations suggests the contrary. In fact, different annotators draw opposite pictures of their presence and importance. For example, in Figure 9, we can notice that Ann A weighs more over-segmentation than under-segmentation errors in RO transcription, while Ann B does the opposite. The system results on punctuation marking, and full stop identification, in particular, seems to be below state-of-the-art performance [16]. Ann C (Figures 10–13) seems to be more severe about morpho-syntactic errors in ASR. The same errors are annotated as lexical substitutions by the other annotators, as in Example 1.

(1) **REF:** Protéger les citoyens de la haine en ligne, voici un bel usage **des** technologies les plus avancées. **Et** voici aussi un usage très approprié de l'Union européenne.
"Protecting citizens from online hate, here is a good use **of the** most advanced technologies. **And** here a very appropriate use by the European Union."
ASR: Protéger les citoyens de la haine en ligne. Voici un bel usage**.** **Les** technologies les plus avancées. **En** voici aussi un usage très approprié de l'Union européenne.
"Protecting citizens from online hate. Here is a good use. The most advanced technologies. Here also a very appropriate use by the European Union."
FR-IT: Proteggere i cittadini dall'odio online. Qui è un buon uso. Le tecnologie più avanzate. Anche questo è un uso molto appropriato da parte dell'Unione europea.
"Protecting citizens from online hate. Here is a good use. The most advanced technologies. This is also a very appropriate use by the European Union."

Ann C marked the errors in bold in Example 1 as morpho-syntactic errors of a major nature, Ann D as lexical substitution of a minor nature. This is a blurry area, if you consider that both are functional words and in other languages could be rendered morphologically. We think that in a multilingual perspective, these should be treated as morphological being functional words. However, probably they are not major errors, as they do not affect a main idea of the speech (decision tree in Figure 1).

As far as MT output is concerned, we report the results in Figures 14– 20 in Appendix A. We notice the inappropriate use of the unintelligible category. Unintelligible should mark segments in which it is impossible to understand the message and to identify all the errors that led to the incomprehensible segment. The fact that on the same set, different annotators used it or not, it is a clear sign of misunderstanding (Figures 15, 17, 18 and 19). In fact, in Example 2, unintelligible is used in a segment in which it is possible to understand the meaning, although there is a minor grammatical error (*attaccano* 'they attack') and a minor accuracy error (relative clause instead of adverbial clause, *che* 'that' substituting *per* 'to').

(2) **REF:** [. . .] state precum Federația Rusă utilizează instrumentele moderne pentru **a ataca** state, pentru a ataca entități, pentru a pune în pericol democrația europeană**, acest** lucru necesită un răspuns **rapid** și unit.
"[. . .] countries like the Russian Federation use modern tools to attack states, to attack entities, to endanger European democracy, this requires a rapid and united response."
ASR: State precum Federația Rusă utilizează instrumentele moderne pentru **a. Ataca** state pentru a ataca entități pentru a pune în pericol democrația europeană**. Acest** lucru necesită un răspuns**. Rapid** și unit, [. . .]
FR-IT: Paesi come la Federazione Russa usano strumenti moderni per **Attaccano** gli Stati per attaccare entità **che** mettono in pericolo la democrazia europea**. Ciò** richiede una risposta**. Veloce** e unito,

In Example 2 we also notice over-segmentation errors in the ASR transcription cascading in MT (*Acest lucru necesită un răspuns. Rapid și unit*). In addition, it seems that Ann B, and in other examples Ann C, annotated the output as if it was a written text and not an oral text transposed in written. Thus, the reference text is only one of the possible transpositions. This is evident looking at punctuation. In Example 2, in fact, Ann B not only marked the over-segmentation error dividing the noun *răspuns* from its modifiers (*rapid și unit*), but also another over-segmentation error (despite marked as minor) because in the reference this sentence is joint to the preceding one with a comma and not divided by a full stop. However, it must be noted that a full stop there is perfectly acceptable.

Unintelligible errors were also marked when the other annotator only noticed punctuation issues, as shown in Example 3.

(3) **REF:** Putin has thrown the world and Europe back to a time we had hoped never to experience again**. A crisis** of such dignity shows our true **colours** – if we are on the right side of history or choose the [path] path of destruction.
ASR: Putin has thrown the world and Europe back to a time, we would hope to never experienced again **crisis** of such dignity shows our true **colours** if we are on the right side of history or choose the path path of destruction.
EN-FR: Poutine a renvoyé le monde et l'Europe à une époque, nous espérons ne plus avoir connu **de crise** de cette dignité montre nos vraies **couleurs** si nous sommes du bon côté de l'histoire ou si nous choisissons le chemin de la destruction.

(4) **REF:** [...] Ceux qui ont harcelé et appelé au meurtre **sur Internet Samuel Paty**, sont-ils, étaient-ils, des vecteurs de liberté d'expression? Poser la question, c'est déjà y apporter une réponse.
"Were those who harassed and called for the murder of Samuel Paty on the Internet vectors of freedom of expression? To ask the question is to answer it."
ASR: Ceux qui ont harcelé. **Su Internet. Internet. Jsem jej petic.** Et appelé au meurtre sur Internet, Samuel Paty. Sont-ils, étaient-ils des vecteurs de liberté d'expression. Poser la question c'est déjà y apporter une réponse.
EN-IT: Coloro che hanno molestato. **Su internet. Internet. Sono una petizione.** E ha chiesto omicidio su Internet, Samuel Paty. Sono loro, erano vettori della libertà di espressione. Fare la domanda è già fornire una risposta.
"Those who harassed. **On the Internet. Internet. They are a petition.** And called for murder on the internet, Samuel Paty. It's them, they were vectors of freedom of expression. To ask the question is already to provide an answer."

An actual unintelligible error is instead reported in Example 4. LID errors in this case caused unintelligibility in the translation because the same portions of audio were transcribed in different languages (transcribed as IT, PL, and CS). Perhaps including the information about the source language in the translated output could be useful to reduce the impact that LID errors like these have in the MT understanding. Morpho-syntactic errors annotated in the ASR are frequently correct in the MT output. Over-segmentation, instead, in particular when involves a full stop, remains unchanged in the MT output, as MT models usually mirror the punctuation of the source text.

4. Conclusion

We presented a quantitative and qualitative evaluation of the tool that has been developed in the context of a EP's Innovation Partnership. We used WER score and human manual evaluation to evaluate the quality of ASR, and only human evaluation for MT quality. The average WER is 6.43% in the multilingual test set made of 19 languages deployed by November 2022, which is very low but it does not take into account segmentation issues. Human evaluation highlighted the need for refining sentence segmentation, especially in languages in which the WER was very low (e.g. RO and IT). This could indicate that WER by itself is not enough to have a clear picture of the quality of the transcription. However, human evaluation remains a highly subjective task which attains all categories, also those considered clear-cut categories (e.g. sentence segmentation). The annotators' background has an influence on error severity perception and error identification, and should be investigated in detail. In line with what found in [17], we also found that annotators' sensitivity in deepening the error annotation is a main cause of disagreement, in this case due to the attempt to annotate also the consequences of the error. Quantitative results of human evaluation considering the ASR output and its translation into IT (except for IT translated into EN) indicate ES as qualitatively better output, followed by IT, FR and EN, and RO as the worse output. In general, annotators rated ASR output worse than the MT output. However, this might be a consequence of the attitude of annotators putting too much emphasis on the provided reference transcription of the speech, not considering that, especially if punctuation is concerned, it is only one of the possible accepted transpositions. Qualitative results highlighted that different annotators draw different pictures of the same speeches and that a second round of annotations would be necessary to reduce disagreement and to clarify the use of error categories, like unintelligible, frequently improperly applied.

Acknowledgments

I want to thank the European Parliament for giving me the opportunity to conduct this study and the annotators for participating and giving me the authorisation to use their annotations for research purposes. I thank the anonymous reviewers for their precious comments, and I apologies if not all of them have been addressed.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: 31st Conference on Neural Information Processing Systems (NIPS 2017), California, USA, 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[2] O. Bojar, C. Buck, C. Federmann, B. Haddow, P. Koehn, J. Leveling, C. Monz, P. Pecina, M. Post, H. Saint-Amand, R. Soricut, L. Specia, A. Tamchyna, Findings of the 2014 workshop on statistical machine translation, in: Proceedings of the Ninth Workshop on Statistical Machine Translation, Association for Computational Linguistics, Baltimore, Maryland, USA, 2014, pp. 12–58. URL: https://aclanthology.org/W14-3302. doi:10.3115/v1/W14-3302.

[3] L. Bentivogli, M. Cettolo, M. Gaido, A. Karakanta, A. Martinelli, M. Negri, M. Turchi, Cascade versus direct speech translation: Do the differences still make a difference?, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 2873–2887. URL: https://aclanthology.org/2021.acl-long.224. doi:10.18653/v1/2021.acl-long.224.

[4] M. Sperber, M. Paulik, Speech translation and the end-to-end promise: Taking stock of where we are, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7409–7421. URL: https://aclanthology.org/2020.acl-main.661. doi:10.18653/v1/2020.acl-main.661.

[5] A. Anastasopoulos, L. Barrault, L. Bentivogli, M. Zanon Boito, O. Bojar, R. Cattoni, A. Currey, G. Dinu, K. Duh, M. Elbayad, C. Emmanuel, Y. Estève, M. Federico, C. Federmann, S. Gahbiche, H. Gong, R. Grundkiewicz, B. Haddow, B. Hsu, D. Javorský, V. Kloudová, S. Lakew, X. Ma, P. Mathur, P. McNamee, K. Murray, M. Nǎdejde, S. Nakamura, M. Negri, J. Niehues, X. Niu, J. Ortega, J. Pino, E. Salesky, J. Shi, M. Sperber, S. Stüker, K. Sudoh, M. Turchi, Y. Virkar, A. Waibel, C. Wang, S. Watanabe, Findings of the IWSLT 2022 evaluation campaign, in: Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022), Association for Computational Linguistics, Dublin, Ireland (in-person and online), 2022, pp. 98–157. URL: https://aclanthology.org/2022.iwslt-1.10. doi:10.18653/v1/2022.iwslt-1.10.

[6] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318. URL: https://aclanthology.org/P02-1040. doi:10.3115/1073083.1073135.

[7] B. Dorr, J. Olive, J. McCary, C. Christianson, Machine translation evaluation and optimization, in: J. Olive, C. Christianson, J. McCary (Eds.), Handbook of Natural Language Processing and Machine Translation, Springer, 2011, pp. 745–843.

[8] J. Moorkens, S. Castilho, F. Gaspari, S. Doherty, Translation quality assessment, Machine translation: Technologies and applications, Springer, 2018.

[9] E. Chatzikoumi, How to evaluate machine translation: A review of automated and human metrics, Natural Language Engineering 26 (2020) 137–161.

[10] M. Freitag, G. Foster, D. Grangier, V. Ratnakar, Q. Tan, W. Macherey, Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation, in: Transactions of the Association for Computational Linguistics, volume 9, 2021, pp. 1460–1474. URL: https://doi.org/10.1162/tacl_a_00437. doi:10.1162/tacl_a_00437.

[11] H. Hassan, A. Aue, C. Chen, V. Chowdhary, J. Clark, C. Federmann, X. Huang, M. Junczys-Dowmunt, W. Lewis, M. Li, S. Liu, T.-Y. Liu, R. Luo, A. Menezes, T. Qin, F. Seide, X. Tan, F. Tian, L. Wu, S. Wu, Y. Xia, D. Zhang, Z. Zhang, M. Zhou, Achieving Human Parity on Automatic Chinese to English News Translation, 2018. arXiv:1803.05567.

[12] A. Toral, S. Castilho, K. Hu, A. Way, Attaining the unattainable? reassessing claims of human parity in neural machine translation, in: Proceedings of the Third Conference on Machine Translation: Research Papers, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 113–123. URL: https://aclanthology.org/W18-6312. doi:10.18653/v1/W18-6312.

[13] S. Läubli, R. Sennrich, M. Volk, Has machine translation achieved human parity? a case for document-level evaluation, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 4791–4796. URL: https://aclanthology.org/D18-1512. doi:10.18653/v1/D18-1512.

[14] S. Nießen, F. J. Och, G. Leusch, H. Ney, An evaluation tool for machine translation: Fast evaluation for MT research, in: Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00), European Language

Resources Association (ELRA), Athens, Greece, 2000. URL: http://www.lrec-conf.org/proceedings/lrec2000/pdf/278.pdf.
[15] J. C. López Otero, On the acceptability of the spanish dom among romanian-spanish bilinguals, in: A. Mardale, S. Montrul (Eds.), The Acquisition of Differential Object Marking Trends in Language Acquisition Research, John Benjamins Publishing Company, 2020, pp. 161–181.
[16] O. Guhr, A.-K. Schumann, F. Bahrmann, H.-J. Böhme, FullStop: Multilingual Deep Models for Punctuation Prediction, in: Swiss Text Analytics Conference, 2021.
[17] E. Di Nuovo, Introducing VALICO-UD: a parallel, learner Italian treebank for language learning research , Pàtron Editore, 2023.

A. Figures

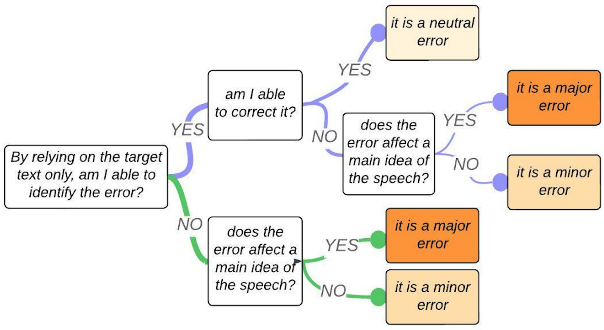


Figure 1: Decision tree used to annotate error severity.

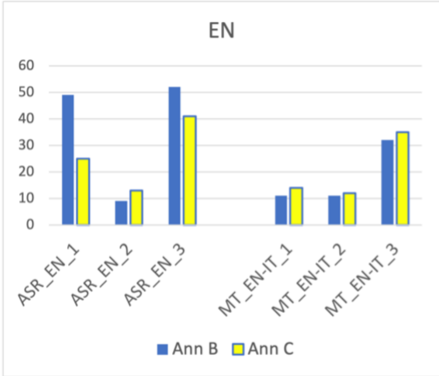


Figure 4: Ann B and C annotations of EN ASR and EN-IT MT.

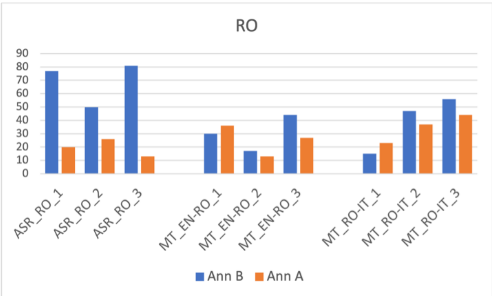


Figure 2: Ann A and B annotations of RO ASR, EN-RO MT and RO-IT MT.

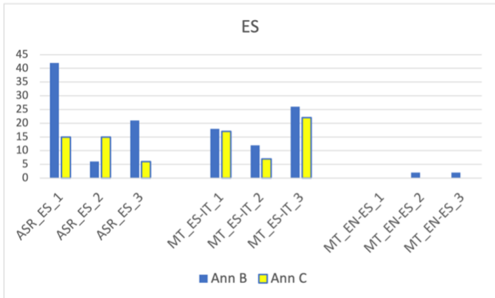


Figure 5: Ann B and C annotations of ES ASR and ES-IT MT; Ann B annotations of EN-ES MT.

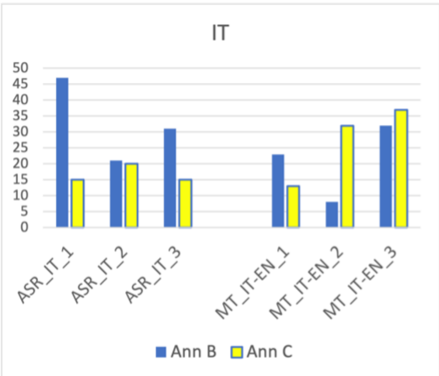


Figure 3: Ann B and C annotations of IT ASR and IT-EN MT.

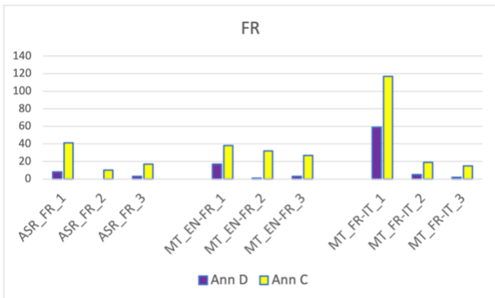


Figure 6: Ann C and D annotations of FR ASR, EN-FR MT and FR-IT MT.

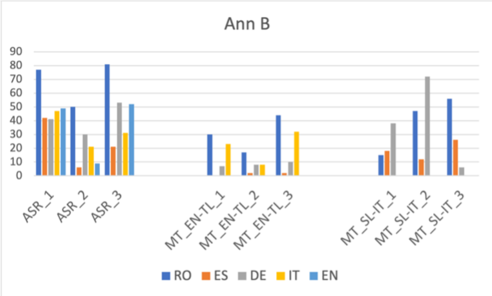


Figure 7: RO, IT, EN, ES and DE annotated by Ann B. In the figure, SL stands for Source Language, TL for Target Language.

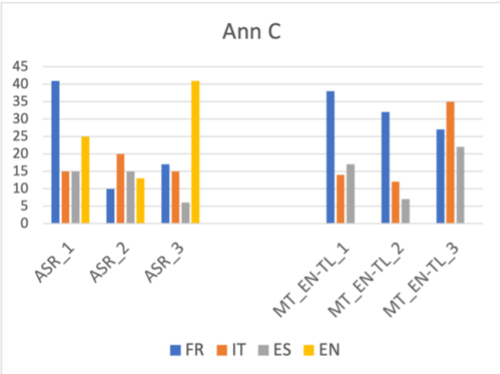


Figure 8: IT, EN, ES, and FR annotated by Ann C. In the figure, SL stands for Source Language, TL for Target Language.

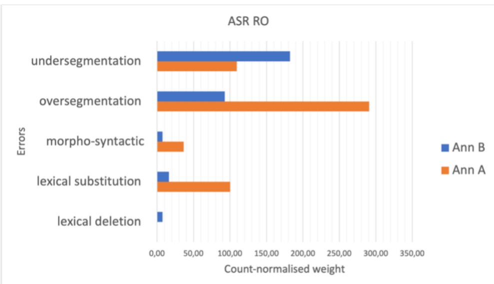


Figure 9: Error categories as annotated by Ann A and B in ASR RO.

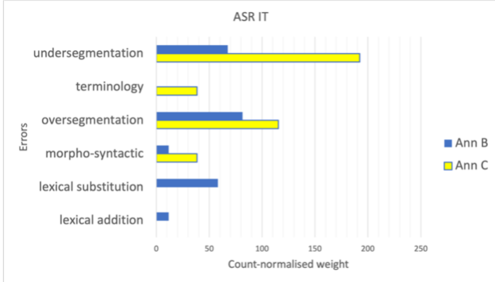


Figure 10: Error categories as annotated by Ann B and C in ASR IT.

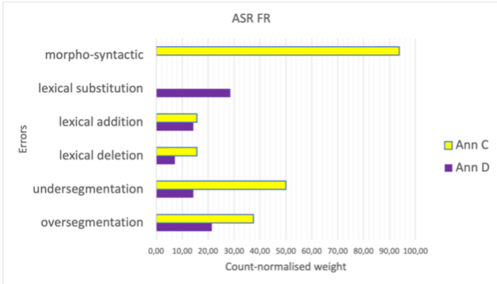


Figure 11: Error categories as annotated by Ann C and D in ASR FR.

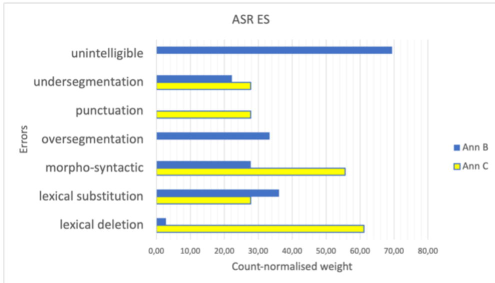


Figure 12: Error categories as annotated by Ann B and C in ASR ES.

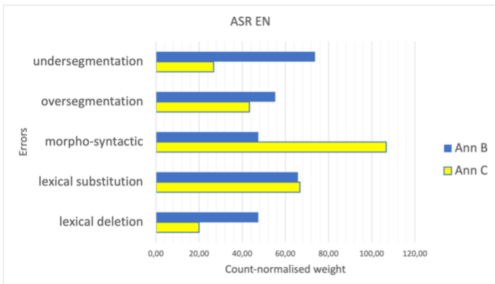


Figure 13: Error categories as annotated by Ann B and C in ASR EN.

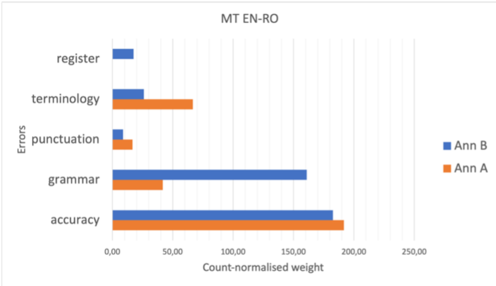


Figure 14: Error categories as annotated by Ann A and B in MT EN-RO.

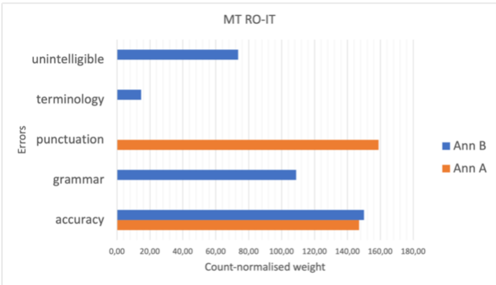


Figure 15: Error categories as annotated by Ann A and B in MT RO-IT.

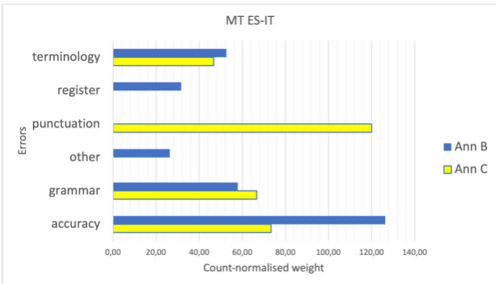


Figure 16: Error categories as annotated by Ann B and C in MT ES-IT.

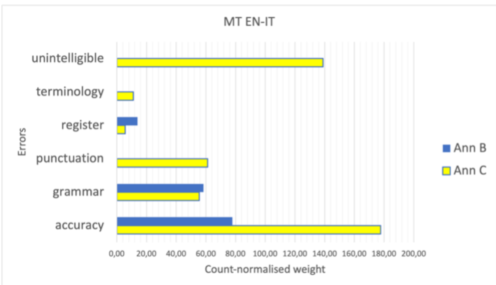


Figure 17: Error categories as annotated by Ann B and C in MT EN-IT.

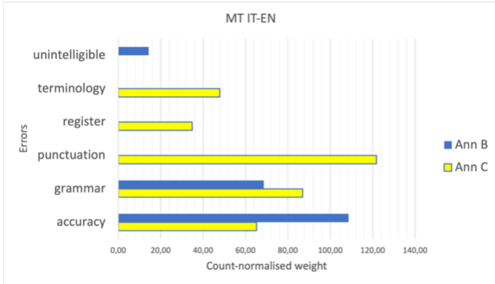


Figure 18: Error categories as annotated by Ann B and C in MT IT-EN.

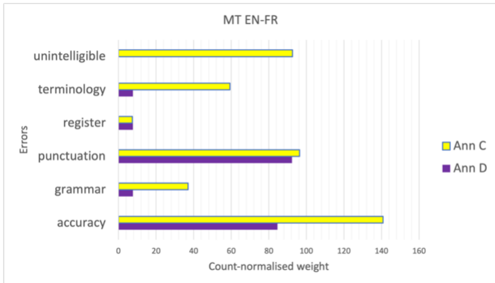


Figure 19: Error categories as annotated by Ann C and D in MT EN-FR.

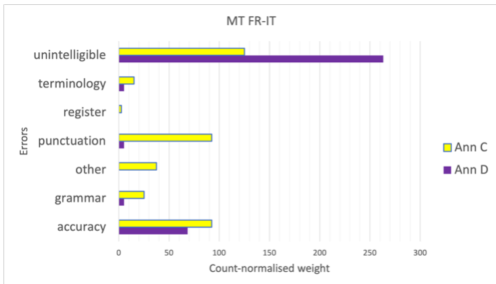


Figure 20: Error categories as annotated by Ann C and D in MT FR-IT.

A Cognitive Linguistics analysis of Phrasal Verbs' representation in Distributional Semantics

Melissa Donati[1], Carlo Strapparava[2,*]

[1]Alma Mater Studiorum - University of Bologna, Italy
[2]FBK-irst, Trento, Italy

Abstract
Phrasal Verbs (PVs) constitute a peculiar feature of the English language and represent a challenge for both language learners and computational models because of their complex and idiomatic nature, which has made them appear unsystematic and unpredictable. Recently, Cognitive Linguistics has offered a more systematic explanation of the semantics of PVs by relating their non-compositional meanings to the metaphorical extensions of the particle's meaning. In order to assess the computational suitability of this approach using Distributional Semantics, we analyzed three different semantic spaces to understand how PVs and particles are represented and whether any of the embeddings capture the significance of particles in the semantics of the entire construction. The results indicate that phrase embeddings are effective in representing the meanings of PV constructions, while word embeddings excel at capturing particle meanings and additionally support the Cognitive Linguistics hypothesis. Since improving the semantic representation of PVs can benefit various NLP applications, further research is necessary to validate these findings.

Keywords
Phrasal Verbs, Cognitive Linguistics, Distributional Semantics

1. Introduction

Phrasal Verbs (PVs) represent a distinctive peculiarity of the English language and are defined as a lexicon unit composed of a verb (e.g. *look*) and a particle (e.g. *out*), whose meaning is often non-compositional (e.g. *look out* means 'to beware') [1, 2]. PVs comprise a significant portion of the verb vocabulary [3] and are commonly used in everyday language, particularly in spoken and informal contexts [4, 5]. In addition, they are highly productive, with new ones continually being coined to reflect societal changes (e.g. *google up*) [6]. They are also characterized by their polysemy, with each phrasal verb having multiple meanings on average [5]. To further enhance their complexity, for a long time, linguists and grammarians have claimed that the selection of verb and particle in the PV construction is totally unsystematic and unpredictable [6, 7, 8], therefore the traditional pedagogical approach to PVs has always been based on memorization of the verb-particle combination and the corresponding meaning causing a general discouragement of learners around PVs. Recent research in Cognitive Linguistics, however, has proposed a more systematic explanation of the association between these verb-particle combinations and their apparently randomly assigned idiomatic meaning, by suggesting that it is the particle (in particular its metaphorically extended meaning [9]) that plays a crucial role in shaping the overall meaning of the PV [2]. Given that this approach has shown promising results in language learning [10], and that computational models of language face difficulties that are similar to English as a Second Language (ESL) learners in understanding the semantic complexity of PVs, we wanted to examine whether such Cognitive Linguistics account holds also from a Distributional Semantics perspective, where the representation of words' meanings and semantic relationship have repeatedly been proved to be similar to the way they are represented in the human cognitive system [11]. With this aim, we analysed three different semantic spaces –word embeddings, phrase embeddings and POS-tagged embeddings– to determine the most accurate way of representing PVs and particles and whether the Cognitive Linguistics hypothesis was accounted for in any of them. The importance of the particle in shaping the meaning of PVs was confirmed but the results appeared to vary across the different semantic spaces, suggesting the need for further and more detailed research.

2. Related Works

Phrasal Verbs have for long been a hot topic among linguists and lexicographers who have largely debated on their definition and classification, proposing various theories based on their syntactic and semantic features [6, 12, 4, 13, 14]. Corpus linguistics has also played a crucial role in studying PVs, providing insights into their frequency and meaning distribution, thus aiding language

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
melissa.donati@studio.unibo.it (M. Donati); strappa@fbk.eu (C. Strapparava)

teacher by identifying the most frequently used PVs and their meanings [4, 5, 15, 16, 17]. However, to develop effective teaching strategies for PVs, it is essential to consider the cognitive processes involved in storing and retrieving these structures from the mental lexicon and that is where the Cognitive Linguistics approach comes into play offering a new perspective that considers PVs as conceptually motivated constructions rather than arbitrary combinations [18, 19, 20, 2, 21]. This account is based on one of the cornerstones of Cognitive Linguistics which is Metaphor Theory. According to this view, metaphors play a fundamental role in conceptualization and thinking as they allow us to understand and experience abstract concepts by mapping them onto concrete entities that we can bodily perceive [9]. In the case of PVs, the Cognitive Linguistics account considers the metaphorical extension of the particle's literal meaning as responsible for the idiomatic meaning of the entire construction [2, 22], unlike traditional approaches that often neglect the semantic role of particles, and function words more in general [23]. According to this view, the prototypical meaning of particles, which is usually related to spatial locations and orientations, can be metaphorically extended to abstract non-physical domains that are thought of in terms of space, such as attitudes, knowledge, completion, or increase [2]. For example, the particle "UP" literally denotes a physical upward motion (e.g. *to pick up*), but can be used to denote a number of abstract domains that we categorize assigning (spatial) values along a vertical line such as temperature, ranks, attitudes, knowledge etc. Therefore, the metaphorical extension of 'UP' can indicate improvement (e.g. *to brush up*), higher visibility and accessibility (e.g. *to turn up*), completion (e.g. *to fill up*), and reaching a a boundary (e.g. *to be fed up*) [2]. This perspective allows to get insights into the systematicity and predictability of PVs' semantics and demonstrates that idiomatic and polysemous meanings of PVs are connected through a network of senses derived from the prototypical meaning of the particle. Empirical studies have shown that adopting this more cognitively plausbile approach in the instruction of PVs can benefit ESL learners as it helps them grasp the relationship between idiomatic and literal meanings of PVs thus facilitating the processing and acquisition of these constructions [24, 25, 26, 27, 10]. Having briefly discussed the Cognitive Linguistics perspective concerning the semantics of PVs, we now briefly explore how the meaning of PVs is processed computationally, adopting a Distributional Semantics approach. Distributional Semantics is a computational approach to language meaning where words are represented as distributional vectors in a semantic space based on their contextual usage. The underlying assumption, referred to as the Distributional Hypothesis [28], is that words that occur in similar contexts tend to purport similar meanings and that the semantic similarity between two words can be measured as the geometrical distance between the vectors representing such words (for a more detailed explanation of the different frameworks see [11]; [29]). Even though this approach works extremely well for representing the meaning of single words, it faces some challenges when representing the meaning of PVs and multi-word expressions more in general due to their non-compositional and often polysemous meaning. Most studies addressed this issue by developing strategies to detect compositionality using dictionary-based [30, 31] and distributional similarity methods [32, 33, 34]. While providing effective working solutions, these compensation strategies do not fix the root problem at the level of semantic representation. Recently, DS has been extended to incorporate larger units, such as multi-word expressions and phrases, thus creating more informative embeddings and leading to better performance in NLP tasks [35, 36, 37, 38]. It is against this background that we framed our research questions and decided to investigate which type of embeddings could better represent the complex semantics of PVs and whether the distributional semantic space manages to capture the role of the particle's meaning in shaping the meaning of the entire construction, as posited by the Cognitive Linguistics account.

3. Methodology

In order to investigate how the semantics of PVs is represented within the Distributional Semantics framework, and more specifically to test whether distributed representations can capture the importance of particles in PV constructions, as suggested by the Cognitive approach, we analyzed three types of embeddings:

Word embeddings: we selected the pre-trained word vectors released by Google[1], which are 300-dimensional vectors trained using a Skip-gram model on a portion of the Google News dataset, with a window-size of 5. The Skip-gram model was selected because it has shown superior performance in semantic tasks compared to other models like CBOW, NNLM, and RNNLM [39]. The window-size of 5 allows capturing broader semantic information beyond immediate context, which is suitable for investigating the semantics of PVs with separable particles [40]. The vector size of 200-300 dimensions strikes a balance between informativeness and computational complexity [41].

Phrase embeddings: we selected the embeddings for generalized phrases introduced by [38]. They collected two-word phrases, categorized them as continuous or discontinuous, and trained a Skip-gram model to learn

[1]https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTT-lSS21pQmM/edit?resourcekey=0-wjGZdNAUop6WykTtMip30g

embeddings for both words and phrases. They showed that phrase embeddings outperformed word embeddings in semantic tasks, demonstrating their better representative power for such multi-word expressions, because considering them as linguistic units allows to capture the attributes of their real contexts of usage and thus to create accurate semantic representations that account for their non-compositional meaning.

POS-tagged embeddings: we included POS-tagged embeddings in our analysis to test whether building isolate representations for particles, distinguishing their occurrences in PV constructions from those of the same words used as prepositions or adverbs, would lead to more accurate semantic representation of the entire PV and could account for the Cognitive Linguistics hypothesis. We selected the English TreeTagger [42] trained on the PENN Treebank [43], which includes a specific tag for particles (RP) and used the BNC corpus for training. To ensure accurate identification of particles, we cross-checked the TreeTagger annotations comparing it with the dependency parsing annotation performed with SpaCy Dependency Parser [44] and corrected any misclassifications. Finally, we trained Word2vec on the double-checked POS-tagged BNC data with a Skip-gram architecture, a window size of 5, and 300-dimensional vectors. In compliance with the hyperparameter setting of word embeddings, we adopted the Skip-gram architecture, setting the window-size to 5 and the vectors' dimension to 300 [39].

4. Experiments

We conducted exploratory analyses on the three different semantic spaces to investigate how PV constructions and particles are distributionally represented. We selected as target verbs the 150 most frequent PVs identified by previous corpus-based studies [15]. To determine the meanings of these PVs, we referred to the PHaVE List [16], which provides key meaning senses based on frequency distributions. Similarly, we addressed the challenge of representing particle meanings, which are often overlooked in Distributional Semantics, by turning to [2] work for detailed meaning descriptions of particles and we selected simplified synonyms to evaluate the accuracy of particles' semantic representations. For both PVs and particles, we evaluated their meaning representation using cosine similarity measures. The final lists of particle meanings used in our analysis can be found in Appendix A, while in Appendix B we reported a sample of PVs that were used in our analysis with the corresponding meanings selected from the PHaVE List [16]. Having defined the meaning of reference for PVs and particles, we designed three types of analyses to answer our research questions.

4.1. Semantic representation of PVs

In order to examine how the meaning of PVs is represented in the different semantic spaces, we measured the cosine similarity between the PV vectors and the vectors representing their possible meanings, as adapted from the PHaVE List [16] (see Appendix B for a few examples). If a single meaning consisted of multiple synonyms (e.g. for the PV *look out* the first meaning is represented by the pair *observe/contemplate*), we obtained the similarity score by summing the similarities of the PV with each synonym (i.e. sim(*look_out* - meaning 1) = sim(*look_out* - *observe*) + sim(*look_out* - *contemplate*)). In cases where the selected synonyms were multi-word expressions (e.g. for the PV *give out* the meaning 2 *make public*), in word and POS-tagged embeddings we obtained the vectors by summing the individual word vectors (i.e. vect(*make_public* = vect(*make)* + vect(*public*)), while for phrase embeddings, we checked if the multi-word expression was represented as a generalized phrase in the embeddings. If so, we used the corresponding embedding; otherwise, we obtained the vector through summation, similar to word and POS-tagged embeddings (for an explanation of the additive property of vectors see [39]).

4.2. Semantic representation of particles

In order to examine if and how the meaning of particles is represented in the different semantic spaces, similarly to the analysis conducted for PVs, we computed the cosine similarity between the vectors representing each particle and their possible meanings, adapted from [2]. If the meanings of particles consisted of multi-word expressions (e.g., for the particle *up* the corresponding meaning *positive verticality*), we obtained the vectors for the entire phrase by summing the individual word vectors (e.g., vect(*positive_verticality* = vect(*positive*) + vect(*verticality*)). Likewise, if the meaning of a particle was described by multiple synonyms (e.g., for the particle *on* the pair *contact/continuation*), also in this case we computed the similarity score by summing the similarity of the particle with each synonym (i.e., sim(*on* - meaning) = sim(*on* - *contact*) + sim(*on* - *continuation*)) [39].

4.3. Verb vs particle in the semantic representation of PVs

In order to test whether the distributional representations of meanings successfully capture the cognitive peculiarity of PVs' semantics, specifically the fact that particles play a significant role in shaping the meaning of the entire construction compared to the verb proper [2], we compared the similarity between the particle and the verb proper with the whole PV construction. For instance, considering the PV *set up* if the distributional representation effectively captures the Cognitive Linguistics account of PVs' semantics, we would expect the cosine similarity score between the entire PV and the particle (sim(*set up*–*up*)) to be higher than that between the PV and the verb proper (sim(*set up*–*set*)). In other words, the vector representing the meaning of the PV should be more similar, and therefore closer, to the vector representing the particle than to the vector representing the verb proper.

5. Results

Since one of the primary objectives of this work was to understand what could be the most appropriate way to build a distributional semantic representation of PVs that truthfully accounts for their complex semantics, we will now briefly[2] present and discuss the results of the three types of analyses that were carried out comparing the results obtained with the three semantic spaces.

5.1. Semantic representation of PVs

When evaluating the similarity between the PVs and their meanings to assess the quality of the semantic representation, overall significance was not high in any embeddings. However, phrase embeddings outperformed the others as they showed higher similarity scores for PV-meaning pairs (32%) compared to word (27%) and POS-tagged (17%) embeddings (see Table 1 for an example). This might be explained by the fact that phrase embeddings treat PVs as a single unit, capturing their real contexts of usage and therefore represent their semantic complexity more accurately. Conversely, word and POS-tagged embeddings, which summed the verb and particle vectors, fell short in capturing the full meaning of PVs. In conclusion, phrase embeddings proved to be the most suitable for representing PV semantics.

Table 1
Sample of our results showing the similarity scores obtained with the three different embeddings for the PV *take out* and its corresponding three meanings *remove, invite, obtain*. The highest similarity scores highlighted in bold are obtained, in all three cases, with phrase embeddings.

Meanings	Similarity scores		
take out	Word-e	Phrase-e	POS tag-e
remove	0.2918	**0.3894**	0.2422
invite	0.2773	**0.3491**	0.0819
obtain	0.2153	**0.2319**	0.0467

[2]For the sake of brevity, just a sample of the results is reported herein, for a comprehensive overview and a more detailed analysis see [45].

5.2. Semantic representation of particles

When evaluating the similarity between the particles and their meanings, we obtained similarity scores that were overall below the 0.5 significance threshold across the three types of embeddings. Surprisingly, comparing the results, word embeddings performed best, followed by phrase embeddings, while POS-tagged embeddings showed very poor performance (see Table 2 for a sample of the results). This was unexpected because it is only in POS-tagged embeddings that we could represent particles in isolated vectors, therefore they were supposed to capture more precisely their meaning. Conversely, in word embeddings the meaning representation of particles was collapsed in a single vector that included also occurrences of the same words when used with different syntactic functions (i.e. prepositions or adverbs), while in phrase embeddings the vectors representing the particle was actually built excluding the occurrences of the words as particles because those where captured within the phrase-vectors themselves. These findings point to two possible conclusions which are not mutually exclusive but rather complementary: on the one hand, they suggest that for building distributional representation of words that occur frequently in different syntactic roles –such as particles, and function words more in general– collapsing all the occurrences within a single vector representation might lead to better capturing their meaning, and on the other, they hint that particles used in PVs may have a core (prototypical) meaning that transcends their syntactic role, aligning with the Cognitive Linguistics hypothesis [2].

Table 2
Sample of our results showing the similarity scores obtained with the three types of embeddings for the particle *on* and its corresponding two meanings *contact* (m1) and *continuation* (m2), as well as the total meaning obtained by computing the similarity between the particle and the collapsed vector of the two meanings (see Section 4.2) The highest similarity scores, highlighted in bold, are obtained with word embeddings.

Meanings	Similarity scores		
on	Word-e	Phrase-e	POS tag-e
contact	**0.0388**	0.0284	-0.0122
continuation	**0.1421**	0.1075	-0.0345
tot meaning v(m1) + v(m2)	**0.1809**	0.1359	-0.0467

5.3. Verb vs particle in the semantic representation of PVs

In the final analysis, we compared the similarity between the PV and the particle and the PV and the verb across the three types of embeddings to test whether any of them

Table 3
Sample of our results showing the similarity scores obtained with the three types of embeddings for the verb *take out* compared to the verb proper (*take*) and the particle (*out*). The highest similarity scores, highlighted in bold, are obtained with word embeddings.

		Similarity scores	
PV	Embeddings	PV – v	PV – prt
take out	word	**0.784**	0.714
	phrase	**0.476**	0.373
	POS-tagged	0.645	**0.685**

supported the Cognitive Linguistics hypothesis about the role of particles in PV semantics. Broadly speaking, as expected word and POS-tagged embeddings performed better than phrase embeddings. They both gave overall significant similarity scores but showed the opposite patterns of similarity (see Table 3 for an example). Indeed, while in word embeddings it is the verb proper that resulted to be more similar to the PV compared to the particle, the opposite is true for POS-tagged embeddings where it is the particle that resulted to be more similar to the PV compared to the verb proper. These contrasting patterns align respectively with the traditional view and with the Cognitive Linguistics view (Section 2) on PVs meaning, and suggest that in a semantic space in which particles are accurately (i.e. separately) represented, the Cognitive Linguistic view claiming the higher significance of the particle (vs the verb proper) in shaping the PVs meaning, is supported and accounted for.

6. Conclusion and Future Directions

The aim of this work was to analyze the distributional representation of PVs from a Cognitive Linguistics perspective. More specifically we wanted to examine three different semantic spaces (word embeddings, phrase embeddings and POS-tagged embeddings) using simple vector combination (sum) and mathematical computations (cosine similarity) to evaluate whether: 1) the meaning of the PV construction is properly represented; 2) the particles' embeddings truthfully capture their meaning; 3) the greater role of the particle in shaping the semantics of the PV, as posited by the Cognitive Linguistics approach, is accounted for. The current results showed that, as expected, phrase embeddings performed best in capturing the complex semantics of PVs, supporting the idea of treating PVs as single tokens when training the embeddings so as to capture the true context of occurrence and obtain more accurate meaning representation that account also for the less compositional meanings.

As far as particles are concerned, word embeddings outperformed both phrase and POS-tagged embeddings. These results were unexpected because we anticipated better performance from POS-tagged embeddings, which were designed to isolate particle occurrences. We identified different possible explanations for these results, including limitations in integrating POS-tag information (that reduces the number of occurrences) compared to including occurrences with different syntactic functions, the issues related to the general difficulty of representing function words and also the possibility that words used as particles carry a unique core meaning regardless of their syntactic functions. Further research is needed, along these lines, to disentangle these factors and understand what could be the most effective way to represent particles and function words in Distributional Semantics.

Finally, the results of the third and last type of analysis, that was designed precisely to test the Cognitive Linguistics claim on the role of the particle, showed that when separate vector representations are built for particles, i.e. distinguishing the occurrences of the same words with other syntactic functions, as was the case with POS-tagged embeddings, particles do appear to play a greater role than the verbs proper in the semantics of the PVs. Conversely, when the vector representation of particles is less accurate, i.e. includes occurrences of the same words with other syntactic functions, as it was the case with word embeddings, it is the verb proper that appears to be more crucial in the semantics of the PV in most cases.

Overall our findings align with the literature, in that they support the idea that vectors for PVs should be treated as single tokens rather than splitting them into individual words [38] and that representing the meaning of particles is challenging, justifying their removal in many NLP applications [46]. However, we believe that understanding how to build appropriate semantic representations for particles is crucial for analyzing their contribution to larger constructions, such as PVs. In order to do so, future studies can explore different types of embeddings and testing whether refining POS-tagged embeddings (for example by weighting each POS-tag feature according to the task or alternatively using Neural Networks for combining these features into a unique meaningful hidden representation) could improve representation accuracy and thus lead to better performance of the models in specific semantic tasks.

Last but not least, our results provide initial evidence supporting the Cognitive Linguistics account of PV semantics from a Distributional Semantic perspective, although further confirmation is needed. Adopting the Cognitive Linguistics approach to PVs in education and leveraging NLP applications in this direction can facilitate the acquisition of this complex English structure for ESL learners. Additionally, capturing and representing PV meanings and the semantic roles of their components can benefit NLP tasks involving semantic and morphosyntactic relations (such as machine translation, question-answering, summarization, automatic synonym detection, etc.). For these reasons, we hope this work stimulates further advanced research in this area, leveraging the insights from Cognitive Linguistics and Computational Linguistics.

A. Appendix A. Particle's meanings

List of particles' meanings adapted from [2] that were used for the analysis.

Particles	Meanings
on	contact/continuation
up	positive verticality/increasing/completing
back	returning/past
out	leaving/exhaustion
in	entering/being inside
down	negative verticality/decreasing/ending
off	separation
ahead	progressing
over	crossing/overcoming
a/round	vicinity/proximity
through	crossing/completing
about	dispersion
along	parallel/accompanying

B. Appendix B. Sample of PVs' meanings

Small sample of PVs-meanings pairs as extracted from the PHaVE list [16].

PV	Meaning 1	Meaning 2	Meaning 3
get up	rise		
take out	remove	invite	obtain
go down	move	decrease	go
look out	observe / contemplate	take care / protect	
give out	give	make public	collapse / fail

Acknowledgements

We acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU.

References

[1] S. Thornbury, How to teach vocabulary, Pearson Education India, 2006.

[2] B. Rudzka-Ostyn, Word power: Phrasal verbs and compounds: A cognitive approach, Walter de Gruyter, 2008.

[3] W. Li, X. Zhang, C. Niu, Y. Jiang, R. K. Srihari, An expert lexicon approach to identifying english phrasal verbs, in: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, 2003, pp. 513–520.

[4] D. Biber, S. Johansson, G. Leech, S. Conrad, E. Finegan, Longman grammar of spoken and written english, 1999.

[5] D. Gardner, M. Davies, Pointing out frequent phrasal verbs: A corpus-based analysis, TESOL quarterly 41 (2007) 339–359.

[6] D. L. M. Bolinger, The phrasal verb in English, Cambridge University Press, 1971.

[7] L. Lipka, Semantic structure and word-formation (1972).

[8] K. A. Sroka, The syntax of English phrasal verbs, volume 129, Walter de Gruyter GmbH & Co KG, 2016.

[9] G. Lakoff, M. Johnson, Metaphors we live by, University of Chicago press, 2008.

[10] B. J. White, A conceptual approach to the instruction of phrasal verbs, The Modern Language Journal 96 (2012) 419–438.

[11] A. Lenci, et al., Distributional semantics in linguistic and cognitive research, Italian journal of linguistics 20 (2008) 1–31.

[12] R. Quirk, S. Greenbaum, G. Leech, J. Svartvik, A comprehensive grammar of the english language: Longman group limited (1985).

[13] M. C. Murcia, D. L. Freeman, The grammar book, Heinle & Heinle Publishers, 1999.

[14] C. M. Darwin, L. S. Gray, Going after the phrasal verb: An alternative approach to classification, Tesol Quarterly 33 (1999) 65–83.

[15] D. Liu, The most frequently used english phrasal verbs in american and british english: A multicorpus examination, Tesol Quarterly 45 (2011) 661–688.

[16] M. Garnier, N. Schmitt, The PHaVE list: A pedagogical list of phrasal verbs and their most frequent meaning senses, Language Teaching Research 19 (2015) 645–666.

[17] D. Liu, D. Myers, The most-common phrasal verbs with their key meanings for spoken and academic written english: A corpus analysis, Language Teaching Research 24 (2020) 403–424.

[18] S. J. Lindner, A lexico-semantic analysis of English verb particle constructions with out and up, Uni-

versity of California, San Diego, 1981.
[19] A. Tyler, V. Evans, The semantics of English prepositions: Spatial scenes, embodied meaning, and cognition, Cambridge University Press, 2003.
[20] G. Lakoff, Women, fire, and dangerous things: What categories reveal about the mind, University of Chicago press, 2008.
[21] É. Kovács, The traditional vs. cognitive approach to english phrasal verbs, Journal of Linguistics 1 (2011) 141–160.
[22] D. Thom, How to teach phrasal verbs using conceptual metaphors (2017).
[23] T. Baldwin, V. Kordoni, A. Villavicencio, Prepositions in applications: A survey and introduction to the special issue, Computational Linguistics 35 (2009) 119–149.
[24] A. Kurtyka, M. Putz, S. Niemeier, R. Dirven, Teaching english phrasal verbs: A cognitive approach, Applied cognitive linguistics 2 (2001) 29–54.
[25] N. Condon, How cognitive linguistic motivations influence the learning of phrasal verbs, Applications of cognitive linguistics 6 (2008) 133.
[26] S. Yasuda, Learning phrasal verbs through conceptual metaphors: A case of japanese efl learners, Tesol Quarterly 44 (2010) 250–273.
[27] H. Lee, Concept-based approach to second language teaching and learning: Cognitive lingusitics-inspired instruction of English phrasal verbs, The Pennsylvania State University, 2012.
[28] Z. S. Harris, Distributional structure, Word 10 (1954) 146–162.
[29] B. H. Partee, Formal semantics: Origins, issues, early impact, Baltic International Yearbook of Cognition, Logic and Communication 6 (2011) 13.
[30] D. Lin, Automatic identification of non-compositional phrases, in: Proceedings of the 37th annual meeting of the Association for Computational Linguistics, 1999, pp. 317–324.
[31] C. Bannard, T. Baldwin, A. Lascarides, A statistical approach to the semantics of verb-particles, in: Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment, 2003, pp. 65–72.
[32] D. McCarthy, B. Keller, J. A. Carroll, Detecting a continuum of compositionality in phrasal verbs, in: Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment, 2003, pp. 73–80.
[33] S. N. Kim, T. Baldwin, Detecting compositionality of english verb-particle constructions using semantic similarity, in: Proceedings of the 7th Meeting of the Pacific Association for Computational Linguistics (PACLING 2007), Citeseer, 2007, pp. 40–48.
[34] D. Kiela, S. Clark, Detecting compositionality of multi-word expressions using nearest neighbours in vector space models, in: Proceedings of the 2013 conference on empirical methods in natural language processing, 2013, pp. 1427–1432.
[35] M. Baroni, R. Zamparelli, Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space, in: Proceedings of the 2010 conference on empirical methods in natural language processing, 2010, pp. 1183–1193.
[36] J. Mitchell, M. Lapata, Composition in distributional models of semantics, Cognitive science 34 (2010) 1388–1429.
[37] R. Socher, B. Huval, C. D. Manning, A. Y. Ng, Semantic compositionality through recursive matrix-vector spaces, in: Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning, 2012, pp. 1201–1211.
[38] W. Yin, H. Schütze, An exploration of embeddings for generalized phrases, in: Proceedings of the ACL 2014 Student Research Workshop, 2014, pp. 41–47.
[39] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, Advances in neural information processing systems 26 (2013).
[40] O. Levy, Y. Goldberg, Dependency-based word embeddings, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2014, pp. 302–308.
[41] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.
[42] H. Schmid, Probabilistic part-ofispeech tagging using decision trees, in: New methods in language processing, 2013, p. 154.
[43] M. Marcus, B. Santorini, M. A. Marcinkiewicz, Building a large annotated corpus of english: The penn treebank (1993).
[44] M. Honnibal, M. Johnson, An improved non-monotonic transition system for dependency parsing, in: Proceedings of the 2015 conference on empirical methods in natural language processing, 2015, pp. 1373–1378.
[45] M. Donati, Phrasal verbs in distributional semantics: a cognitive linguistics analysis, 2022.
[46] R. Feldman, J. Sanger, The text mining handbook: advanced approaches in analyzing unstructured data, Cambridge university press, 2007.

How To Build Competitive Multi-gender Speech Translation Models For Controlling Speaker Gender Translation

Marco Gaido[1], Dennis Fucci[1,2], Matteo Negri[1] and Luisa Bentivogli[1]

[1]*Fondazione Bruno Kessler*
[2]*University of Trento*

Abstract
When translating from notional gender languages (e.g., English) into grammatical gender languages (e.g., Italian), the generated translation requires explicit gender assignments for various words, including those referring to the speaker. When the source sentence does not convey the speaker's gender, speech translation (ST) models either rely on the possibly-misleading vocal traits of the speaker or default to the masculine gender, the most frequent in existing training corpora. To avoid such biased and not inclusive behaviors, the gender assignment of speaker-related expressions should be guided by externally-provided metadata about the speaker's gender.[1] While previous work has shown that the most effective solution is represented by separate, dedicated gender-specific models, the goal of this paper is to achieve the same results by integrating the speaker's gender metadata into a single "multi-gender" neural ST model, easier to maintain. Our experiments demonstrate that a single multi-gender model outperforms gender-specialized ones when trained from scratch (with gender accuracy gains up to 12.9 for feminine forms), while fine-tuning from existing ST models does not lead to competitive results.

Keywords
gender bias, gradient reversal, speech translation

1. Introduction

Spurred by growing concerns about fairness in language technologies, research on understanding and mitigating gender bias in automatic translation is gaining traction [1]. The bias of automatic systems is extremely evident when it comes to ambiguous sentences or expressions, where there are no explicit cues in the source content about the correct gender[1] assignment of a referent (e.g., en: *The doctor arrived* – it: *Il/La dottore/essa è arrivato/a*). In this setting, the state-of-the-art neural models often choose the masculine forms or perpetuate stereotypical assignments, as they reflect the condition statistically more likely based on their (biased) training data [2, 3].

This situation frequently occurs when the source language is genderless or employs notional gender, expressing gender in a limited set of parts of speech, and the target language follows a grammatical gender system, embedding gender distinctions throughout a broad inventory of parts of speech. Focusing on the case in which the source language is English, a notional gender language, and the target language is Italian, a grammatical gender language, a frequent instance of this condition is represented by first-person references, i.e. by the words and expressions referred to the speaker (e.g., en: *I am a young researcher* – it: *Sono una/un giovane ricercatrice/tore*). In this case, text-to-text machine translation (MT) models mostly output masculine forms, while direct (or end-to-end) speech-to-text translation (ST) systems partly rely on the biological cue of the speaker's vocal traits to assign gender [4, 5]. However, direct ST models are still largely biased toward producing masculine forms, and, most importantly, biological aspects are related to the sex rather than to the gender of an individual. Hence, their exploitation is not inclusive of all people, harming several groups such as transgenders [6].

As a solution, [7] proposed to leverage external metadata about the speaker's gender to control the gender assignment of words referred to the speaker. Specifically, they investigated two approaches: *i)* the development of two separate *gender-specialized models*, fine-tuned on gender-specific data as also proposed later in MT [8], and *ii)* a single *multi-gender model*, where the speaker gender is a tag fed to a single model as in multilingual systems [9]. While the second solution would be preferable (as the specialized solution involves the higher cost of maintaining two separate models), the experiments in [7] demonstrate that specialized models outperform the multi-gender approach by a large margin in terms of gender accuracy.

In light of the above, in this paper we address the following research questions: *i)* why do specialized models outperform multi-gender ones? *ii)* Can we build competitive multi-gender systems? Through experiments on English-Italian translation of TED talks, we show that the low accuracy of multi-gender models comes from the initialization with the weights of a gender-unaware ST

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
mgaido@fbk.eu (M. Gaido); dfucci@fbk.eu (D. Fucci); negri@fbk.eu (M. Negri); bentivo@fbk.eu (L. Bentivogli)

[1]Throughout the paper, we use the word *gender* to indicate the preferred linguistic expression of gender and not the gender identity.

system and the inability to override the behavior of the base ST model (i.e., the reliance on vocal cues) during the fine-tuning stage. We also try to address this problem with two solutions: *i)* a contrastive loss that penalizes the extraction of gender cues from speech input, and *ii)* altering vocal properties of training data to misalign gender cues with gender tags and gender translations.[2] Despite the slight improvements brought by these solutions in gender accuracy and overall translation quality, none of them effectively close the performance gap with the specialized solution. However, training multi-gender models from scratch yields competitive results, outperforming the specialized approach with gender accuracy gains of up to 12.9 points for feminine translations. Therefore, we recommend building multi-gender models from scratch, while building them on top of existing systems remains an open research question.

2. Background

In this section, we introduce the basic concepts useful for understanding the rest of the paper. First, we provide an overview of the methods proposed in the literature to integrate language tags into neural multilingual translation models (§2.1), from which multi-gender models draw inspiration. Then, we present how gender information has been removed from neural representations through adversarial training in previous works (§2.2), from which we derive our solution presented in §3.1.

2.1. Tags Integration in Multilingual Models

State-of-the-art models in MT and ST are sequence-to-sequence models made of an encoder and an autoregressive Transformer decoder [10]. The autoregressive decoder predicts the next-token probability over a predefined vocabulary at every iteration by looking at the encoder output and at the previously generated tokens, which are pre-pended a special token named *beginning of sentence* (<*bos*>). Formally, the probability $p_V(y_t)$ over the vocabulary V at time step t is:

$$\text{softmax}(D(E(X); \textit{<bos>}, y_0, ..., y_{t-1})) \quad (1)$$

where *E* is the encoder, *D* is the decoder, *X* is the input sequence, and y_i is the token generated at the i-th time step.

While early attempts to build multilingual MT models were based on training dedicated encoders and decoders for each language [11, 12], nowadays the preferred solution is a model made of a single universal encoder and decoder where the language is represented as a tag pre-pended to the text [13, 14, 9]. In the case of one-to-many multilingual models, this means that the <*bos*> token is replaced with a token that indicates the language, so that Eq. 1 becomes:

$$\text{softmax}(D(E(X); \text{LID}, y_0, ..., y_{t-1})) \quad (2)$$

where *LID* is the identifier of the desired target language.

In direct ST, [15] demonstrated the effectiveness of this solution, also known as "target forcing", while [16] proposed other methods to integrate the language information into the architecture. Thanks to its simplicity and effectiveness, target forcing is currently the most widespread method to build multilingual ST systems [17, 18], also when using large pre-trained textual models such as mBART [19] to initialize the ST decoder [20, 21]. In line with this trend, [7] obtained their best multi-gender models with target forcing. As such, we build multi-gender models using target forcing with the *F* and *M* tags representing the two grammatical genders instead of the language identifiers.

2.2. Gender Information Removal

With the goal of fairer technology that does not rely on spurious cues reflecting stereotypical biases in the available data, researchers have tried to build systems that achieve "equalized odds" among different demographic groups [22]. Formally, this means that, given an attribute z representing the belonging to one of the Z demographic groups, the predicted probability $p(\hat{Y})$ of a fair system should be independent of the variable z, i.e. $p(\hat{Y}) = p(\hat{Y}|z), \forall z \in Z$. The variable z is named the *protected attribute* and in the context of gender bias literature represents the gender of the involved person. So in this work we consider $Z = \{F, M\}$.[3]

The first attempts to achieve equalized odds across genders in neural systems have focused on deep neural network (DNN) classifiers [23, 24, 25, 26]. In this line of work, the last hidden representation of the DNN is passed both to a linear layer that predicts the classification scores $\hat{Y}$ and to a linear layer (the *discriminator*) devoted to predicting the protected attribute z. The DNN is then trained in an adversarial manner [27], i.e. it is alternatively trained to *i)* predict z (while keeping the shared DNN freezed) and *ii)* predict $\hat{Y}$ while minimizing the ability to predict z (keeping the protected attribute classification layer freezer). As this training procedure is often unstable, similar practices based on minmax optimizations have been proposed [28], even with discriminators made of functions different from linear projections

[2]Our code is released open source under Apache 2.0 Licence at: https://github.com/hlt-mt/FBK-fairseq/

[3]Although this paper does not aim at perpetuating a binary vision of gender, in this work we limit to the feminine and masculine categories for the sake of simplicity, as the available benchmarks currently do not cover the non-binary case.

[29], or using more than one discriminator [30]. [31] also proposed methods to automatically extract the protected attributes in case they have not been provided.

Such adversarial training procedures can be seen as an extension of the *gradient reversal* [32], where the training alternatively freezes the base model (to refine the discriminator) and the discriminator (inverting its loss to train the base model to be unable to discriminate). In fact, the gradient reversal layer, applied to the hidden representations before feeding them to the discriminator, is an identity function in the forward pass, while inverts the gradient in the backward pass, scaling it by a positive factor λ. By naming x the hidden representation, D the discriminator, and GRL the gradient reversal layer, this means that:

$$GRL(x) = x, \nabla D(GRL(x)) = -\lambda \nabla D(x) \quad (3)$$

where λ is a hyperparameter that can either be fixed or updated according to the following equation:

$$\lambda = \frac{2}{1 + e^{-\gamma p}} - 1 \quad (4)$$

where p is the ratio between the number of parameter updates performed and the total number of updates needed to complete the training.

3. Solutions for Multi-gender Models

To create multi-gender ST models that solely rely on the gender tag, ignoring spurious cues related to speakers' vocal traits, we test two approaches. First, we try to create gender-invariant encoder representations by adding a gradient-reverted discriminator on the speaker's gender (§3.1). Second, we manipulate the input audio by altering the speaker's pitch, so that the correlation between the gender tag (and output text) and the speaker's vocal traits is lost (§3.2).

3.1. Gradient Reversal

As seen in §2.1, the decoder of a multi-gender model has three inputs: the encoder output, a tag representing the speaker's gender, and the previously generated tokens. As we want the decoder to have the tag as the only source of information about the speaker's gender, we propose to create encoder outputs that do not convey any information regarding the speaker's gender by adding a gradient-reverted discriminator on top of the encoder, motivated by the success of this approach in MT with sentences where there is a single referent whose gender has to be determined [33]. The discriminator is made of two fully-connected layers with ReLU activation function [34], whose output is averaged over the temporal dimension to obtain a single vector representing the logit[4] of the discriminator.

Furthermore, we experiment with assigning dedicated class weights to the loss of the discriminator, as a countermeasure to the class imbalance between female and male speakers in the training data. Specifically, we assigned the weights (w_f, w_m) proportionally to the inverse of the frequency of each class (f_f, f_m) in the training data:

$$\begin{cases} w_f \propto \frac{1}{f_f}, w_m \propto \frac{1}{f_m} \\ w_f * \frac{f_f}{f_m + f_f} + w_m * \frac{f_m}{f_m + f_f} = 1 \end{cases}$$

resulting in $w_f = 1.4$, $w_m = 0.8$ in our case.

3.2. Audio Manipulation

Our second approach aims to break the correlation between the vocal characteristics of the speaker on one side and the gender tag and target translation on the other. To this aim we manipulate part of the training data using the *Opposite* pitch manipulation strategy by [35]. The amount of data that is manipulated at each iteration (epoch) is controlled with a hyperparameter, p, which determines the probability of altering an utterance, regardless of whether it is produced by a male or female speaker. The manipulation is performed by altering two crucial acoustic parameters distinguishing between male and female voices [36, 37]: $f0$ and formants. In particular, we first estimate the $\tilde{f0}$ median of the $f0$ contour of the considered speech segment. Then, we sample a new $\tilde{f0}'$ median of the desired output audio from a normal distribution whose mean and standard deviation depend on the target gender: for feminine voices, we use 250 Hz as the mean and 17 as the standard deviation so that the sampled value is between 199 Hz and 301 Hz with 99.7% probability; for masculine voices, the mean is 140 Hz and the standard deviation is 20 to obtain a 99.7% probability range within 80 Hz and 200 Hz. Once $\tilde{f0}$ and $\tilde{f0}'$ are defined, we compute a scaling factor α as the ratio $\tilde{f0}'/\tilde{f0}$. Lastly, the original $f0$ contour is scaled by the α factor, while the formants are scaled by 1.2 when converting from male to female voices, or by 0.8 otherwise. This perturbation is applied independently to each sample during each training epoch, so as to maximize the variability of the training data.

4. Experimental Settings

Our ST models are composed of a Conformer [38] encoder with 12 layers and a Transformer [10] decoder with 6 layers. We used the Conformer implementation by [39],

[4]The *logit* is the vector of raw predictions before a function (commonly, the softmax) that maps it into probabilities.

Model	BLEU (↑)	Gender Accuracy (↑)			
		1F	1M	1F-Tag M	1M-Tag F
Fine-tuning					
Specialized	**27.4**	73.3	92.5	80.9	56.1
Multi-gender	26.0	66.8	78.0	64.6	47.1
+ gradient reversal	26.7	60.7	85.9	77.0	43.3
+ gradient reversal weighted	26.3	62.4	83.5	77.7	45.9
+ audio manipulation (50%)	26.4	56.0	82.6	60.7	33.9
+ audio manipulation (80%)	26.3	69.3	81.1	69.0	47.7
Training from scratch					
Multi-gender	27.2	**84.0**	92.7	93.4	**69.0**
+ gradient reversal	24.9	70.9	**93.2**	**94.1**	58.7
+ gradient reversal weighted	24.2	75.8	92.6	92.8	63.5
+ audio manipulation (50%)	26.2	79.6	92.4	91.5	67.9
+ audio manipulation (80%)	25.7	81.7	92.6	91.0	65.3

Table 1
BLEU and gender accuracy scores for the specialized models (Specialized) and the multi-gender models (Multi-gender) both trained from scratch and fine-tuned, also with gradient reversal and audio manipulation.

which does not contain bugs related to the presence of padding. The embedding size was 512, and the dropout was set to 0.1. We optimized label-smoothed cross entropy using Adam. The learning rate followed the Noam scheduler with 25,000 warmup updates and a maximum value of $2e^{-3}$. We train for 50,000 updates and average the last 7 checkpoints.

We train our models on MuST-C [40], an ST corpus built from TED data, for which is also available the annotation of the gender of the speaker [7]. We extract 80 features with log mel-filterbank from the input audio and normalize them with cepstral mean and variance [41]. The target text is encoded into subwords with 8,000 BPE merge rules [42] learned on the training set. We evaluate on the MuST-SHE benchmark [4], which contains a section ("Category 1") dedicated to assessing the gender assignment of words referring to the speaker. We compute SacreBLEU[5] [43] on the whole MuST-SHE test set to evaluate the translation quality of our models and gender accuracy [7] on the feminine and masculine sections of "Category 1" to evaluate the ability of each model to correctly assign gender to words referring to the speaker.

Gradient Reversal. The loss of the auxiliary speaker-classification task is summed to the loss on the decoder output scaling it by a 0.5 factor. For the gradient reversal layer, we tested both fixed values of λ and controlling its value with γ. For fine-tunings, we set $\lambda = 10$, so as to give similar weight to the gender classification loss and the cross-entropy loss for the translation. When training from scratch, instead, despite different attempts the training is unstable and diverges unless lambda is set to a fixed, small value, where its contribution is negligible. We report results for $\lambda = 0.5$, which is the highest λ value for which the loss on the validation set does not explode during training.

[5] case:mixed|eff:no|tok:13a|smooth:exp|version:2.0.0

Audio Manipulation. In our experiments, we tested two values (0.5 and 0.8) for the hyperparameter p, which controls the probability of manipulating a speech segment. In the first case, 50% of the data is manipulated, leading to a complete loss of correlation between gender tags and vocal traits (50% of the samples with the F tag would exhibit frequency characteristics typical of masculine voices, and 50% of the samples with the M tag would have frequency characteristics typical of feminine voices). In the second case, instead, the correlation between the gender tag and the vocal traits is negative, to counteract the patterns learned by a gender-unaware ST model. In any case, as the training data is imbalanced (70% of the samples are uttered by male speakers, and 30% by female speakers), and the manipulation probability is the same for segments uttered by male and female speakers, the gender imbalance in the training data is not mitigated.

5. Results

We investigate the performance of multi-gender models trained in two different ways: *i)* fine-tuning a base, gender-unaware ST model, and *ii)* training from scratch. In both cases, we study the effect of the introduction of the discriminator with gradient reversal and of the audio manipulation techniques. Table 1 presents BLEU and gender accuracy scores (separately for segments spoken by female (1F) and male (1M) speakers) for all the models, comparing them with the specialized models. To assess the inclusivity of our solution in cases where speakers exhibit vocal traits that do not conform with traditional gender perceptions, we also report gender accuracy for tests where the gender tag is inverted compared to the original audio segment (1F-Tag M and 1M-Tag F). In these instances, the gender translation is expected to align with the gender tag, and we use the "wrong" reference of MuST-SHE, which swaps the speaker's references to the opposite gender.

Fine-tuning. The fine-tuned models from a base ST system consistently yield lower scores compared to the specialized systems. The simple multi-gender model performs 1.4 BLEU points worse than the specialized models in terms of overall translation quality. However, when audio manipulation and especially gradient reversal techniques are employed during fine-tuning, the performance gap is reduced by up to half. Regarding gender accuracy, the multi-gender model achieves considerably lower scores than the specialized models, confirming previous findings from [7]. This indicates that a fine-tuned multi-gender model struggles to accurately follow the gender tag for gender translation. The accuracy gap is particularly high when the tag conflicts with vocal traits (-16.3 in 1F-Tag M, -9.0 in 1M-Tag F), where multi-gender models show below-chance accuracy for feminine forms, being below 50%. Both gradient reversal and audio manipulation techniques seem to further bias the model towards masculine forms. This likely indicates that the reduced ability to rely on the speakers' vocal traits is not compensated by the model looking at the gender tag, rather it strengthens its tendency to default to the most frequent masculine forms. The only technique that consistently improves both masculine and feminine translations compared to the simple fine-tuned multi-gender model is the introduction of audio manipulation with high probability (80%). However, the gains (2.5 for 1F and 3.1 for 1M, and 4.4 for 1F-Tag M and 0.7 for 1M-Tag F) are limited and the gap with specialized models remains large.

Training from scratch. Unlike the fine-tuned models, the multi-gender models trained from scratch yield comparable or even higher results than the specialized models. This suggests that when an ST model is trained from scratch with gender tags, it learns to effectively follow them. Specifically, the simple multi-gender model trained from scratch achieves comparable translation quality to the specialized system (-0.2 BLEU) and significantly outperforms it in gender accuracy, with gains ranging from 0.2 (1M) to 12.5 (1F-Tag M). As in the fine-tuning case, neither gradient reversal nor audio manipulations increase the reliance of the model on the tag and the resulting models are more biased toward masculine forms. In fact, the multi-gender with gradient reversal reaches the highest accuracies in producing masculine forms (93.2 in 1M and 94.1 in 1F-Tag M), while suffering substantial drops in feminine accuracy. This effect is reduced when the weight of the F class is increased in the discriminator. In addition, in both cases the translation quality suffers a considerable drop. In the case of audio manipulation, the translation quality drop is lower (although still present), as well as the differences in terms of gender accuracy. They do not provide, though, any benefit compared to the simple multi-gender training.

In summary, the results demonstrate that the previous finding about the low performance of multi-gender models is due to the adoption of a fine-tuning strategy. In this setting, the model cannot effectively override the reliance on speakers' vocal traits of the gender-unaware base ST model. In addition, techniques aimed at avoiding the exploitation of speakers' vocal traits seem ineffective. However, training the multi-gender model effectively solves the problem and the model is capable of following the indication given by the gender tag, outperforming even the specialized strategy by up to 12.9 gender accuracy (1M-Tag F).

6. Conclusions

In this paper, we studied the effect of different training strategies to build multi-gender ST models, i.e. models that are informed of the gender of the speaker by an explicit gender tag. Focusing on English-Italian translations, we demonstrated that the low accuracy of multi-gender models shown by previous work stems from the their initialization with gender-unaware ST system weights and the inability of effectively overriding the reliance on vocal cues during fine-tuning. On the other hand, training multi-gender models from scratch proved to be an effective solution, outperforming the approach based on the creation of two gender-specialized models. As training from scratch is not always feasible, we also experimented with two methods to enhance the reliance on the gender tag in fine-tuned multi-gender models: penalizing the extraction of gender cues from speech input, and altering the vocal properties of the speakers in the training data to avoid the alignment between biological cues and gender tags and translations. While these solutions partially improved gender accuracy and overall translation quality in fine-tuned multi-gender models, they did not close the gap with specialized models. Therefore, further research is needed in this direction.

Acknowledgments

This work is part of the project "Bias Mitigation and Gender Neutralization Techniques for Automatic Translation", which is financially supported by an Amazon Research Award AWS AI grant. We acknowledge the support of the PNRR project FAIR - Future AI Research

(PE00000013), under the NRRP MUR program funded by the NextGenerationEU.

References

[1] B. Savoldi, M. Gaido, L. Bentivogli, M. Negri, M. Turchi, Gender bias in machine translation, Transactions of the Association for Computational Linguistics 9 (2021) 845–874. URL: https://aclanthology.org/2021.tacl-1.51. doi:10.1162/tacl_a_00401.

[2] M. O. R. Prates, P. H. C. Avelar, L. C. Lamb, Assessing gender bias in machine translation: a case study with google translate, Neural Computing and Applications 32 (2020) 6363–6381. doi:10.1007/s00521-019-04144-6.

[3] W. I. Cho, J. W. Kim, S. M. Kim, N. S. Kim, On Measuring Gender bias in Translation of Gender-neutral Pronouns, in: Proceedings of the First Workshop on Gender Bias in Natural Language Processing, Association for Computational Linguistics, Florence, Italy, 2019, pp. 173–181. URL: https://www.aclweb.org/anthology/W19-3824. doi:10.18653/v1/W19-3824.

[4] L. Bentivogli, B. Savoldi, M. Negri, M. A. Di Gangi, R. Cattoni, M. Turchi, Gender in Danger? Evaluating Speech Translation Technology on the MuST-SHE Corpus, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 6923–6933. URL: https://www.aclweb.org/anthology/2020.acl-main.619.

[5] M. Gaido, M. A. Di Gangi, M. Negri, M. Turchi, On Knowledge Distillation for Direct Speech Translation, in: Proceedings of CLiC-IT 2020, Online, 2021. URL: http://ceur-ws.org/Vol-2769/paper_28.pdf.

[6] L. Zimman, Transgender language, transgender moment: Toward a trans linguistics, in: K. Hall, R. Barrett (Eds.), The Oxford Handbook of Language and Sexuality, 2020. doi:10.1093/oxfordhb/9780190212926.013.45.

[7] M. Gaido, B. Savoldi, L. Bentivogli, M. Negri, M. Turchi, Breeding Gender-aware Direct Speech Translation Systems, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 3951–3964. URL: https://www.aclweb.org/anthology/2020.coling-main.350. doi:10.18653/v1/2020.coling-main.350.

[8] P. K. Choubey, A. Currey, P. Mathur, G. Dinu, GFST: Gender-filtered self-training for more accurate gender in translation, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 1640–1654. URL: https://aclanthology.org/2021.emnlp-main.123. doi:10.18653/v1/2021.emnlp-main.123.

[9] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, M. Hughes, J. Dean, Google's multilingual neural machine translation system: Enabling zero-shot translation, Transactions of the Association for Computational Linguistics 5 (2017) 339–351. URL: https://aclanthology.org/Q17-1024. doi:10.1162/tacl_a_00065.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc., Long Beach, USA, 2017. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[11] O. Firat, K. Cho, Y. Bengio, Multi-way, multilingual neural machine translation with a shared attention mechanism, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, San Diego, California, 2016, pp. 866–875. URL: https://aclanthology.org/N16-1101. doi:10.18653/v1/N16-1101.

[12] O. Firat, B. Sankaran, Y. Al-onaizan, F. T. Yarman Vural, K. Cho, Zero-resource translation with multi-lingual neural machine translation, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas, 2016, pp. 268–277. URL: https://aclanthology.org/D16-1026. doi:10.18653/v1/D16-1026.

[13] R. Sennrich, B. Haddow, A. Birch, Improving Neural Machine Translation Models with Monolingual Data, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 86–96. URL: https://aclanthology.org/P16-1009. doi:10.18653/v1/P16-1009.

[14] T.-L. Ha, J. Niehues, A. Waibel, Toward multilingual neural machine translation with universal encoder and decoder, in: Proceedings of the 13th International Conference on Spoken Language Translation, International Workshop on Spoken Language Translation, Seattle, Washington D.C, 2016. URL: https://aclanthology.org/2016.iwslt-1.6.

[15] H. Inaguma, K. Duh, T. Kawahara, S. Watanabe, Multilingual end-to-end speech translation, in: 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2019, pp. 570–577. doi:10.1109/ASRU46091.2019.9003832.

[16] M. A. Di Gangi, M. Negri, M. Turchi, One-to-many multilingual end-to-end speech translation, in: 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2019, pp. 585–592. doi:10.1109/ASRU46091.2019.9004003.

[17] C. Wang, Y. Tang, X. Ma, A. Wu, D. Okhonko, J. Pino, Fairseq S2T: Fast speech-to-text modeling with fairseq, in: Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Suzhou, China, 2020, pp. 33–39. URL: https://aclanthology.org/2020.aacl-demo.6.

[18] E. Salesky, M. Wiesner, J. Bremerman, R. Cattoni, M. Negri, M. Turchi, D. W. Oard, M. Post, The Multilingual TEDx Corpus for Speech Recognition and Translation, in: Proc. Interspeech 2021, 2021, pp. 3655–3659. doi:10.21437/Interspeech.2021-11.

[19] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, L. Zettlemoyer, Multilingual denoising pre-training for neural machine translation, Transactions of the Association for Computational Linguistics 8 (2020) 726–742. URL: https://aclanthology.org/2020.tacl-1.47. doi:10.1162/tacl_a_00343.

[20] D. Liu, T. Binh Nguyen, S. Koneru, E. Yavuz Ugan, N.-Q. Pham, T. Nam Nguyen, T. Anh Dinh, C. Mullov, A. Waibel, J. Niehues, KIT's multilingual speech translation system for IWSLT 2023, in: Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023), Association for Computational Linguistics, Toronto, Canada (in-person and online), 2023, pp. 113–122. URL: https://aclanthology.org/2023.iwslt-1.6.

[21] E. Gow-Smith, A. Berard, M. Zanon Boito, I. Calapodescu, NAVER LABS Europe's multilingual speech translation systems for the IWSLT 2023 low-resource track, in: Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023), Association for Computational Linguistics, Toronto, Canada (in-person and online), 2023, pp. 144–158. URL: https://aclanthology.org/2023.iwslt-1.10.

[22] M. Hardt, E. Price, N. Srebro, Equality of opportunity in supervised learning, in: Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16, Curran Associates Inc., Red Hook, NY, USA, 2016, p. 3323–3331.

[23] A. Beutel, E. H. Chi, J. Chen, Z. Zhao, Data decisions and theoretical implications when adversarially learning fair representations, 2017. URL: https://arxiv.org/pdf/1707.00075.pdf.

[24] B. H. Zhang, B. Lemoine, M. Mitchell, Mitigating unwanted biases with adversarial learning, in: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 335–340. URL: https://doi.org/10.1145/3278721.3278779. doi:10.1145/3278721.3278779.

[25] Y. Elazar, Y. Goldberg, Adversarial removal of demographic attributes from text data, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 11–21. URL: https://aclanthology.org/D18-1002. doi:10.18653/v1/D18-1002.

[26] L. Gao, H. Zhan, A. Chen, V. Sheng, Mitigate gender bias using negative multi-task learning, 2022. URL: https://doi.org/10.21203/rs.3.rs-2024101/v1. doi:10.21203/rs.3.rs-2024101/v1.

[27] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Commun. ACM 63 (2020) 139–144. URL: https://doi.org/10.1145/3422622. doi:10.1145/3422622.

[28] S. Ravfogel, M. Twiton, Y. Goldberg, R. D. Cotterell, Linear adversarial concept erasure, in: K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, S. Sabato (Eds.), Proceedings of the 39th International Conference on Machine Learning, volume 162 of *Proceedings of Machine Learning Research*, PMLR, 2022, pp. 18400–18421. URL: https://proceedings.mlr.press/v162/ravfogel22a.html.

[29] S. Ravfogel, F. Vargas, Y. Goldberg, R. Cotterell, Adversarial concept erasure in kernel space, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 6034–6055. URL: https://aclanthology.org/2022.emnlp-main.405.

[30] X. Han, T. Baldwin, T. Cohn, Diverse adversaries for mitigating bias in training, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, 2021, pp. 2760–2765. URL: https://aclanthology.org/2021.eacl-main.239. doi:10.18653/v1/2021.eacl-main.239.

[31] S. Shao, Y. Ziser, S. B. Cohen, Erasure of unaligned attributes from neural representations, Transactions of the Association for Computational Linguistics 11 (2023) 488–510. URL: https://aclanthology.

org/2023.tacl-1.29. doi:10.1162/tacl_a_00558.

[32] Y. Ganin, V. Lempitsky, Unsupervised Domain Adaptation by Backpropagation, in: F. Bach, D. Blei (Eds.), Proceedings of the 32nd International Conference on Machine Learning, volume 37 of *Proceedings of Machine Learning Research*, PMLR, Lille, France, 2015, pp. 1180–1189. URL: https://proceedings.mlr.press/v37/ganin15.html.

[33] E. Fleisig, C. Fellbaum, Mitigating Gender Bias in Machine Translation through Adversarial Learning, 2022. arXiv:2203.10675.

[34] V. Nair, G. E. Hinton, Rectified linear units improve restricted boltzmann machines, in: Proceedings of the 27th International Conference on International Conference on Machine Learning, ICML'10, Omnipress, Madison, WI, USA, 2010, p. 807–814.

[35] D. Fucci, M. Gaido, M. Negri, M. Cettolo, L. Bentivogli, No Pitch Left Behind: Addressing Gender Unbalance in Automatic Speech Recognition through Pitch Manipulation, in: IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Taipei, Taiwan, 2023.

[36] R. O. Coleman, A comparison of the contributions of two voice quality characteristics to the perception of maleness and femaleness in the voice, Journal of Speech & Hearing Research 19(1) (1976) 168–180. doi:https://doi.org/10.1044/jshr.1901.168.

[37] J. M. Hillenbrand, M. J. Clark, The role of f0 and formant frequencies in distinguishing the voices of men and women, Attention Perception & Psychophysics 71(5) (2009) 1150–1166. doi:10.3758/APP.71.5.1150.

[38] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, R. Pang, Conformer: Convolution-augmented Transformer for Speech Recognition, in: Proceedings of the 21st Annual Conference of the International Speech Communication Association, International Speech Communication Association, Shanghai, China (Online), 2020, pp. 5036–5040. doi:10.21437/Interspeech.2020-3015.

[39] S. Papi, M. Gaido, A. Pilzer, M. Negri, When Good and Reproducible Results are a Giant with Feet of Clay: The Importance of Software Quality in NLP, 2023. arXiv:2303.16166.

[40] R. Cattoni, M. A. Di Gangi, L. Bentivogli, M. Negri, M. Turchi, Must-c: A multilingual corpus for end-to-end speech translation, Computer Speech & Language 66 (2021) 101–155. URL: https://www.sciencedirect.com/science/article/pii/S0885230820300887. doi:https://doi.org/10.1016/j.csl.2020.101155.

[41] O. Viikki, K. Laurila, Cepstral domain segmental feature vector normalization for noise robust speech recognition, Speech Communication 25 (1998) 133–147. URL: https://www.sciencedirect.com/science/article/pii/S0167639398000338. doi:https://doi.org/10.1016/S0167-6393(98)00033-8.

[42] M. A. Di Gangi, M. Gaido, M. Negri, M. Turchi, On target segmentation for direct speech translation, in: Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track), Association for Machine Translation in the Americas, Virtual, 2020, pp. 137–150. URL: https://aclanthology.org/2020.amta-research.13.

[43] M. Post, A call for clarity in reporting BLEU scores, in: Proceedings of the Third Conference on Machine Translation: Research Papers, Association for Computational Linguistics, Belgium, Brussels, 2018, pp. 186–191. URL: https://www.aclweb.org/anthology/W18-6319.

Hate Speech Detection in an Italian Incel Forum Using Bilingual Data for Pre-Training and Fine-Tuning

Paolo Gajo[1], Silvia Bernardini[1], Adriano Ferraresi[1] and Alberto Barrón-Cedeño[1]

[1]Department of Interpreting and Translation, Università di Bologna, Corso della Repubblica, 136, 47121, Forlì, FC, Italy

Abstract

English. In this study, we aim to enhance hate speech detection in Italian incel posts. We pre-train monolingual (Italian) and multilingual Transformer models on corpora built from two incel forums, one in Italian and one in English, using masked language modeling. Then, we fine-tune the models on combinations of English and Italian corpora, annotated for hate speech. Experiments on a hate speech corpus derived from the Italian incel forum show that the best results are achieved by training multilingual models on bilingual data, rather than training monolingual models on Italian-only data. This emphasizes the importance of using training and testing data from a similar linguistic domain, even when the languages differ.

Italiano. *In questo studio, ci proponiamo di migliorare il rilevamento dei discorsi d'odio in post tratti da un forum italiano di incel. Addestriamo modelli Transformer mono (italiano) e multilingue su corpora ottenuti da due forum di incel, uno in italiano e uno in inglese, con il masked language modeling. Facciamo quindi il fine-tuning dei modelli su corpora in italiano e inglese con annotazioni indicanti se un post esprime odio. Sperimentando su un corpus annotato per i discorsi di odio ottenuto da un forum italiano di incel mostriamo che i risultati migliori si ottengono addestrando modelli multilingue su combinazioni bilingue di corpora e non con modelli italiani e dati monolingue. Ciò sottolinea l'importanza di utilizzare dati di addestramento appartenenti a un contesto linguistico simile a quello dei dati di valutazione, anche con lingue differenti.*

Keywords

incels, hate speech, masked language modeling, transformers, bert, multilingual mlm, multilingual masked language modeling

1. Introduction

While there is no scarcity of English-language models and training resources for the detection of hate speech (HS), especially with the recent rise in popularity of this research topic [1], much work can still be carried out on this problem in other languages. For less-resourced languages, such as Italian, one of the main difficulties of combating this phenomenon is the lack of annotated data [2]. The problem is even more severe when considering the detection of hate speech in niche contexts, such as in forums frequented by incels, short for "involuntary celibates", a community known for its hateful language [3, 4] and use of specific misogynous and racist lexicon [5, 6]. In particular, it seems no work has yet been done on the detection of hate speech in Italian incel forums.

In this paper, we present a simple approach to improve the performance of hate speech detection models in Italian forums frequented by incels. Our contribution is two-fold:

(*i*) Masked language modeling. We adapt monolingual Italian models to the linguistic domain of Italian incel forums by training them on the masked language modeling (MLM) task. As training material, we use an unlabelled corpus compiled from an Italian incel forum. We also adopt an existing multilingual model, already domain-adapted to the incel domain in both English and Italian. We release these novel models, which can be used for further research on the topic.[1]

(*ii*) Hate speech detection. We fine-tune the vanilla and domain-adapted models on the downstream task of detecting hate speech in Italian incel posts. Monolingual models are trained on Italian-only combinations of corpora binary-annotated for hate speech, while Italian–English combinations are used for the multilingual models.

Testing the performance of the models on a labelled hate speech corpus, obtained by annotating posts from the Italian incel forum, shows that the best results are obtained by first training the base multilingual model on bilingual data taken from both the Italian and English incel forums, using the MLM task, and then fine-tuning it on combinations of Italian and English corpora, annotated for hate speech. In the approached scenarios, pre-training and fine-tuning on bilingual in-domain incel annotated data may therefore be more effective than

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

paolo.gajo2@unibo.it (P. Gajo); silvia.bernardini@unibo.it (S. Bernardini); adriano.ferraresi@unibo.it (A. Ferraresi); a.barron@unibo.it (A. Barrón-Cedeño)
https://www.unibo.it/sitoweb/paolo.gajo2 (P. Gajo); https://www.unibo.it/sitoweb/silvia.bernardini (S. Bernardini); https://www.unibo.it/sitoweb/adriano.ferraresi (A. Ferraresi); https://www.unibo.it/sitoweb/a.barron (A. Barrón-Cedeño)
0009-0009-9372-3323 (P. Gajo); 0000-0003-0750-4861 (S. Bernardini); 0000-0002-6957-0605 (A. Ferraresi); 0000-0003-4719-3420 (A. Barrón-Cedeño)

[1]Links to the models, released on HuggingFace: https://github.com/paolo-gajo/clic23.

training on general target-language labelled corpora, despite part of the training data not being in the language of the downstream task. In addition, the results show that this strategy can be used to improve model performance when in-domain target-language data is scarce, by using in-domain data from other languages.

The rest of the paper is organized as follows: Section 2 presents related work on hate speech detection in Italian and English, as well as multilingual approaches to the problem. Section 3 describes the corpora used in this study. Section 4 presents the employed models. Section 5 describes the experiments conducted and discusses the results. Section 6 closes the contribution with conclusions and future work.

2. Related Work

Prior work on Italian hate speech detection has been conducted chiefly within the context of EVALITA. The 2018 edition hosted a shared task on hate speech detection [7] based on two corpora, one comprising tweets and one Facebook posts. The participating teams experimented with a variety of algorithms, with the top team relying on an SVM and a BiLSTM [8]. The 2020 edition hosted a shared task on the detection of hate speech, stereotypes, and nominal utterances, especially against migrants, focusing on tweets and news headlines [9]. In this case, the best team's approach for the hate speech detection sub-task [10] was to fine-tune BERT_{base} [11] along with AlBERTo [12] and UmBERTo [13], two BERT models pre-trained on Italian data.

As regards misogyny in particular, EVALITA 2018 hosted for the first time a shared task on automatic misogyny identification (AMI), where the top performing teams used a combination of TF-IDF and SVD for the Italian scenario, and TF-IDF with logistic regression for the English one [14]. EVALITA 2020 hosted the second edition of the AMI shared task, focusing on Italian tweets [15], where an ensemble of BERT models obtained the top performance [16]. In EVALITA 2023, Di Bonaventura et al. [17] used triple verbalisation, prompting and majority vote to improve the performance of an AlBERTo model on the tasks of homotransphobia and hate speech detection.

English-language hate speech detection has been conducted in a variety of ways. Among others, Davidson et al. [18] build a corpus of tweets annotated with multi-class labels ("hate speech", "offensive", "neither") and train logistic regression and linear SVM models on it. Mathew et al. [19] build a corpus called HateXplain from Twitter and Gab posts, annotated with multi-class labels based on whether the post is "offensive", expresses "hate", or is "normal", which they use to fine-tune a BERT hate speech classifier. Caselli et al. [20] retrain BERT_{base} on the MLM task using an unlabelled corpus built from hateful and offensive Reddit messages, obtaining a model called HateBERT, capable of outperforming BERT_{base} on hate speech identification on various benchmark datasets.

In multilingual settings, Pelicon et al. [21] use a multilingual combination of corpora annotated for hate speech to improve the performance of classifiers in zero-shot, few-shot and well-resourced settings. Gokhale et al. [22] use MLM training to improve the hate speech detection performance of BERT in Hindi and Marathi, separately. We follow such approaches in improving the performance of our models, with a specific focus on monolingual vs. bilingual pre-training, compared to Gajo et al. [23].

3. Corpora

We leverage three labelled Italian-language corpora from past EVALITA campaigns, along with two labelled corpora compiled from two incel forums.

EVALITA corpora The first Italian corpus we use was compiled for the first edition of the Hate Speech Detection (HaSpeeDe) shared task, from EVALITA 2018 [7] (henceforth HSD-FB), by annotating Facebook posts for hate speech. The second one is from the 2020 edition of HaSpeeDe [24] (HSD-TW), compiled by adding new data to the HaSpeeDe 2018 Twitter corpus. The third corpus is the one compiled for the Automatic Misogyny Identification (AMI) shared task [15] (AMI-20), hosted at EVALITA 2020. AMI-20 is annotated with misogyny labels, which we use as hate speech labels to train our classifiers. Where the corpora were not partitioned, we split them 70/30 between training and development sets. We do not use the test partitions, as we are interested in maintaining consistency with the use of the original splits of these corpora.

Incel corpora We use two unlabelled corpora compiled by scraping two incel forums [23]: *Incels.is*[2] and *Il forum dei brutti*[3], respectively in English and Italian.

A subset of the two corpora was annotated for both misogyny and racism.[4] The annotated partitions are referred to as IFS-EN and IFS-IT ("Incel Forum, Supervised, English" and "Italian").

We keep the training, development and testing partitions as in the released corpora. IFS-IT is used in its entirety solely as a test set, due to the unavailability of additional annotated Italian incel data for training. Said scarcity prompted us to leverage the available data in order to conduct cross-lingual experiments for the incel domain, for which Italian is a low-resource language.

[2]https://incels.is (Last access: 11 Aug 2023)
[3]https://ilforumdeibrutti.forumfree.it (Last access: 11 Aug 2023)
[4]Refer to Gajo et al. [23] for details on the annotation process.

Table 1
Existing corpora class distribution.

Corpus	HS	Non-HS
AMI-20 [15]	2,337	2,663
HSD-FB [7]	1,382	1,617
HSD-TW [7]	971	2,028

Table 2
Incel corpora class distribution [23].

Corpus	Misogyny	Racism	Both	Neither
IFS-EN$_{tr}$	806	630	46	2,160
IFS-EN$_{de}$	173	130	13	464
IFS-EN$_{te}$	160	125	7	489
IFS-IT$_{te}$	187	8	5	300

Table 1 shows the class distribution of all three EVALITA corpora, whereas Table 2 shows the distribution for the incel corpora, where posts are considered hateful if they are either labeled as misogynous or racist. As can be inferred from the statistics, while misogynous instances comprise around 39% of the instances in both IFS-EN and IFS-IT, the same cannot be said for the racist ones, which are much more prevalent in IFS-EN (13% vs. 0.03%). This shows a clear difference in terms of the hate speech produced by the two incel communities.

4. Models

With relation to the Italian-only scenario, we use UmBERTo and AlBERTo for our baseline models. We choose these models because they achieved the best performance in previous EVALITA shared tasks on hate speech [9] and misogyny [15] identification. In order to improve the performance of the two models on the task of identifying hate speech in Italian incel forums, we train them on the MLM task on posts extracted from *Il forum dei brutti*. We follow this approach because it has been shown to work in English both for general hateful content [20] and incel forums [23]. For training data, we use the entirety of the contents of the forum, for a total of $627k$ posts. The intersection between the unlabelled incel corpora and the annotated corpora listed in Table 2 is void. That is, none of the data contained in IFS-IT was obtained from the Italian data scraped from *Il forum dei brutti* and used for MLM pre-training. The same is true for IFS-EN and the English MLM pre-training data taken from *Incels.is*. Doing this, we obtain two new models which we refer to as "Incel UmBERTo" and "Incel AlBERTo".

The MLM pre-training process is carried out in all cases by tokenizing post contents using each model's own tokenizer and masking tokens with a probability of 15%. We use a batch size of 32 samples and train the models for one epoch on one Tesla P100 GPU with 16 GB of VRAM.

As regards the bilingual setting, we use mBERT$_{base}$ as our baseline. We also use an MLM-enhanced version of it, "Incel mBERT",[5] obtained by further pre-training mBERT$_{base}$ on $500k$ posts sampled from *Il forum dei brutti* and $500k$ posts sampled from *Incels.is*, for a total of $1M$ posts in Italian and English [23].

[5]https://huggingface.co/pgajo/incel-mbert

5. Experiments and Results

We approach the task of identifying hate speech as a binary classification problem, where a post can either be hateful or not. We train each model five times on all possible combinations of the corpora listed in Tables 1 and 2 in order to make our results more reliable and diminish the effect of the random initialization of the models. In the monolingual Italian setting we never use IFS-EN, while it is always included when training the multilingual models in the bilingual setting. We select the number of epochs based on the convergence of the performance on the validation set, in terms of F_1-measure on the positive class. For each corpus combination, the training and validation sets are the union of the individual training and validation sets of each merged corpus. The models are then evaluated on the IFS-IT test set.

Monolingual setting Table 3 shows the performance in terms of precision, recall and F_1-measure for the Italian-only models and corpora combinations. The top-performing model is Incel AlBERTo, which achieves a test F_1 of 0.707 when training solely on HSD-FB. Compared to AlBERTo, this represents an improvement of 2.4 points. To a lesser degree, the same can be observed with regard to Incel UmBERTo and UmBERTo (+0.9 F_1 points), when using the same combination. In both cases, this shows that pre-training AlBERTo and UmBERTo using MLM on Italian posts extracted from *Il forum dei brutti* is effective in improving their performance.

The worst results are obtained when training solely on HSD-TW, with Incel AlBERTo and Incel UmBERTo performing worse than UmBERTo and AlBERTo, showing an opposite trend to the one observed when training on HSD-FB. The validation scores are also lower for HSD-TW combinations, compared to combinations including HSD-FB, showing that the models have a harder time learning from HSD-TW. This is coherent with the results obtained by teams participating in the two HaSpeeDe shared tasks [7, 9] and with the fact that HSD-FB's messages are "longer and more correct than those in Twitter, allowing systems (and humans too) to

Table 3
Performance when fine-tuning the monolingual models on Italian-only corpora. Epochs (e) selected based on validation F_1. Best scores in bold, second-best underlined; ■ = corpus used for training.

	HSD-FB	HSD-TW	AMI-20	(e)	$F_{1\ val}$	R_{val}	P_{val}	$F_{1\ test}$	R_{test}	P_{test}
UmBERTo	■			5	0.855±0.003	0.868	0.843	0.696±0.010	**0.879**	0.576
UmBERTo		■		4	0.754±0.004	0.800	0.713	0.432±0.060	0.319	0.685
UmBERTo			■	4	0.914±0.004	0.931	**0.899**	0.569±0.031	0.520	0.631
UmBERTo	■	■		4	0.788±0.006	0.824	0.755	0.666±0.024	0.758	0.595
UmBERTo	■		■	5	0.883±0.004	0.900	0.867	0.697±0.019	0.747	0.653
UmBERTo		■	■	5	0.828±0.003	0.844	0.814	0.596±0.017	0.526	0.688
UmBERTo	■	■	■	5	0.822±0.003	0.836	0.808	0.680±0.016	0.692	0.671
Incel UmBERTo	■			5	0.867±0.006	0.887	0.848	**0.705±0.009**	0.870	0.593
Incel UmBERTo		■		4	0.756±0.002	0.810	0.708	0.403±0.024	0.285	0.692
Incel UmBERTo			■	4	**0.918±0.001**	**0.946**	0.891	0.652±0.031	0.608	0.705
Incel UmBERTo	■	■		4	0.790±0.003	0.831	0.754	0.660±0.014	0.696	0.627
Incel UmBERTo	■		■	5	0.886±0.002	0.901	0.872	0.704±0.005	0.732	0.678
Incel UmBERTo		■	■	2	0.831±0.003	0.866	0.799	0.648±0.011	0.544	**0.802**
Incel UmBERTo	■	■	■	5	0.828±0.003	0.853	0.804	0.699±0.029	0.718	0.682
AlBERTo	■			4	0.850±0.003	0.899	0.807	0.683±0.006	**0.941**	0.537
AlBERTo		■		1	0.752±0.006	0.817	0.698	0.520±0.089	0.426	0.716
AlBERTo			■	2	0.907±0.004	**0.952**	0.866	0.528±0.022	0.517	0.542
AlBERTo	■	■		2	0.775±0.003	0.803	0.750	0.695±0.007	0.786	0.623
AlBERTo	■		■	3	0.879±0.003	0.918	0.843	0.705±0.011	0.803	0.629
AlBERTo		■	■	3	0.820±0.001	0.888	0.762	0.652±0.018	0.645	0.660
AlBERTo	■	■	■	2	0.808±0.011	0.872	0.753	0.684±0.015	0.821	0.587
Incel AlBERTo	■			5	0.847±0.005	0.863	0.831	**0.707±0.007**	0.791	0.639
Incel AlBERTo		■		1	0.748±0.002	0.785	0.715	0.506±0.035	0.370	**0.805**
Incel AlBERTo			■	5	**0.912±0.003**	0.930	**0.895**	0.617±0.018	0.562	0.685
Incel AlBERTo	■	■		2	0.771±0.004	0.791	0.752	0.673±0.016	0.721	0.632
Incel AlBERTo	■		■	5	0.873±0.003	0.888	0.858	0.668±0.014	0.663	0.674
Incel AlBERTo		■	■	1	0.818±0.004	0.864	0.776	0.656±0.007	0.593	0.736
Incel AlBERTo	■	■	■	4	0.800±0.009	0.828	0.773	0.688±0.017	0.747	0.639

find more and more clear indications of the presence of HS" [7]. The fact that messages in HSD-FB are longer is also coherent with the Italian incel models performing better than the vanilla models when training on HSD-FB, since *Il forum dei brutti* on average contains rather long posts (∼53 avg. tokens),[6] unlike Twitter corpora, which were limited to 280 characters per tweet prior to 2023. Finally, another element which might explain the lower performance when training on HSD-TW is that it contains hate speech against migrants, which might not be as relevant when it comes to *Il forum dei brutti*, since racism is not all that prevalent in this forum, compared to misogyny.

As regards combining different Italian corpora, the strategy yields the highest performance for AlBERTo and UmBERTo when training on both HSD-FB and AMI-20. However, once the models are MLM-trained on *Il forum dei brutti*, the performance decreases for some combinations, with MLM pre-training seemingly nullifying the improvements obtained by merging different corpora. Therefore, while some improvement can be observed by merging different corpora, MLM appears to be a more effective strategy for improving the performance of the models, although it requires greater computational resources.

Bilingual setting Table 4 reports the results for the bilingual setting. Compared to the best combination using mBERT_{base}, which achieves a test F_1 of 0.688, the best combination using Incel mBERT achieves a test F_1 of 0.722 (+3.4 F_1 points), which is also the highest score across both language settings. Just like in the monolingual setting, mBERT_{base} performs better when only training it on HSD-FB (in addition to IFS-EN). Conversely, Incel mBERT performs better when training on AMI-20 and IFS-EN. This is interesting, since the incorporation of the AMI-20 corpus lowered the performance of all Italian-only models, compared to only training on HSD-FB. Since misogyny is the main way hate speech is expressed in *Incels.is* (39.44% of the instances in IFS-EN

[6] Obtained with BertTokenizer: https://huggingface.co/docs/transformers/model_doc/bert#transformers.BertTokenizer

Table 4
Performance when fine-tuning the multilingual models on mono- and bilingual corpora combinations. Epochs (e) selected based on validation F_1. Best scores in bold; ■ = corpus used for training.

		HSD-FB	HSD-TW	AMI-20	IFS-EN	(e)	$\mathbf{F}_{1\ val}$	$\mathbf{R}_{val}$	$\mathbf{P}_{val}$	$\mathbf{F}_{1\ test}$	$\mathbf{R}_{test}$	$\mathbf{P}_{test}$
mBERT	Monolingual	■				4	0.807±0.008	0.846	0.775	0.651±0.013	0.891	0.516
			■			4	0.717±0.012	0.745	0.693	0.493±0.037	0.455	0.540
				■		5	**0.885±0.003**	**0.890**	**0.881**	0.425±0.034	0.384	0.479
		■	■			2	0.732±0.014	0.743	0.724	0.619±0.034	0.711	0.552
		■		■		3	0.844±0.002	0.873	0.818	0.639±0.023	0.747	0.560
			■	■		3	0.789±0.008	0.830	0.754	0.560±0.023	0.621	0.516
		■	■	■		5	0.784±0.002	0.793	0.775	0.600±0.014	0.634	0.569
					■	5	0.846±0.010	0.854	0.837	0.465±0.046	0.345	**0.725**
	Bilingual	■			■	3	0.841±0.004	0.852	0.832	**0.688±0.006**	**0.921**	0.549
			■		■	3	0.846±0.002	0.866	0.827	0.529±0.041	0.482	0.588
				■	■	5	0.844±0.007	0.849	0.838	0.634±0.023	0.587	0.692
		■	■		■	4	0.844±0.006	0.844	0.844	0.616±0.028	0.675	0.568
		■		■	■	5	0.842±0.004	0.844	0.840	0.676±0.012	0.801	0.585
			■	■	■	4	0.847±0.008	0.841	0.853	0.570±0.048	0.535	0.620
		■	■	■	■	5	0.837±0.008	0.829	0.845	0.613±0.019	0.639	0.590
Incel mBERT	Monolingual	■				1	0.817±0.009	0.865	0.777	0.668±0.016	0.841	0.557
			■			5	0.727±0.006	0.757	0.701	0.461±0.032	0.379	0.589
				■		4	**0.895±0.007**	**0.912**	**0.879**	0.574±0.047	0.511	0.666
		■	■			3	0.747±0.007	0.774	0.723	0.605±0.015	0.640	0.574
		■		■		3	0.854±0.004	0.896	0.816	0.654±0.025	0.752	0.583
			■	■		4	0.802±0.003	0.840	0.767	0.645±0.019	0.655	0.639
		■	■	■		3	0.799±0.004	0.825	0.774	0.648±0.022	0.644	0.653
					■	3	0.855±0.003	0.877	0.834	0.516±0.071	0.386	**0.807**
	Bilingual	■			■	5	0.859±0.010	0.853	0.864	0.708±0.007	**0.889**	0.588
			■		■	2	0.861±0.015	0.866	0.856	0.615±0.025	0.558	0.690
				■	■	4	0.853±0.009	0.863	0.844	**0.722±0.028**	0.704	0.746
		■	■		■	5	0.857±0.007	0.855	0.859	0.679±0.014	0.731	0.635
		■		■	■	4	0.856±0.009	0.857	0.856	0.689±0.011	0.707	0.673
			■	■	■	5	0.850±0.007	0.839	0.860	0.644±0.010	0.580	0.725
		■	■	■	■	5	0.869±0.003	0.878	0.861	0.700±0.013	0.702	0.698

are misogynous) and Incel mBERT was pre-trained using posts extracted from this forum, the performance boost could be due to the fact that the model is better at learning about misogynous language compared to mBERT$_{base}$ and the Italian-only models.

On average, the lowest performance is achieved when training separately on IFS-EN and on the Italian corpora (monolingual rows in Table 4). When using bilingual data, the worst results are obtained when training on HSD-TW and combinations containing it, coherently with the results in the monolingual settings shown in Table 3.

For almost all combinations of Italian corpora, performance increases once IFS-EN is added to the training data, i.e. bilingual data leads to better performance.

Monolingual vs. bilingual The results of our experiments show that the highest performance is not obtained by fine-tuning on the Italian-only corpus combinations, but on the bilingual ones. Indeed, for four bilingual corpus combinations out of seven, Incel mBERT's performance is higher than all other models. The combinations for which Incel mBERT does not beat all the others are HSD-FB+HSD-TW, HSD-FB+AMI-20 and HSD-TW+AMI-20.

Since mBERT was originally pre-trained in 104 languages and AlBERTo and UmBERTo were pre-trained only on Italian corpora, the fact that Incel mBERT can outperform them by pre-training on just $1M$ bilingual instances is rather unexpected. Even more interesting is the fact that, although we are testing on an entirely Italian corpus, Incel mBERT also outperforms Incel AlBERTo and Incel UmBERTo. Therefore, in the approached scenarios, using bilingual instances to pre-train a multilingual model using MLM yields higher performance than pre-training Italian models only on Italian posts. Furthermore, the number of Italian posts used to train Incel AlBERTo and Incel UmBERTo is $627k$, which is greater than the $500k$ Italian posts used for Incel mBERT.

As such, we could arguably conclude that the model

Table 5
Performance of our best model on on the different subclasses of the IFS-IT test corpus.

Class	Precision	Recall	F_1	Inst.
None	0.806	0.857	0.830	300
Misogyny	0.746	0.674	0.708	187
Racism	1.000	0.875	0.933	8
Both	1.000	1.000	1.000	5
Macro	0.888	0.851	0.868	500

is learning to spot hate speech more effectively in IFS-IT by learning language-agnostic incel concepts, since Incel mBERT is pre-trained on posts extracted from two incel forums in two different languages. Although the two considered incel communities are distinct, the hateful Red Pill ideology has spread internationally and is shared by both. This could explain why Incel mBERT performs better than the Italian-only models: the model might be learning about incel hate speech by paying more attention to the sociological concepts underlying the language, and putting less focus on purely linguistic features, ultimately improving its performance.

Performance on misogyny/racism subclasses Table 5 reports the performance of the best model —the one obtained by fine-tuning Incel mBERT on IFS-EN ∪ AMI-20— on the individual misogyny and racism labels of the IFS-IT test set. When looking at the difference between the performance on misogyny vs. racism, we notice a stark difference, with racism having perfect precision and a much higher F_1. Expectedly, this also translates into the instances that are both misogynous and racist, with perfect precision and recall. The explanation for the racist instances being much easier to detect is twofold: (*i*) the number of instances which are only racist is much smaller (8 vs. 187) and (*ii*) compared to the misogyny expressed by *Il forum dei brutti* users, the racism is much more explicit and simpler to identify. This can be seen in the examples in Table 6, which display explicit language in the first four instances, which contain racism. Here, the model can easily detect hate, even though users might even attempt to auto-censor themselves by substituting letters with numbers, as in example #1 (most likely in order to bypass automatic forum filters). Conversely, the misogyny in the last two samples is much more implicit, with the model failing to detect misogyny in sample #5.

6. Conclusions

In this paper, we have presented an approach to improve the performance of hate speech detection models in Italian incel posts. Our experiments show that domain-adapting transformer models to the contents of incel forums boosts their performance when predicting the hatefulness of incel forum posts, both when using Italian-only and multilingual models. The increase in performance obtained through MLM pre-training is particularly high when using bilingual training data with mBERT_{base}, which might indicate that the model is learning about incel hate speech by learning language-agnostic incel concepts. We have also shown that for the base Italian models (AlBERTo and UmBERTo) fine-tuning on combinations of different Italian corpora can lead to a boost in performance. However, this performance boost is nullified after MLM pre-training, which appears to be a more effective strategy for improving the performance of the models. When looking at racism vs. misogyny identification in posts extracted from *Il forum dei brutti*, the former appears to be much easier to detect. This seems due to the fact that racist language is much more explicit than misogynous language in the scrutinized forum, but further research is needed to ascertain such a supposition.

In future work, we plan to experiment with different resources for MLM pre-training, using corpora in different languages, since it seems multilingual models such as mBERT_{base} are capable of learning about hate speech in a language-agnostic way from multiple languages. In addition, with more computational resources, larger corpora and more training epochs could be used to further improve the performance of the models. Lastly, further experiments can be carried out as regards the

Table 6
Examples of racist and misogynous posts from IFS-IT, with gold annotation and Incel mBERT prediction labels.

Post	Gold	Pred.
Compagno le n3gr3 sono oggettivamente brutte, le asiatiche lo sono in media - ma quelle belle lo sono davvero e staccano di misura le cosiddette belle nostrane.	Both	Both
No perchè sessualmente mi fanno schifo le negre e i trans (più questi ultimi eh).	Both	Both
Ora capisco perché non scopava, curry percepito anche se non è curry, currycel in pratica.	Rac.	Rac.
Che dire allora dei terroni quasi tutti arabi quindi negri e di quei rari bianchi europei che provengono da altre nazioni europee? La mafia cioè i terroni stanno importando queste merde in massa per farci terronizzare e negrizzare come loro.	Rac.	Rac.
le 8+ sono davvero rare. tuttavia, dal 5 in su si atteggiano tutte come fossero modelle...	Mis.	None
Probabilmente la vedrai tra qualche settimana ad ipergamare con qualche architetto	Mis.	Mis.

performance of the scrutinized models on the individual sub-tasks of misogyny and racism identification, respectively.

References

[1] F. Alkomah, X. Ma, A Literature Review of Textual Hate Speech Detection Methods and Datasets, Information 13 (2022). doi:10.3390/info13060273.

[2] H. Van, Mitigating Data Scarcity for Large Language Models, 2023. doi:10.48550/arXiv.2302.01806.

[3] A. Nagle, Kill All Normies: Online Culture Wars from 4chan and Tumblr to Trump and the Alt-Right, Zero Books, Winchester, Hampshire, UK, 2017. doi:10.5817/PC2018-3-270.

[4] S. Jaki, T. De Smedt, M. Gwóźdź, R. Panchal, A. Rossa, G. De Pauw, Online Hatred of Women in the Incels.me Forum: Linguistic Analysis and Automatic Detection, Journal of Language Aggression and Conflict 7 (2019) 240–268. doi:10.1075/jlac.00026.jak.

[5] T. Farrell, M. Fernandez, J. Novotny, H. Alani, Exploring Misogyny Across the Manosphere in Reddit, in: Proceedings of the 10th ACM Conference on Web Science, ACM, Boston, MA, 2019, pp. 87–96. doi:10.1145/3292522.3326045.

[6] K. C. Gothard, Exploring Incel Language and Subreddit Activity on Reddit, Honors college senior thesis, UVM Honors College, 2020. URL: https://scholarworks.uvm.edu/hcoltheses/408.

[7] C. Bosco, F. Dell'Orletta, F. Poletto, M. Sanguinetti, M. Tesconi, Overview of the EVALITA 2018 Hate Speech Detection Task, in: [25], 2018, pp. 67–74.

[8] A. Cimino, L. De Mattei, F. Dell'Orletta, Multi-task Learning in Deep Neural Networks at EVALITA 2018, in: [25], 2018, pp. 86–95.

[9] V. Basile, D. M. Maria, C. Danilo, L. C. Passaro, et al., EVALITA 2020: Overview of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, in: [26], 2020, pp. 1–7.

[10] E. Lavergne, R. Saini, G. Kovács, K. Murphy, TheNorth @ HaSpeeDe 2: BERT-based Language Model Fine-tuning for Italian Hate Speech Detection, in: [26], 2020.

[11] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, Association for Computational Linguistics, Minneapolis, MN, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[12] M. Polignano, P. Basile, M. De Gemmis, G. Semeraro, V. Basile, AlBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets, in: R. Bernardi, R. Navigli, G. Semeraro (Eds.), Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019), volume 2481, CEUR-WS, Bari, Italy, 2019.

[13] L. Parisi, S. Francia, P. Magnani, UmBERTo: An Italian Language Model Trained with Whole Word Masking, https://github.com/musixmatchresearch/umberto, 2020.

[14] E. Fersini, D. Nozza, P. Rosso, Overview of the EVALITA 2018 Task on Automatic Misogyny Identification (AMI), in: [25], 2018.

[15] E. Fersini, D. Nozza, P. Rosso, AMI @ EVALITA 2020: Automatic Misogyny Identification, in: [26], 2020.

[16] A. Muti, A. Barrón-Cedeño, UniBO @ AMI: A Multi-Class Approach to Misogyny and Aggressiveness Identification on Twitter Posts Using AlBERTo, in: [26], 2020.

[17] C. Di Bonaventura, A. Muti, M. A. Stranisci, O-Dang at HODI and HaSpeeDe3: A Knowledge-Enhanced Approach to Homotransphobia and Hate Speech Detection in Italian, in: M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi (Eds.), Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2023), volume 3473, CEUR-WS, Parma, Italy, 2023.

[18] T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated Hate Speech Detection and the Problem of Offensive Language, in: Proceedings of the 11th International AAAI Conference on Web and Social Media, ICWSM '17, Montreal, Canada, 2017, pp. 512–515.

[19] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, A. Mukherjee, HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection, Proceedings of the AAAI Conference on Artificial Intelligence 35 (2021) 14867–14875. doi:10.1609/aaai.v35i17.17745.

[20] T. Caselli, V. Basile, J. Mitrović, M. Granitzer, HateBERT: Retraining BERT for Abusive Language Detection in English, in: Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021), Association for Computational Linguistics, Online event, 2021, pp. 17–25. doi:10.18653/v1/2021.woah-1.3.

[21] A. Pelicon, R. Shekhar, B. Škrlj, M. Purver, S. Pollak, Investigating Cross-Lingual Training for Offensive Language Detection, PeerJ Computer Science 7 (2021) e559. doi:10.7717/peerj-cs.559.

[22] O. Gokhale, A. Kane, S. Patankar, T. Chavan,

R. Joshi, Spread Love Not Hate: Undermining the Importance of Hateful Pre-training for Hate Speech Detection, 2022. URL: http://arxiv.org/abs/2210.04267.
[23] P. Gajo, A. Muti, K. Korre, S. Bernardini, A. Barrón-Cedeño, On the Identification and Forecasting of Hate Speech in Inceldom, in: G. Angelova, M. Kunilovskaya, R. Mitkov (Eds.), Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2023), volume 2263, INCOMA Ltd., Varna, Bulgaria, 2023, pp. 373–384.
[24] M. Sanguinetti, G. Comandini, E. Di Nuovo, S. Frenda, M. Stranisci, C. Bosco, T. Caselli, V. Patti, I. Russo, HaSpeeDe 2 @ EVALITA2020: Overview of the EVALITA 2020 Hate Speech Detection Task, in: [26], 2020.
[25] T. Caselli, N. Novielli, V. Patti, P. Rosso (Eds.), Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2018), volume 2263, CEUR-WS, Turin, Italy, 2018.
[26] V. Basile, D. Croce, M. Di Maro, L. C. Passaro (Eds.), Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020), volume 2765, CEUR-WS, Online event, 2020.

Linking the Dictionary of Medieval Latin in the Czech Lands to the LiLa Knowledge Base

Federica Gamba[1], Marco C. Passarotti[2] and Paolo Ruffolo[2]

[1]Charles University, Faculty of Mathematics and Physics, Malostranské náměstí 25, 118 00 Prague, Czechia
[2]Università Cattolica del Sacro Cuore, largo A. Gemelli 1, 20123 Milan, Italy

Abstract
The paper presents the process of linking the *Dictionary of Medieval Latin in the Czech Lands* to the LiLa Knowledge Base, which adopts the Linked Data paradigm to make linguistic resources for Latin interoperable. An overview of the Dictionary and of the architecture of the LiLa Knowledge Base is first provided; then, the stages of the process of linking the *Dictionary* to LiLa's collection of lemmas are described. In conclusion, a query illustrates how interoperability allows for full exploitation of Latin resources.

Keywords
Linked Data, LiLa Knowledge Base, dictionary, Medieval Latin

1. Introduction

Many resources are available for Latin, making it a particularly privileged language among the historical ones. However, most often those resources are scattered, with their sparsity representing a substantial hindrance to the full exploitation of the information they contain. To overcome the sparsity of resources, stored in separate silos, the CIRCSE Research Center in Milan, Italy, started the LiLa - Linking Latin project[1] (2018-2023), which built a Knowledge Base to make all existing textual and lexical resources for Latin interoperable by adopting the four principles of the Linked Open Data (LOD) paradigm [1]: 1) use URIs as names for things; 2) use HTTP URIs so that people can look up those names; 3) when someone looks up a URI, provide useful information; 4) include links to other URIs. so that they can discover more things.[2]

The LiLa Knowledge Base has already a wide coverage in terms of interlinked resources. Classical Latin is naturally well-represented, as proved by the LASLA corpus, which includes 130 Classical Latin texts [2], and by the Lewis and Short dictionary [3], whose primary focus is on Classical Latin. Later stages of Latin are found as well in the Knowledge Base; for instance, the *Index Thomisticus* Treebank [4] comprises texts by Thomas Aquinas (1225–1274), the UDante treebank [5] encompasses Medieval Latin works written by Dante Alighieri, and the Computational Historical Semantics Corpus [6] includes e.g. the *Decretum Gratiani*, a collection of canon law compiled in the XII century.

However, while the LiLa Knowledge Base already extends over a large temporal range, its spatial coverage is not as wide. So far, no resource from the Eastern Europe areas where Latin was spoken has been linked. For this reason, we decided to link to LiLa the *Dictionary of Medieval Latin in the Czech Lands*, a lexical resource that aims at collecting the Latin vocabulary as it emerged in that area during the Middle Ages. The resource encompasses a late variety of Latin (1000-1500 CE), strongly tied to a specific geographical area. These two levels of variability, along the temporal and spatial axes, make it extremely interesting to link such a resource to the Knowledge Base, as we expect it to contribute to enlarge the amount of lemmas stored in the large collection of Latin lemmas that represents the core part of the whole architecture of LiLa.

The paper is organised as follows. Section 2 introduces the LiLa Knowledge Base. Section 3 describes the *Dictionary*. Section 4 outlines the process of linking the *Dictionary* to LiLa. Section 5 shows the added value of interoperability of Latin resources in LiLa by presenting a query on the *Dictionary* interlinked.

2. The LiLa Knowledge Base

The LiLa Knowledge Base [7] 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 [8], Ontolex-Lemon for lexical data [9, 10] and POWLA for corpus data [11]. As for the latter, the Resource Description Framework (RDF) [12] is a data model

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
gamba@ufal.mff.cuni.cz (F. Gamba); marco.passarotti@unicatt.it (M. C. Passarotti); paolo.ruffolo@unicatt.it (P. Ruffolo)
0000-0003-3632-0594 (F. Gamba); 0000-0002-9806-7187 (M. C. Passarotti); 0000-0002-9120-0846 (P. Ruffolo)

[1]https://lila-erc.eu/.
[2]https://www.w3.org/wiki/LinkedData.

used to describe information in terms of triples, consisting of: (1) a predicate-property that connects (2) a subject (i.e. a resource) with (3) its object (another resource or a literal). Data recorded in the form of RDF triples are queried via the SPARQL query language [13].

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 Bank, a collection of around 200,000 lemmas taken from the database of the morphological analyser LEMLAT [14] and constantly extended. A `lila:Lemma`[3] is a subclass 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 lemmatisation 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 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.

3. The Dictionary of Medieval Latin in the Czech Lands

The *Dictionary of Medieval Latin in the Czech Lands*[10] is a lexical resource developed at the Department of Medieval Lexicography of the Institute of Philosophy of the Czech Academy of Sciences. It aims to collect the vocabulary of Medieval Latin as it was used in the Czech lands from about 1000 CE, when Latin writing began in the area, to 1500 CE. In light of this aim, the *Dictionary* features three types of entries:

- Vocabulary taken from Classical Latin without any semantic change in the Middle Ages. Only source citations with a translation illustrate its meaning. E.g., *labellum* 'small lip'.
- Vocabulary taken from Classical Latin with changes. This type of entry is composed of two parts: first, ancient meanings are listed; then, the + sign introduces Medieval developments (syntactical alternations, new phrases, meanings of the word coined in Medieval times). E.g., *falcatus* 'curved' + 'shod'.
- Vocabulary that emerged during the Middle Ages. Such entries are marked with an asterisk (∗). Square brackets [ ] with etymology and references to other dictionaries including the word follow the heading of the entry. E.g., *emicamen* 'splendour, clarity'.

Moreover, the *Dictionary* relies on a differential method to capture all divergences – at several linguistic layers – of Medieval Latin vocabulary inherited from the ancient era as compared with the Classical norms. Indeed, it records language phenomena not attested in the 8th edition (and later unchanged editions) of Georges' Latin-German Lexicon [15].

The material the *Dictionary* is built upon amounts today to ca. 800,000 excerpt sheets, assembled from various sources of Czech provenance (diplomatical, official, belles-lettres, scientific literature, etc.). What is particularly valuable is that not only edited texts served as a source to build the *Dictionary*, but also several manuscripts and old prints from Czech and foreign libraries were used. The excerption of sources has been carried out from 1934, when the project of the *Dictionary* started, until the 1970s. In 1977 the first fascicle was published, illustrating editorial principles and lists of sources and abbreviations. Overall, the electronic database [16] is built upon, and comprises, the three volumes prepared by Silagiová and colleagues ([17], [18], [19]).

So far, letters A-M are covered, for a total of 48,452 entries. 24,943 out of these are full entries (provided with meanings, definitions, grammatical information, examples), whereas 23,509 are references that point to full entries (see 3.1). Fascicle 24, encompassing entries beginning with N, is currently under preparation.

The *Dictionary* is accessible through a dedicated website[11] and can be downloaded from the LINDAT/CLARIAH-CZ research infrastructure[12] as a compressed set of XML files.

3.1. XML Files

We provide a brief overview of the structure of the XML files of the *Dictionary*, as those data are relevant for the process of modeling information and linking the entries

[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] The Czech title is *Slovník středověké latiny v českých zemích*; the Latin one *Latinitatis medii aevi lexicon Bohemorum*.
[11] http://lb.ics.cas.cz.
[12] http://hdl.handle.net/11234/1-4792.

to the LiLa Knowledge Base. The lexical entry for the adjective *exquisitus* 'exquisite' (Figure 1) will serve as an example of the XML files of the resource.

The whole entry is encoded as the value of an `entryFree` element, which contains a single unstructured entry in any kind of lexical resource, such as a dictionary or lexicon. Core information about the entry is provided through attributes: the lemma is given, together with a numerical unique identifier assigned to it; `georges='True'` or `'False'` specifies whether an entry for the same lemma is found or not in Georges' dictionary. Optionally, `hom_nr` distinguishes homographs, and `type='reference'` denotes that the entry is just a reference to a different one; for instance, the dummy entry for *geniculor* 'to bend the knee' is just a reference to its active counterpart *geniculo*, which, in light of that, is the only full entry of the two (with meanings, grammatical information, etc.). Then, in the `orth` element the lemma is stated once again as a value; `orth` includes the attribute `type` either with value 'lemma', if it is a full entry, or with value 'ref_all', if it is a reference.

Following the lemma, the `gramGrp` element encodes grammatical information about the lexical item, roughly corresponding to its Part of Speech (POS) and (possibly) its inflectional category. In the case of *exquisitus*, the value `<gramGrp> 3. </gramGrp>` indicates that it is an adjective of the first class, i.e. with three distinct endings for the three genders (*exquisitus, -a, -um*, respectively for the forms of masculine, feminine and neuter singular nominative).

The `sense` elements (possibly more than one for a same entry) capture the different meanings of a lexical item. For each `sense`, a definition `def` is provided both in Latin and in Czech, with the Czech one corresponding to a translation of the Latin counterpart. Some examples are listed as well, together with their source. The label *script. et form.* is used to record orthographic and morphological variants (e.g., *exequisitus* for *exquisitus*), while the label *metr.* for metrical ones.

4. Linking the Dictionary to LiLa

This section describes the process of linking the *Dictionary of Medieval Latin in the Czech Lands* to the LiLa Knowledge Base. The coverage of the linking task is not yet complete, as, so far, we have been working only with full entries (i.e., excluding those with `type='reference'`).

As mentioned in Section 2, in LiLa the lemma works as interface between interlinked resources. In light of the pivotal use of lemmas, the core operation at the base of the linking process is to perform a string match between the tuples *(lemma, POS)* in the resource to be linked and the lemmas and their POS in the LiLa Lemma Bank. The goal is to retrieve the correct lemma in the Lemma Bank corresponding to the lemma/POS used in the entry of the *Dictionary*.

The string match results in three possible outcomes: a) only one matching lemma/POS is found in the Lemma Bank; b) more than one matching lemma/POS is found, resulting in an ambiguity due to homography; c) no matching lemma/POS is found, as the couple is not present in the Lemma Bank.

The first outcome is overall straightforward and does not raise particular issues. The second one, i.e. multiple matches found, requires disambiguation to be performed. To this aim, grammatical information (inflectional classes) can be exploited, although they do not always guarantee a full resolution of the ambiguity; Subsection 4.1 elaborates on this. The third possibility, i.e. missing matches, represents the most interesting outcome; firstly, because it entails enlarging the Lemma Bank with new canonical forms of citation, and secondly because it allows to reflect about the peculiar aspects of the variety of the Latin vocabulary represented in the *Dictionary*, by focusing on those lexical items provided by the *Dictionary* that result as out-of-vocabulary with respect to the current Lemma Bank of LiLa.

4.1. Aligning Grammatical Information

In order to automatically disambiguate multiple matches, we exploit the grammatical information provided by the *Dictionary* in the `gramGrp` element. However, this information is not encoded in a fully standardised way, thus requiring an alignment to be performed. Indeed, we need to define a set of heuristics to align grammatical categories as they are encoded in the *Dictionary* and the set of tags employed in LiLa, which is based on the Universal POS tagset [20] and expanded with inflectional categories. As an illustration, the word *acus* 'needle' has *-us, f.* as `gramGrp`, i.e. the genitive ending and the gender. From that we can generalise and establish a correspondence between the genitive ending in *-us* together with the gender, as found in the *Dictionary*, and a NOUN with inflectional class `n4`[13] in LiLa.

In most cases, grammatical information provided by the *Dictionary* is sufficiently fine-grained to provide all elements needed to disambiguate the multiple linking to the Lemma Bank, as it roughly consists of POS and inflection class, like in the case of *acus*. Yet, sometimes only information corresponding to POS is available. Several substantives are marked just as *subst.* (e.g., *deptar*, type of medicinal plant), which makes it non-trivial, if possible at all, to infer an inflectional category.

[13] n4 corresponds to fourth declension nouns.

```xml
<?xml version='1.0' encoding='utf8'?>
<entryFree   georges='True' lemma='exquisitus' n='263390'>
<orth type='lemma'>exquisitus</orth>
<gramGrp><norm/>3.<norm_end/></gramGrp>

<form><norm/>exequ- <bibl type='source' index_as='source'>LupCus 44</bibl><norm_end/></form>

<sense georges='false'>
<sense type='hier' n='a'>
<sense type='expl'>
<def lang='lat' index_as='definition-lat'><i/>electus, egregius <i_end/></def>
<def lang='cs' index_as='definition-cze'><i/> - vybraný, vynikající<i_end/></def>

</sense>
</sense>
<sense type='hier' n='b'>
<sense type='expl'>
<def lang='lat' index_as='definition-lat'><i/>quaesitus, singularis, insolitus <i_end/></def>
<def lang='cs' index_as='definition-cze'><i/> - hledaný, zvláštní, neobvyklý:<i_end/></def>

<div type='examples'>
<cit type='example' index_as='example'><norm/>deliciosi cibi e-i wymysleny <bibl type='source' index_as='source'>HusBethl II 75</bibl>.<norm_end/></cit>

</div>

</sense>
</sense></sense>

<ab type='formated'><b/>exquisitus<b_end/><norm/>  3.   <norm_end/><i/>script. et form.:<i_end/><norm/> exequ- |LupCus 44| <norm_end/><b/>  a<b_end/><no:

</entryFree>
```


Figure 1: XML file of the *Dictionary* entry *exquisitus* 'exquisite'.

4.2. Linking to the Lemma Bank

After aligning the two tagsets, we proceed to link the *Dictionary* entries to the Lemma Bank. The one-to-one matches, i.e. lemmas in the *Dictionary* that match with just one lemma in the LiLa Lemma Bank with respect to both lemma and POS, have been considered validated. The following subsections discuss the two other scenarios, namely one-to-many and one-to-zero matches.

4.2.1. One-to-Many

The string match on lemma and POS results in 827 ambiguous matches. Therefore, we add inflectional class as a further constraint; as a result, 303 lemmas are disambiguated automatically, while 445 still remain ambiguous and need to be inspected manually. For instance, for *lacertus* a correspondence in the Lemma Bank is found with *lacertus* 'upper arm' and *lacertus* 'lizard; a seafish', both NOUNs of the second declension (inflectional class n2). Only the manual checking of the meaning can thus allow to retrieve the correct match.

4.2.2. One-to-Zero

After performing the string match on lemma and POS, no match is found in the LiLa Lemma Bank for 10,278 lemmas. Among those, we automatically handle adverbs, verbs and *pluralia tantum* to find out whether they could be linked to the Lemma Bank respectively as hypolemmas of an adjective, lemma variants of a corresponding verb with opposite voice (active if deponent and vice versa) or lemma variant of a noun in singular form. By defining a set of heuristics applied automatically, we find that: (a) 92 adverbs can be linked to the adjective they are derived from (e.g., *homagialiter* - *homagialis* 'of homage'); (b) 18 verbs can be linked to their counterpart with opposite voice (e.g., *attaedio* - *attaedior* 'to bore'); (c) 80 plural forms can be linked to their singular equivalent (*moscilli* - *moscillus* 'little habit').

A closer look at lemmas that remain unmatched (10,088) raises interesting insights, allowing for some linguistic considerations. First, clear evidence of areal contact is provided by forms like *bosako, -onis* and *kamennikko, -onis*. As the spelling reveals, these forms are the result of a contact with the language that was spoken in the area at that time, namely Old Czech. Indeed, *bosako* comes from the Czech form *bosák*, denoting a monk that by virtue of the rule has to walk barefoot, while *kamenniko* 'stonemason' derives from *kameník*. Additionally, several lemmas pertain to very specific domains. Consider e.g. *ascoa*, a sea animal, *igenecha*, a type of quadruped[14], or *cinapus*, a species of fish, as an example of vocabulary of fauna. Flora is found as well: e.g., *elipurgis*, corresponding to *Cynoglossum officinale*, *bulboquilon*, 'mandrake', and *atomana*, a herb. Similar forms evidently display the specificity of some domains covered by the *Dictionary of Medieval Latin in the Czech Lands*.

4.3. Results

The string match on lemmas and POS tags results in 55.5% one-to-one mappings; for 3.3% of entries more than one possible match was found, while for 41.2% no match was retrieved. The amount of lemmas that are not found in the Lemma Bank reflects the nature of the *Dictionary*, and especially its temporal, geographical and domain specificity. For comparison purposes, consider, for instance, that the process of linking the bilingual Latin-English

[14]Possibly the common genet.

dictionary by Lewis and Short, which is focused on Classical Latin, resulted in only 9% of unmatched lemmas [21]. The percentage of no-match entries increases to 70% in the case of the *Neulateinische Wortliste* by Ramminger [22], which covers a time range spanning between 1300 and 1700 and features entries mirroring contemporary changes in the society, e.g. *typographus* ‘typographer’.

Figure 2 shows an example of an entry of the *Dictionary* (*exquisitus*) linked to the LiLa Knowledge Base. The (yellow) node in the center of Figure 2 is the `ontolex:lexicalEntry` for *exquisitus*, which is linked via the property `lime:entry`[15] to the node that represents the entire *Dictionary* (an individual of the class `lime:lexicon`[16]) and to the corresponding lemma in the Lemma Bank via the property `ontolex:canonicalForm`. The lexical entry works as gateway to all information associated to it in the resource. For instance, Figure 2 shows how the two meanings associated to *exquisitus* in its entry in the *Dictionary* are modeled. The two definitions provided by the resource (in Latin and in Czech) are linked to the lexical entry as individuals of the class `ontolex:lexicalSense`[17] via the property `ontolex:sense`[18]. Each sense is the specific lexicalisation of a more general `ontolex:lexicalConcept`[19] to which the sense is linked via the property `ontolex:isLexicalizedSenseOf`[20].

Although not visible in Figure 2, the lemma *exquisitus* in the Lemma Bank is linked via `ontolex:canonicalForm` to the entries for *exquisitus* in several other lexical resources and to its occurrences (tokens) in the textual resources interlinked in LiLa[21] .

5. Querying the Dictionary in LiLa

This Section presents a query to exemplify the added value of interoperability between the resources linked to LiLa[22]. The query, available within a set of preocompiled queries in the SPARQL endpoint of LiLa, retrieves all those lemmas whose entries in the *Dictionary* include the word *natura* ‘nature’ in their definition(s) and do not occur also in the Lewis and Short dictionary, and returns the number of their occurrences in the textual corpora linked to LiLa.

[15] http://www.w3.org/ns/lemon/lime#entry.

[16] http://www.w3.org/ns/lemon/lime#Lexicon.

[17] http://www.w3.org/ns/lemon/ontolex#LexicalSense.

[18] http://www.w3.org/ns/lemon/ontolex#sense.

[19] http://www.w3.org/ns/lemon/ontolex#LexicalConcept.

[20] http://www.w3.org/ns/lemon/ontolex#isLexicalizedSenseOf.

[21] For the full list of the resources currently made interoperable through LiLa, see https://lila-erc.eu/data-page/.

[22] The linguistic resources for Latin linked in LiLa can be queried either via a query graphical interface (https://lila-erc.eu/query/) or through a SPARQL endpoint (https://lila-erc.eu/sparql/).

The 11 retrieved lemmas[23] occur in 5 corpora, for a total of 132 occurrences, 5 out of which are found in the Computational Historical Semantics corpus[24] , 104 in the *Index Thomisticus* Treebank,[25] 4 in UDante,[26] 1 in the CIRCSE Latin Library[27] (specifically, in Augustine’s *Confessiones*) and 18 in the LASLA corpus[28] [2]. The results of the query confirm once again the specificity of the *Dictionary of Medieval Latin in the Czech Lands*. Having excluded Classical lemmas that can also be found in the Lewis and Short dictionary, what remains are mostly lemmas that occur in corpora featuring texts of later stages of Latin: for instance, the texts from the *Index Thomisticus* Treebank and UDante date back respectively to XIII and XIV centuries. The only exception is represented by the LASLA corpus, which includes Classical Latin. Yet, occurrences in LASLA are limited to the lemma *mollitia* ‘softness, weakness’, which is therefore attested in Classical times as well, while all the other lemmas appear to have originated later.

6. Conclusions

Linking the *Dictionary* to the LiLa Knowledge Base not only was a further step towards the full exploitation of linguistic resources for Latin, thanks to their interoperability, but also contributed to improve the degree of linguistic diversity represented in LiLa as for three aspects, that are particularly relevant for Latin as a language that was used for centuries all over Europe: (a) diachronic diversity: the *Dictionary* collects a portion of the Latin vocabulary that emerged in Medieval times; (b) diatopic diversity: the lexical resource includes items from a specific area, namely the Czech lands; (c) domain-based diversity: quite frequently the entries of the *Dictionary* belong to very specific domains (e.g., flora and fauna; see Section 4.2.2). The contribution of the lemmas from the *Dictionary* in enlarging the LiLa Lemma Bank is thus considerable both in terms of quantity and in terms of quality, and highlights the importance of linking to the Knowledge Base also resources that feature non-standard varieties of Latin.

In the near future, we intend to finalise the linking, by disambiguating ambiguous matches and adding missing lemmas to the Lemma Bank, as well as by including referencing lemmas besides full entries (see Section 3.1). We also intend to model citations of attestations, i.e. refer-

[23] *Accidentalis*, *bestialitas*, *connaturalis*, *connaturalitas*, *contingentia*, *eligibilis*, *finitas*, *fumositas*, *leuiathan*, *materialitas*, *mollitia*.

[24] http://lila-erc.eu/data/corpora/CompHistSem/id/corpus.

[25] http://lila-erc.eu/data/corpora/ITTB/id/corpus.

[26] http://lila-erc.eu/data/corpora/UDante/id/corpus.

[27] http://lila-erc.eu/data/corpora/CIRCSELatinLibrary/id/corpus. Collection of Latin texts enhanced with different layers of linguistic annotation.

[28] http://lila-erc.eu/data/corpora/Lasla/id/corpus.

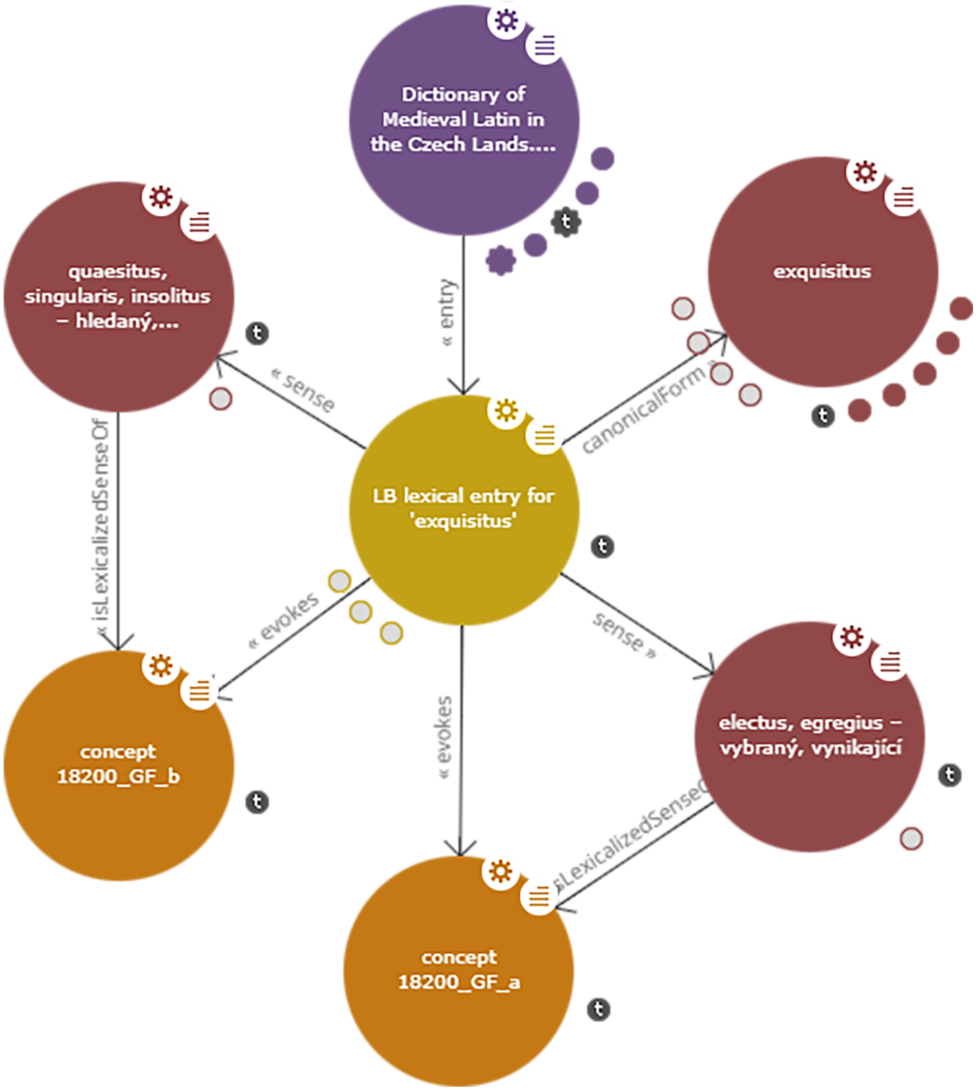


Figure 2: The entry for *exquisitus* after being linked to LiLa.

ences to other dictionaries where an entry is found and to sources of examples. Moreover, we plan to link to the Knowledge Base some documents from the same area and period as the *Dictionary*, such as the Czech Medieval sources from the AHISTO project[29]. However, these documents are currently available only as raw texts, and would need to be lemmatised before the linking. Given the peculiar nature of their Latin variety, conditioned by the Czech language and rich of local proper names, lemmatisation with the currently available trained models will probably provide low accuracy rates. Once again, this proves the importance of collecting non-standard Latin data (and resources) and investigating to what ex-

[29]https://nlp.fi.muni.cz/projekty/ahisto/portal.

tent Latin varieties differ.

Acknowledgments

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. This work was partially supported by the Grant No. 20-16819X (LUSyD) of the Czech Science Foundation (GACR).
We want to thank Pavel Nývlt for his collaboration in providing the database “The Dictionary of Medieval Latin in Czech Lands”, available via the LINDAT/CLARIAH-CZ Research Infrastructure, supported by the Ministry of Education, Youth, and Sports of the Czech Republic (Project No. LM2018101).

References

[1] T. Berners-Lee, J. Hendler, O. Lassila, The Semantic Web, Scientific american 284 (2001) 34–43.

[2] M. Fantoli, M. Passarotti, F. Mambrini, G. Moretti, P. Ruffolo, 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, European Language Resources Association, Marseille, France, 2022, pp. 26–34. URL: https://aclanthology.org/2022.ldl-1.4.

[3] C. T. Lewis, C. Short, A Latin Dictionary, Clarendon Press, Oxford, 1879.

[4] M. Passarotti, The Project of the Index Thomisticus Treebank, in: M. Berti (Ed.), Digital Classical Philology. Ancient Greek and Latin in the Digital Revolution, De Gruyter, Berlin, 2019, pp. 299–319.

[5] F. M. Cecchini, R. Sprugnoli, G. Moretti, M. Passarotti, UDante: First Steps Towards the Universal Dependencies Treebank of Dante’s Latin works, in: Seventh Italian Conference on Computational Linguistics, CEUR-WS.org, Bologna, 2020, pp. 1–7.

[6] T. Geelhaar, A. Mehler, B. Jussen, A. Henlein, G. Abrami, D. Baumartz, T. Uslu, et al., The Frankfurt Latin Lexicon from Morphological Expansion and Word Embeddings to Semiographs, Studi e saggi linguistici 58 (2020) 45–81. doi:10.4454/ssl.v58i1.276.

[7] M. Passarotti, F. Mambrini, G. Franzini, F. M. Cecchini, E. Litta, G. Moretti, P. Ruffolo, R. Sprugnoli, Interlinking through Lemmas. The Lexical Collection of the LiLa Knowledge Base of Linguistic Resources for Latin, Studi e Saggi Linguistici 58 (2020) 177–212.

[8] C. Chiarcos, M. Sukhareva, Olia–ontologies of linguistic annotation, Semantic Web 6 (2015) 379–386.

[9] P. Buitelaar, P. Cimiano, J. McCrae, E. Montiel Ponsoda, T. Declerck, Ontology lexicalisation: The lemon perspective, in: Proceedings of theWorkshops 9th International Conference on Terminology and Artificial Intelligence, 2011, pp. 33–36. URL: http://tia2011.crim.fr/Workshop-Proceedings/TIAW-2011.pdf, ontology Engineering Group - OEG.

[10] J. P. McCrae, J. Bosque-Gil, J. Gracia, P. Buitelaar, P. Cimiano, The Ontolex-Lemon model: development and applications, in: Proceedings of eLex 2017 conference, 2017, pp. 19–21.

[11] C. Chiarcos, 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 9, Springer, 2012, pp. 225–239.

[12] O. Lassila, R. Swick, Resource Description Framework (RDF) model and syntax specification, 1998. Available online at https://www.w3.org/TR/1999/REC-rdf-syntax-19990222/.

[13] E. Prud’Hommeaux, A. Seaborne, SPARQL query language for RDF, W3C working draft 4 (2008).

[14] M. Passarotti, M. Budassi, E. Litta, P. Ruffolo, The Lemlat 3.0 package for morphological analysis of Latin, in: Proceedings of the NoDaLiDa 2017 workshop on processing historical language, 2017, pp. 24–31.

[15] K. E. Georges, H. Georges, Ausführliches lateinisch-deutsches Handwörterbuch, 2 vols, Hannovre/Leipzig: Hahnsche Buchhandlung (1913).

[16] J. Ctibor, P. Nývlt, On-line Dictionary of medieval Latin in the Czech lands, 2021. URL: http://hdl.handle.net/11234/1-4792, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

[17] Z. Silagiová, J. Černá, H. Florianová, P. Nývlt, H. Šedinová, K. Vršecká, Latinitatis medii aevi lexicon Bohemorum – The Dictionary of Medieval Latin in Czech Lands, Volume I (A-C), second, revised edition, 2018.

[18] Z. Silagiová, P. Nývlt, J. Černá, H. Florianová, B. Kocánová, H. Šedinová, K. Vršecká, Latinitatis medii aevi lexicon Bohemorum – The Dictionary of Medieval Latin in Czech Lands, Volume II (D-H), second, revised edition, 2019.

[19] Z. Silagiová, J. Černá, D. Martínková, B. Kocánová, M. Koronthályová, K. Vršecká, R. Mašek, J. Matl, H. Miškovská, P. Nývlt, H. Šedinová, I. Zachová, Latinitatis medii aevi lexicon Bohemorum – The Dictionary of Medieval Latin in Czech Lands, Vol-

ume III (I-M), 1995 to 2016. The electronic version has been created by Jan Ctibor.
[20] S. Petrov, D. Das, R. McDonald, A Universal Part-of-Speech Tagset, in: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey, 2012, pp. 2089–2096. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/274_Paper.pdf.
[21] F. Mambrini, E. Litta, M. Passarotti, P. Ruffolo, Linking the Lewis & Short Dictionary to the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin, in: Proceedings of the Eighth Italian Conference on Computational Linguistics (CLiC-it 2021). Milan, Italy, January 26-28, 2022, 2021, pp. 214–220.
[22] F. Iurescia, E. Litta, M. Passarotti, M. Pellegrini, G. Moretti, P. Ruffolo, Linking the Neulateinische Wortliste to the LiLa Knowledge Base of Interoperable Resources for Latin, in: Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 82–87. URL: https://aclanthology.org/2023.latechclfl-1.9.

CheckIT!: A Corpus of Expert Fact-checked Claims for Italian

Jacopo Gili[1], Lucia Passaro[2] and Tommaso Caselli[3]

[1]*Department of Computer Science, University of Turin, Italy*
[2]*Department of Computer Science, University of Pisa, Italy*
[3]*CLCG, University of Groningen, Netherlands*

Abstract
This paper introduces *CheckIT!*, a resource of expert fact-checked claims, filling a gap for the development of fact-checking pipelines in Italian. We further investigate the use of three state-of-the-art generative text models to create variations of claims in zero-shot settings as a data-augmentation strategy for the identification of previously fact-checked claims. Our results indicate that models struggles in varying the surface forms of the claims.

Keywords
Fact-checking, Corpora, Data augmentation, Generative AI Model Evaluation

1. Introduction

The pollution of the information ecosystem by means of misleading or false information has reached unprecedented levels at a global scale. This has been possible thanks to a combination of multiple factors, among which the collapse of (local and national) journalism; an increasing sense of distrust in science and evidence-based facts; and the presence of computational amplification tools such as bots [1].

Manually fact-checking claims is an expensive operation (in terms of time and effort) and in many cases, it comes too late. Authors in [2] have shown how false and inaccurate information propagates online eight times faster than true and reliable information. Letting this kind of information free to circulate may have harmful impacts on groups and individuals as well as threaten the texture of democratic societies. It is thus urgent and critical to implement automatic solutions that can assist content moderators and information professionals to promptly react in presence of false or misleading information.

In Figure 1, we present the full fact-checking verification pipeline [3]. As it appears, multiple steps are involved: (i) assessing whether a claim is worth of being fact-checked; (ii) checking whether the claim has been previously fact-checked; (iii) if this is not the case, then evidence to evaluate the veracity of the claim must be gather (usually using reliable sources online); and (iv) finally, assessing its veracity status. Having access to a database of previously fact-checked claims is a valuable resource for fact-checkers because claims tend to be repeated (even if with small variations) over time, and this is particularly true for politicians [4, 5, 6]. The availability of such a resource can save time and contribute to mitigate the effects of misinformation.

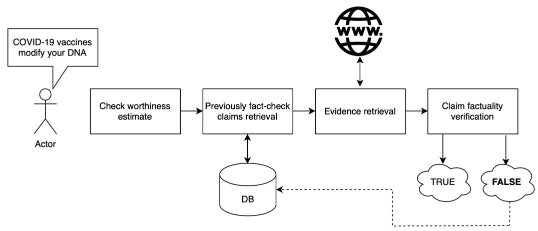


Figure 1: Fact-checking pipeline: (i) check-worthiness; (ii) previously verified claims retrieval; (iii) claim evidence retrieval; (iv) claim veracity assessment. The figure is an adaptation from [5] and [7].

This paper presents *CheckIT!* the first corpus of previously fact-checked claims for Italian. In its current version, *CheckIT!* is based on a collection of 3,577 claims of 317 Italian politicians and public figures, provided with evidences and veracity labels.

Contributions Our main contributions can be summarized as follows: (i) **we introduce *CheckIT!***, a fact-checking resource filling a gap in the language resource panorama for Italian for claim verification and, more generally, for misinformation detection and countering; (ii) we conducted a **feasibility study on automatic paraphrasing** in Italian, exploring the potential of leveraging advanced Natural Language Processing (NLP) techniques

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
jacopo.gili584@edu.unito.it (J. Gili); lucia.passaro@unipi.it (L. Passaro); t.caselli@rug.nl (T. Caselli)
0009-0007-1343-3760 (J. Gili); 0000-0003-4934-5344 (L. Passaro); 0000-0003-2936-0256 (T. Caselli)

for generating high-quality texts that preserve the original meaning of the claims while introducing linguistic variations; (iii) we propose an **initial framework for the automatic extension of fact-checking resources**, which enables the continuous growth and enrichment of *CheckIT!* with additional fact-checked claims and related evidence.

The remainder of the paper is structured as follows: Section 2 describes the data collection, the veracity label harmonization process, and presents an analysis of the dataset. In Section 3, we discuss the results of our paraphrases experiments with text generation tools for Italian (`mT5`, `Camoscio`, `ChatGPT`) as a strategy to extend the variability of expressions of previously fact-checked claims. Our efforts have been mainly focused on assessing the quality of these generative tools. Related work is discussed in Section 4. Finally, Section 5 concludes the paper and draws directions of further development.[1]

2. *CheckIT!*: Data Collection and Analysis

CheckIT! has been obtained by collecting all available fact-checked claims from *Pagella Politica*[2] and structuring them into a unified representation format. *Pagella Politica* is a web-based news outlet fully dedicated to fact-checking and analysis of political news in Italy since October 2012. *Pagella Politica* aims to provide accurate information and it aims to empower readers with knowledge, fostering a deeper and more informed engagement with the political landscape. To gather all claims we have obtained access to *Pagella Politica*'s public APIs, and scraped claims covering a period from October 3rd 2012 to April 26th 2023. In our harmonization process, we retained 3,577 claims out of 4,547, with 17 common attributes (see Table A in Appendix A for details). For each claim, the evidence text has been split into sentences and all hyperlinks have been extracted and stored separately.

As for the veracity verdicts, *Pagella Politica* has changed its labelling scheme since its firsts appearance: they have moved from a five-label scheme ("Vero" [True], "Ni" [Mostly true], "C'eri quasi" [Half True], "Pinocchio andante" [False], "Panzana pazzesca" [Pants on fire]) to verbose verdicts with explanations (e.g., "the politician is right"). However, mapping the verbose verdicts to the original five labels was impossible, especially for non-experts and for the very nuanced difference between the labels "Ni" [Mostly true], "C'eri quasi" [Half True]. In addition to this, verbose verdicts are not optimal for training machine learning classifiers. To avoid losing 270 of the most recent claims, we decided to simplify the labeling scheme. We thus reduced the label granularity from five to three, by collapsing "Ni" [Mostly true] and "C'eri quasi" [Half True] into "Impreciso" [Imprecise], and "Pinocchio andante" [False] and "Panzana pazzesca" [Pants on fire] into "Falso" [False] to separate certainly true and false information from the imprecise one. Subsequently, we manually analyzed the verbose verdicts and assigned the corresponding label.

At the end of this operation, we have the following label distributions: **1,255** claims labelled as "**Vero**" [True], **1,512** labelled as "**Impreciso**" [Imprecise], and **810** labelled as "**Falso**" ["False"]. The label distribution is not perfectly balanced, with the majority class being "Impreciso". While on the one hand it is comforting to see that, in absolute terms, politicians do not overtly lie, on the other hand it is not surprising to observe that politicians may manipulate data and news as a propaganda strategy to convince the audience of their arguments. The tendency of the last 16 months is worrying as 61.14% (192 out of 314) of the claims have been fact-checked as false.

The distribution of the claims over time is not well balanced as illustrated in Table 1. Early years see a rich activities, while this diminishes in more recent times (see also Figure 1 in Appendix B. Election years (2013, 2018, and 2022 for national Parliament elections; 2014 and 2019 for European Parliament elections) contain the majority of the fact-checked claims.

		Verdict		Year		Verdict	
Year	True	Imprecise	False	Year	True	Imprecise	False
2012∗	143	121	54	2018	25	88	29
2013	315	294	97	2019	79	180	76
2014	263	255	83	2020	63	153	92
2015	144	191	60	2021	59	84	95
2016	40	73	24	2022	83	16	136
2017	22	42	28	2023∗	19	4	46

Table 1
CheckIT!: Distribution of verdict labels per year. Years marked with ∗ cover less than 12 months.

When focusing on the most debated topics, the large majority of the claims (79.54%) concern four main areas: Social Issues (983), Economics (919), Institutions (599), Foreign Affairs (350), clearly corresponding to topics of public interest as they directly affect the lives of citizens and the working of the democratic institutions. Some of these topics face peaks of fact-checking in correspondence of relevant events. For example, 21.71% of the claims concerning Foreign Affairs are registered in 2015 during the European migrant crisis[3]; 12.71% of claims related to Social Issues are in 2020 during the first phases of the COVID-19 pandemics; 10.63% of claims for Economics are in 2019, when the citizens' income ("Reddito di Cittadinanza") was introduced. An overview of the distribution of the topics and the corresponding verdict

[1] Code and data: https://github.com/Jj-source/Check-It.
[2] https://pagellapolitica.it/fact-checking
[3] https://en.wikipedia.org/wiki/2015_European_migrant_crisis

labels is presented in Table 2.

Topic	True	Verdict Imprecise	False
Environment	44	59	18
Social Issues	291	402	290
Economics	291	459	169
Justice (Civil and Criminal)	62	63	30
Foreign Affairs	124	163	63
Institutions	266	233	91
Other	50	65	34
Not Specified	127	64	214

Table 2
CheckIT!: Distribution of verdict labels per topic.

CheckIT! contains 317 unique Italian politicians/public figures. The corpus has a very long tail, with the large majority of politicians being attributed only one claim. An aspect to consider in this dataset concerns the popularity and the roles that politicians have. The top 10 politicians are all prominent figures in the Italian political sphere. They are (former) secretary of major political parties, Prime Ministers, ministers, or popular party leaders. This top 10 covers 52.80% of all the fact-checked claims. On the other hand, only 18.92% (60) of the politicians appear in at least 10 claims. A statistic we are not able to provide in full given the current version of *CheckIT!* is the distribution of the claims per political party. Although we know that there are 16 political parties, more than 1,000 claims lack this information, i.e., it was not available through the APIs. Table 3 shows the distributions of the verdict labels for the top 10 political figures.

Politician	True	Verdict Imprecise	False
Matteo Renzi	169	186	73
Matteo Salvini	68	137	109
Beppe Grillo	62	125	55
Giorgia Meloni	39	69	64
Silvio Berlusconi	32	60	52
Luigi Di Maio	40	64	37
Renato Brunetta	39	46	27
Enrico Letta	52	38	12
Alessandro Di Battista	31	40	13
Laura Boldrini	52	26	1

Table 3
CheckIT!: Distribution of verdict labels for the top 10 politicians.

Are the claims biased? Documentation of potential biases in datasets has gained increasing awareness in the NLP community. From what we have seen so far, the dataset does not seem to present major biases in terms of political orientations, i.e., sovra-representation of a political party or side. The top 10 politicians (Table 3) are quite evenly distributed among the three major political areas that characterizes Italy in the past 10 years: three for the center-left/left, three for the M5S area, and four for the center-right/right. As a way to estimate the presence of potential biases, we have run a simple machine learning experiment to estimate the prediction of the veracity labels from the claims themselves. Previous work has shown that this is not an easy task (if even possible) [8, 9]. We have thus split *CheckIT!* into a Train (80%) and Test (20%) and trained two linear Support Vector Machine (SVM) models. We have used a simple TF-IDF vectorization[4] in both cases. In the second experiment, we have concatenated the names of the politicians to the text of their claims. Both SVMs are further compared with a Dummy classifier implementing majority voting. Results are summarized in Table 4.

Model	Label	P	R	Macro-F1
Dummy	True	0.0	0.0	0.195
	Imprecise	0.416	1.0	
	False	0.0	0.0	
SVM - claims only	True	0.458	0.392	0.422
	Imprecise	0.457	0.573	
	False	0.387	0.294	
SVM - claims & politicians	True	0.456	0.411	<u>0.425</u>
	Imprecise	0.449	0.553	
	False	0.411	0.300	

Table 4
Claim veracity prediction. Underscore figures indicate the best result.

As expected, the results are way far from being satisfying. Although the SVMs seem to learn something, when compared to the Dummy classifier, their overall macro-F1 is well below 0.5. A slight improvement in the False class can be observed when the names of the politicians are concatenated with the claims. However, this appears to be an effect of the data split (out of 317 unique entities, 121 appear both in our train and test splits). While on one hand, these results further confirm a limited presence of bias in the data, they further support previous results on the difficulty of assessing the veracity of a claim from the claim itself, especially when it is uttered using formally correct language [10].

3. Automatic Paraphrases of Fact-checked Claims

This battery of experiments is devoted to evaluate the use of generative language models to enrich fact-checking datasets by varying the expression of the claims. This

[4] We have used word uni- and bigrams, character n-grams, with a range of 2-5, and stop-word removals.

data augmentation approach plays a pivotal role for the development of robust systems for the identification of previously fact-checked claims (step (ii) in Figure 1), and thus reducing the manual workload of professional fact-checkers. In particular, we generate five alternative versions of the *CheckIT!* claims using three generative models, namely `mt5`, `Camoscio`, and `ChatGPT`.

mt5 The only available Italian model for paraphrase generation is `aiknowyou/mt5-base-it-paraphraser`. This model is based on `mt5` and fine-tuned on Tapaco and STS Benchmark datasets for Paraphrasing. `mt5` [11] is a multilingual variant of `T5` [12] that was pre-trained on a new Common Crawl-based dataset covering 101 languages. The TaPaCo Corpus, used for fine-tuning, is a freely available paraphrase corpus for 73 languages extracted from the Tatoeba database.

Camoscio The second method we used to generate paraphrases is based on instruction-based models. Specifically, we used `Camoscio` [13], an Italian version of Alpaca [14] obtained by instruction-tuning LlAMA on Italian data automatically translated with `ChatGPT`. To obtain the paraphrases, we used the following prompt: "Scrivi 5 parafrasi di questa frase: *claim*" ('Write 5 paraphrases of this sentence: *claim*') where *"claim"* is one of the original claims belonging to *CheckIT!*.

ChatGPT The third method consists in directly prompting `ChatGPT` APIs[5] with the following text: "Parafrasa le seguenti frasi: *claims*" ('Paraphrase the following: *claims*') where "claims" are the original claims belonging to *CheckIT!*.

For all models, we have used the default parameters. For `ChatGPT`, the temperature was left to 1 and `max_token` to 2,000.

3.1. Evaluation Metrics

To assess the goodness of the generated texts, we conducted a comprehensive evaluation encompassing comparisons between the model-generated paraphrases, the original sentences, and paraphrases by three human annotators.

In all evaluation settings, we use four automatic metrics, Cosine Similarity (*Cos*), BLEU [15], ROUGE [16], and BERTScore [17], to gain multiple perspectives on the models' performance and gauge both the fidelity and the variations with respect to the original claims exhibited by the models. In particular, *Cos* will return the the semantic similarity between the two texts based on word frequency distributions. BLEU, although commonly used for Machine Translation, will assess the overlap of n-grams (word sequences) between the claim and the paraphrases as a proxy for text variation. Similarly, ROUGE,[6] which returns the overlap of n-grams and the longest common subsequence, will also assess the variations of the generated text with respect to the original claims. Finally, BERTScore, which calculates the similarity between two sentences or texts by utilizing contextualized embeddings from pre-trained language models and comparing the embeddings of overlapping words between the candidate and reference sentences, will help us to better assess the semantic similarity.

3.2. Evaluation Settings and Results

Overall, we have four evaluation blocks. The first block is based on 10% (i.e., 357) of the claims in *CheckIT!*. In this case, we compared the automatically-generated paraphrases against the original claims.

The latter three are based on a subset of 50 claims that have been independently paraphrased by the human annotators.[7] Annotators were given basic instructions which closely resembled the prompts of `Camoscio` and `ChatGPT`: "Provide a paraphrase for each of the following sentences." In the second evaluation block, we compare human-generated paraphrases (a total of 150 instances corresponding to 3 different variants per claim) with the original claims. In the third evaluation block, we evaluate the human-generated paraphrases with respect to each other: for each data point, we compared the four metrics between all the combinations of annotators (e.g., A1 *vs.* A2; A2 *vs.* A3; A1 *vs.* A3, and so on). Note that some of the metrics (i.e., ROUGE and BERTScore) are not symmetric, thus results may vary. In the fourth evaluation block, we compared the automatically-generated paraphrases against human-generated ones.

Block I: Machines *vs.* Claims The summary of the results is in Table 5. `Camoscio` produced a considerable number of empty paraphrases. To ensure fair comparisons, we excluded these empty paraphrases from the metrics calculation. Overall, we notice a trend of higher variation in generation for `ChatGPT`. Despite the high average cosine similarity with the original texts, `ChatGPT` displayed better performances for creative rephrasing. Surprisingly, `mT5` does not perform very well, as indicated by the high scores across all metrics. Differences between the training materials and the *CheckIT!* data may have had an impact. Finally, `Camoscio` is the worst performing models. Out of 1,785 possible paraphrases for the 357 claims considered, it fails to generate an output 1,320 times. The few successful cases are almost exact

[5] We used `GPT 3.5-turbo`.

[6] ROUGE is a set of metrics: ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-LSum.

[7] All annotators are also the author of this paper.

Metric	ChatGPT	mt5	Camoscio*
BERT-P	0.80	0.88	0.91
BERT-R	0.79	0.82	0.85
BERT-F1	0.79	0.85	0.88
BLEU	0.13	0.27	0.59
Cos	0.93	0.92	0.95
ROUGE-1	0.56	0.71	0.87
ROUGE-2	0.32	0.58	0.82
ROUGE-L	0.48	0.68	0.86
ROUGE-LS	0.48	0,68	0.86

Table 5
Generated paraphrases *vs.* claims.

Metric	ChatGPT	mt5	Camoscio*
BERT-P	0.85	0.80	0.86
BERT-R	0.87	0.81	0.86
BERT-F1	0.86	0.80	0.85
BLEU	0.25	0.16	0.29
Cos	0.84	0.81	0.87
ROUGE-1	0.60	0.53	0.62
ROUGE-2	0.40	0.31	0.41
ROUGE-L	0.54	0.49	0.56
ROUGE-LS	0.54	0.49	0.56

Table 8
Generated paraphrases *vs.* human.

Metric	A1	A2	A3
BERT-P	0.76	0.78	0.83
BERT-R	0.71	0.72	0.80
BERT-F1	0.73	0.75	0.81
BLEU	0.05	0.07	0.16
Cos	0.83	0.86	0.93
ROUGE-1	0.35	0.44	0.61
ROUGE-2	0.16	0.22	0.38
ROUGE-L	0.28	0.35	0.56
ROUGE-LS	0.28	0.35	0.56

Table 6
Human paraphrases *vs.* claims.

Metric	A1-A2	A1-A3	A2-A1	A2-A3	A3-A1	A3-A2
BERT-P	0.80	0.78	0.81	0.80	0.76	0.76
BERT-R	0.81	0.76	0.80	0.76	0.78	0.80
BERT-F1	0.80	0.77	0.80	0.78	0.77	0.78
BLEU	0.10	0.06	0.10	0.07	0.05	0.07
Cos	0.89	0.85	0.89	0.87	0.85	0.87
ROUGE-1	0.45	0.36	0.45	0.42	0.36	0.42
ROUGE-2	0.22	0.17	0.22	0.19	0.17	0.19
ROUGE-L	0.37	0.29	0.37	0.35	0.29	0.35
ROUGE-LS	0.37	0.29	0.37	0.35	0.29	0.35

Table 7
Comparison across annotators.

repetitions of the original claims, as highlighted by the scores of the various measures and a manual inspection. Clear evidence of this parroting behavior is shown by the BLEU score.

Block II: Humans *vs.* Claims Scores are reported in Table 6. In general, it seems that humans introduce more superficial variations, as highlighted by BLEU and ROUGE. However, there is an increasing adherence to the original formulation of the claim among the annotators. Notably, A1 exhibited a greater propensity for variation in their paraphrasing, while A3 tended to produce paraphrases closer to the original texts, as evidenced by the higher BLEU and ROUGE-LS. Clearly, the closer in wording to the original claim, the bigger the impact also on the more semantic oriented measures such as BERTScore and *Cos*. While A1 and A2 present close performances, A3 achieves the highest results. It appears that divergent interpretations of what a paraphrases of a claim is and how to do it have affected the results, suggesting that more precise instructions will be needed in the future to achieve more varied results.

Block III: Human *vs.* Human As we delved into the comparison among the annotators (Table 7), we found that A1 and A2 produced paraphrases that were notably more similar to each other in comparison to those produced by A3. This clearly indicates that distinct stylistic preferences have been adopted.

Block IV: Machines *vs.* Humans We evaluated the quality of the generated paraphrases by comparing them to the three human-produced paraphrases, considering the latter as references. A summary of these results is presented in Table 8. Surprisingly, the automatically generated paraphrases have a higher degree of similarity and lexical overlap with the manually generated ones. The results for `Camoscio` are quite unexpected, as it seems to qualify as the best second system after `ChatGPT`. However, this is a distortion due to the measures and the manual paraphrases. As we have seen in Table 6, A3 is very conservative, generating paraphrases close to the original claim. This is also the behavior of `Camoscio`, as observed in Table 5. On the other hand, `mT5` and `ChatGPT` appears to be more suitable candidates for this task.

4. Related Work

Automatic fact-checking is a growing field of research and previous work has already investigated multiple aspects. Early work has focused on detecting rumors in Social Media [18, 19], or on the identification of the stance of a document with respect to a claim [20, 21, 22]. Following Figure 1, the claim detection step is one of the easiest and one of the most controversial subtask. While the identification of claims is comparable to Attribution Detection [23, 24], the check-worthiness status of claims is challenging since it involves some level of subjectivity. To address this issue, previous work has collected data from authoritative sources run by professional fact-checkers (e.g., PolitiFact, Snopes) or have seen the direct involvement of human experts for the veracity labelling [3, 4, 25, 26, 27, 28, 29, 30].

Evidence retrieval requires the identification of relevant passages from external resources that can be used to verify the claim. Two mainstream automatic verification methods are employed: Stance Detection and Natural Language Inference (NLI) [25, 31, 32]. They make use of unstructured data (i.e., textual sources) and assume that evidence is available for every claim and make a closed world world assumption, i.e., evidence is available only in one source. Complimentary methods make use of structured data, where evidence can be retrieved inside a knowledge graph [33].

Each of the subtasks involved in the fact-checking pipeline is framed as a classification task, with a varying number of labels: from a binary classification for the check-worthiness, to rich multi-class classification tasks for the veracity of the claim. For *CheckIT!* , we have opted for a three-way classification of the claim, in line with most of the previous work. The advantage of (more) fine-grained veracity classifiers is that it allows to capture also misleading or imprecise information and avoiding to reduce the world into a black or white picture.

5. Conclusions and Future Work

This work has introduced *CheckIT!*, an expert-curated fact-checked repository of claims by politicians and prominent public figures in Italy. *CheckIT!* covers 10 years of claims and it is the first publicly available dataset for fact-checking in Italian. In our analysis of *CheckIT!*, we have observed a drop in the numbers of fact-checked claims suggesting that manual fact-checking is increasingly difficult to conduct and that automated assisted tools are more and more needed.

We have conducted a preliminary investigation of three state-of-the-art automatic text generation tools for claim paraphrases. By combining multiple automatic measures, it appears that `ChatGPT` and `mT5` are the two best candidate to further explore, while `Camoscio` presents non-trivial issues with respect to failure to produce an output and variations of the generated texts.

Future work will focus on three aspects: conduct a qualitative (human-based) evaluation of the two best models; evaluate the generated paraphrases for previously fact-checked claim retrieval on the line of [5]; finally evaluate the generated paraphrasis against the topics.

Acknowledgments

The authors want to thank *Pagella Politica* and its director, Giovanni Zagni, for the access to the APIs that made the data collection for *CheckIT!* possible.

This work has been partially supported by the EU H2020 TAILOR project, GA n. 952215.

References

[1] C. Wardle, H. Derakhshan, Information disorder: Toward an interdisciplinary framework for research and policymaking, volume 27, Council of Europe Strasbourg, 2017.

[2] S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online, Science 359 (2018) 1146–1151.

[3] A. Vlachos, S. Riedel, Fact checking: Task definition and dataset construction, in: Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science, Association for Computational Linguistics, Baltimore, MD, USA, 2014, pp. 18–22. URL: https://aclanthology.org/W14-2508. doi:10.3115/v1/W14-2508.

[4] N. Hassan, G. Zhang, F. Arslan, J. Caraballo, D. Jimenez, S. Gawsane, S. Hasan, M. Joseph, A. Kulkarni, A. K. Nayak, et al., Claimbuster: The first-ever end-to-end fact-checking system, Proceedings of the VLDB Endowment 10 (2017) 1945–1948.

[5] S. Shaar, N. Babulkov, G. Da San Martino, P. Nakov, That is a known lie: Detecting previously fact-checked claims, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 3607–3618. URL: https://aclanthology.org/2020.acl-main.332. doi:10.18653/v1/2020.acl-main.332.

[6] S. Shaar, F. Alam, G. Da San Martino, P. Nakov, The role of context in detecting previously fact-checked claims, in: Findings of the Association

for Computational Linguistics: NAACL 2022, Association for Computational Linguistics, Seattle, United States, 2022, pp. 1619–1631. URL: https://aclanthology.org/2022.findings-naacl.122. doi:10.18653/v1/2022.findings-naacl.122.

[7] P. Nakov, A. Barrón-Cedeño, G. da San Martino, F. Alam, J. M. Struß, T. Mandl, R. Míguez, T. Caselli, M. Kutlu, W. Zaghouani, C. Li, S. Shaar, G. K. Shahi, H. Mubarak, A. Nikolov, N. Babulkov, Y. S. Kartal, M. Wiegand, M. Siegel, J. Köhler, Overview of the clef–2022 checkthat! lab on fighting the covid-19 infodemic and fake news detection, in: A. Barrón-Cedeño, G. Da San Martino, M. Degli Esposti, F. Sebastiani, C. Macdonald, G. Pasi, A. Hanbury, M. Potthast, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer International Publishing, Cham, 2022, pp. 495–520.

[8] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, Y. Choi, Truth of varying shades: Analyzing language in fake news and political fact-checking, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 2931–2937. URL: https://aclanthology.org/D17-1317. doi:10.18653/v1/D17-1317.

[9] S. Volkova, K. Shaffer, J. Y. Jang, N. Hodas, Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on Twitter, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 647–653. URL: https://aclanthology.org/P17-2102. doi:10.18653/v1/P17-2102.

[10] T. Schuster, R. Schuster, D. J. Shah, R. Barzilay, The Limitations of Stylometry for Detecting Machine-Generated Fake News, Computational Linguistics 46 (2020) 499–510. URL: https://doi.org/10.1162/coli_a_00380. doi:10.1162/coli_a_00380.

[11] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mt5: A massively multilingual pre-trained text-to-text transformer, arXiv preprint arXiv:2010.11934 (2020).

[12] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, The Journal of Machine Learning Research 21 (2020) 5485–5551.

[13] A. Santilli, Camoscio: An italian instruction-tuned llama, https://github.com/teelinsan/camoscio, 2023.

[14] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, T. B. Hashimoto, Stanford alpaca: An instruction-following llama model, 2023.

[15] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318. URL: https://aclanthology.org/P02-1040. doi:10.3115/1073083.1073135.

[16] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013.

[17] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with BERT, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net, 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.

[18] L. Derczynski, K. Bontcheva, M. Liakata, R. Procter, G. Wong Sak Hoi, A. Zubiaga, SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours, in: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 69–76. URL: https://aclanthology.org/S17-2006. doi:10.18653/v1/S17-2006.

[19] G. Gorrell, E. Kochkina, M. Liakata, A. Aker, A. Zubiaga, K. Bontcheva, L. Derczynski, SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 845–854. URL: https://aclanthology.org/S19-2147. doi:10.18653/v1/S19-2147.

[20] D. Küçük, F. Can, Stance detection: A survey, ACM Computing Surveys (CSUR) 53 (2020) 1–37.

[21] M. Hardalov, A. Arora, P. Nakov, I. Augenstein, A survey on stance detection for mis- and disinformation identification, in: Findings of the Association for Computational Linguistics: NAACL 2022, Association for Computational Linguistics, Seattle, United States, 2022, pp. 1259–1277. URL: https://aclanthology.org/2022.findings-naacl.94. doi:10.18653/v1/2022.findings-naacl.94.

[22] J. Zheng, A. Baheti, T. Naous, W. Xu, A. Ritter, Stanceosaurus: Classifying stance towards multicultural misinformation, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 2132–2151. URL: https://aclanthology.org/2022.emnlp-main.138.

[23] S. Pareti, PARC 3.0: A corpus of attribution rela-

tions, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), European Language Resources Association (ELRA), Portorož, Slovenia, 2016, pp. 3914–3920. URL: https://aclanthology.org/L16-1619.

[24] C. Scheible, R. Klinger, S. Padó, Model architectures for quotation detection, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 1736–1745. URL: https://aclanthology.org/P16-1164. doi:10.18653/v1/P16-1164.

[25] A. Hanselowski, C. Stab, C. Schulz, Z. Li, I. Gurevych, A richly annotated corpus for different tasks in automated fact-checking, in: Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 493–503. URL: https://aclanthology.org/K19-1046. doi:10.18653/v1/K19-1046.

[26] I. Augenstein, C. Lioma, D. Wang, L. Chaves Lima, C. Hansen, C. Hansen, J. G. Simonsen, MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims, 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), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 4685–4697. URL: https://aclanthology.org/D19-1475. doi:10.18653/v1/D19-1475.

[27] P. Atanasova, P. Nakov, L. Màrquez, A. Barrón-Cedeño, G. Karadzhov, T. Mihaylova, M. Mohtarami, J. Glass, Automatic fact-checking using context and discourse information, Journal of Data and Information Quality (JDIQ) 11 (2019) 1–27.

[28] N. Kotonya, F. Toni, Explainable automated fact-checking for public health claims, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 7740–7754. URL: https://aclanthology.org/2020.emnlp-main.623. doi:10.18653/v1/2020.emnlp-main.623.

[29] L. C. Passaro, A. Bondielli, P. Dell'Oglio, A. Lenci, F. Marcelloni, In-context annotation of topic-oriented datasets of fake news: A case study on the notre-dame fire event, Inf. Sci. 615 (2022) 657–677. URL: https://doi.org/10.1016/j.ins.2022.07.128. doi:10.1016/j.ins.2022.07.128.

[30] A. Bondielli, P. Dell'Oglio, A. Lenci, F. Marcelloni, L. C. Passaro, M. Sabbatini, Multi-fake-detective at EVALITA 2023: Overview of the multimodal fake news detection and verification task, in: M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi (Eds.), Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), Parma, Italy, September 7th-8th, 2023, volume 3473 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023. URL: https://ceur-ws.org/Vol-3473/paper32.pdf.

[31] J. Maillard, V. Karpukhin, F. Petroni, W.-t. Yih, B. Oguz, V. Stoyanov, G. Ghosh, Multi-task retrieval for knowledge-intensive tasks, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 1098–1111. URL: https://aclanthology.org/2021.acl-long.89. doi:10.18653/v1/2021.acl-long.89.

[32] M. Arana-Catania, E. Kochkina, A. Zubiaga, M. Liakata, R. Procter, Y. He, Natural language inference with self-attention for veracity assessment of pandemic claims, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 1496–1511. URL: https://aclanthology.org/2022.naacl-main.107. doi:10.18653/v1/2022.naacl-main.107.

[33] J. Kim, K.-s. Choi, Unsupervised fact checking by counter-weighted positive and negative evidential paths in a knowledge graph, in: Proceedings of the 28th international conference on computational linguistics, 2020, pp. 1677–1686.

Appendix A: *CheckIT!* Attribute Descriptions

Attribute	Value	Attribute	Value
`id`	unique id of the claim	`date`	timestamp of fact-checking
`link`	Pagella Politica URL	`content`	fact-checking evidence
`statement_date`	timestamp of the claim	`source`	URL of the news outlet/platform where the claim has appeared
`statement`	the claim	`verdict`	veracity label of the claim
`verdict_ext`	verbose veracity judgment of the claim	`politician`	full name of the politician or public figure owning the claim
`political_party`	Political party membership at the time of the claim	`platform`	Name of the news outlet/platform where the claim has appeared
`politicians_in`	the name(s) of any politician(s) mentioned in the claim (other than the owner of the claim)	`macro_area`	broader topic of the claim
`tags`	keywords to describe the content of the claim	`links`	list of URLs used to retrieve evidence, write the `content`, and the verdict
`versione`	versioning of the dataset		

Table A
CheckIT!: List of the attributes used to represent the data.

Appendix B: Verdict distribution overview

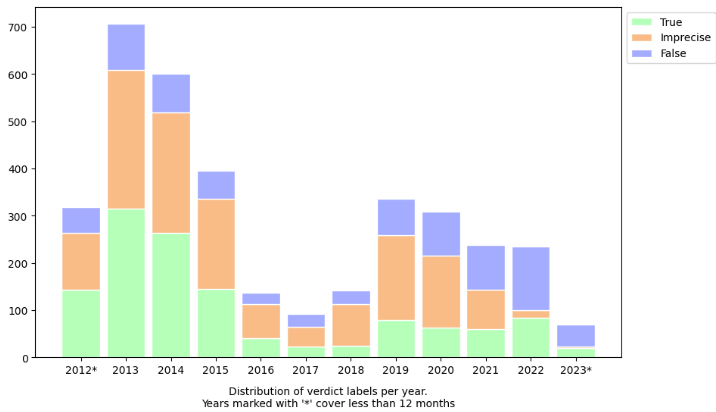


Figure 1: *CheckIT!*: Distribution of verdict labels per year (histogram)

End-to-end Dependency Parsing via Auto-regressive Large Language Models

Claudiu Daniel Hromei[1], Danilo Croce[1] and Roberto Basili[1]

[1]University of Rome Tor Vergata, Italy

Abstract
This paper presents a straightforward application of Large Language Models (LLMs) for Dependency Parsing. The parsing process is approached as a sequence-to-sequence task, where a language model takes a sentence as input and generates a bracketed form, allowing for the deterministic derivation of the dependency graph. The experimental evaluation explores the feasibility of utilizing LLMs for this purpose, while also assessing the process's sustainability with modest parameter sizes (training on a single GPU with limited resources) and investigating the impact of incorporating multilingual data during training. The results demonstrate that an end-to-end dependency parsing process can indeed be formulated using a task-agnostic architecture.

Keywords
Dependency Parsing, Large Language Models, End-to-End Parsing, Sequence-to-Sequence, Auto-Regressive Models

1. Introduction

Dependency parsing is a crucial component of natural language processing that plays a significant role in capturing the syntactic intricacies within sentences [1]. The primary objective of dependency parsing is to establish dependency relations among words. This allows humans to understand how words are connected and how they depend on one another in the sentence's structure [2]. Such understanding is instrumental in a wide range of applications, including semantic interpretation, machine translation, relation extraction, and various other linguistic tasks.

One notable parsing technique is the shift-reduce method, as exemplified in [3, 4]. This parser processes sentences from left to right, word by word while maintaining a buffer for words that are yet to be fully processed. Other approaches have been proposed, based on machine learning techniques, such as the biaffine neural networks, as in [5]. These networks, based on Bi-LSTMs, have proven effective in capturing complex dependencies between words. Another intriguing parsing approach is UDPipe [6], focused on parsing with the Universal Dependency Framework [7]. UDPipe stands out because it performs dependency parsing and other essential tasks like tokenization, morphological analysis, part-of-speech tagging, and lemmatization for multiple languages. UDPipe performs all these tasks without relying on external data. It employs a Bi-LSTM architecture fed with end-to-end, character-level, pre-trained, and contextualized embeddings. The model was trained on an extensive dataset of over a million sentences across different languages to capture cross-lingual relations effectively. The system was later extended [8] as UDPipe+ by incorporating multilingual BERT [9] in its token representations. However, while these methods have achieved state-of-the-art results in various languages, they are tailored specifically for the structure prediction problem based on ad hoc methods.

More recently, models based on the Transformer architecture [10] have gained popularity for their ability to perform classification, regression, and rewriting tasks. These models operate on a sequence and they output another sequence. For instance, the work in [11] introduced an end-to-end seq2seq method for dependency parsing, where the model directly predicts the relative position of the head for each word in the sentence. It also utilized a beam search decoder with tree constraints and sub-root decomposition to improve the results. Moreover, in [12] the authors have experimented with a multi-task, multilingual version of BERT [9]. This model was pre-trained on 104 languages and could predict not only dependency parsing trees but also lemmas, part-of-speech tags, and more for each word in an input sentence. One notable Transformer-based architecture is the LLaMA [13] foundational models. LLaMA is a large model with billions of parameters that generates output sequences in an auto-regressive manner based on the input and previously generated output tokens. It has been recently applied in [14] to a variety of linguistic tasks by instruction-tuning a monolithic architecture to solve them all.

In this work, we raise a crucial question about the applicability of models like LLaMA for predicting tree-

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
hromei@ing.uniroma2.it (C. D. Hromei); croce@info.uniroma2.it (D. Croce); basili@info.uniroma2.it (R. Basili)
0009-0000-8204-5023 (C. D. Hromei); 0000-0001-9111-1950 (D. Croce); 0000-0002-1213-0828 (R. Basili)

like structures. Specifically, we seek to explore whether such models can be used to define an end-to-end parsing process without relying on architecture choices that are task-dependent. We envision a system that, given an input sentence, predicts an output sequence in a parenthetical form as in [15]. This output sequence allows for the reconstruction of the dependency tree of the original sentence. The experimental results on three Italian treebanks demonstrate that such an approach is not only feasible but also capable of achieving results comparable to the state-of-the-art while requiring minimal training resources, such as training on a single GPU with modest memory.

In the rest of the paper, Section 2 described the proposed approach, Section 3 presents and discusses the experimental evaluations, while Section 4 derives the conclusions.

2. Dependency Parsing via Auto-regressive Language Model

The architecture of Transformers [10] has revolutionized Natural Language Processing (NLP), achieving increasingly higher results. In fact, the Architecture can be divided into three major families: Encoder-only models like BERT [9], RoBERTa [16], and DeBERTa [17] that are responsible for encoding input sequences and generating meaningful representations (embeddings) using the self-attention mechanism; Encoder-Decoder models, such as T5 [18] and BART [19], able to combine the strengths of both the encoder and decoder components and to maintain the integration of the two aforementioned blocks and typically used in tasks like machine translation, summarization, or question-answering, where complex input understanding and transduction are required; Decoder-only models like GPT [20], GPT3 [21], and LLaMA [13], that generate output sequences in an auto-regressive manner based on the input and previously generated output tokens. Recently, approaches based on Large Language Models (LLMs) have shown state-of-the-art performance in countless scenarios and tasks. LLMs excel at understanding language and following instructions, with ChatGPT[1] being a prime example.

However, training and fine-tuning such models require heavy computational resources, i.e. countless GPUs. Recently, a method for efficient training has been introduced, called Low-Rank Adaptation (LoRA [22]). LoRA involves freezing the weights of the pre-trained model and introducing trainable rank decomposition matrices into each layer of the Transformer architecture. This approach significantly limits the number of trainable parameters for downstream tasks while avoiding additional inference latency. Additionally, [23] introduces Quantized-LoRA, an optimization that further reduces memory usage enough to finetune a 65B parameter model on a single 48GB GPU while preserving full 16-bit finetuning task performance. QLoRA backpropagates gradients through a frozen, 4-bit quantized pre-trained language model into LoRA.

One of the challenges in modeling tasks with LLMs (Large Language Models) is that these models take sequences as input and produce sequences as output. For instance, consider this (Italian) sentence[2]:

"*Tutti gli esseri umani hanno capacità non sfruttate, non utilizzate.*" (1)

The Dependency Graph of this sentence is represented in Figure (1). In this graph, each node represents a word, and the arcs define the syntactic relationships among them. Additionally, each arc is labeled to indicate the type of dependency. A special node labeled ROOT is included to mark the root of the sentence, typically the main verb.

By converting the sentence into an arboreal structure, a Dependency Tree (Figure (2)) can be obtained. This tree illustrates the hierarchical structure of the sentence, with the main verb (*hanno*) serving as the ROOT and all other words depending on it. Non-terminal nodes in the tree represent the labels of the dependencies from the Dependency graph in Figure (1), while terminal nodes represent the words from the original sentence. For example, the NSUBJ arc indicates that the word *esseri* is the subject of the sentence, and the OBJ arc shows that the word *capacità* is the object of the verb. Furthermore, the ADVMOD label indicates that the word *sfruttate* is modified by the word *non*, negating its meaning. Both the Dependency Graph and Tree representations are equivalent, and it has been demonstrated in [15] that the Dependency Tree can be transformed into a linguistic representation, e.g. for computational purposes. The linguistic representation of the sentence corresponds to:

[ROOT [NSUBJ [DET:PREDET [Tutti]] [DET [gli]] [esseri] [AMOD [umani]]] [hanno] [OBJ [capacità] [ACL [ADVMOD [non]] [sfruttate] [CONJ [PUNCT [,]] [ADVMOD [non]] [utilizzate]]]] [PUNCT [.]]] (2)

Finally, it is worth noting that the process is reversible, meaning the DP tree can be constructed from the linguistic representation and vice versa. This ability facilitates

[1] https://openai.com/blog/chatgpt

[2] In English: "*All human beings have untapped, unused capacities.*".

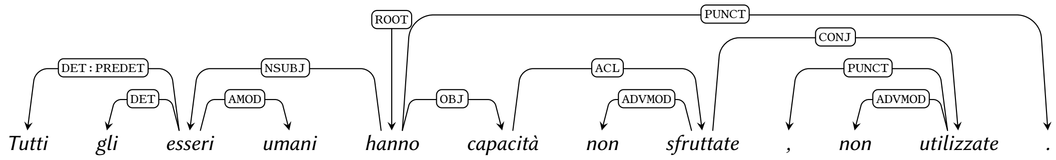


Figure 1: Example of a dependency graph associated to the sentence "*Tutti gli esseri umani hanno capacità non sfruttate, non utilizzate.*"

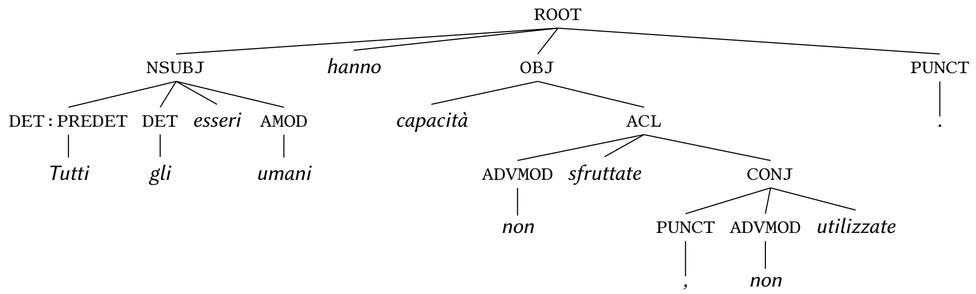


Figure 2: The syntactic parse tree associated with the dependency graph from Figure (1).

various computational tasks involving language modeling and analysis and, more importantly, allows the usage of such LLMs for predicting the Dependency Tree of sentences by training on the linguistic representation

In fact, several studies, such as [21], have highlighted the impressive few-shot learning capabilities of Language Models. These models can generalize information from only a limited number of input examples provided through prompting, producing coherent and accurate output. In this paper, we explore the application of the LLaMA $7B$ and $13B$ foundational models to Italian sentences with the goal of extracting DP Trees in parenthetical form. LLaMA is one of the Large Language Models that operates by taking a sequence of words as input and predicting the next word to generate text recursively. The model is built on the popular Transformers architecture [10], with several key differences. Firstly, to enhance training stability, the RMSNorm function [24] is applied before each layer for normalization. Secondly, the SwiGLU activation function [25] is utilized. Lastly, Rotary Embeddings (RoPE) [26] replace absolute positional embeddings. The combination of these modifications, along with the vast size of parameters and training data (trillions of tokens), makes LLaMA a highly promising model for various natural language processing tasks.

It's important to stress that LLaMA operates as a sequence-to-sequence model, following an auto-regressive approach, where text is fed as input, and text is generated as output. This allows the model to capture complex linguistic structures and dependencies in the input sentences and produce corresponding DP Trees in parenthetical form, showing the effectiveness of the approach in parsing Italian sentences. The input/output pairs used for LLaMA consist of sentences (as in Eq. (1)) and the linguistic representation of the DP Trees (as in Eq. (2)). During training and inference, the model is prompted with a simple instruction ("*Parse this sentence.*") to guide it in generating the desired output.

3. Experimental Evaluation

The training of the model utilized PyTorch and the Huggingface library, along with the Peft packages, to implement the Q-LoRA technique. The LLaMA models underwent 3 epochs of training with a learning rate of $3 \cdot 10^{-4}$ and a batch size of 32. To optimize the model's performance, a linear scheduler with warmup was utilized, using a warmup ratio of 0.1. The training process employed Q-LoRA 4-bit to refine the transformer's W_q and W_v modules, as in [23]. The LoRA matrices had a matrix rank R of 8 and a parameter α of 16. The training was performed on a single Tesla T4 GPU with 16GB of memory. This is particularly interesting as we have implied the two smallest available models, i.e. with 7 and 13-billion-parameters, to demonstrate that it can be used even on standard architectures. It doesn't rule out the possibility of evaluating larger models like LLaMA $65B$, but currently, they require such computational power that would limit their applicability in real-world scenarios, due to their extensive training duration and memory requirements.

We used the Universal Dependency Parsing dataset

Table 1
UAS using the Gold Standard tokenization provided.

Model	IT-ISDT	IT-ParTUT	IT-PoSTWITA
UDPipe	93.49%	92.64%	86.03%
UDPipe+	94.97%	95.36%	87.25%
7B_ita_b1	90.52%	93.00%	83.27%
7B_ita_b4	92.04%	93.18%	83.96%
13B_ita_b1	91.51%	93.76%	84.34%
13B_ita_b4	92.38%	94.01%	85.53%
7B_multi_b1	93.06%	94.22%	84.89%
7B_multi_b4	93.30%	94.55%	85.45%

and, to align with [6], we utilized version 2.3 of the dataset and focused on the same subsets of examples in the Italian language, i.e., three Treebanks: IT-ISDT obtained by conversion from ISDT (Italian Stanford Dependency Treebank), IT-ParTUT a conversion of a multilingual parallel treebank and consisting of a variety of text genres, and IT-PoSTWITA a collection of Italian tweets. The neural architecture was trained on the union of these three datasets, comprising $20,270$ training examples, $1,391$ development examples, and $1,309$ test sentences. This initial set of experiments, referred to as ita, involved training and evaluating the neural architecture using examples from the same language.

Objectives. In this experimentation, our primary objective is to address three crucial experimental questions related to natural language processing using the LLaMA model. First and foremost, we seek to determine if this process effectively works and if LLaMA is capable of achieving state-of-the-art performance. Secondly, we aim to explore the potential advantages of employing larger architectures (from $7B$ to $13B$ parameters): traditional large models have been criticized for their considerable computational and environmental costs. By investigating the use of bigger architectures in the LLaMA model, we strive to determine if advancements in performance can be achieved without compromising sustainability. Furthermore, we want to investigate the significance of multilingual data in enhancing the LLaMA model's performance. We draw inspiration from previous works such as UDPipe [6], which have demonstrated the positive impact of multilingual training on various natural language processing tasks. As the LLaMA model supports multiple languages, we aim to analyze whether incorporating multilingual data leads to improved overall performance on the Dependency Parsing task.

In a second set of experiments (referred to as $multi$), we trained the system by incorporating data from English, French, and Spanish datasets. Specifically, we added training examples from the English-EWT, English-GUM, English-LinES, English-ParTUT, French-GSD, French-ParTUT, French-Sequoia, French-Spoken, French-Old, Spanish-AnCora, and Spanish-GSD datasets to the training material. In this case, while the test data remained unchanged, being in Italian, the training dataset consisted of 101,284 examples. The development dataset was also kept in Italian for comparison purposes. To evaluate the performance of the LLaMA model, we have selected two key metrics: *UAS* (Unlabeled Attachment Score), and *LAS* (Labeled Attachment Score). *UAS* assesses the accuracy of the model's dependency tree structure by verifying if the correct head and dependency arcs are generated. On the other hand, *LAS* provides a more comprehensive evaluation by measuring the accuracy of the dependency labels assigned to each arc in the dependency tree.

Table 2
LAS using the Gold Standard tokenization provided.

Model	IT-ISDT	IT-ParTUT	IT-PoSTWITA
UDPipe	91.50%	90.50%	81.80%
UDPipe+	93.40%	93.40%	83.10%
7B_ita_b1	87.40%	89.50%	77.80%
7B_ita_b4	89.00%	90.00%	78.60%
13B_ita_b1	88.87%	90.61%	79.09%
13B_ita_b4	89.81%	90.87%	80.46%
7B_multi_b1	90.42%	91.19%	79.63%
7B_multi_b4	90.80%	91.61%	80.37%

Results Discussion. The experimental results are reported in Table 1, and 2 for the *UAS* and *LAS* metrics, respectively. Here we compare our approach with UDPipe [6] and the subsequent extension UDPipe+ [8], as these are the state-of-the-art systems for the Italian Treebanks. Notice that our LLaMA-based models fail in 0.5-1% of the times to correctly rewrite the whole sentence, i.e. they sometimes skip a word and do not produce any label, differently from UDPipe which covers 100% of the words in a sentence. Please note that, for the purpose of comparison, we have applied gold-standard tokenization in these initial experiments, as done in [8].

Our models are divided into two categories: *ita*, which is trained exclusively on Italian data, and *multi*, trained using material from other languages. From Tables 1 and 2, it is evident that the $7B$ models fall short of achieving state-of-the-art performance; however, they only slightly lag behind UDPipe. Advancing to the $13B$ models shows a modest performance increase, but considering their larger size, their practicality may be limited. Furthermore, we observe a performance boost when incorporating multilingual data during fine-tuning, as the LLaMA models support multiple languages. By enriching the Italian training set with data from other languages, we effectively leverage valuable relations and structures from diverse linguistic sources.

Moreover, since the dependency parsing process of a sentence is a global property of the entire sentence, we have also investigated more complex decoding processes, such as adopting deterministic beam search [27] during the decoding process. In a nutshell, beam search involves exploring up to b possible sequences during decoding until the completion of individual generations. This way, we believe that the generated sequence is not penalized by locally optimal choices for the decoding process but rather optimized at the sentence level. This enables us to enhance decoding strategies by adopting a larger beam search size (*b4*[3]) instead of relying solely on greedy search (*b1*). As a result, the adoption of beam search systematically improves performances for both the $7B$ and $13B$ parameter models. The results indicate that the model excels in generating the syntactic structure of sentences, with comparable performance to UDPipe. Remarkably, the neural model remains task-agnostic. Additionally, the study suggests that incorporating data from other languages, if possible, is more beneficial than merely scaling up to larger architectures.

Notice that these evaluations used the Gold Standard tokenization of the sentences available from the Treebanks, both during training and inference. For this reason, we trained and evaluated the same models using the "raw" sentences, without any tokenization and requiring the model to produce the resulting DP tree. For instance, the sentence from Eq (1) will be given to the model without any additional spaces for the punctuation, but the resulting output should remain the same, i.e. the one from Eq (2). The models, thus, are required to learn the tokenization during the training phase. The results are in Tables 3 and 4, where the performance for the UDPipe models is not available as they rely completely on the Gold Standard tokenization. For both *UAS* and *LAS* metrics, there is a loss in performance: every model drops around 1% of accuracy with respect to the GS tokenization, with $7B$*_ita_bs1* losing almost 2% on the *IT-PoSTWITA* treebank. Intuitively, this drop is due to the fact that the model is required to learn the tokenization and the majority of the errors are because of missing punctuation and so on. This result shows the robustness of the LLaMA model even on "un-tokenized" data.

Table 3
UAS computed on the end-to-end process, where the tokenization is performed by the model.

Model	IT-ISDT	IT-ParTUT	IT-PoSTWITA
7B_ita_b1	90.13%	91.19%	81.36%
7B_ita_b4	91.27%	91.96%	82.18%
13B_ita_b1	92.31%	93.80%	82.90%
13B_ita_b4	92.57%	94.34%	83.59%
7B_multi_b1	92.65%	94.12%	83.31%
7B_multi_b4	93.22%	94.31%	83.99%

[3]We experimented with different values for the beam search parameter '*b*', but none of them yielded significant performance improvements except for when '*b*' was set to 4.

Table 4
LAS computed on the end-to-end process, where the tokenization is performed by the model.

Model	IT-ISDT	IT-ParTUT	IT-PoSTWITA
7B_ita_b1	87.27%	87.77%	76.26%
7B_ita_b4	88.51%	88.45%	77.13%
13B_ita_b1	89.81%	90.52%	77.90%
13B_ita_b4	90.10%	91.04%	78.79%
7B_multi_b1	89.93%	91.32%	78.04%
7B_multi_b4	90.57%	91.62%	78.78%

4. Conclusions

In this paper, we investigate the application of recent popular LLMs, specifically the LLaMA foundational models, to address the Dependency Parsing problem. Our exploration aimed to answer three key questions: **Can we utilize LLaMA in a sequence-to-sequence scenario to rewrite Dependency Parsing Trees from input sentences?** The answer is affirmative. Although LLaMA did not achieve a new state-of-the-art performance, our results demonstrate that the adopted model and approach are competitive with UDPipe, the current leading model. **Can we scale up LLaMA by increasing the number of parameters while ensuring sustainability?** Our evaluation reveals that almost doubling the model's parameters leads to little or no significant gain in performance. However, we found a notable performance increment by leveraging the beam search technique instead of the greedy search. This aspect could be explored further in the future. **Does the inclusion of multilingual data improve the LLaMA model?** Our findings in this paper support the initial hypothesis that using multilingual data enhances the LLaMA model's performance. Every model trained with multilingual data consistently outperforms those trained solely on Italian data.

As a future work, it would be interesting to exploit data from all available languages and evaluate the model's capabilities across a broader linguistic spectrum. This approach could lead to the development of a Universal Dependency Parsing Model as a unified architecture, which holds significant promise in advancing the field of dependency parsing.

Acknowledgments

We would like to thank the "Istituto di Analisi dei Sistemi ed Informatica - Antonio Ruberti" (IASI) for supporting the experimentations through access to dedicated computing resources. Claudiu Daniel Hromei is a Ph.D. student enrolled in the National Ph.D. in Artificial Intelligence, XXXVII cycle, course on *Health and life sciences*, organized by the Università Campus Bio-Medico di Roma. We acknowledge financial support from the PNRR MUR project PE0000013-FAIR.

References

[1] S. Kübler, R. McDonald, J. Nivre, Dependency Parsing, Springer Cham, 2009. URL: https://doi.org/10.1007/978-3-031-02131-2. doi:10.1007/978-3-031-02131-2.

[2] L. Tesnière, Éléments de syntaxe structurale, Klincksieck, Paris, 1959.

[3] J. Nivre, Algorithms for deterministic incremental dependency parsing, Computational Linguistics 34 (2008) 513–553. URL: https://aclanthology.org/J08-4003. doi:10.1162/coli.07-056-R1-07-027.

[4] D. Chen, C. Manning, A fast and accurate dependency parser using neural networks, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 740–750. URL: https://aclanthology.org/D14-1082. doi:10.3115/v1/D14-1082.

[5] T. Dozat, C. D. Manning, Deep biaffine attention for neural dependency parsing, CoRR abs/1611.01734 (2016). URL: http://arxiv.org/abs/1611.01734. arXiv:1611.01734.

[6] M. Straka, J. Straková, Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe, in: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 88–99. URL: https://aclanthology.org/K17-3009. doi:10.18653/v1/K17-3009.

[7] M.-C. de Marneffe, T. Dozat, N. Silveira, K. Haverinen, F. Ginter, J. Nivre, C. D. Manning, Universal Stanford dependencies: A cross-linguistic typology, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), European Language Resources Association (ELRA), Reykjavik, Iceland, 2014.

[8] M. Straka, J. Straková, J. Hajic, Evaluating contextualized embeddings on 54 languages in POS tagging, lemmatization and dependency parsing, CoRR abs/1908.07448 (2019). URL: http://arxiv.org/abs/1908.07448. arXiv:1908.07448.

[9] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the NAACL 2019, 2019, pp. 4171–4186. URL: https://doi.org/10.18653/v1/n19-1423. doi:10.18653/v1/n19-1423.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, CoRR abs/1706.03762 (2017). URL: http://arxiv.org/abs/1706.03762. arXiv:1706.03762.

[11] Z. Li, J. Cai, S. He, H. Zhao, Seq2seq dependency parsing, in: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 3203–3214. URL: https://aclanthology.org/C18-1271.

[12] D. Kondratyuk, 75 languages, 1 model: Parsing universal dependencies universally, CoRR abs/1904.02099 (2019). URL: http://arxiv.org/abs/1904.02099. arXiv:1904.02099.

[13] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, 2023. arXiv:2302.13971.

[14] C. D. Hromei, D. Croce, V. Basile, R. Basili, ExtremITA at EVALITA 2023: Multi-task sustainable scaling to large language models at its extreme, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[15] D. Croce, A. Moschitti, R. Basili, Structured lexical similarity via convolution kernels on dependency trees, in: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Edinburgh, Scotland, UK., 2011, pp. 1034–1046. URL: https://aclanthology.org/D11-1096.

[16] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[17] P. He, X. Liu, J. Gao, W. Chen, Deberta: decoding-enhanced bert with disentangled attention, in: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, 2021. URL: https://openreview.net/forum?id=XPZIaotutsD.

[18] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, J. Mach. Learn. Res. 21 (2020) 140:1–140:67. URL: http://jmlr.org/papers/v21/20-074.html.

[19] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, CoRR abs/1910.13461 (2019). URL: http://arxiv.org/abs/1910.13461. arXiv:1910.13461.

[20] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., Improving language understanding by generative pre-training, 2018.

[21] T. B. Brown, B. Mann, N. Ryder, M. Subbiah,

J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, CoRR abs/2005.14165 (2020). URL: https://arxiv.org/abs/2005.14165. arXiv:2005.14165.

[22] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, W. Chen, Lora: Low-rank adaptation of large language models, CoRR abs/2106.09685 (2021). URL: https://arxiv.org/abs/2106.09685. arXiv:2106.09685.

[23] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, Qlora: Efficient finetuning of quantized llms, 2023. arXiv:2305.14314.

[24] B. Zhang, R. Sennrich, Root mean square layer normalization, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 32, Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/1e8a19426224ca89e83cef47f1e7f53b-Paper.pdf.

[25] N. Shazeer, GLU variants improve transformer, CoRR abs/2002.05202 (2020). URL: https://arxiv.org/abs/2002.05202. arXiv:2002.05202.

[26] J. Su, Y. Lu, S. Pan, B. Wen, Y. Liu, Roformer: Enhanced transformer with rotary position embedding, CoRR abs/2104.09864 (2021). URL: https://arxiv.org/abs/2104.09864. arXiv:2104.09864.

[27] Y. Shao, S. Gouws, D. Britz, A. Goldie, B. Strope, R. Kurzweil, Generating high-quality and informative conversation responses with sequence-to-sequence models, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 2210–2219. URL: https://www.aclweb.org/anthology/D17-1235. doi:10.18653/v1/D17-1235.

DisaggregHate It Corpus: A Disaggregated Italian Dataset of Hate Speech

Marco Madeddu[1], Simona Frenda[1,2], Mirko Lai[1,2], Viviana Patti[1] and Valerio Basile[1]

[1]*Università di Torino, Italy*
[2]*Aequa-tech srl, Turin, Italy*

Abstract
Recent studies in Machine Learning advocate for the exploitation of disagreement between annotators to train models in line with the different opinions of humans about a specific phenomenon. This means that datasets where the annotations are aggregated by majority voting are not enough. In this paper, we present an Italian disaggregated dataset concerning hate speech and encoding some information about the annotators: the DisaggregHate It Corpus. The corpus contains Italian tweets that focus on the topic of racism and has been annotated by native Italian university students. We explain how the dataset was gathered by following the recommendation of the *perspectivist* approach [1], encouraging the annotators to give some socio-demographic information about them. To exploit the disagreement in the learning process, we proposed two types of *soft labels*: softmax and standard normalization. We investigated the benefit of using disagreement by creating a baseline binary model and two regression models that were respectively trained on the 'hard' (aggregated label by majority voting) and the two types of 'soft' labels. We tested the models in an in-domain and out-of-domain setting, evaluating their performance using the cross-entropy as a metric, and showing that the models trained on the soft labels performed better.

Keywords
hate speech, perspectivism, disagreement

1. Introduction

The rise of the Internet and social media platforms has given many users the opportunity to express their opinion online. Unfortunately, this leads to the diffusion of a new online phenomenon: the hate speech. To prevent the viral spread of this kind of expressions on social media, hate speech detection became a popular task in Natural Language Processing (NLP). A lot of tools have been created to detect and counter hate speech[2, 3, 4].

Recently, there have been studies that suggest trying to shift away from the golden standard approach in Machine Learning, especially in tasks partly subjective and influenced by the social and cultural context, like hate speech [5, 1]. These works advocate that different opinions given in the annotation process are not a noise factor but can be used to make better systems [6]. This shift inspires scholars to try different techniques to train models using datasets where the target label is not simply determined by majority voting on the annotations. In this line, two theoretical paradigms have been established, both looking for the inclusion of different perspectives: the *learning from disagreement* and *perspectivism.* The former could be considered like a 'soft perspectivist approach' because it takes into account the presence of disagreement in the annotated data, while the latter, overcomes the idea of "ground truth" in the construction of datasets and on the creation and evaluation of NLP models, focusing more on who the annotators are.

Our work could be considered a tentative to approach hate speech detection, exploiting the possible disagreement among the annotators. Usually, models are trained on data associated to a 'hard' label. In the case of binary classification, each item is assigned a label whose value is either 0 or 1. The hard label value is commonly obtained through majority voting, therefore this implies that controversial instances have the same label as the ones that saw all annotators in agreement. This may be thought of like a loss of valuable information that can be used in the training phase of the models [7]. On the other hand, 'soft' labels approaches try to avoid this waste of data by assigning a real number to the label. Different functions can be used in the process of determining the value of the soft label, such as standard normalization or a softmax function [8].

In this context, we created the DisaggregHate It Corpus, a new disaggregated dataset about hate speech in the Italian language that incorporates some socio-demographic information about annotators[1]. A corpus like this could be beneficial in exploring how different segments of population are sensitive to certain social issues like hate speech, and how this information can be used to create better systems.

After explaining the different characteristics of the

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
marco.madeddu@edu.unito.it (M. Madeddu); simona.frenda@unito.it (S. Frenda); mirko.lai@unito.it (M. Lai); viviana.patti@unito.it (V. Patti); valerio.basile@unito.it (V. Basile)

[1]The corpus is available here: https://github.com/madeddumarco/DisaggregHateIt

dataset in section 3 we will validate the corpus by using it as the training set of different models in section 4. The performed experiments show that training models on a soft label rather than a hard label leads to better results. As suggested by [7], we used the cross entropy metric for evaluating the models.

2. Related Work

The past years have seen an increase in using different paradigms that try to model the different opinions of human annotators, especially in cases of recognition of subjective phenomena, like hate speech. Adopting a *soft perspectivist approach*, recent challenges like Le.Wi.Di (*Learning with disagreement*) shared task were proposed at SemEval 2021 and 2023 [8, 9]. In particular, this shared task asked participants to model various phenomena, such as humor and hate speech detection, exploiting the *soft labels*. These, contrary to the *hard labels* (simple labels), are obtained computing a sort of distribution of the labels chosen by annotators. Modelling this distribution, the systems are able to approximate the probability distribution of the opinions about the specific phenomenon. A *strong perspectivist approach*, instead, looks at whom the annotators are and how to model their opinion [1].

In the experimental part of this work, we focused especially on the use of soft labels to model the different labels without considering the information available about annotators, on the example of Uma et al. [7]. In this study, the authors experimented the application of the soft labels to detect various phenomena, employing a standard and a softmax normalization of the labels. They proved that in both hard and soft evaluation settings, respectively using accuracy and cross-entropy metrics, the use of soft labels in the modelling leads to better results.

Following their example, we evaluated the new disaggregated dataset on hate speech, DisaggregHate It Corpus, composed of Italian tweets, and enriched with some socio-demographic information about the annotators. Our idea was to create a dataset according perspectivist recommendations provided by Cabitza et al. [1] to ensure the transparency of the created perspective dataset. Among these recommendations, the authors mention the involvement of enough and heterogeneous annotators, and the collection of information about them. Moreover, with our work we meet also other their recommendations such as the report about the annotation process, the use of hard labels (computed by majority voting) and the soft labels (to represent the distribution of the decisions provided by annotators), and finally, we validated our models in an out-of-domain setting.

The works on hate speech that comply to some of these recommendations and release disaggregated datasets, are few, and to our knowledge, are only in other languages. One of the most famous is the Measuring of Hate Speech corpus [10][2] available only in English, that encodes various dimensions of hate speech (with disaggregated labels) and also different information about annotators. Follow: the HS-Brexit disaggregated dataset created by Akhtar et al. [11], ToxCR dataset [12] and JSRPData [13], on hate speech and toxic language. All of these datasets are in English and contain little information about the annotators. About Italian language, to our knowledge, only IMSyPP-IT dataset [14] have been released with disaggregated labels but without information about annotators.

In this context, a dataset like DisItaggragated released with disaggregated labels about hate speech, and that encodes also some information about annotators, contributes to enrich the resources for Italian community and to encourage the modeling of perspectives and different opinions in a very subjective phenomenon like hate speech.

3. Dataset

In this section, we first introduce our dataset by illustrating the context of the annotation process and secondly the general statistics about the corpus. Further, we will analyze the distribution of the positive and negative label for both the hard and soft label.

3.1. Corpus Creation

The DisaggregHate It Corpus used for this work is composed of 1100 tweets extracted from Contro L'Odio [15], an Italian corpus that focuses on racist hate and in particular on discrimination towards immigrants. The annotation process carried out as part of a master degree course, so the participants are all university students aged between 21 and 30, and native of the Italian language. A specific educational web platform has been realized on the example of the one developed by [16], for allowing the annotation process and the collection of some basic information about the annotators. For each tweet, annotators have been asked to decide the presence hate speech (yes or no), irony (yes or no) and the stance of the author of the message towards immigration issues (positive, neutral, or negative)[3]. For our experiment, we only considered the hate speech annotations, so from this point forward when we will talk about the target label we are referring to the **hate speech** one.

[2]This dataset is released on HuggingFace: https://huggingface.co/datasets/ucberkeley-dlab/measuring-hate-speech

[3]The used guidelines are the ones adopted to annotate data in the HaSpeeDe context [17] (for hate speech), in the context of IronIta [18] (for irony), and in the context of SardiStance [19] (for stance). Especially the last guidelines have been adapted to the context of immigration.

Profile	Annotators	Tweets	Krippendorff's α
City <50k	11	300	0.32
City >50k	8	300	0.40
TSCI	4	100	0.19
Humanistic	1	100	-
Men	12	403	0.28
Women	11	400	0.24
Low SM	30	700	0.32
High SM	36	700	0.34

Table 1
Information about annotators

Annotators provided basic information about their gender, how many social media platforms they use, if they live in a city with more than 50 thousands residents and their school background (TSCI or Humanistic). The participants could choose to give one or more information about them.

In order to collect as many annotations as possible for each tweet, students have been grouped in teams of minimum 5 components, and each annotator was asked to annotate at least 100 tweets per group. However, some sets of data have been annotated by more than 1 group of students and others only by few annotators. Therefore, every tweet has a number of annotation in a range from 1 to 13.

In this context, we computed the agreement among the annotators, taking into account the information that they provided, using Krippendorff's Alpha [20]. This metric, indeed, allows evaluating agreement when the matrix of annotations is sparse (i.e., the number of annotators is not constant for each tweet and, thus, some values could miss). We did not report the value of Krippendorff's Alpha for the 'humanistic' profile as the function requires at least two annotators (see Table 1). The value of Krippendorff's Alpha for the whole dataset is 0.34.
In table 1 we can observe that the agreement intra-group is quite low as Krippendorff's alpha values that are equal to 0 indicate absence of reliability meanwhile values that are equal to 1 show perfect agreement [20]. It means that annotators have different perception of hate speech even if they share the same socio-demographic trait. The only profile that shows a fair agreement is the 'City >50K' (living in a city with more than 50 thousands residences). However, the scores are low, motivating an approach based on *learning with disagreement.*

3.2. Hard and Soft Label Distribution

We assigned each tweet three different labels: a hard label and two soft labels. The hard label matches the majority vote of annotations, while the two soft labels, respectively, employ a standard and a softmax normalisation. Using the generalization of Uma et al. [7], given an instance i and C classes , we can determine a vector $[d_i^1, d_i^2, .., d_i^C,]$ where d_i^j is the number of votes given by the annotators for class j to the instance i. Softmax normalization determines the value of the soft label l_i^j for each example i and class j with the following formula:

$$l_i^j = \frac{exp(d_i^j)}{\sum_a exp(d_i^a)}$$

Meanwhile standard normalization is obtained by applying:

$$l_i^j = \frac{d_i^j}{\sum_a d_i^a}$$

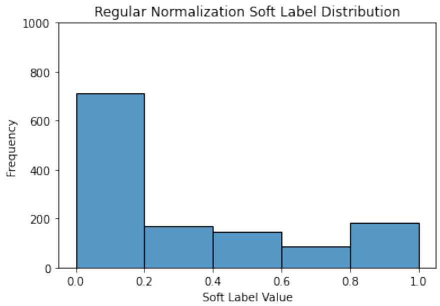
This case study addresses the hate speech annotation as a binary problem: $j \in [HS, \neg HS]$. We computed standard and a softmax normalisations l_i^{HS} for the sole positive class.

Addressing the data labelling with a soft label approach prevents discarding annotations and allows for the creation of a more informative annotated corpus. As Uma et al. [7] pointed out the softmax normalization, unlike the standard one, assigns to an instance a non zero value even where a class received zero votes. Therefore, the softmax normalisation could be seen as a way to smooth the label distribution, but it could also cause some side effects. Indeed, whenever $d_i^c \simeq \sum_a d_i^a$, i.e. there is complete agreement among annotators but the there is only a very small number of annotators, $l_i^j \forall j \neq c$ will be however sensibly larger than 0. Therefore, the use of standard normalization would be preferable in the presence of many classes and few annotators.

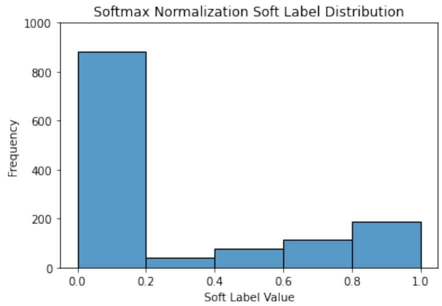
In table 2 we can observe how many ties, positive and negative instances are present in our dataset when we apply a majority voting to obtain a hard label. We can also observe how many tweets had an even number of annotators resulting in possible ties. We can see that there is different percentages of positive instances in some demographic division criteria like gender (Men and Women). Other category distinctions, like the one based on social media usage, show little difference between the two groups (Low SM and Hight SM). The number of ties is very different between the various categories

Category	Examples	Ties %	Pos. %	Neg. %	Tie Chance %
Whole Dataset	1100	3.7%	15.3%	80.9%	66%
City <50k	300	0%	9.6%	90.3%	0%
City >50k	300	6.6%	11.6%	81.6%	33.3%
TSCI	100	15%	9%	76%	100%
Humanistic	100	0%	19%	81%	0%
Men	403	18.11%	10.17%	71.71%	74%
Women	400	9.25%	23.25%	67.5%	31%
Low SM	700	6%	12.14%	81.85%	71%
High SM	700	5.85%	12%	82.14%	42%

Table 2
Dataset Composition



(a) Distribution for standard normalization



(b) Distribution for softmax normalization

Figure 1: Histograms for distributions of soft labels

ranging from 0% to more than 18% of the total instances. A very high number of ties indicates the presence of controversial instances that could be very important in the training phase of a model. The Krippedndorff's alpha values paired with the number of ties show that the DisaggregHate It Corpus contains a not neglectable level of disagreement between the annotators. Overall, we can see that the DisaggregHate It Corpus is unbalanced towards the negative class; therefore, in Section 4, we proposed to train the models using weighted labels.

In Figure 1, we observe the label distribution using standard or softmax normalization. We can observe that there are more negative instances than positive ones as the most represented bin is the one with $l_i^{HS} < 0.2$. We can observe a mostly similar tendency comparing the Figures 1a and 1b even if the standard normalization has more examples in the bins for the middle values. Overall, we can observe that annotators usually tend to be in agreement when there is a clear signal of hate speech, indeed the bin with values $l_i^{HS} > 0.8$ has more instances compared to other ones.

4. Experiments

The DisaggregHate It Corpus has been used to carry out two main settings of experiments: in-domain (test set of DisaggregHate It Corpus corpus) and out-of-domain (two test sets of two new shared tasks at EVALITA 2023). The tested models are: a standard model trained on aggregated labels (called here *Binary*), and two new models trained on soft labels (called here *Regression*) computed in two different manners. The former trained to detect the presence or absence of hate speech in the tweets, the latter trained to give a probability about the presence of hate speech in the tweets in line with the distribution of labels provided by annotators.

4.1. Models Description

We built all of our models by fine-tuning an already existing BERT (Bidirectional Encoder Representations from Transformers) based model for Italian. BERT is the state-of-the-art family of Large Language Models based on the transformer architecture [21]. There are a lot of BERT models that have been trained on large amount of data, thus they can be easily fine-tuned to perform in other tasks by fine-tuning them with smaller data sets. The model we chose to use is the uncased Italian BERT model with the Huggingface identifier: dbmdz/bert-base-italian-uncased created by the MDZ Digital Library team [22]. We accessed it through the Huggingface platform and the Python library Transformers which offers easy to use functions to design a simple architecture for fine-tuning the pre-trained models for specific tasks like the one of classification (i.e., *BertForSequenceClassification*). Considering the characteristics of our dataset and the kind of experiments that we wanted to perform, we designed some specific techniques.

The first regards the output of the network. We created three different models: one trained for binary classification with the hard label of the dataset, and two regression models respectively trained on the soft label computed with standard normalization and the softmax normalization. Taking into account the need of using a *soft metric* (cross-entropy) to compare the performance of our models, as suggest by [7, 8, 9], for the binary classifier we obtained soft label predictions by applying the softmax function to the logit outputs. The probabilities from the regression models are simply obtainable thanks to the Transformers library by setting the number of labels parameter to 1 of a classification model. As the outputs of the regression models are not bounded, we applied the clip function to limit their value between 0 and 1.

The second is about the different balance of the classes in our dataset. The DisaggregHate It Corpus contained, indeed, more negative label examples than positive ones (see Table 2). To deal with this, we experimented by assigning different weights to the positive and negative label. We obtained these different weights through the *compute_class_weight* function present in the scikit learn Python library. These weights were used in the calculation of the loss function for each model. The binary model was trained with a weighed cross-entropy loss function and given that the training set contained hard labels, we easily assigned different weights to each label. The regression models were trained with a weighed Mean Squared Error loss function and as the label values were real number, we assigned the postive binary label weight to examples with a soft label value ≥ 0.5 and we assigned the negative binary label weight to the rest. The models were trained for 5 epochs, each with a learning rate parameter equal to $2e^{-5}$.

4.2. In and Out-of-Domain Testing

The **in-domain test set** has been extracted from the DisaggregHate It Corpus, selecting 20% of the entire dataset, while the rest was used for the training and validation sets. As **out-of-domain test sets**, we used two datasets in the Italian language that have been released in the occasion of the 2023 edition of the EVALITA campaign. The first one is the corpus regarding the second task of the HaSpeeDe3 (Hate Speech Detection) shared task [23] annotated in regard to political and religious hate. The used test set from HaSpeeDe3 is composed of 5600 tweets, containing 2144 positive examples. The second dataset is the corpus from the HoDI (Homotransphobia Detection in Italian) shared task [24] containing 5000 tweets about homophobia. The test set of HoDI is composed of 5000, containing 2008 positive examples. So after training our models with the in-domain training sets we tested them on the in-domain tests, the entire HaSpeeDe 3 and HoDI training sets.

4.3. Results

In table 3 we report all the results in terms of cross-entropy (CE) for the in-domain and out-of-domain experiments. We decided to only report the CE scores with certain test sets to avoid an unfair comparison. Therefore, we excluded testing the regression model trained on the standard normalization soft labels with the softmax normalized test set, and vice versa. As the binary model soft label predictions are obtained by applying the softmax function, thus we decided it is adequate to calculate the CE with the softmax normalized test set. About the out-of-domain testing, we calculated the CE between the soft label predictions and the hard label versions of the test sets, as the disaggregated annotations are not available.

We can observe in table 3 that both regression models report better scores than the binary models in all tests both in-domain and out-of-domain. When we compare the CE score obtained with the binary model with the ones obtained with regression models, we can see a very significant difference in favor of the regression model in both scenarios. Observing in details the standard normalization and softmax normalization regression models, we notice that the softmax normalization seems works better in general, in both experimental settings. However, if in the in-domain setting, the scores report a difference of 5% in terms of Δ, in the out-of-domain setting, the results from both regression models are similar. These results are in line with the ones obtained in the study of Uma et al. [7].

Moreover, we can observe that both regression models score slightly worse when compared to the in-domain setting, and this could have been expected as the cross-domain task is difficult. Another factor of this drop in performance could be that the target label of the cross-domain datasets was binary and not a real number. This encourages the releasing of datasets with disaggregated labels.

Model Type	Train Set	In-Domain Test CE Std.	In-Domain Test CE Softmax	HaSpeeDe 3 CE	HoDI CE
Binary	Hard Label	-	1.084	0.814	0.851
Regression	Standard Norm. Label	**0.616**	-	0.668	**0.674**
Regression	Softmax Norm. Label	-	**0.588**	**0.662**	0.678

Table 3
Cross Entropy Test Results

5. Conclusion

In this work, we presented the DisaggregHate It Corpus, a new disaggregated dataset in the Italian language of hate speech. To our knowledge, it is the first dataset released with disaggregated labels and some socio-demographic information about the annotators. Computing the agreement among annotators with the same profile, we noticed that the Krippendolf'α is very low. Moreover, this information, paired with the number ties obtained by majority voting, showed us how disagreement is a real factor in corpora. That motivates the need to approach the hate speech detection task with models that encode the different opinions of humans annotators. To this purpose, we experimented with the use of a soft label, exploring two different computation of soft labels: standard and softmax normalization.

To continue our study on the usage of disagreement as a factor in learning we carried out different experiments testing the performance of our models in two specific settings: in-domain and out-of-domain. We created a binary model based on the hard labels and two regression models trained on the soft labels (computed with the two different normalization, regular and softmax). Inspired by previous works [7, 8, 9], we evaluated the models, employing the cross-entropy between the soft labels of annotations and the model predictions. Observing the results, we noticed that the regression models perform better both when considering in-domain and out-of-domain test sets. This implies that a soft label is helpful to integrate annotators disagreement inside our models in order to be more in line with the distribution of the opinions of human annotators.

Taking into account these results, we plan to use the same DisaggregHate It Corpus, to explore a stronger perspectivist approach modelling the perspectives of different groups of annotators on the basis of their socio-demographic traits or other commonalities.

Ethics Statement

The annotation process involved students of the Politecnico di Torino, who performed this task in an educational environment. The guidelines and the information about the annotation task have been shared via the educational platform exploited for implementing the annotation process, and discussed during the lessons. The efforts required to the students has been limited to their time and oriented to complete a project work being part of the exam of *Internet e social media: tecnologie e derive della comunicazione in rete*. This annotation task has been used, first of all, to give the students the opportunity to discuss the disagreement, encouraging a deep reflection on the importance of developing high quality annotated resources, to train and evaluate machine learning models.

References

[1] F. Cabitza, A. Campagner, V. Basile, Toward a perspectivist turn in ground truthing for predictive computing, Proceedings of the AAAI Conference on Artificial Intelligence 37 (2023) 6860–6868. URL: https://ojs.aaai.org/index.php/AAAI/article/view/25840. doi:10.1609/aaai.v37i6.25840.

[2] A. Schmidt, M. Wiegand, A survey on hate speech detection using natural language processing, in: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 1–10. URL: https://aclanthology.org/W17-1101.

[3] P. Fortuna, S. Nunes, A survey on automatic detection of hate speech in text, ACM Computing Surveys 51 (2018) 85:1–85:30. URL: https://doi.org/10.1145/3232676.

[4] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, V. Patti, Resources and benchmark corpora for hate speech detection: A systematic review, Language Resources and Evaluation 55 (2021) 477–523. URL: https://rdcu.be/cCdaB.

[5] B. Plank, The "problem" of human label variation: On ground truth in data, modeling and evaluation, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 10671–10682. URL: https://aclanthology.org/2022.emnlp-main.731.

[6] V. Basile, M. Fell, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, A. Uma, We need to con-

sider disagreement in evaluation, in: Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future, Association for Computational Linguistics, Online, 2021, pp. 15–21. URL: https://aclanthology.org/2021.bppf-1.3. doi:10.18653/v1/2021.bppf-1.3.

[7] A. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, A case for soft loss functions, Proceedings of the AAAI Conference on Human Computation and Crowdsourcing 8 (2020) 173–177. URL: https://ojs.aaai.org/index.php/HCOMP/article/view/7478. doi:10.1609/hcomp.v8i1.7478.

[8] A. Uma, T. Fornaciari, A. Dumitrache, T. Miller, J. Chamberlain, B. Plank, E. Simpson, M. Poesio, SemEval-2021 task 12: Learning with disagreements, in: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Association for Computational Linguistics, Online, 2021, pp. 338–347. URL: https://aclanthology.org/2021.semeval-1.41. doi:10.18653/v1/2021.semeval-1.41.

[9] E. Leonardelli, A. Uma, G. Abercrombie, D. Almanea, V. Basile, T. Fornaciari, B. Plank, V. Rieser, M. Poesio, Semeval-2023 task 11: Learning with disagreements (lewidi), 2023. arXiv:2304.14803.

[10] C. J. Kennedy, G. Bacon, A. Sahn, C. von Vacano, Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application, ArXiv abs/2009.10277 (2020). URL: https://api.semanticscholar.org/CorpusID:221836648.

[11] S. Akhtar, V. Basile, V. Patti, Modeling annotator perspective and polarized opinions to improve hate speech detection, Proceedings of the AAAI Conference on Human Computation and Crowdsourcing 8 (2020) 151–154. URL: https://ojs.aaai.org/index.php/HCOMP/article/view/7473. doi:10.1609/hcomp.v8i1.7473.

[12] D. Kumar, P. G. Kelley, S. Consolvo, J. Mason, E. Bursztein, Z. Durumeric, K. Thomas, M. Bailey, Designing toxic content classification for a diversity of perspectives, in: Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021), 2021, pp. 299–318.

[13] N. Goyal, I. D. Kivlichan, R. Rosen, L. Vasserman, Is your toxicity my toxicity? exploring the impact of rater identity on toxicity annotation, Proceedings of the ACM on Human-Computer Interaction 6 (2022) 1–28.

[14] M. Cinelli, A. Pelicon, I. Mozetič, W. Quattrociocchi, P. Kralj Novak, F. Zollo, Italian YouTube Hate Speech corpus, 2021. URL: http://hdl.handle.net/11356/1450, slovenian language resource repository CLARIN.SI.

[15] A. T. Capozzi, M. Lai, V. Basile, F. Poletto, M. Sanguinetti, C. Bosco, V. Patti, G. Ruffo, C. Musto, M. Polignano, et al., Computational linguistics against hate: Hate speech detection and visualization on social media in the" contro l'odio" project, in: CEUR Workshop Proceedings, volume 2481, CEUR-WS, 2019, pp. 1–6.

[16] S. Frenda, A. T. Cignarella, M. A. Stranisci, M. Lai, C. Bosco, V. Patti, et al., Recognizing hate with nlp: The teaching experience of the# deactivhate lab in italian high schools, in: CEUR WORKSHOP PROCEEDINGS, volume 3033, CEUR-WS. org, 2021, pp. 1–7.

[17] C. Bosco, D. Felice, F. Poletto, M. Sanguinetti, T. Maurizio, et al., Overview of the evalita 2018 hate speech detection task, in: Ceur workshop proceedings, volume 2263, CEUR, 2018, pp. 1–9.

[18] A. T. Cignarella, S. Frenda, V. Basile, C. Bosco, V. Patti, P. Rosso, Overview of the EVALITA 2018 task on irony detection in Italian tweets (IronITA), in: Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2018) co-located with the Fifth CLiC-it, volume 2263, 2018, pp. 1–6.

[19] A. T. Cignarella, M. Lai, C. Bosco, V. Patti, R. Paolo, et al., Sardistance@ evalita2020: Overview of the task on stance detection in italian tweets, in: Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Ceur, 2020, pp. 1–10.

[20] K. Krippendorff, Computing Krippendorff's Alpha-Reliability, Technical Report, University of PennSylvania, 2011. URL: https://www.asc.upenn.edu/sites/default/files/2021-03/Computing%20Krippendorff%27s%20Alpha-Reliability.pdf.

[21] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[22] K. Clark, M.-T. Luong, Q. V. Le, C. D. Manning, Electra: Pre-training text encoders as discriminators rather than generators, in: International Conference on Learning Representations, 2020, pp. 1–14. URL: https://openreview.net/forum?id=r1xMH1BtvB.

[23] M. Lai, F. Celli, A. Ramponi, S. Tonelli, C. Bosco, V. Patti, Haspeede3 at evalita 2023: Overview of the political and religious hate speech detection task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023, pp. 1–8. URL: https://ceur-ws.org/Vol-3473/paper22.pdf.

[24] D. Nozza, A. T. Cignarella, G. Damo, T. Caselli,

V. Patti, HODI at EVALITA 2023: Overview of the Homotransphobia Detection in Italian Task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023, pp. 1–8. URL: https://ceur-ws.org/Vol-3473/paper26.pdf.

Automatic Detection of Parkinson's Disease with Connected Speech Acoustic Features: towards a Linguistically Interpretable Approach

Marta Maffia[1], Loredana Schettino[2] and Vincenzo Norman Vitale[2,3]

[1]*Dept. of Literary, Linguistics and Comparative Studies, University of Naples L'Orientale, Italy*
[2]*Interdepartmental Research Center Urban/Eco, University of Naples Federico II, Italy*
[3] *DIETI - University of Naples Federico II, Italy*

Abstract
Alterations in speech and voice are among the earliest symptoms of Parkinson's Disease (PD). Nevertheless, the rich information carried by patients' speech and voice is only partially used for diagnosis and clinical decision-making that is currently based on holistic ratings of speech intelligibility. An accurate diagnosis could be supported by the application of fully automated analytic methods and machine learning techniques on speech recordings. However, most of the proposed procedures were designed for highly functional but "artificial" vocal paradigms such as sustained phonation and consider all the considerable amount of features that can be extracted using automatic systems. In this work, we perform PD detection trials using features extracted from connected speech rather than isolated speech units. Moreover, we support the adopted machine learning-based methods with linguistic considerations so as to reduce the number of features to some meaningful ones. The main findings highlight that this procedure allows more accurate, economical and, most importantly, interpretable discrimination.

1. Introduction

[1] Parkinson's Disease (PD) is the most common movement disorder and the second most common neurodegenerative disorder worldwide after Alzheimer disease. It affects more than 2-3% of the population aged 65 and over [1, 2].

Caused by the deterioration or loss of dopaminergic neurons in the *substantia nigra* of basal ganglia, PD is generally diagnosed based on clinical criteria, by using a medical individual's history and a physical/neurological exam. The loss of dopamine in the central nervous system, along with the anatomical and physiological changes related to the disease, has an impact on laryngeal, respiratory and articulatory functions of Persons with PD (PwPD). Alterations in speech and voice are in fact among the earliest symptoms of PD, which results in a motor speech disorder called hypokinetic dysarthria [3, 4]. Nevertheless, the rich information carried by patients' speech and voice is only partially used for diagnosis and clinical decision-making, since the Unified Parkinson's Disease Rating Scale (UPDRS), a standardized rating tool used to assess the severity and progression of the pathology, only presents one item (item 3.1) that concerns the evaluation of speech [5]. This item is based on the clinician's perception and mostly considers speech in terms of intelligibility. A deeper understanding of speech and voice phenomena by advanced data analytics methods could be therefore very useful in both the diagnostic phase and in the monitoring of therapy response in PwPD.

2. Speech in Parkinson's Disease

PD-related dysarthria, caused by poor activation and coordination of the muscles involved in speech production, includes a range of alterations, extensively described in experimental studies on different languages [6].

As for the voice quality, a breathy, husky-semiwhisper and hoarse voice is often reported in PwPD, accompanied by vocal tremor, an increase in nasality, reduced voice intensity and constant loudness [7]. Voice quality spectrum was also studied using a deep learning approach applied to differential phonological posterior features for the characterization of pathological PD speech, collected through different tasks and compared to healthy non-modal phonation. [8].

At the segmental level, the decreased amplitude of motility of lips, tongue, and jaw provokes imprecision in the production of consonantal sounds, with the so-called spirantization phenomenon or occlusive weakening [9, 10]. A reduction in the vowel space area and an im-

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
maffia@unior.it (M. Maffia); loredana.schettino@unina.it (L. Schettino); vincenzonorman.vitale@unina.it (V. N. Vitale)
0000-0002-4913-374X (M. Maffia); 0000-0002-3788-3754 (L. Schettino); 0000-0002-0365-8575 (V. N. Vitale)

[1]This article is the result of the collaboration among the authors. However, for academic purposes, Marta Maffia is responsible for sections 1 and 2, Loredana Schettino for section 3 and Norman Vincenzo Vitale for section 4. All the authors are responsible for section 5.

paired and less distinctive formant generation in speech of PwPD have also been described, both in sustained prolongation of single vowels [11] and in continuous speech, such as sentence repetition [12] or reading passage [13]. The centralization of formant values, measured by the Vowel Articulation Index (VAI), was also proposed as a potential early marker of PD, especially when observed in spontaneous speech [14].

As for the suprasegmental aspects, PwPD often report a significantly narrower tonal range (monopitch) or an abnormal pitch variability, along with a compromised ability to consciously manipulate intonation [15, 4]. Articulation and speech rate are also altered in PD, although previous findings do not highlight a uniform pattern of variation in the speech of PwPD: in some studies a reduction in speech rate was observed in PD patients [16], while some reported the opposite effect [17, 18] and other found no intergroup differences between pathological and healthy speech [19]. Furthermore, different rhythmic metrics were used to describe the alteration of rhythm in PD speech, as part of a more "general dysrhythmia" [20]. In recent studies on Italian PD patients, the percentage of vocalic intervals (%V) was found to be effective in characterizing pathological speech, when compared to that of healthy individuals, both in read and spontaneous conditions and even at a very early stage of the disease [21, 22].

In the last decades, in line with the growing interest and efforts in the identification of reliable linguistic and acoustic biomarkers of PD, some studies demonstrated that an accurate diagnosis could be supported by the application of fully automated analytic methods and machine learning techniques on speech recordings [23]. However, most of the proposed procedures were designed for highly functional (but "artificial") vocal paradigms such as sustained phonation, diadochokinetic tasks, syllable repetition, short sentences [24, 25, 26, 27, 28, 29]. These kinds of elicitation techniques indeed provide highly controlled signals, but such control affects phonation and may even mask features that may emerge in less controlled semi(spontaneous) connected speech. In addition, previous studies often achieve high levels of accuracy in the detection of PD speech by taking into account a very large number of features, and the classification focuses on computational aspects rather than linguistic ones [30].

In this contribution, we address the following issues:

- investigate the role of acoustic features, usually overlooked or, however, not always or directly taken into account by specialists for PD diagnosis;
- consider patterns that emerge from connected (read) speech rather than isolated speech units (phones, syllables, words) productions;
- support machine learning-based methods with linguistic considerations so as to reduce the size of the big sets of features automatically extracted to some meaningful ones and provide an effective linguistic interpretation of the results.

3. Method

3.1. Data and Annotation

The study has been conducted on data from the *Italian Parkinson's Voice and Speech* corpus [31, 32], which consists of speech data collected through different speech production tasks from three groups of Italian (Apulian) speakers: PD patients, age-matched healthy control (HC) speakers and younger HC speakers.

In particular, we considered a subset of this corpus, consisting in 25 speech samples elicited through a reading task[2] from 15 PD patients and 10 age-matched healthy speakers. Subjects in the PD group are classified by the specialists as <4 on the modified Hoehn and Yahr scale, which stands for a non-severe stage of the severity of their disease. The patients' speech ability is evaluated following the tips provided in section 3.1 (eloquence) of the Unified Parkinson's Disease Rating Scale (UPDRS) as minimally/slightly impaired (maximum score is 4 = severe impairment). Demographic and clinical features of patients with PD and HC speakers are resumed in Table 1.

	HC (n=10)	**PD (n=15)**
Age (m±SD)	68±6	64±9
Sex (M/F)	4/6	11/4
H&Y	-	<4
UPDRS (Item 3.1)	-	1.07±1.18

Table 1
Biographical (Sex and Age) characteristics of the PD and HC speakers and clinical data (H&Y: Hoehn & Yahr scale; UPDRS: Unified Parkinson's Disease Rating Scale) of PD speakers [32].

The considered dataset had already been the object of a spectroacoustic analysis in a previous study [22] and the acoustic signal had been therefore manually segmented and annotated into vowel (V) and consonantal (C) intervals (see Figure 1). Main descriptive statistics of the dataset are reported in Table 2.

3.2. Analysis

In this study, we intend to use the described continuous speech data for PD detection based on a reduced set of interpretable features of the acoustic signal. To this aim,

[2]The reading task was based on a phonemically balanced text [31].

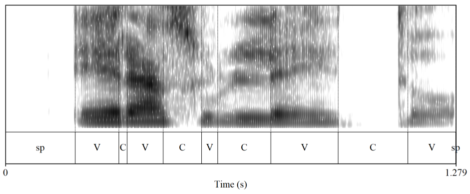


Figure 1: Spectrogram and annotation of the utterance "*era sul letto*", "(Dad) was on the bed". C: consonantal interval, V: vowel interval, sp: silent pause.

	Tot	HC	PD
Total duration (s)	1765	614	1151
Duration of speech portions (s)	1206	455	751
Duration of samples (s) (m±SD)	71±17	61±4	77±20
n. of V intervals	4664	1761	2903
n. of CV intervals	5260	2107	3153
n. of Phonetic Chains	910	312	598

Table 2
Descriptive statistics of the considered subset.

three trials of PD detection were conducted, each time considering a different basic unit, namely:

- Vowels (V) - in previous studies, the percentage of vocalic interval in the speech signal was demonstrated to be informative in PD detection. So we investigate whether vowels alone contain enough information for the detection task;
- Consonant and Vowels (CV) - we extend the context of vowels to the previous consonants, obtaining a wider feature extraction window to evaluate the influence of consonants preceding vowels on PD detection;
- Phonetic Chains (PC) - lastly we employ the phonetic chain, namely the sequence of vowels and consonants between two silent pauses. On the one hand, such units provide the most comprehensive automatically detectable window for feature extraction. On the other hand, being a larger unit of speech production, it should provide far enough features to discriminate speaker status.

Based on the OpenSmile toolkit [33], we selected the eGeMAPSv02 [34] as the basic feature set, and then investigated which features could be considered as the most relevant for discrimination considering previous literature [35] and inspection of the data with the Orange software [36].

Then, the impact of the selected features was evaluated by employing two unsupervised machine-learning techniques:

- The **K-Means**[2] [37] a vector-quantization method which divides n objects in k clusters based on their mean distance.
- **Hierarchical Agglomerative Clustering (HAC)**[2] [38] is a greedy technique that aims at grouping (or splitting) clusters based on a similarity measure. The final output is a clusters hierarchy which could be divided based on the number of desired clusters.

These simple yet efficient techniques were employed to obtain explainable and interpretable results.

The PD detection trials were conducted considering the following sets of features:

- a full feature set, i.e. the eGeMAPSv02 complete feature set (88 features) [34] plus the speakers' sex.
- a subset feature set, i.e. **18** features from the eGeMAPSv02 feature set, plus the sex (see Appendix A).

In both cases, features were normalized at zero mean and unitary variance.

4. Results

The inspection conducted with the Orange software highlighted that the most relevant features for discriminating between PwPD and HC speakers are those concerning the spectral distribution (i.e., slope, alpha ratio, Hammarberg index), followed by those concerning energy and amplitude (i.e. loudness, shimmer), and frequency (MFCC). The observed features were included in the subset employed for the discrimination trials (as reported in Appendix A. Also, the table in Appendix C shows the Mean values and Standard Deviation of these features in PC units per speaker).

Results show that classification based on the Phonetic Chain (see Figure 4) outperforms by far classifiers based on both V and CV. On the one hand, the HAC classifier with the full feature set reaches nearly 99% of true positive detection and 85% of true negative detection. On the other hand, the K-means performs at its best with the feature subset with an 89% of true positive and a 72% of true negative. This means that by reducing the number of features of 75% with respect to the original feature set, the K-means has a 10% reduction in true positive (i.e., PD) detection and a 13% reduction in true negative (i.e., HC) detection, with respect to HAC on the full feature set.

The vowels-based setting (see Figure 2) shows better performances with the feature subset with both K-means

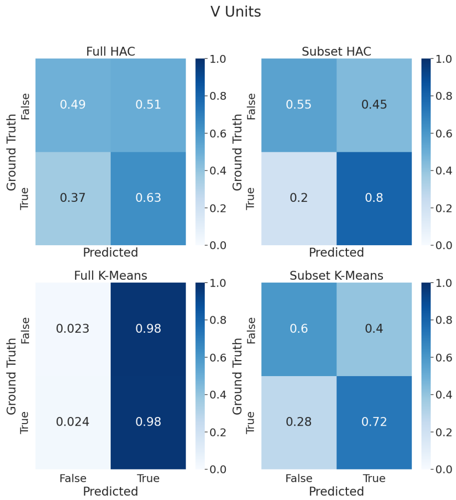


Figure 2: V Clustering.

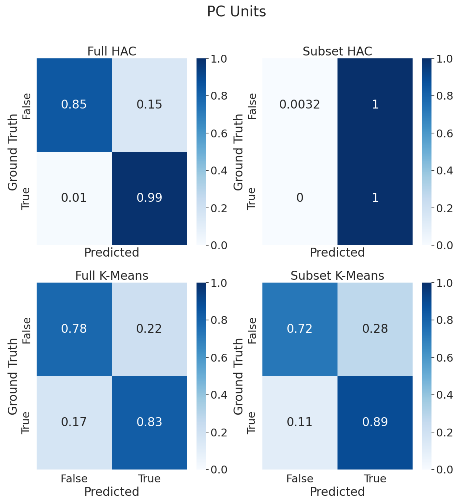


Figure 4: PC Clustering.

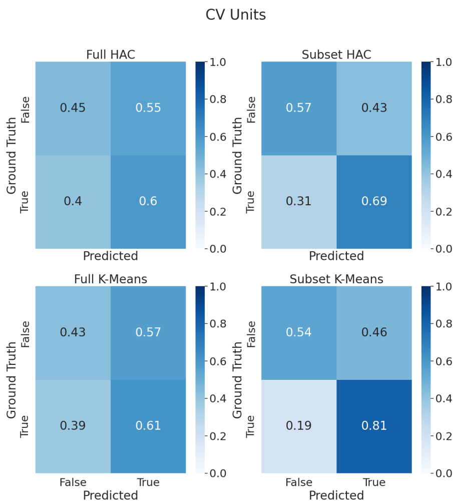


Figure 3: CV Clustering.

and HAC. However, the True negative detection rate is near 60% in the best case, while the true positive rate is at 80% in the best case.

Finally the CV setting (see Figure 3) shows performances which are comparable to a coin toss in most of cases. Only the K-means based on feature subset reaches a true positive detection rate of 81%, with a true negative detection rate of 54%.

In light of these results, we decided to also investigate the correlation between the considered features and the intelligibility score (from the above-described UPDRS) given by the specialists. As illustrated in figure 5, no strong correlation emerges between UPDRS scores and the analysed acoustic features with the exception of slopeV0-500 that negatively correlates with the specialists' ratings (see Appendix B for the correlation matrices concerning the features extracted from V and VC intervals, Figure 7).

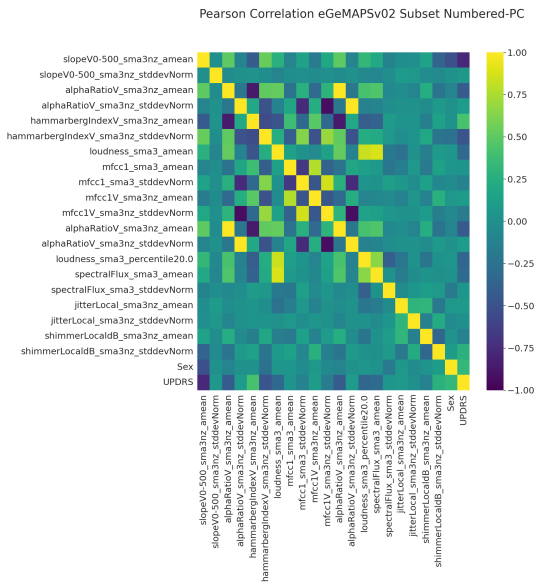


Figure 5: Feature correlation considering PC units.

5. Discussion and Conclusion

The present study provides relevant findings both for the development of PD detection systems and the analysis of Parkinsonian speech characteristics by integrating computational methods with domain-specific linguistic knowledge.

The correlation data between the UPDRS ratings concerning PD speakers' speech ability and the acoustic features automatically extracted from the speech signal corroborate the observation that the specialists' holistic assessment overlooks, or at least only partially and indirectly considers, acoustic features, which, nonetheless, prove to provide crucial information for the diagnosis. In fact, the speech signal is affected by the condition of the muscles involved in phonation. So, if the vocal apparatus is somewhat compromised as an effect of the muscular impairment due to the disease (dysarthria), the signal should show this. Hence, the relevance of including acoustic features in the assessment of the outbreak and severity of PD.

However, fully automated extraction and treatment of speech acoustic features is usually achieved with highly complex systems whose interpretation is quite difficult for both computational scientists, who might be not familiar with PD symptoms and the linguistic value of the features of the speech signal, and for domain experts, who might not be familiar with machine learning methods. Therefore, the design of models in a way that their predictions can be explainable and easily interpretable may actually be most sensible and economical. In fact, this study highlights that not all the possibly considerable acoustic features provide the same amount of information and are actually relevant for discrimination. Moreover, their contribution may vary as a result of the type and span of the linguistic unit used for the feature extraction.

More specifically, the classification results show that considering vowel intervals as units of reference for the features extraction is already quite effective. Most effective is, however, considering wider contexts as provided by the inter-pausal phonetic chain intervals, whereas enlarging the vocalic intervals only to the previous consonant (CV intervals as a basic unit) turns out to be noisy rather than informative.

Then, on average, the feature subset proved to be most informative, carrying out sufficient information to let the classifiers reach a reasonable detection rate in the considered medical scenario. In particular, the subset mainly includes features concerning spectral distribution, followed by those involving energy and amplitude and finally frequency features (MFCC above all).

It is worth noticing that the study has been conducted on continuous speech rather than on isolated phones, syllables or words, to get closer to the normal working dynamic of the vocal apparatus during utterance phonation and avoid artificial effects that may arise when producing single short items.

To conclude, supporting automated analytic methods and machine learning techniques with linguistic considerations allows for more accurate, economical and, most importantly, interpretable discrimination. Future work will be devoted to delving deeper into the linguistic analysis of the way the emergent features characterize PD speech and the investigation of the explainability of classification methods based on deep neural networks.

References

[1] L. M. De Lau, M. M. Breteler, Epidemiology of parkinson's disease, The Lancet Neurology 5 (2006) 525–535.

[2] W. Poewe, K. Seppi, C. M. Tanner, G. M. Halliday, P. Brundin, J. Volkmann, A.-E. Schrag, A. E. Lang, Parkinson disease, Nature reviews Disease primers 3 (2017) 1–21.

[3] F. L. Darley, A. E. Aronson, J. R. Brown, Clusters of deviant speech dimensions in the dysarthrias, Journal of speech and hearing research 12 (1969) 462–496.

[4] S. Pinto, A. Chan, I. Guimarães, R. Rothe-Neves, J. Sadat, A cross-linguistic perspective to the study of dysarthria in parkinson's disease, Journal of Phonetics 64 (2017) 156–167.

[5] C. G. Goetz, B. C. Tilley, S. R. Shaftman, G. T. Stebbins, S. Fahn, P. Martinez-Martin, W. Poewe, C. Sampaio, M. B. Stern, R. Dodel, et al., Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (mds-updrs): scale presentation and clinimetric testing results, Movement disorders: official journal of the Movement Disorder Society 23 (2008) 2129–2170.

[6] J. R. Duffy, Defining, understanding, and categorizing motor speech disorders, Motor speech disorders—substrates, differential diagnosis, and management. 3rd edn. Saint Louis: Elsevier Mosby (2013) 3–13.

[7] D. G. Hanson, B. R. Gerratt, P. H. Ward, Cinegraphic observations of laryngeal function in parkinson's disease, The Laryngoscope 94 (1984) 348–353.

[8] M. Cernak, J. R. Orozco-Arroyave, F. Rudzicz, H. Christensen, J. C. Vásquez-Correa, E. Nöth, Characterisation of voice quality of parkinson's disease using differential phonological posterior features, Computer Speech & Language 46 (2017) 196–208.

[9] H. Ackermann, W. Ziegler, Articulatory deficits in parkinsonian dysarthria: an acoustic analysis., Journal of Neurology, Neurosurgery & Psychiatry 54 (1991) 1093–1098.

[10] D. Duez, Acoustic analysis of occlusive weakening in parkinsonian french speech, in: International Congress of Phonetic Sciences, Université de Saarebrücken, 2007, pp. 1–4.
[11] I. Eliasova, J. Mekyska, M. Kostalova, R. Marecek, Z. Smekal, I. Rektorová, Acoustic evaluation of short-term effects of repetitive transcranial magnetic stimulation on motor aspects of speech in parkinson's disease, Journal of Neural Transmission 120 (2013) 597–605.
[12] S. Sapir, L. O. Ramig, J. L. Spielman, C. Fox, Formant centralization ratio: A proposal for a new acoustic measure of dysarthric speech (2010).
[13] S. Skodda, W. Visser, U. Schlegel, Vowel articulation in parkinson's disease, Journal of voice 25 (2011) 467–472.
[14] J. Rusz, R. Cmejla, T. Tykalova, H. Ruzickova, J. Klempir, V. Majerova, J. Picmausova, J. Roth, E. Ruzicka, Imprecise vowel articulation as a potential early marker of parkinson's disease: Effect of speaking task, The Journal of the Acoustical Society of America 134 (2013) 2171–2181.
[15] A. M. Goberman, C. A. Coelho, M. P. Robb, Prosodic characteristics of parkinsonian speech: The effect of levodopa-based medication, Journal of medical speech-language pathology 13 (2005) 51–69.
[16] C. L. Ludlow, N. P. Connor, C. J. Bassich, Speech timing in parkinson's and huntington's disease, Brain and language 32 (1987) 195–214.
[17] H. Hirose, S. Kiritani, M. Sawashima, Velocity of articulatory movements in normal and dysarthric subjects, Folia Phoniatrica et Logopaedica 34 (1982) 210–215.
[18] H. Ackermann, J. Konczak, I. Hertrich, The temporal control of repetitive articulatory movements in parkinson's disease, Brain and language 56 (1997) 312–319.
[19] S. Skodda, U. Schlegel, Speech rate and rhythm in parkinson's disease, Movement disorders: official journal of the Movement Disorder Society 23 (2008) 985–992.
[20] J. M. Liss, L. White, S. L. Mattys, K. Lansford, A. J. Lotto, S. M. Spitzer, J. N. Caviness, Quantifying speech rhythm abnormalities in the dysarthrias (2009).
[21] M. Maffia, R. De Micco, M. Pettorino, M. Siciliano, A. Tessitore, A. De Meo, Speech rhythm variation in early-stage parkinson's disease: a study on different speaking tasks, Frontiers in Psychology 12 (2021) 668291.
[22] M. Maffia, M. Pettorino, Voce, età e parkinson: questioni di ritmo, in: CLUB Working Papers in Linguistics, volume 6, Alma Mater Studiorum Università di Bologna, 2022, pp. 66–78.
[23] J. Hlavnička, R. Čmejla, T. Tykalová, K. Šonka, E. Ržička, J. Rusz, Automated analysis of connected speech reveals early biomarkers of parkinson's disease in patients with rapid eye movement sleep behaviour disorder, Scientific reports 7 (2017) 12.
[24] J. I. Godino-Llorente, S. Shattuck-Hufnagel, J.-Y. Choi, Moro-Velázquez, J. A. Gómez-García, Towards the identification of idiopathic parkinson's disease from the speech. new articulatory kinetic biomarkers, PloS one 12 (2017) e0189583.
[25] M. Chronowski, M. Klaczynski, M. Dec-Cwiek, K. Porebska, Parkinson's disease diagnostics using ai and natural language knowledge transfer, arXiv preprint arXiv:2204.12559 (2022).
[26] L. Ali, C. Zhu, M. Zhou, Y. Liu, Early diagnosis of parkinson's disease from multiple voice recordings by simultaneous sample and feature selection, Expert Systems with Applications 137 (2019) 22–28.
[27] J. S. Almeida, P. P. Rebouças Filho, T. Carneiro, W. Wei, R. Damaševičius, R. Maskeliūnas, V. H. C. de Albuquerque, Detecting parkinson's disease with sustained phonation and speech signals using machine learning techniques, Pattern Recognition Letters 125 (2019) 55–62.
[28] R. Lamba, T. Gulati, H. F. Alharbi, A. Jain, A hybrid system for parkinson's disease diagnosis using machine learning techniques, International Journal of Speech Technology 3 (2021) 583–593.
[29] C. O. Sakar, G. Serbes, A. Gunduz, H. C. Tunc, H. Nizam, B. E. Sakar, M. Tutuncu, T. Aydin, M. E. Isenkul, H. Apaydin, A comparative analysis of speech signal processing algorithms for parkinson's disease classification and the use of the tunable q-factor wavelet transform, Applied Soft Computing 74 (2019) 255–263.
[30] F. Cordella, A. Paffi, A. Pallotti, Classification-based screening of parkinson's disease patients through voice signal, in: 2021 IEEE International Symposium on Medical Measurements and Applications (MeMeA), IEEE, 2021, pp. 1–6.
[31] G. Dimauro, F. Girardi, Italian parkinson's voice and speech, IEEE Dataport (2019).
[32] G. Dimauro, V. Di Nicola, V. Bevilacqua, D. Caivano, F. Girardi, Assessment of speech intelligibility in parkinson's disease using a speech-to-text system, IEEE Access 5 (2017) 22199–22208.
[33] F. Eyben, M. Wöllmer, B. Schuller, Opensmile: the munich versatile and fast open-source audio feature extractor, in: Proceedings of the 18th ACM international conference on Multimedia, 2010, pp. 1459–1462.
[34] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, et al., The geneva minimalistic

acoustic parameter set (gemaps) for voice research and affective computing, IEEE transactions on affective computing 7 (2015) 190–202.

[35] L. Moro-Velazquez, J. A. Gomez-Garcia, J. D. Arias-Londoño, N. Dehak, J. I. Godino-Llorente, Advances in parkinson's disease detection and assessment using voice and speech: A review of the articulatory and phonatory aspects, Biomedical Signal Processing and Control 66 (2021) 102418.

[36] J. Demšar, T. Curk, A. Erjavec, Črt Gorup, T. Hočevar, M. Milutinovič, M. Možina, M. Polajnar, M. Toplak, A. Starič, M. Štajdohar, L. Umek, L. Žagar, J. Žbontar, M. Žitnik, B. Zupan, Orange: Data mining toolbox in python, Journal of Machine Learning Research 14 (2013) 2349–2353. URL: http://jmlr.org/papers/v14/demsar13a.html.

[37] J. A. Hartigan, M. A. Wong, Algorithm as 136: A k-means clustering algorithm, Journal of the royal statistical society. series c (applied statistics) 28 (1979) 100–108.

[38] F. Murtagh, A survey of recent advances in hierarchical clustering algorithms, The computer journal 26 (1983) 354–359.

Appendix A: Further Features Analysis

List of the features included in the considered subset of the eGeMAPSv02 features.

Features concerning the spectral distribution:

- slopeV0-500_sma3nz_amean
- slopeV0-500_sma3nz_stddevNorm
- alphaRatioV_sma3nz_amean
- alphaRatioV_sma3nz_stddevNorm
- hammarbergIndexV_sma3nz_amean
- hammarbergIndexV_sma3nz_stddevNorm
- spectralFlux_sma3_amean
- spectralFlux_sma3_stddevNorm

Features concerning energy and amplitude:

- loudness_sma3_amean
- loudness_sma3_percentile20.0
- shimmerLocaldB_sma3nz_amean
- shimmerLocaldB_sma3nz_stddevNorm

Features concerning frequency:

- mfcc1_sma3_amean
- mfcc1_sma3_stddevNorm
- mfcc1V_sma3nz_amean
- mfcc1V_sma3nz_stddevNorm
- jitterLocal_sma3nz_amean
- jitterLocal_sma3nz_stddevNorm

Appendix B: Further Results

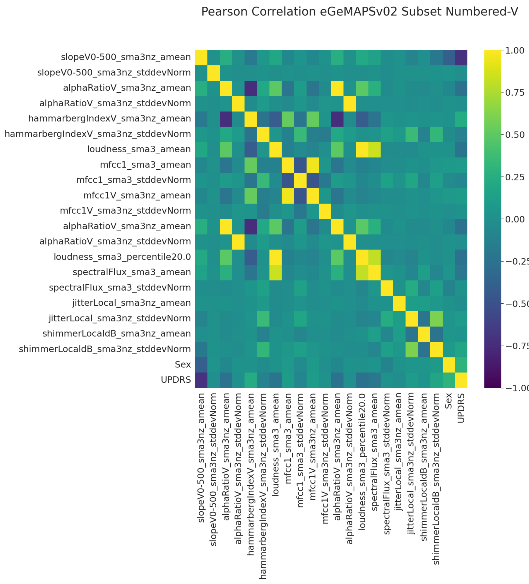


Figure 6: Feature correlation considering V units.

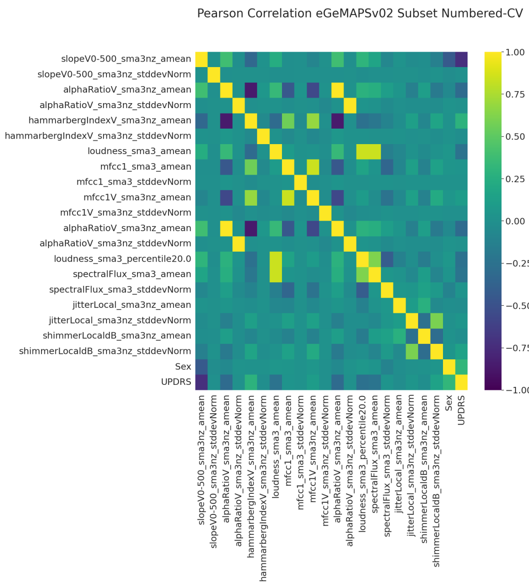


Figure 7: Feature correlation considering V and CV units.

Speaker	Slope	alphaRatio	H-Index	Shimmer	Loudness	MFCC
01PDm	0,0003 ± 0,0159	-21,2933 ± 2,9959	32,5541 ± 3,1311	1,2979 ± 0,2976	0,3965 ± 0,1292	36,3328 ± 5,8431
02PDm	0,0074 ± 0,0111	-21,3531 ± 2,7926	32,3354 ± 3,5172	1,0792 ± 0,2899	0,4698 ± 0,1215	34,4790 ± 5,6093
03PDm	0,0012 ± - 0,0124	-21,6803 ± 3,7769	29,2366 ± 4,5719	1,0642 ± 0,1860	0,5931 ± 0,1269	38,3075 ± 5,1538
04PDf	0,0083 ± 0,0107	-25,8529 ± 3,5913	37,0023 ± 3,0799	0,7232 ± 0,1815	0,3373 ± 0,0763	33,3890 ± 5,8998
05PDf	0,0283 ± 0,0102	-16,0985 ± 2,3628	26,4895 ± 1,8925	0,9988 ± 0,1635	0,6615 ± 0,1167	29,3979 ± 5,7566
06PDf	0,0080 ± 0,0077	-19,9178 ± 3,7987	32,1550 ± 3,5414	1,2773 ± 0,2916	0,2556 ± 0,0906	30,4934 ± 6,5089
07PDf	0,0322 ± 0,0088	-18,0467 ± 1,7098	28,5596 ± 2,2559	0,9679 ± 0,2658	0,2350 ± 0,0626	30,9480 ± 4,0652
08PDm	0,0020 ± 0,0113	-22,3596 ± 3,0759	31,7079 ± 2,7711	1,4473 ± 0,2784	0,2608 ± 0,0790	34,6625 ± 5,7257
09PDm	0,0002 ± 0,0074	-20,2667 ± 6,1511	29,4351 ± 8,4180	1,0641 ± 0,4419	0,3720 ± 0,1468	33,0758 ± 8,7907
10PDm	0,0006 ± 0,0117	-22,4290 ± 3,8050	31,4671 ± 3,5929	1,1519 ± 0,2819	0,1562 ± 0,0651	26,6755 ± 6,4457
11PDm	0,0074 ± 0,0134	-22,3154 ± 6,3170	31,3600 ± 8,7320	1,0987 ± 0,4362	0,1010 ± 0,0460	24,7227 ± 7,9015
12PDm	0,0000 ± 0,0115	-29,8513 ± 3,0748	40,0691 ± 3,5827	1,1172 ± 0,3162	0,1761 ± 0,0754	30,2178 ± 6,2936
13PDm	0,0051 ± 0,0098	-18,6693 ± 6,1645	25,7068 ± 8,2491	0,9146 ± 0,4247	0,2866 ± 0,1100	33,8904 ± 8,3735
14PDm	0,0185 ± 0,0111	-23,1260 ± 2,3925	32,0836 ± 2,9519	1,2185 ± 0,2678	0,4287 ± 0,1531	32,0228 ± 4,6341
15PDm	0,0090 ± 0,0108	-21,9545 ± 3,0236	30,2927 ± 3,3761	1,1309 ± 0,3094	0,2977 ± 0,0723	32,9187 ± 6,1951
16HCf	0,0633 ± 0,0096	-18,6750 ± 3,0968	27,9920 ± 3,7985	1,2668 ± 0,2783	0,2887 ± 0,1005	29,5724 ± 7,6025
17HCf	0,0926 ± 0,0076	-18,0667 ± 3,9360	25,8280 ± 3,6499	1,1462 ± 0,2630	0,2700 ± 0,0491	30,1793 ± 7,9545
18HCm	0,0711 ± 0,0060	-10,3776 ± 2,8562	21,0355 ± 3,7580	1,2385 ± 0,2646	0,7188 ± 0,1946	22,0280 ± 5,7435
19HCf	0,0783 ± 0,0091	-16,3902 ± 3,5428	27,4936 ± 5,7825	1,2023 ± 0,2160	0,4559 ± 0,1197	30,9898 ± 7,4353
20HCf	0,0846 ± 0,0059	-21,5727 ± 3,8459	34,0136 ± 4,6024	1,6192 ± 0,2362	0,1724 ± 0,0470	33,4514 ± 6,8876
21HCf	0,0716 ± 0,0097	-17,2437 ± 3,2172	26,4872 ± 4,1833	1,0707 ± 0,3338	0,4527 ± 0,1017	29,4803 ± 6,0164
22HCm	0,0662 ± 0,0110	-14,8324 ± 4,1724	25,3376 ± 4,3715	1,5356 ± 0,3035	0,4509 ± 0,1726	32,5912 ± 8,1394
23HCf	0,0948 ± 0,0073	-20,3646 ± 4,4190	29,3404 ± 4,9695	1,1668 ± 0,2565	0,4223 ± 0,0784	33,8704 ± 6,1713
24HCm	0,0715 ± 0,0090	-11,9822 ± 3,3146	22,3366 ± 4,3487	1,3189 ± 0,2434	0,4978 ± 0,1417	33,4114 ± 7,8817
25HCf	0,0876 ± 0,0066	-16,7485 ± 4,4084	27,8112 ± 5,7054	1,2287 ± 0,2130	0,4543 ± 0,0990	27,9522 ± 8,3363

Table 3
Mean value and Standard Deviation of the most relevant features in PC units per speaker.

Appendix C: Individual Variability

Introducing Deep Learning with Data Augmentation and Corpus Construction for LIS

Manuela Marchisio[1,*,†], Alessandro Mazzei[1,†] and Dario Sammaruga[2,†]

[1] *Università degli Studi di Torino - Corso Svizzera 185, 10149, Torino, Italy*
[2] *Orbyta Tech S.r.l. - Piazza Castello 113, 10121 Torino, Italy.*

Abstract

The development of home video recording has had a big impact in the development of video documents containing Italian Sign Language (LIS) sentences. LIS2SPEECH is an ongoing project by Orbyta Tech s.r.l. to build a complete translation chain from LIS to speech. The idea is to build a free software framework to transform video containing LIS sentence into Italian vocal sentences. In this way, LIS signers can indirectly produce Italian vocal sentences. In this paper we describe two milestones for LIS2SPEECH, that are: i. the development of some deep neural models trained by using data augmentation technique, and ii. the construction of a new dataset for LIS to Italian. Referring to the first point, a number of deep learning models were developed and tested. Then data augmentation was performed by using some geometric transformations to the videos belonging to the original training set. With reference to the second point, we constructed the TGLIS-227 dataset by using video and audio segmentation techniques, starting from a corpus of RAI newscasts. This dataset is a novelty in the current research panorama as there are no public datasets for LIS with sentence-level granularity.

Keywords

Sign Language Recognition, LIS language, Deep Learning, RNNs, CNN, LIS dataset, Data Augmentation,

1. Introduction

In this paper, we present the architecture of a real-time translation system from Italian Sign Language (henceforth LIS) to Italian speech. An ideal system platform for this task should be composed of three main modules:

- a first module which, starting from an input video, returns the glosses[1] contained in the video. So, this module performs a *Sign Language Recognition* Task.
- a second module that translates the glosses into the Italian language. So, this module performs a *Sign Language Translation* task.
- a third module for text-to-speech system, that is for pronouncing the sentences in Italian. So, this module performs a *Text to Speech* task.

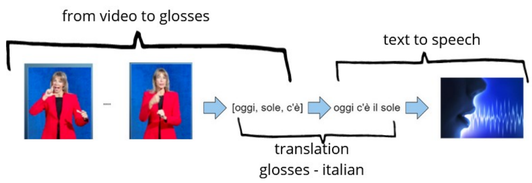


Figure 1: High level architecture for LIS2SPEECH

The task of Sign Language Recognition (SLR) is a classification task that allows to automatically obtain the glosses corresponding to the signs performed by a signer in a video. In general, the SLR is approached as a multiclass classification problem, i.e. a type of supervised learning which, on the basis of a (statistical) model, can associate the correct gloss among the k, where k is the cardinality of the LIS dictionary considered. In the specific settings of the LIS2SPEECH[2] project, the input is a sequence of signs encoded in a video while the labels are the corresponding glosses. So, this is a case of supervised learning, where the models are trained on a dataset containing numerous examples (signs) labeled with the relative class (gloss).

Considering the module for translation from LIS to Italian, this implement a task where is true the rule "more data is better data". In this paper we follow this prescription in two distinct directions. In Section 2, we describe a number of experiments with deep neural models trained

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
*Corresponding author.
† These authors contributed equally.
manuela.marchisio@edu.unito.it (M. Marchisio); alessandro.mazzei@unito.it (A. Mazzei); dario.sammaruga@orbyta.it (D. Sammaruga)
https://github.com/march2345 (M. Marchisio); https://github.com/alexmazzei (A. Mazzei); https://github.com/BeanRepo (D. Sammaruga)
0009-0002-6658-9340 (M. Marchisio); 0000-0003-3072-0108 (A. Mazzei); 0009-0005-4276-9269 (D. Sammaruga)

[1] We write "gloss" to denote a naming system for signs together with the encoding of the relevant morpho-syntactic features.

[2] Copyright for LIS2SPEECH project by Orbyta Tech s.r.l.

by using a data augmentation technique. We enlarge the possibility given by a relatively small initial dataset, by using a number of geometrical transformations to the original videos. We describe these transformations and experiment their impact on the performances of the Isolated Sign Language Recognition (ISLR) task, i.e. when each video contains a single sign. Moreover, in Section 3, we describe the initial steps toward the release of a new dataset for LIS in the news domain. We provide a description of the algorithmic process used to provide a sentence level segmentation of the original videos. In Section 4, we summarizing the contributions of this paper and provide a brief description of the ongoing work.

2. Deep Learning models for Italian SLR

In all experiments concerning neural learning there are two crucial ingredients, these are *the dataset* and the *neural architecture*. In this Section, we describe a number of work related to our project (Section 2.1), we describe the dataset employed in our experiments (Section 2.2), we describe the details of the preprocessing and training applied (Section 2.3) and, finally we report results of our experiments (Section 2.4).

2.1. Related Works on SLR

The major related works in sign language recognition consider 2 important features: the input and neural models used. First, the input could be static (an image for each sign) or dynamic (a video for each sign). The different granularity of the input allows the SLR to be divided into two different tasks: isolated sign language recognition (ISLR) and continuous sign language recognition (CSLR). The former takes a single sign as input and outputs the corresponding gloss. The latter instead takes as input a sentence or a sequence of signs and returns the correct sequence of glosses. [1] is a quantitative analysis of the state of the art on SLR based on more than 400 results from 1983 until today. In this analysis you note that the number of publications on isolated sign recognition is always greater than on a continuous one. Moreover, there are some works that use a *computer vision* approach to detect information from the input and others that use an *electronic* approach by using some gloves with electronic sensor. By limiting to neural networks models, [2] reports an analysis of the main models used in automatic sign language recognition (SLR) up to now, and includes too traditional machine learning classifiers such as SVM (support vector machine), HMM (hidden markov model), K-NN (k-nearest neighbours), ensemble learning and systems based on fuzzy logics. Recently, some other research studies have also been based on Transformers architectures [3, 4] and attention-based models [5]. However, most studies in this SLR task uses two specific neural architectures, that are convolutional [6, 7, 8] and recurrent neural networks [9, 10, 11]. On this basis, we have chosen to develop and train five different NN models: LSTM, GRU, BILSTM, BIGRU and CONVNET[3] (architecture's details in appendix A).

2.2. The LIS Dataset employed in the experiments

For the LIS there are few datasets (see Section 3), and as a consequence we used the only one suitable for SLR, that is the A3LIS-147 dataset [12]. A3LIS-147 was built by the A3LAB research group of the Università politecnica delle Marche, in collaboration with the ENS (Ente Nazionale Sordi, the Italian National Deaf institution) of Ancona. The dataset is composed of 1480 video. The corpus contains 147 standard (*natural*) signs, plus one special (*artificial*) sign for representing the "silence" (*sil sign*). The latter is not a sign belonging to an LIS natural dictionary, but it encodes the common resting position in corpus conversations.

Crucially, and in contrast to most electronic LIS dictionaries, all the signs of A3LIS-147 have been performed by 10 different signers. This peculiar property of the corpus allows us to use it as a training set for the isolated sign language recognition task (ISLR henceforth). For this specific task, each video represents a single sign preceded and followed by the sil sign (or rest).

2.3. Preprocessing and Training

In this Section, we describe the development of a deep neural system for ISLR trained on the A3LIS-147.

First of all, we have a preprocessing phase for converting videos into numerical data suitable for learning. In preprocessing, we extracted a total of 543 keypoints for each frame of the A3LIS-147 videos using the google model Mediapipe Holistics [13]. We decided to reduce this number to 535 since we eliminated 8 keypoints representing lower limbs. Indeed, very often the LIS signers in the videos are framed from the hip up.

Second, we used these keypoints as input for neural networks trainings. We splitted this set of data into two parts: in the initial phase we[4] use 70% in the training set and 30% in the test set. The split is stratified, i.e. it

[3]All these models have been developed using the Keras API, using a GPU NVIDIA-GeForce GTX 1650 with 4GB of RAM and a CPU Intel i7-9750H with 6 core and 16GB of RAM.

[4]In the final part of this work we use k-fold cross-validation to determine what is the best split between test and training. We obtain the best results with k=5, so best split is 80% in training set and 20% for test set

maintains the proportions of the classes in the training and in the test. After a number of initial experiments in training, we observed two emergent issues:

1. The size of the input was too high when the number of signs increased, provoking an out of memory error.
2. The results had a very bad recognition accuracy when considering the entire 148 signs dataset.

For these two reasons, we decided to perform two other preprocessing steps on the data for solving these issues.

These two steps work, in some senses, in two opposite directions. On the one hand, we performed data reduction (Section 2.3.1) for optimizing the number of features given in input to the neural ISLR classification model. On the other hand, we performed data augmentation (Section 2.3.2) on the number of videos for each sign. Indeed, we realized that 10 videos for each sign are too small number to allow the network to correctly classify. We discuss the impact of these two steps in Section 2.4 where we report the results of the experiments with various neural models.

2.3.1. Data Reduction

We reduced the number of keypoints extracted for each frame with the Google model by considering:

- the number of keypoints of the face is higher than other parts of the body and this could negatively affect the training of the model giving too much importance to this part of the body. For this reason we developed a function which allows us to go from 468 to 128 representative keypoints on the contours of the face, eyes, eyebrows and mouth.
- The Mediapipe documentation recommends discarding the z dimension because the Google system still has low performances in predicting the depth. For this reason, in some tests we discarded the z of each keypoint. In other words, we converted the original 3D data produced by mediapipe into 2D by just removing the z value[5].
- Finally, we applied the principal component analysis (PCA) to reduce the total number of keypoints to four-six components, which represents the 95% of explained variance.

2.3.2. Data Augmentation

We applied a Data Augmentation technique by increasing the number of videos for each sign by making some geometric transformations to the originals. In particular we performed: translation, rotation, flip and smoothing.

[5]In a different trial, we have tried to set the z coordinate to zero.

Translation The following transformation was applied to each original keypoint (x,y,z):

$$x_{trasl}, y_{trasl}, z_{trasl} = (x + \Delta x, y + \Delta y, z)$$
$$\Delta x = np.random.uniform(-max_{sx}, max_{dx})$$
$$\Delta y = np.random.uniform(-max_{giu}, max_{su})$$

where Δx represents the displacement on the x axis, while Δy represents the displacement on the y axis. Both these delta were randomly extracted from an uniform distribution by using, as the range, the values representing the maximum translation downwards, upwards, rightwards, leftwards. Using these limits, we guarantee that all the keypoint coordinates are still values between 0 and 1. The randomly extracted values are the same for all frames of a video. In Figure 2c, we report an example of applying this transformation.

Rotation The following transformation was applied to each original keypoint (x,y,z):

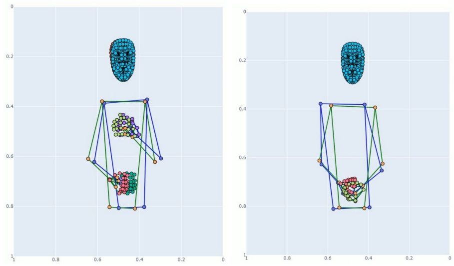
$$x_{rot} = (x - x_{centro})cos\theta - (y - y_{centro})sin\theta + x_{centro}$$
$$y_{rot} = (y - y_{centro})cos\theta + (x - x_{centro})sin\theta + y_{centro}$$
$$z_{rot} = z$$

The rotation was performed with respect to the center of all keypoints $centro = (x_{centro}, y_{centro})$. θ is the angle of rotation: the value is randomly extracted from the uniform distribution between $(-20, 20)$ and is the same for all the keypoints of the frames of one video. In Figure 2a we can see an example of applying this transformation to the keypoints of a video frame.

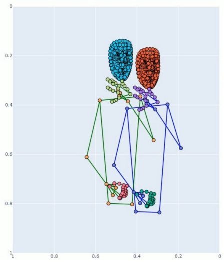
Flip The following transformation was applied to each original keypoint (x,y,z):

$$x_{flip}, y_{flip}, z_{flip} = (2k + x, y, z)$$

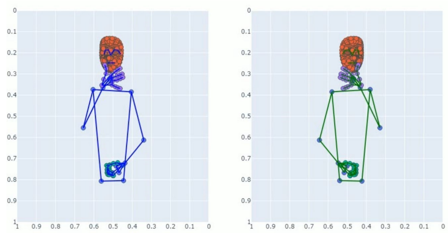
It is an axial symmetry with respect to the straight line $x = k$ parallel to the y axis. In our case k corresponds to the x coordinate of the center of all keypoints. This type of transformation is important because the same sign can also be performed symmetrically, because there are right-handed and left-handed signers. Without this transformation, symmetrical signs cannot managed



(a) original and rotated (b) original and smooth



(c) original and translated



(d) original and flipped

Figure 2: Frame's plot

properly by the neural models, since it recognizes them as different stimulus. In Figure 2d, we can see an example of applying this transformation to the keypoints of a video frame.

Smooth This type of transformation was implemented by considering the study presented in [3]. It consists in applying a different random rotation for each single keypoint up to a maximum of 13 degrees. Note that this transformation was not applied to every part of the body, but only to the keypoints related to the pose. Indeed, keypoints of pose have greater variations when the body moves independently by the head, which in most cases remains in a static position. Applying a different rotation to each point allows to capture variations in the execution of a sign due to a different signer: a slightly more bent elbow, one shoulder lower than the other, different proportions between body parts, etc. This is a crucial transformation because it really produces a kind of totally *new* keypoints, that is really different from those in the original dataset. In Figure 2b, we can see an example of applying this transformation to the keypoints of a video frame.

2.4. Experiments and results

We performed 800 tests divided into two different groups. The first group, called the general group, contains all the possible combinations (480 tests) of the parameters shown in Table 2 (Appendix B). In this group all the values of the parameters were tested. The second group, called specific data augmentation group, is designed to test the impact of each data augmentation transformation on results. It contains some combinations of all the possible combinations in Table 3 (Appendix B)

The results of each test is reported with the results of accuracy, precision and recall curve, confusion matrix, F1-score.

2.4.1. Test Evaluation

We upload the results of the test on the online platform QlikSense[6] in order to build a dashboard that allows us to visualize them. In this Section, we use the these graphical representations to comment results.

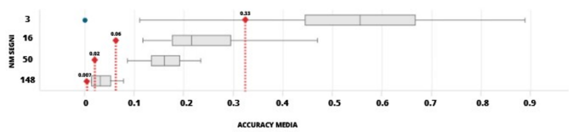


Figure 3: Box plot avg accuracy VS number of signs - without data augmentation

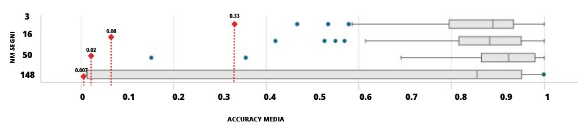


Figure 4: Box plot avg accuracy VS number of signs - with data augmentation

In Figure 3 we observed that if we consider the first group of tests, without data augmentation, the average accuracy decreases if the number of signs increases. The accuracy of a naive model that selects a class at random among the N possible ones using a uniform probability distribution are indicated in red. We use this baseline for accuracy for the models that we have trained. Moreover,

[6]https://www.qlik.com/it-it/products/qlik-sense

to correctly evaluate the generalization power, we consider for accuracy, the precision, the recall, the F1-score and the confusion matrix. In the Figure 3 and 4, we have that the baselines are 33% for 3 signs, 6% for 16 signs, 2% for 50 signs, 0.6% for 148 signs. In general, considering N signs, the baseline is equal to $\frac{1}{N}$ since the classes are balanced in the training dataset.

In Figure 5, we consider average accuracy for different neural models, and we have a very wide range of values for accuracy (from 0 to 0.8). In contrast, in the data

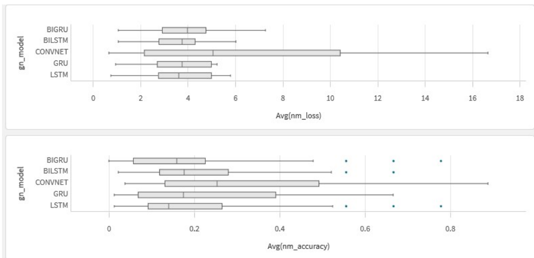


Figure 5: Box plot avg accuracy VS models - without data augmentation

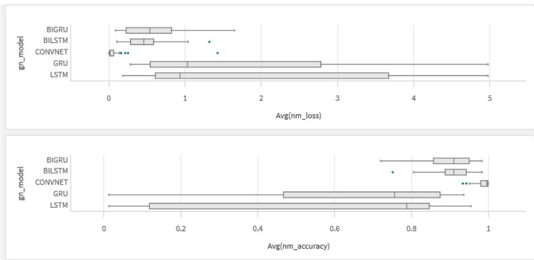


Figure 6: Box plot avg accuracy VS models - with data augmentation

augmentation experimentation, there is no decrease in accuracy when the number of signs increases (Figure 4) and there is a significant increase in the accuracy with respect to the models (Figure 6). Furthermore from the second group of tests we understood that the smooth is the best transformation. This demonstrates that increasing the number of videos for each sign, that is simulating the generation of new videos, had an important impact on the results. Moreover, the results show that in Figure 7 it seems that there is no real difference in using all the coordinates for each keypoint (that are x,y,z) or only two (that are x,y). From experimentations, the best neural model seems to be CONVNET (configured with the specific parameters in Table 4, Appendix B). Indeed, we achieved 100% accuracy (Figure 8) and 100% precision, recall, F1-score on each class. However, we are awzre that these impressive results can be due to overfitting,

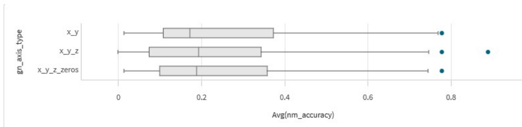


Figure 7: Box plot avg accuracy VS keypoint's coordinates

since the classification task is evaluated on a relatively small dataset of distinct signs (148 signs).

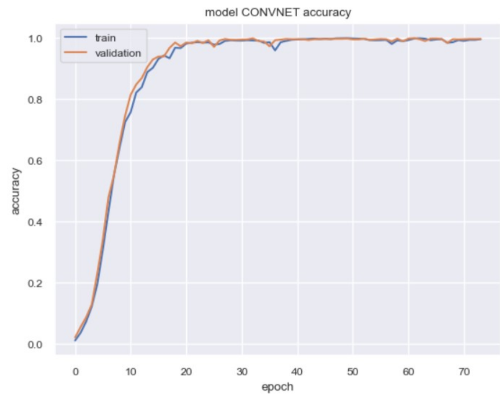


Figure 8: Best model accuracy for CONVNET

2.4.2. A Specific Test for Real Time computation

A small prototype system was developed to test the performances of the neural models implemented in real-time. A number of issues arose for this specific context. As we discussed, all the neural models were trained by applying PCA to the data, but this specific preprocessing step creates problems for a real time application. Indeed, for each prediction of the input, it is necessary to apply PCA to the data collected slowing the real-time performance. To solve this specific issue, we used a specific parallel thread, and by running the PCA on another parallel thread: in this way we have been able to reduce the impact of this problem. By using the OpenCV library we have developed a function that allowed us to read the video from the webcam frame by frame. So, for each frame, the detection of the mediapipe keypoints was performed and saved them in an array. The first prediction took place when we have collected this information for at least N frames, where N is the number of frames of the input shape of the model being tested. After that, each extraction would corresponded to a prediction which is still based on the last N frames present in the array of extracted keypoints. Predictions that exceeded a certain threshold of probability fixed α (in tests performed α = 0.7), are shown in a bar at the top of the window.

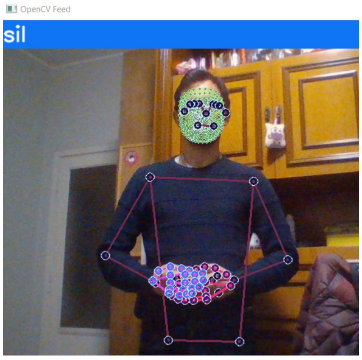


Figure 9: Example of real time detection

3. TGLIS-227: A new dataset for LIS

According to [1] the number of datasets for sign languages is proportional to the number of available datasets of the corresponding national vocal language. As noted in Section 2, the only dataset that can be used in a SLR task for LIS is A3LIS-147 [12], that is composed of 147 signs/videos, performed by 10 different signers. There are other linguistic resources for LIS, that are SpreadTheSign [14], Segni in movimento and SIGNHUB [15]. However, these datasets cannot be effectively used in SLR neural training because they contain only one video for each sign. Moreover, there are no public datasets for LIS with sentence-level granularity.

3.1. Towards a New Dataset

In order to build a new dataset, we considered the available sources of LIS videos. We decided to use the RAI newscasts[7] for three reasons: 1. the quality of LIS production, 2. the availability of many videos, 3. the continuous production of new videos (at least three daily editions). All newscasts have the same video format:

Figure 10: RAI LIS newscast example

[7]https://www.rai.it/dl/easyweb/LIS-e2a267d2-e9a0-4af7-b2ff-baad1f5d060e.html

1. On the left box there is the signer, on the right there is the speaker or, alternatively, the images related to the news.
2. The duration is comprised of between 2 and 5 minutes.
3. Each video contains 5 or 6 different news items.
4. Each news item is preceded and concluded by the sign of silence.
5. The speaker waits for the signer to finish before moving to the next news item.
6. Each news item is accompanied by a subtitle representing the topic.

So, exploiting these features, we developed a system to automatically segment each newscasts video. By identifying the parts of the newscast video containing the silence-LIS in the images and the silence in the audio, we produced a number of videos containing a single news item for each one. The working hypothesis is that the the silences correspond to the transition from one news to an another.

To perform silence-LIS detection we use YOLOv7 (an object detector). To train a YOLO model we needed many images that represented the object to detect, that is the silence-LIS. Since no silence datasets exist, we built it by extracting frames from RAI newscasts. So, we obtained a silence dataset that contained 8000 silence LIS images with resolution 640x640 and with two annotations: only hands or hands+elbows.

3.2. Pipeline construction

We downloaded 20 newscasts from Rai Play and we did a pre-processing step by removing the theme song and by cropping the video to focus on the LIS signer. Then we detected silence-LIS using the YOLO model trained on the silence dataset. Thereafter we detected silences audio for each video: first we extracted audio from video (using moviepy library functions) then we detected only silence that were at least 2 seconds long (using PyDub library function). All detections were recorded into tabular CSV format. Moreover, we built a filter algorithm that took in input detections and returned the ranges that corresponds to the transition from one news item to an other. By using this information, we splitted the newscast in the corresponding news by obtaining a new video file and an audio file for each news. Finally, we annotated each news item (in CSV) with two extra fields:

1. the topic: we applied an optical character recognition (pytesseract) to crop the title of the news items shown in the video
2. the transcript: we applied speech recognition (the Azure SDK speech to text)

These annotations are important in a translation context because the topic represents the context of translation and the transcript represents the target of translation.

3.3. Dataset Final Structure

The final dataset is called TGLIS-227 since it is composed of 227 distinct news items extracted from LIS newscasts editions. For each of them we have:

1. a video (mp4): containing the LIS news item.
2. an audio (wav): that is the audio of the news item in the Italian (vocal) language.
3. topic (in csv): containing the topic of the news item (Appendix E),
4. transcript (in csv): that is the automatic transcription of the news item(Appendix F).

Note that we built an automatic procedure that could be applied several times in order to increase the size of the dataset.

A crucial weakness of the actual dataset is the lack of a standard for LIS transcription in some written form. This linguistics issue requires the collaboration with Deaf organizations and could be performed by using annotation tools for videos such as ELAN [16].

Finally note that for copyright issues we cannot distribute the audio/video content of news items directly, but only the annotations (Appendixes C, D, E, F)[8]. However, by using the timestamps of each news item (Appendix C and Appendix D), and requesting access to the "Teche Rai", it is possible to extract the video and the audio from original newscasts[9].

3.4. Testing deep learning for ISLR on TGLIS-227 videos

We tried to test the best model described in section 2 on the TGLIS-227 dataset (section 3). Since we do not have an LIS annotation (e.g in glosses) for each video, we did a very raw evaluation of the correctness of the ISLR predictions by using the lemma corresponding to the Italian news transcript. In particular, we counted the number of matches between predictions and lemmas, obtaining around 33% of correct matches. This low value is consequence of the the different size of the training dataset, containing only 147 signs, with respect to the size of the TGLIS-227 dataset, containing around 5000 lemma.

[8]data are available at this GitHub: https://github.com/BeanRepo/TGLIS-227

[9]Note that the timestamps are calculated on the videos without start-end theme songs

4. Conclusion and future works

In this work we have presented two main results obtained in the LIS2SPEECH project. First, we have described the application of a number of data augmentation techniques to some deep neural models in the task of ISLR. We proved with experiments that some of these transformations have a strong impact on the final performance of the classification task. Second, we built the TGLIS-227, a new sentence-level dataset for LIS, applying a new procedure for the automatic segmentation of the newscasts.

In future work we intend to develop the following two ideas:

1. to annotate TGLIS-227 video with the glosses that they contain;
2. to develop a system like Common Voice [17] to collect more data to build an open source dataset for LIS;

Moreover, a more challenging development could be to encode additional two video features that are: the lips and facial expressions. Finally, we noted that very often in the news the signers "read" the gloss by using their lips and, moreover, express an emotion related with the gloss by using their facial expression.

References

[1] O. Koller, Quantitative survey of the state of the art in sign language recognition, CoRR abs/2008.09918 (2020). URL: https://arxiv.org/abs/2008.09918. arXiv:2008.09918.

[2] I. Adeyanju, O. Bello, M. Adegboye, Machine learning methods for sign language recognition: A critical review and analysis, Intelligent Systems with Applications 12 (2021) 200056. URL: https://www.sciencedirect.com/science/article/pii/S2667305321000454. doi:https://doi.org/10.1016/j.iswa.2021.200056.

[3] M. Bohacek, M. Hruz, Sign pose-based transformer for word-level sign language recognition, 2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW) (2022) 182–191.

[4] R. Rastgoo, K. Kiani, S. Escalera, Zs-slr: Zero-shot sign language recognition from rgb-d videos, ArXiv abs/2108.10059 (2021).

[5] J. Huang, W. gang Zhou, H. Li, W. Li, Attention-based 3d-cnns for large-vocabulary sign language recognition, IEEE Transactions on Circuits and Systems for Video Technology 29 (2019) 2822–2832.

[6] G. A. Rao, K. Syamala, P. V. V. Kishore, A. S. C. Sastry, Deep convolutional neural networks for

sign language recognition, 2018 Conference on Signal Processing And Communication Engineering Systems (SPACES) (2018) 194–197.

[7] J. Huang, W. gang Zhou, H. Li, W. Li, Sign language recognition using 3d convolutional neural networks, 2015 IEEE International Conference on Multimedia and Expo (ICME) (2015) 1–6.

[8] R. Kumar, A. Bajpai, A. Sinha, Mediapipe and cnns for real-time asl gesture recognition, 2023. arXiv:2305.05296.

[9] M. Borg, K. P. Camilleri, Sign language detection "in the wild" with recurrent neural networks, in: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 1637–1641. doi:10.1109/ICASSP.2019.8683257.

[10] G. Samaan, A. Wadie, A. Attia, A. Asaad, A. Kamel, S. Slim, M. Abdallah, Y.-I. Cho, Mediapipe's landmarks with rnn for dynamic sign language recognition, Electronics 11 (2022) 3228. doi:10.3390/electronics11193228.

[11] D. Kothadiya, C. Bhatt, K. Sapariya, K. Patel, A.-B. Gil-González, J. M. Corchado, Deepsign: Sign language detection and recognition using deep learning, Electronics 11 (2022). URL: https://www.mdpi.com/2079-9292/11/11/1780. doi:10.3390/electronics11111780.

[12] M. Fagiani, S. Squartini, E. Principi, F. Piazza, A new italian sign language database, 2012. doi:10.1007/978-3-642-31561-9_18.

[13] V. Grishchenko, V. Bazarevsky, R. Engineers, G. Research, Mediapipe holistic — simultaneous face, hand and pose prediction, on device, 2020. URL: https://ai.googleblog.com/2020/12/mediapipe-holistic-simultaneous-face.html.

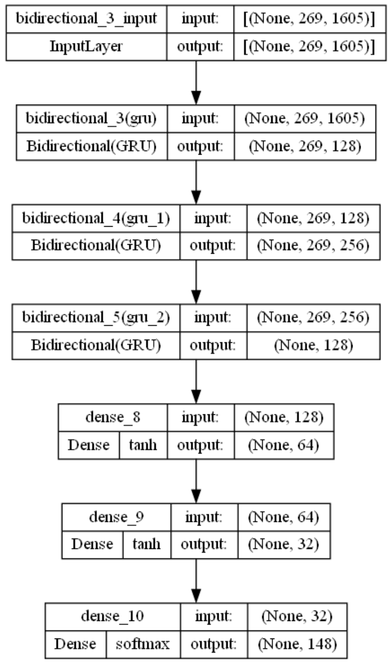
[14] A. Cardinaletti, Il progetto spread the sign, BLITYRI (2016). doi:https://hdl.handle.net/10278/3691616.

[15] Sign-Hub, Sign-hub: Wp 2.4, 2020. URL: https://hdl.handle.net/11403/sign-hub-wp-24/v1, ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

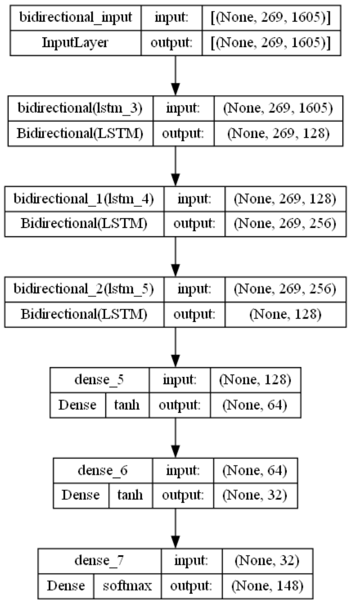
[16] T. L. A. R. f. h. Nijmegen: Max Planck Institute for Psycholinguistics, Elan[computer software], 2023. URL: https://archive.mpi.nl/tla/elan.

[17] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, G. Weber, Common voice: A massively-multilingual speech corpus, in: Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), 2020, pp. 4211–4215.

Appendix A. Neural Networks Architectures

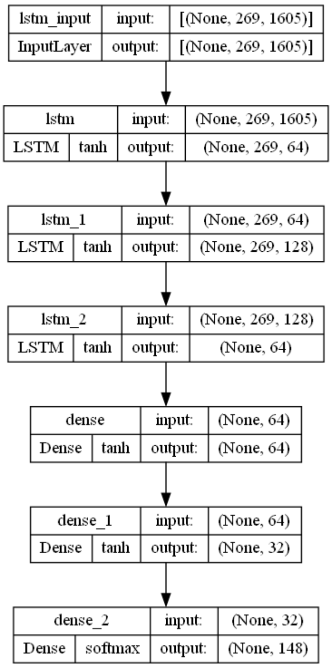


(a) BIGRU

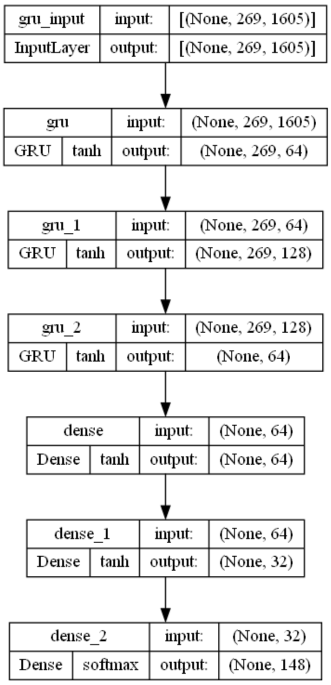


(b) BILSTM

Figure 11: Biderectional Recurrent Architectures



(a) LSTM



(b) GRU

Figure 12: Recurrent architectures

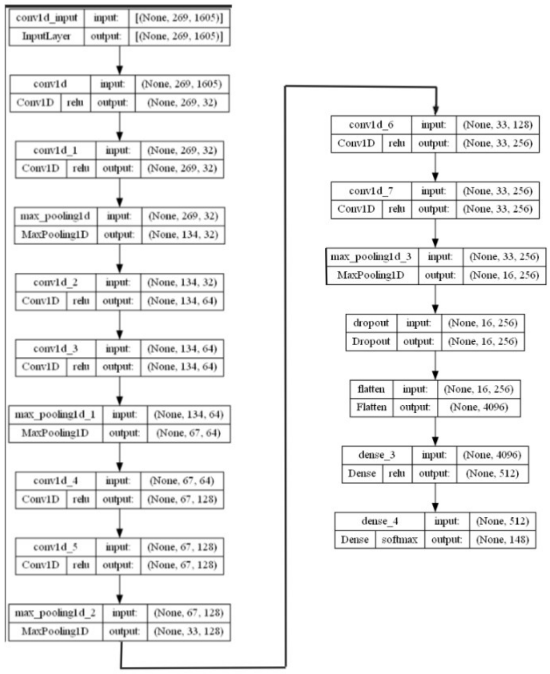


Figure 13: CONVNET Architecture

Appendix B. Test documentation

Table 1
Test documentation: all parameters

info_salvate	type
cd_test_code	integer
cd_train_log_name	string
cd_test_log_name	string
dt_testing_timestamp	datetime
nm_signs	integer
nm_avg_sign_videos	integer
nm_kps	integer
nm_kps_face	integer
nm_kps_pose	integer
nm_kps_lh	integer
nm_kps_rh	integer
gn_axis_type	string
fl_is_normalized	boolean
fl_is_face_red	boolean
gn_face_red_type	string
fl_is_data_aug	boolean
fl_rotated_data	boolean
fl_flipped_data	boolean
fl_smoothed_data	boolean
fl_translated_data	boolean
fl_is_pca	boolean
gn_model	string
gn_model_layers	string
nm_model_parameters	integer
nm_epochs	integer
nm_batch_size	integer
fl_is_early_stop	boolean
nm_patience	integer
gn_test_path	string
gn_optimizer_type	string
gn_accuracy_type	string
gn_loss_type	string
nm_accuracy	float
nm_loss	float

Table 2
General group of tests: parameters

parameters	values
nm_signs	[3,16,50,148]
nm_avg_sign_videos	[10]
gn_axis_type	['x_y','x_y_z', 'x_y_z_zeros']
fl_is_face_red	[True, False]
fl_is_normalized	[True, False]
fl_is_data_aug	[True, False]
fl_is_pca	[True]
gn_model	['GRU','LSTM', 'BIGRU','BILSTM', 'CONVNET']

Table 3
Specific data augmentation group of tests: parameters

parameters	values
nm_signs	[3,16,50,148]
nm_avg_sign_videos	[10]
gn_axis_type	['x_y']
fl_is_face_red	[True]
fl_is_data_aug	[True, False]
fl_is_pca	[True]
fl_is_normalized	[True]
fl_rotated_data	[True, False]
fl_flipped_data	[True, False]
fl_smoothed_data	[True, False]
fl_translated_data	[True, False]
gn_model	['GRU','LSTM', 'BIGRU','BILSTM', 'CONVNET']

Table 4
Best model parameters for the best neural model, i.e. CONVNET

parametro	valore
gn_axis_type	['x_y']
fl_is_face_red	[True]
fl_is_normalized	[True]
fl_is_data_aug	[True]
fl_is_pca	[True]

Appendix C. Video Timestamps[10] This appendix is a sample of video timestamps file. The full version is available on github: https://github.com/BeanRepo/TGLIS-227

Table 5
video range for each newscast

video	range_notizia
01_25_2022.mp4	"['00:00', '00:41'], ['00:44', '01:13'], ['01:15', '01:40'], ['01:44', '02:06'], ['02:09', '02:26'], ['02:28', '02:37']"
01_26_2022.mp4	"['00:00', '00:34'], ['00:39', '01:00'], ['01:06', '01:36'], ['01:40', '02:09'], ['02:14', '02:36'], ['02:38', '02:47']"
01_27_2022.mp4	"['00:00', '00:35'], ['00:39', '01:16'], ['01:20', '01:55'], ['01:59', '02:31'], ['02:36', '03:14'], ['03:16', '03:24']"
01_28_2022.mp4	"['00:00', '00:45'], ['00:48', '01:09'], ['01:13', '01:40'], ['01:44', '02:03'], ['02:05', '02:22'], ['02:24', '02:32']"
02_16_2022.mp4	"['00:00', '00:45'], ['00:50', '01:22'], ['01:27', '01:59'], ['02:04', '02:39'], ['02:42', '03:04'], ['03:05', '03:15']"
02_17_2022.mp4	"['00:00', '00:41'], ['00:45', '01:36'], ['01:40', '02:15'], ['02:21', '02:48'], ['02:52', '03:09'], ['03:12', '03:21']"
...	...

Appendix D. Audio Timestamps[11] This appendix is a sample of audio timestamps file. The full version is available on github: https://github.com/BeanRepo/TGLIS-227

Table 6
audio range for each newscast

audio	range_notizia
01_25_2022.wav	"['00:00', '00:38'], ['00:41', '01:09'], ['01:13', '01:35'], ['01:40', '02:03'], ['02:07', '02:22'], ['02:27', '02:37']"
01_26_2022.wav	"['00:00', '00:29'], ['00:36', '00:55'], ['01:03', '01:29'], ['01:38', '02:06'], ['02:13', '02:32'], ['02:37', '02:47']"
01_27_2022.wav	"['00:00', '00:30'], ['00:37', '01:14'], ['01:18', '01:51'], ['01:57', '02:28'], ['02:34', '03:10'], ['03:14', '03:24']"
01_28_2022.wav	"['00:00', '00:41'], ['00:47', '01:04'], ['01:11', '01:36'], ['01:42', '01:58'], ['02:04', '02:18'], ['02:22', '02:32']"
02_16_2022.wav	"['00:00', '00:43'], ['00:48', '01:18'], ['01:25', '01:55'], ['02:02', '02:36'], ['02:41', '03:00'], ['03:03', '03:15']"
02_17_2022.wav	"['00:00', '00:37'], ['00:43', '01:32'], ['01:39', '02:13'], ['02:18', '02:45'], ['02:51', '03:07'], ['03:11', '03:21']"
...	...

[10] All data in this appendix is protected by Creative Commons Licence CC BY-NC-SA 4.0.

[11] All data in this appendix is protected by Creative Commons Licence CC BY-NC-SA 4.0.

Appendix E. Topic News[12]

This appendix is a sample of topic file. The full version is available on github: https://github.com/BeanRepo/TGLIS-227

Table 7
topic extracted for each news

01_25_2022_chunk_1	I RISULTATI DELLE ELEZIONI
01_25_2022_chunk_2	USA: ITALIA PARTNER IMPORTANTE
01_25_2022_chunk_3	FUGA DALLA RUSSIA PER NON ARRUOLARSI
01_25_2022_chunk_4	IL PRIMO ESPERIMENTO DI DIFESA PLANETARIA
01_25_2022_chunk_5	L'ITALIA BATTE L'UNGHERIA ED È NELLE FINAL FOUR
01_25_2022_chunk_6	
01_26_2022_chunk_1	ELEZIONE PRESIDENTE, IERI FUMATA NERA
01_26_2022_chunk_2	NUOVO IMPULSO AL CONFRONTO TRA I PARTITI
01_26_2022_chunk_3	CRISI UCRAINA, ALTA TENSIONE
01_26_2022_chunk_4	RALLENTA LA CURVA DELL'EPIDEMIA
01_26_2022_chunk_5	OGGI BERRETTINI GIOCA I QUARTI DI FINALE
01_26_2022_chunk_6	
01_27_2022_chunk_1	L'ELEZIONE DEL PRESIDENTE, CONTATTI TRA I PARTITI
01_27_2022_chunk_2	COVID, ALLO STUDIO ESTENSIONE GREEN PASS
01_27_2022_chunk_3	CRISI UCRAINA, DIPLOMAZIA AL LAVORO
01_27_2022_chunk_4	STRAGE DI LICATA, UN PAESE IN LUTTO
01_27_2022_chunk_5	GIORNO DELLA MEMORIA, PAPA : MAI PIÙ QUESTI ORRORI
01_27_2022_chunk_6	
01_28_2022_chunk_1	QUIRINALE, ALLE 11 COMINCIA LA QUINTA VOTAZIONE
01_28_2022_chunk_2	UCRAINA, TELEFONATA BIDEN-ZELENSKY
01_28_2022_chunk_3	OK DELL'EMA ALLA PILLOLA ANTI-COVID
01_28_2022_chunk_4	
01_28_2022_chunk_5	TENNIS, SEMIFINALE BERRETTINI-NADAL
01_28_2022_chunk_6	
...	...

[12] All data in this appendix is protected by Creative Commons Licence CC BY-NC-SA 4.0.

Appendix F. Transcript News[13] This appendix is a sample of transcript file. The full version is available on github: https://github.com/BeanRepo/TGLIS-227

Table 8
transctipt extracted for each news

transcript	
01_27_2022_chunk_6.wav	Ed è tutto grazie per averci seguito. Il tg uno torna alle 8, buona giornata.
01_28_2022_chunk_1.wav	Un giorno dal tg uno la corsa al Quirinale comincerà alle 11, il quinto giorno di votazioni. Il centrodestra sarebbe orientato a votare uno dei nomi proposti nei giorni scorsi. Contrario a questa scelta il centrosinistra, che per protesta potrebbe uscire dall'Aula al momento del voto.
01_28_2022_chunk_1.wav	Intanto il presidente della Camera Roberto Fico.
01_28_2022_chunk_1.wav	Ha convocato alle 10:15 la Conferenza congiunta dei capigruppo di Camera e Senato per decidere se procedere a una doppia votazione giornaliera.
01_28_2022_chunk_2.wav	Cresce la tensione tra Stati Uniti e Russia sulla questione Ucraina telefonata tra Zelensky e Biden.
01_28_2022_chunk_2.wav	Per il Presidente americano c'è la possibilità concreta che i russi invadano l'Ucraina nel mese di Febbraio.
01_28_2022_chunk_3.wav	La situazione Covid in Italia rallenta la curva dei contagi, calano recovery e terapie intensive e si discute della possibilità di cambiare il sistema delle fasce a colori delle regioni e anche le regole che riguardano la scuola.
01_28_2022_chunk_3.wav	Intanto è arrivato l'OK dell'EMA alla pillola anti COVID di freezer.
01_28_2022_chunk_4.wav	Tamponi sospetti e Green pass fasulli chiuso un centro analisi in provincia di Trento e Stop a un secondo punto prelievi nel capoluogo Trentino.
01_28_2022_chunk_4.wav	5 le persone indagate.
01_28_2022_chunk_5.wav	Il tennis nella semifinale degli Australian Open, in campo Matteo Berrettini e Rafa Nadal. Il punteggio al momento è di due set a uno per lo spagnolo.
01_28_2022_chunk_6.wav	Ed è tutto grazie per averci seguito. Il tg uno torna alle 8, buona giornata.
...	...

[13] All data in this appendix is protected by Creative Commons Licence CC BY-NC-SA 4.0.

XL-WA: a Gold Evaluation Benchmark for Word Alignment in 14 Language Pairs

Federico Martelli[11,*], Andrei Stefan Bejgu[2,11], Cesare Campagnano[11], Jaka Čibej[14], Rute Costa[10], Apolonija Gantar[14], Jelena Kallas[6], Svetla Koeva[3], Kristina Koppel[6], Simon Krek[9], Margit Langemets[6], Veronika Lipp[5], Sanni Nimb[12], Sussi Olsen[13], Bolette Sandford Pedersen[13], Valeria Quochi[8], Ana Salgado[1,10], László Simon[5], Carole Tiberius[7], Rafael-J Ureña-Ruiz[4] and Roberto Navigli[11]

[1]*Academia das Ciências de Lisboa, Portugal*
[2]*Babelscape, Italy*
[3]*Bulgarian Academy of Sciences, Bulgaria*
[4]*Centro de Estudios de la Real Academia Española, Spain*
[5]*HUN-REN Hungarian Research Centre for Linguistics, Hungary*
[6]*Institute of the Estonian Language, Estonia*
[7]*Instituut voor de Nederlandse Taal, The Netherlands*
[8]*Istituto di Linguistica Computazionale "A.Zampolli", Consiglio Nazionale delle Ricerche, Italy*
[9]*Jožef Stefan Institute, Slovenia*
[10]*NOVA CLUNL, Portugal*
[11]*Sapienza University of Rome, Italy*
[12]*Society for Danish Language and Literature, Denmark*
[13]*University of Copenhagen, Denmark*
[14]*University of Ljubljana, Slovenia*

Abstract

Word alignment plays a crucial role in several Natural Language Processing tasks, such as lexicon injection and cross-lingual label projection. The evaluation of word alignment systems relies heavily on manually-curated datasets, which are not always available, especially in mid- and low-resource languages. In order to address this limitation, we propose XL-WA, a novel entirely manually-curated evaluation benchmark for word alignment covering 14 language pairs. We illustrate the creation process of our benchmark and compare statistical and neural approaches to word alignment in both language-specific and zero-shot settings, thus investigating the ability of state-of-the-art models to generalize on unseen language pairs. We release our new benchmark at: https://github.com/SapienzaNLP/XL-WA.

Keywords

Word Alignment, Deep Learning, Natural Language Processing, Multilinguality

1. Introduction

Word alignment is the computational task of identifying translation correspondences at word and multi-word level between parallel sentences [1, 2]. Historically, word alignment played a crucial role in Statistical Machine Translation [3, 4, SMT]. However, while SMT has been replaced by end-to-end neural architectures which attain considerably higher performances, word alignment – also thanks to novel neural approaches – still plays a crucial role in many other Natural Language Processing (NLP) tasks, such as lexicon injection and, most importantly, cross-lingual annotation projection [5]. For instance, Procopio et al. [6] recently proposed a state-of-the-art approach to cross-lingual label projection based on word alignment which allows high-quality sense-tagged datasets to be produced automatically. Furthermore, word alignment has also been leveraged effectively

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

*Corresponding author.

martelli@diag.uniroma1.it (F. Martelli); bejgu@babelscape.com (A. S. Bejgu); campagnano@di.uniroma1.it (C. Campagnano); Jaka.Cibej@ijs.si (J. Čibej); costamrv@gmail.com (R. Costa); apolonija.gantar@guest.arnes.si (A. Gantar); jelena.kallas@eki.ee (J. Kallas); svetla@dcl.bas.bg (S. Koeva); kristina.koppel@eki.ee (K. Koppel); simon.krek@ijs.si (S. Krek); margit@eki.ee (M. Langemets); lipp.veronika@nytud.hu (V. Lipp); sn@dsl.dk (S. Nimb); saolsen@hum.ku.dk (S. Olsen); vnb282@ku.dk (B. S. Pedersen); valeria.quochi@ilc.cnr.it (V. Quochi); anasalgado@fcsh.unl.pt (A. Salgado); simon.laszlo@nytud.hu (L. Simon); Carole.Tiberius@ivdnt.org (C. Tiberius); rafa@rae.es (R. Ureña-Ruiz); navigli@diag.uniroma1.it (R. Navigli)

to create silver datasets, not only for Word Sense Disambiguation [7, 8, WSD] but also for other semantic tasks, such as Semantic Role Labeling [9, 10, SRL], thereby addressing the knowledge acquisition bottleneck [11], especially when dealing with mid- and low-resource languages.

While, on the one hand, current architectures for word alignment are achieving increasingly better performance, on the other hand, the lack of high-quality manual data in multiple languages significantly limits their potential and scalability. With a view to addressing the aforementioned drawbacks, our contributions are as follows:

1. We propose a fully manually-annotated evaluation benchmark for word alignment with a total of 14 language pairs, each composed of English and one of the following languages: Arabic, Bulgarian, Chinese, Danish, Dutch, Estonian, Hungarian, Italian, Korean, Portuguese, Russian, Slovenian, Spanish and Swedish.

2. We experiment with statistical and neural approaches to word alignment and evaluate them against our newly created benchmark.

3. We demonstrate that the concatenation of our novel datasets can be exploited effectively to train a neural approach that generalizes on unseen languages in a zero-shot setting, thereby addressing the lack of training data in low-resource languages.

2. Related Work

Approaches Initial approaches to word alignment leveraged statistical and heuristic models [12]. Along these lines, several systems were proposed such as HMM [1], GIZA++[1] [13], PGIZA++, MGIZA++ [14] and FastAlign[2] [15]. Subsequently, statistical approaches were gradually substituted by neural counterparts and the advent of Transformer architectures [16] set a new standard in this task [17, 18, 19, 20, 21]. More recently, Procopio et al. [6] proposed a novel neural discriminative model for word alignment based on multilingual BERT [22], capable of significantly reducing the processing time.

Data Over the course of the last few decades, a number of datasets for word alignment, both manual and automatic, have been created, e.g. Czech-English[3] [23], Dutch-English [24], English-Turkish [25], English-Swedish[4] [26], Chinese-English[5] [27]. Interestingly, Graca et al. [28] proposed a collection of small datasets for word alignment in 6 language combinations; each dataset being composed of 100 sentences derived from the Europarl corpus[6] [29]. Among the currently available resources, we highlight the following contributions which we use in our experiments: the English-French and Romanian-English corpora released during the HLT-NAACL-2003 workshop on Building and Using Parallel Texts[7] [30], and the German-English dataset[8] proposed by Vilar et al. [31]. Finally, Neubig [32] presented a Japanese-English dataset[9] obtained by translating Wikipedia pages. However, despite the preceding efforts undertaken in this direction, to the best of our knowledge, no entirely manually-curated evaluation benchmark, which matches XL-WA in both size and language pairs covered, is currently available.

	# of Sentences		# of Alignments	
Lang.	**Dev**	**Test**	**Dev**	**Test**
EN-AR	90	210	1591	3597
EN-BG	105	245	1719	4179
EN-DA	105	245	1841	4136
EN-ES	105	245	1961	4722
EN-ET	105	245	1614	3722
EN-HU	105	245	1580	3781
EN-IT	103	243	1980	4765
EN-KO	90	210	1277	3007
EN-NL	105	245	1886	4490
EN-PT	105	245	1849	4578
EN-RU	90	210	1114	2582
EN-SL	105	245	1942	4537
EN-SV	90	210	1522	3530
EN-ZH	90	210	1724	4135
$\sum$	1393	3253	23600	55761

Table 1
Composition of XL-WA. We report from left to right: the available language combinations, the number of sentences and alignments divided by data split. In our experiments, we use approximately 30% of our data for development so as to obtain a more representative set.

3. XL-WA

To tackle the aforementioned gap, we introduce XL-WA, a novel entirely manually-curated evaluation benchmark

[1]https://github.com/moses-smt/giza-pp
[2]https://github.com/clab/fast_align
[3]https://ufal.mff.cuni.cz/czech-english-manual-word-alignment
[4]https://www.ida.liu.se/divisions/hcs/nlplab/resources/ges/
[5]https://nlp.csai.tsinghua.edu.cn/~ly/systems/TsinghuaAligner/TsinghuaAligner.html
[6]https://www.statmt.org/europarl/
[7]https://web.eecs.umich.edu/~mihalcea/wpt/
[8]https://www-i6.informatik.rwth-aachen.de/goldAlignment/
[9]http://www.phontron.com/kftt

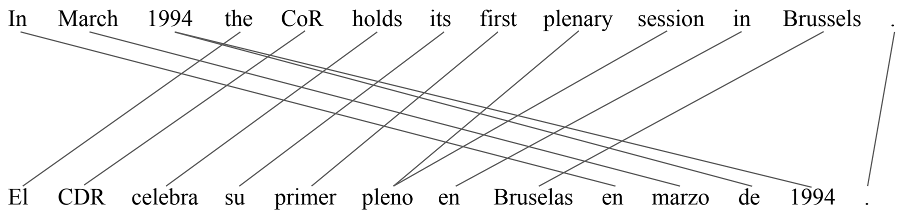


Figure 1: An example of alignment between English and Spanish, derived from the EN-ES dataset in XL-WA.

for word alignment. XL-WA is currently composed of 14 datasets, out of which 9 are parallel. The languages included in XL-WA cover 7 different language families, i.e. Afro-Asiatic, Indo-European (Germanic), Indo-European (Romance), Indo-European (Slavic), Sino-Tibetan, Uralic (Finno-Ugric) and Koreanic.

Importantly, all datasets include English as source language. This choice is motivated by the fact that enabling word alignment from English to multiple target languages is crucial for tasks such as label projection, where the majority of high-quality annotated data whose labels can be propagated is typically available in English.

We show the composition of our dataset in Table 1. Importantly, XL-WA is annotated exclusively by professional mother tongue annotators with a solid academic background and proven experience in linguistic annotation tasks. A detailed description of the format which we adopt is provided in Section 3.2.

3.1. Creation process

In this section, we detail the creation process and illustrate the guidelines adopted during the annotation phase.

The creation of XL-WA can be divided into three steps: i) *automatic extraction* of candidate sentences from a corpus, ii) *manual selection* of sentences satisfying specific linguistic criteria, and iii) *manual alignment.*

In order to obtain a balanced corpus in terms of domains and genres, similarly to the procedure adopted by Martelli et al. [33], we extract our data from WikiMatrix[10] [34], a wide-coverage collection of parallel sentences derived from the Wikipedia[11] corpus using an automatic approach based on multilingual sentence embeddings, covering 1620 different language combinations. First, we consider the WikiMatrix datasets containing English as the source language, and extract the highest number of overlapping source sentences across datasets. To this end, we compute a Boolean matrix $A \in \{0, 1\}^{m \times n}$ where m is the number of English sentences in WikiMatrix and n the number of the target languages other than English covered in XL-WA.

We compute A such that A_{ij} contains: i) 1 if, for the i-th English sentence, a translation into the j-th target language is available, ii) 0 otherwise. We first extract the sentences shared in the highest number of languages. Subsequently, we manually discard sentences which are not well-formatted or contain significant grammatical errors. We then ask annotators to provide the missing translations in order to fill the gaps in our parallel dataset[12]. Finally, we ask our annotators to perform word alignment from scratch.

Guidelines All annotators are required to follow specific annotation guidelines for word alignment inspired by Lambert et al. [35], who provide detailed instructions and suggestions regarding the annotation of datasets for word alignment, including specific cases and exceptions. Importantly, annotators are asked to align source and target words also when these do not share the same part of speech. Furthermore, annotators are required to align complex lexical units such as compounds and multi-word expressions. For instance, given an open compound word c_{en}, e.g. *bus driver* in the English source sentence, translated into Dutch with the compound word c_{nl} *buschauffeur*, each component of c_{en} should be aligned to c_{nl}.

3.2. Alignment Format

We now describe our alignment format and provide an example for the language combination English-Spanish (EN-ES).

We adopt the Pharaoh alignment format [36]. Specifically, we use a Tab-Separated Values (TSV) format, where each row is formatted as follows: `source sentence<tab>target sentence<tab>alignments`. Tokens and alignments are separated by spaces; each alignment is composed of a pair of integers which identify the corresponding positions of source and target tokens, starting from zero. In order to deal with multi-word expressions in which 1:1 alignments are not possible, e.g., due to collocations or idiomatic expressions, we align all components of a given

[10]https://ai.facebook.com/blog/wikimatrix/
[11]https://www.wikipedia.org/

[12]Due to time constraints and the limited availability of professional annotators for specific language combinations, we carry out this step in 9 language combinations only.

multi-word expression in English with all components of the corresponding multi-word expression in the target language.

Below we report an example extracted from the EN-ES dataset:

- **Source:** `In March 1994 the CoR holds its first plenary session in Brussels .`
- **Target:** `El CDR celebra su primer pleno en Bruselas en marzo de 1994 .`
- **Alignments:** `3-0 4-1 5-2 6-3 7-4 8-5 9-5 10-6 11-7 0-8 1-9 2-10 2-11 12-12`

A visual representation of the above example is provided in Figure 1.

3.3. Inter-annotator agreement

Finally, in order to assess the reliability of our manual annotations, we compute the inter-annotator agreement[13]. To this end, we randomly select a sample of approximately 50 sentence pairs in two language combinations, namely EN-DA and EN-IT, and ask new annotators to align these manually. We compute the Cohen's kappa and obtain 0.94 and 0.89 in EN-DA and EN-IT, respectively. Importantly, these results indicate a remarkable level of agreement, which suggests a high degree of annotation consistency across datasets.

4. Experimental Setup

In this section, we illustrate our experimental setup and carry out a performance analysis. To this end, we put forward two different experimental settings. Specifically, we propose a comparison between statistical and neural approaches tested against our novel benchmark in a language-specific setting, i.e. we train and test on the same language pairs (Section 4.1.1). Subsequently, we investigate the behavior of models in a zero-shot setting, thus exploring the ability of state-of-the-art models to deal with languages unseen during the training phase (Section 4.1.2). Finally, we describe the evaluation metrics adopted.

4.1. Settings

We now describe our two experimental settings. Technical details regarding hyperparameters and hardware are reported in Appendix A.

4.1.1. Language-specific setting

Systems In this setting, we experiment with two statistical approaches, namely GIZA++ and FastAlign, and two state-of-the-art neural models, i.e. the SQuAD-style formulation for word alignment[14], which relies on multilingual BERT, proposed by Nagata et al. [20] and the MultiMirror neural word aligner by Procopio et al. [6].

For each language pair, the aforementioned statistical systems are trained on a randomly selected sample of 0.5M parallel sentences concatenated with our test data. Instead, for neural approaches requiring aligned data, which is not available in all our language combinations, we follow Garg et al. [17]. Specifically, we use sentences derived from the aforementioned silver training data, tagged both with GIZA++ and FastAlign, and randomly choose 1,000 sentences with the highest number of overlapping alignments.

Data For this setting, we derive training data from three well-established parallel corpora, namely, Europarl, WikiMatrix and UNPC[15] [37]. Importantly, this choice allows us to cover all language combinations considered. Instead, for validation and evaluation purposes we use the XL-WA datasets whose composition is reported in Table 1. In this case, our goal is to show and analyze the performance achieved by state-of-the-art models on each language pair.

4.1.2. Zero-shot setting

In the zero-shot setting, we experiment with MultiMirror only, since this model shows a reasonable balance between results and processing speed. Specifically, we train MultiMirror on the concatenation of our datasets and evaluate it against unseen language pairs, thus demonstrating the effectiveness of XL-WA when no aligned data is available in a given language combination. In this case, our goal is to determine the extent to which the model is able to generalize on language pairs unseen during training, i.e. EN-DE, EN-FR, EN-JA and EN-RO. The data is split as in Nagata et al. [20].

4.2. Evaluation metrics

As customary in the word alignment task, we adopt the following evaluation metrics: precision, recall and F1. In this work, we do not use the Alignment Error Rate (AER) metric, since previous works argue that AER is unlikely to be a useful metric for word alignment, due to its bias towards precision [4].

[13] Due to time constraints we compute the inter-annotator agreement in two language combinations.

[14] https://github.com/nttcslab-nlp/word_align

[15] https://opus.nlpl.eu/UNPC.php

	GIZA++ [13]			FastAlign [15]			SQuAD BERT [20]			MultiMirror [6]		
Lang.	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
EN-AR	66.0	57.8	61.6	68.6	66.3	67.4	87.3	78.8	82.9	88.3	77.9	82.8
EN-BG	73.6	74.7	74.1	65.7	75.7	70.4	83.3	88.5	85.8	85.5	88.3	86.9
EN-DA	73.6	75.0	74.3	68.7	75.7	72.0	90.6	94.0	92.3	90.8	93.4	92.1
EN-ES	78.1	71.7	74.8	72.3	74.8	73.5	90.7	84.7	87.6	89.5	84.4	86.8
EN-ET	59.6	67.4	63.3	61.1	68.1	64.4	76.3	86.4	81.0	77.2	86.8	81.7
EN-HU	55.9	63.7	59.6	53.3	63.0	57.7	71.4	82.6	76.6	72.4	80.1	76.1
EN-IT	56.3	49.8	52.9	53.7	55.5	54.6	86.9	81.2	84.0	88.2	78.7	83.2
EN-KO	51.2	53.3	52.2	50.7	52.8	51.7	31.4	64.2	42.1	69.5	70.4	69.9
EN-NL	80.3	77.7	79.0	76.2	79.9	78.0	94.9	93.7	94.3	94.2	93.4	93.8
EN-PT	78.4	75.2	76.7	72.6	77.8	75.1	89.2	88.4	88.8	87.9	87.9	87.9
EN-RU	74.0	73.6	73.8	71.8	77.4	74.5	84.1	85.6	84.9	87.6	84.0	85.8
EN-SL	71.0	66.4	68.6	67.9	67.9	67.9	83.5	81.4	82.4	85.4	81.4	83.3
EN-SV	79.4	72.6	75.8	74.7	73.0	73.9	92.1	86.5	89.2	91.5	87.2	89.3
EN-ZH	47.4	41.9	44.5	49.6	48.2	48.9	78.8	70.7	74.5	79.2	71.7	75.3
Avg	67.5	65.8	66.5	64.8	68.3	66.4	81.5	83.3	81.9	84.8	83.3	83.9

Table 2
Comparison between statistical baselines (GIZA++ and FastAlign) and current state-of-the-art approaches (SQuAD BERT and MultiMirror) on our datasets. **P**, **R** and **F1** stand for **P**recision, **R**ecall and **F1**-score, respectively; all the scores are calculated using the Micro average. Note that the neural approaches are trained on silver data generated with the statistical baselines.

Lang.	**P**	**R**	**F1**
EN-DE	89.4	78.5	83.6
EN-FR	94.7	55.7	70.1
EN-JA	79.4	42.5	55.3
EN-RO	86.7	80.7	83.6
Avg	87.5	64.4	73.2

Table 3
Results of MultiMirror trained on all XL-WA datasets and evaluated on unseen data, i.e., in a zero-shot setting. To facilitate analysis and comparison, we keep English as source language.

5. Results

In this section, we discuss the results obtained. As can be seen in Table 2, in the language-specific setting, we observe a remarkable difference between statistical and neural approaches, with the latter outperforming the former by up to 17.5 points in terms of F1 score on average. In this setting, the best results are attained by Nagata et al. [20] in the English-Dutch (EN-NL) combination. Interestingly, we note that even neural models struggle to achieve good results in topic-prominent languages such as Chinese, Hungarian and Korean. In fact, in these languages, both statistical and neural approaches obtain significantly below-average results.

Instead, as far as the zero-shot scenario is concerned, we observe a good generalization capability of MultiMirror when trained on the concatenation of our novel datasets and tested against unseen language pairs, as reported in Table 3. In particular, the language combinations EN-DE and EN-RO attain a remarkable 83.6 F1 score. Importantly, this seems to suggest that the zero-shot paradigm can be employed as a viable approach to compensate effectively for the lack of annotated data in many low-resource languages.

Finally, we investigate the impact of the size of the training data generated with GIZA++ and FastAlign, as described in Section 4.1, on the overall performance achieved by MultiMirror[16]. To this end, we increase the size of the silver training data to 10,000 sentence pairs and compare the results obtained with those achieved in the previous setting where we use 1,000 sentence pairs. As can be seen in Table 4, the greater quantity of data allows us to achieve better results in terms of both precision and F1 score. However, interestingly, when training on 10,000 sentence pairs, MultiMirror reports a slightly inferior performance in terms of recall, with a decrease of 0.3 on average.

[16]As mentioned in Section 4.1.2, we use MultiMirror in this experiment due to a satisfactory trade-off between performance and processing speed.

	MultiMirror 1k sentences			MultiMirror 10k sentences		
Lang.	**P**	**R**	**F1**	**P**	**R**	**F1**
EN-AR	88.3	77.9	82.8	88.9	79.3	83.8
EN-BG	85.5	88.3	86.9	84.7	88.5	86.6
EN-DA	90.8	93.4	92.1	91.3	92.0	91.7
EN-ES	89.5	84.4	86.8	91.8	82.6	86.9
EN-ET	77.2	86.8	81.7	78.0	84.8	81.3
EN-HU	72.4	80.1	76.1	74.2	79.6	76.8
EN-IT	88.2	78.7	83.2	89.5	79.7	84.3
EN-KO	69.5	70.4	69.9	71.1	72.2	71.6
EN-NL	94.2	93.4	93.8	95.9	92.5	94.2
EN-PT	87.9	87.9	87.9	89.0	86.8	87.9
EN-RU	87.6	84.0	85.8	87.5	85.2	86.3
EN-SL	85.4	81.4	83.3	84.4	81.3	82.8
EN-SV	91.5	87.2	89.3	92.3	85.9	89.0
EN-ZH	79.2	71.7	75.3	80.5	72.3	76.2
Avg	84.8	83.3	83.9	85.7	83.0	84.2

Table 4
Comparison between MultiMirror trained on different size silver datasets. P, R and F1 stand for Precision, Recall and F1-score, respectively; all the scores are calculated using the Micro average.

6. Conclusion

In this work, we introduce XL-WA, a novel evaluation benchmark for word alignment in 14 language pairs. We detail the creation process for our novel evaluation suite, as well as our experimental setup in which we compare statistical and neural approaches to word alignment. We investigate the behavior of models in zero-shot scenarios and show that the concatenation of our datasets can be used effectively to align languages unseen during training, thus tackling the paucity or limited availability of data for word alignment in low-resource languages. We release our new benchmark at: https://github.com/SapienzaNLP/XL-WA.

As future work, we intend to investigate the impact of language-specific peculiarities on the overall performance of neural models for word alignment. Furthermore, we plan to increase the language coverage of XL-WA and, importantly, investigate the role played by additional low-resource languages in zero-shot settings. Finally, we aim to explore novel neural approaches to word alignment which can be employed in the field of cross-lingual label projection in order to create multilingual silver training datasets for several Natural Language Understanding tasks, such as WSD, SRL and Semantic Parsing.

Acknowledgments

The authors gratefully acknowledge the support of the ELEXIS project No. 731015 under the European Union's Horizon 2020 research and innovation programme. Furthermore, the authors are sincerely thankful for the support of the Estonian Research Council grant (PRG 1978). Finally, the authors gratefully acknowledge the support of the PNRR MUR project PE0000013-FAIR. This work has been carried out while Andrei Stefan Bejgu was enrolled in the Italian National Doctorate on Artificial Intelligence run by Sapienza University of Rome.

References

[1] S. Vogel, H. Ney, C. Tillmann, Hmm-based word alignment in statistical translation, in: COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics, 1996. URL: https://aclanthology.org/C96-2141.

[2] J. Tiedemann, Word to word alignment strategies, in: COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics, 2004, pp. 212–218. URL: https://aclanthology.org/C04-1031.

[3] F. J. Och, C. Tillmann, H. Ney, Improved alignment models for statistical machine translation, in: 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, 1999. URL: https://aclanthology.org/W99-0604.

[4] A. Fraser, D. Marcu, Measuring word alignment quality for statistical machine translation, Computational Linguistics 33 (2007) 293–303. URL: https://aclanthology.org/J07-3002.

[5] D. Yarowsky, G. Ngai, Inducing multilingual pos taggers and np bracketers via robust projection across aligned corpora, in: Second Meeting of the North American Chapter of the Association for Computational Linguistics, 2001. URL: https://aclanthology.org/N01-1026.

[6] L. Procopio, E. Barba, F. Martelli, R. Navigli, Multimirror: Neural cross-lingual word alignment for multilingual word sense disambiguation, in: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21, 2021, pp. 3915–3921. URL: https://www.ijcai.org/proceedings/2021/0539.pdf.

[7] E. Barba, L. Procopio, N. Campolungo, T. Pasini, R. Navigli, Mulan: Multilingual label propagation for word sense disambiguation, in: Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Arti-

ficial Intelligence, 2021, pp. 3837–3844. URL: https://www.ijcai.org/Proceedings/2020/0531.pdf.

[8] M. Bevilacqua, T. Pasini, A. Raganato, R. Navigli, Recent trends in word sense disambiguation: A survey, in: Z. Zhou (Ed.), Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021, ijcai.org, 2021, pp. 4330–4338. URL: https://doi.org/10.24963/ijcai.2021/593. doi:10.24963/ijcai.2021/593.

[9] S. Padó, M. Lapata, Cross-lingual annotation projection for semantic roles, Journal of Artificial Intelligence Research 36 (2009) 307–340. URL: https://www.jair.org/index.php/jair/article/download/10629/25416.

[10] A. Daza, A. Frank, X-SRL: A parallel cross-lingual semantic role labeling dataset, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 3904–3914. URL: https://aclanthology.org/2020.emnlp-main.321. doi:10.18653/v1/2020.emnlp-main.321.

[11] W. A. Gale, K. W. Church, D. Yarowsky, A method for disambiguating word senses in a large corpus, Computers and the Humanities 26 (1992) 415–439. URL: https://www.jstor.org/stable/30204634.

[12] V. J. Della Pietra, The mathematics of statistical machine translation: Parameter estimation, Using Large Corpora (1994) 223. URL: https://aclanthology.org/J93-2003.pdf.

[13] F. J. Och, H. Ney, A systematic comparison of various statistical alignment models, Computational linguistics 29 (2003) 19–51. URL: https://aclanthology.org/J03-1002.

[14] Q. Gao, S. Vogel, Parallel implementations of word alignment tool, in: Software engineering, testing, and quality assurance for natural language processing, 2008, pp. 49–57. URL: https://aclanthology.org/W08-0509.

[15] C. Dyer, V. Chahuneau, N. A. Smith, A simple, fast, and effective reparameterization of ibm model 2, in: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2013, pp. 644–648. URL: https://aclanthology.org/N13-1073.pdf.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017). URL: http://arxiv.org/abs/1706.03762.

[17] S. Garg, S. Peitz, U. Nallasamy, M. Paulik, Jointly learning to align and translate with transformer models, arXiv preprint arXiv:1909.02074 (2019). URL: https://aclanthology.org/D19-1453.

[18] E. Stengel-Eskin, T.-r. Su, M. Post, B. Van Durme, A discriminative neural model for cross-lingual word alignment, 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), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 910–920. URL: https://aclanthology.org/D19-1084. doi:10.18653/v1/D19-1084.

[19] T. Zenkel, J. Wuebker, J. DeNero, Adding interpretable attention to neural translation models improves word alignment, arXiv preprint arXiv:1901.11359 (2019). URL: http://arxiv.org/abs/1901.11359.

[20] M. Nagata, K. Chousa, M. Nishino, A supervised word alignment method based on cross-language span prediction using multilingual BERT (2020). URL: https://aclanthology.org/2020.emnlp-main.41.

[21] M. J. Sabet, P. Dufter, F. Yvon, H. Schütze, Simalign: High quality word alignments without parallel training data using static and contextualized embeddings, in: Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 1627–1643. URL: https://aclanthology.org/2020.findings-emnlp.147.

[22] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[23] D. Mareček, Automatic Alignment of Tectogrammatical Trees from Czech-English Parallel Corpus, Master's thesis, Charles University, MFF UK, 2008. URL: https://ufal.mff.cuni.cz/pcedt3.0/pubs/Marecek2008_diplomka.pdf.

[24] L. Macken, An annotation scheme and gold standard for dutch-english word alignment, in: 7th conference on International Language Resources and Evaluation (LREC 2010), European Language Resources Association (ELRA), 2010, pp. 3369–3374. URL: http://www.lrec-conf.org/proceedings/lrec2010/pdf/100_Paper.pdf.

[25] M. T. Cakmak, S. Acar, G. Eryigit, Word alignment for english-turkish language pair., in: LREC, 2012, pp. 2177–2180. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/380_Paper.pdf.

[26] M. Holmqvist, L. Ahrenberg, A gold standard for english-swedish word alignment, in: Proceedings

of the 18th Nordic conference of computational linguistics (NODALIDA 2011), 2011, pp. 106–113. URL: https://aclanthology.org/W11-4615.pdf.
[27] Y. Liu, M. Sun, Contrastive unsupervised word alignment with non-local features, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 29, 2015. URL: http://arxiv.org/abs/1410.2082.
[28] J. Graca, J. P. Pardal, L. Coheur, D. Caseiro, Building a golden collection of parallel multi-language word alignment., in: LREC, 2008. URL: http://www.lrec-conf.org/proceedings/lrec2008/pdf/250_paper.pdf.
[29] P. Koehn, Europarl: A parallel corpus for statistical machine translation, in: Proceedings of machine translation summit x: papers, 2005, pp. 79–86. URL: https://aclanthology.org/2005.mtsummit-papers.11.
[30] R. Mihalcea, T. Pedersen, An evaluation exercise for word alignment, in: Proc. of HLT-NAACL, 2003, pp. 1–10. URL: https://aclanthology.org/W03-0301.
[31] D. Vilar, M. Popović, H. Ney, Aer: Do we need to "improve" our alignments?, in: Proc. of Workshop on Spoken Language Translation, 2006. URL: https://aclanthology.org/2006.iwslt-papers.7.pdf.
[32] G. Neubig, The Kyoto free translation task, 2011. URL: http://www.phontron.com/kftt.
[33] F. Martelli, R. Navigli, S. Krek, J. Kallas, P. Gantar, S. Koeva, S. Nimb, B. Sandford Pedersen, S. Olsen, M. Langemets, K. Koppel, T. Üksik, J. Dobrovoljc, R.-J. Ureña-Ruiz, J.-L. Sancho-Sánchez, V. Lipp, T. Váradi, A. Győrffy, S. László, V. Quochi, M. Monachini, F. Frontini, C. Tiberius, R. Tempelaars, R. Costa, A. Salgado, J. Čibej, T. Munda, Designing the ELEXIS parallel sense-annotated dataset in 10 european languages, in: Proceedings of the eLex Conference, 2021. URL: https://elex.link/elex2021/wp-content/uploads/2021/08/eLex_2021_22_pp377-395.pdf.
[34] H. Schwenk, V. Chaudhary, S. Sun, H. Gong, F. Guzmán, Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia, arXiv preprint arXiv:1907.05791 (2019). URL: https://arxiv.org/pdf/1907.05791.pdf.
[35] P. Lambert, A. De Gispert, R. Banchs, J. B. Mariño, Guidelines for word alignment evaluation and manual alignment, Language Resources and Evaluation 39 (2005) 267–285. URL: https://link.springer.com/article/10.1007/s10579-005-4822-5.
[36] P. Koehn, Pharaoh: A beam search decoder for phrase-based statistical machine translation models, in: R. E. Frederking, K. B. Taylor (Eds.), Machine Translation: From Real Users to Research, Springer Berlin Heidelberg, Berlin, Heidelberg, 2004, pp. 115–124. URL: https://aclanthology.org/2004.amta-papers.13/.
[37] M. Ziemski, M. Junczys-Dowmunt, B. Pouliquen, The united nations parallel corpus v1. 0, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 2016, pp. 3530–3534. URL: https://aclanthology.org/L16-1561/.
[38] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, J. Han, On the variance of the adaptive learning rate and beyond, in: International Conference on Learning Representations, 2019. URL: https://iclr.cc/virtual_2020/poster_rkgz2aEKDr.html.

A. Hyperparameters and Hardware

In this appendix, we report the hyperparameters and hardware setup for the experiments described in the paper.

We adopt four approaches with the following hyperparameters:

- We use two statistical approaches, namely GIZA++ [13] and FastAlign [15]. We compile the code downloaded from the original repositories and we run all the experiments on CPU. Neither of the approaches requires any parameter tuning.
- SQuAD mBERT-based model [20], whose code is downloaded from the official repository. We run all the experiments using the default hyperparameters. For the sake of consistency and fairness, we do not tune any hyperparameters and use the optimal ones according to the authors, as specified in their paper. All the experiments run for 2 training epochs with a learning rate of 3×10^{-5} and a batch size of 6. Language-specific experiments run for approximately 20 minutes each. We also experiment with the whole multilingual dataset, which requires 4 hours to complete the training. Inference for the language-specific experiments takes around one minute per language on GPU.
- MultiMirror [6] is an mBERT-based model whose code is obtained from the authors for research purposes. All the experiments run with a patience of 50, using the RAdam [38] optimizer with a learning rate of $1e-05$ and a token batch size of 512. Language-specific experiments run for approximately 10 to 15 minutes each, while the multilingual experiment on the whole XL-WA dataset runs for approximately 1 hour. Inferences time is negligible: a few seconds on CPU for the language-specific data and around one minute for the whole dataset.

All the experiments are conducted on the same hardware, i.e. an Intel Core i7 7800x CPU and NVidia RTX 2080ti GPU with 11GB of VRAM.

Is Change the Only Constant? An Inquiry Into Diachronic Semantic Shifts in Italian and Spanish

Matteo Melis[1], Anastasiia Salova[1] and Roberto Zamparelli[1]

[1]Centre for Mind/Brain Sciences, University of Trento, Rovereto, Italy

Abstract

An increasingly prevalent approach to studying the gradual change of word meanings over time involves using distributional semantics, which is based on neighboring words. This study combines methods from Hamilton et al. (2016) [1] and Uban et al. (2019) [2] to analyze deceptive cognate pairs in historical and contemporary Italian and Spanish corpora. By employing fastText word embeddings and various similarity measures, it aims to investigate the change of word meanings and test two laws of regularity proposed by Hamilton et al. (2016) [1], along with a new hypothesized regularity in language change regarding analogy. The findings show a coherent evolution of deceptive cognates across the two languages. However, no meaningful correlation is found regarding the two aforementioned laws. Nevertheless, the results of the hypothesized regularity offer valuable insight into how the context of word usage shifts along with the word.

Keywords

Diachronic semantics, semantic shifts, distributional semantics, similarity measures, deceptive cognates

1. Introduction

1.1. Background

In recent years, there has been a growing interest in studying the shift of word meanings over time, with word embeddings emerging as a valuable tool for this purpose. Hamilton et al. (2016) [1] conducted research focusing on diachronic word embeddings to uncover specific statistical laws associated with semantic change. They examined the law of conformity, which suggests that words tend to change inversely to their frequency. Additionally, they explored the law of innovation, which proposes that words with greater polysemy tend to undergo semantic changes more frequently, regardless of how often they are used. The findings confirmed the hypothesized statistical laws. The study primarily focused on English, aligning word embeddings from different time periods and measuring semantic similarity using cosine similarity.

Dubossarsky et al. (2017) [3] contested the validity of the reported laws of semantic change based on word representation models. Replicating previous studies, they found that the law of conformity and the law of innovation did not withstand the more rigorous standard. The negative correlation between word frequency and meaning change was weaker than previously claimed, and the positive correlation between polysemy and meaning change was largely dependent on word frequency without independent contribution.

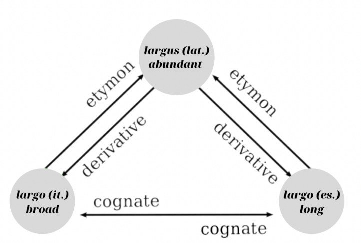


Figure 1: A pair of deceptive cognates in Italian-Spanish, with a shared etymon.

Similarly, to Hamilton et al. (2016) [1], Uban et al. (2019) [2] investigated semantic divergence across languages by examining deceptive cognate sets, which are words with a common origin in different languages. They focused on analyzing modern embeddings to quantify semantic shifts originating from shared etymology, identify false friends (deceptive cognates) in the cognate sets, and measure their score of falseness, namely the dissimilarity between the cognates. The study primarily concentrated on six Romance languages. The authors introduced methodologies such as aligning word embeddings across languages, measuring semantic similarity and divergence between cognate sets, and quantifying the magnitude of semantic changes. Their findings contradict those of Hamilton et al. (2016) [1], who found a negative correlation between frequency and meaning shift. However, they align with their findings regarding the law of innovation.

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
matteo.melis@studenti.unitn.it (M. Melis); anastasiia.salova@studenti.unitn.it (A. Salova); roberto.zamparelli@unitn.it (R. Zamparelli)
https://github.com/matteo-mls (M. Melis); https://github.com/anastasiia-slv (A. Salova)

1.2. Objectives

The primary focus of this study is to investigate the presence of statistical laws governing semantic shifts within the Romance language group, specifically Italian and Spanish. The research questions revolve around exploring the laws of conformity and innovation. It is hypothesized that more frequent words are less likely to undergo semantic shifts, while more polysemous words are more prone to such changes. Additionally, the study introduces a new follow-up analysis on analogy, suggesting that over time periods the meaning of a word which is semantically related to a target (in terms of context-based nearest neighbors), tends to shift in the Euclidean space coherently with the target word.

The study uses distributional semantics as a methodology to explore language change. A crucial part of this research involves analyzing deceptive cognate pairs, which have a similar or the same form in different languages but diverged in meaning over time, unlike true cognates that retain the same meaning. For instance, Figure 1 illustrates how *largo* (broad) in Italian and *largo* (long) in Spanish have diverged in meaning through a semantic shift, despite both words originating from the shared Latin etymon *largo* (abundant). We believe this allows for a robust comparison of semantic changes, especially in related languages, providing illustrative examples and easily interpretable results. Our primary focus is on systematic semantic change that originates from the shared etymon and continues, while also controlling for the random appearance of lexical units in a language. Moreover, this approach would enable cross-language analysis in prospective studies.

Our study aims to expand the current understanding of language change by incorporating cognate comparisons across languages and examining individual changes within specific time periods. To enhance the robustness of our analyses, we introduce various similarity measures.

2. Corpora

2.1. Corpora Selection Criteria

The study uses two different time periods of language usage in its corpora: the 19th and 20th centuries (until 1969) for historical data, and the 21st century for modern data.

To address the size difference between the two datasets, we reduced the modern data to match the historical data's size. This was achieved by counting the number of required tokens and removing the tokens exceeding this number. This allowed for two different training sets for the modern data, enabling comparisons and allowing us to draw conclusions about the minimum amount of data needed for these analyses.

2.1.1. Italian

Four corpora were collected online for this study: Histcorp [4], ChroniclItaly v3.0 [5], Unità corpus [6], and PAISÀ corpus [7]. The first three corpora were merged to form the historical dataset, covering the years 1805-1969, with a total of 545,068,401 tokens. The PAISÀ corpus represented the modern data, containing 1,089,014,748 tokens, while the reduced modern version consisted of 545,106,781 tokens.

2.1.2. Spanish

Similarly, four corpora were collected online for Spanish: Conha19 [8], Impact-es (BVC section) [9], Corpus of Political Speeches [10], and The Large Spanish Corpus [11]. The historical data consists of a merged collection of the first three corpora, covering the period from 1830 to 1969 and containing 204,904,549 tokens. The modern data representation utilizes 'The Large Spanish Corpus' (Wikipedia section), containing 975,251,278 tokens from 2019. Additionally, a reduced version of The Large Spanish Corpus was created, containing 206,900,109 tokens.

2.2. Pre-processing Techniques

The pre-processing for both languages followed the same steps. After collecting the text files for each corpus, we used the NLTK library [12] for tokenization and stopword removal. The files were cleaned by removing URLs, numbers, non-letters, multiple empty spaces, and set to lowercase. For Spanish, diacritic marks were replaced using unicodedata. The spaCy library [13], with its reported accuracy of 0.96 for Spanish and 0.97 for Italian, was employed for lemmatization, and the files were merged into a representative single file for each historical period and language.

2.3. Cognate Dataset

We used an existing resource: an automatically generated multilingual lexicon of false friends [14]. Following the logic that cognate pairs are considered false friends if a word in the second language is closer in meaning to the original word in the shared semantic space than its cognate in that language, a falseness score is provided.

For instance, given the cognate pair (*imbarazzata, embarazada*), where *imbarazzata* (embarassed) is a word in Italian and *embarazada* (pregnant) is a word in Spanish, if there is a word x in Spanish such that for any word w in Spanish the distance (imbarazzata, x) is less than the distance (imbarazzata, w), then the pair is considered a deceptive cognates pair. Since the Spanish word *avergonzada* (embarassed) exists, the pair (*imbarazzata, embarazada*) constitutes a set of false friends, and their arithmetic difference is the score of falseness, which ranges from 0 to 1. It is lower for false friends that are closer in meaning and higher for more distant false friends.

Given this, we decided to extract the 156 deceptive cognate pairs with a falseness value higher than 0.25. This step was taken to ensure the accuracy of the dataset and account for its limitations in the unsupervised data collection method.

3. Methodology

Methodologically, the study can be divided into the following steps[1]:

3.1. FastText Word Embeddings Retrieval

We trained six fastText models [15] in an unsupervised regime using the six corpora that we obtained and prepared. For each model, we employed the skip-gram algorithm, set the vector dimension to 100, and trained for 5 epochs. These parameters are considered default, and as indicated by Mikolov et al. (2013) [16], the algorithm has been found to work well with small datasets. This resulted in three models for each language, trained on historical data, modern data, and modern reduced data, respectively. This produced a total of 6 different vector spaces.

3.2. Embeddings Overview with RSA

In order to obtain a comprehensive overview of the vector spaces and as the initial step of our analysis, we computed Representational Similarity Analysis (RSA) between dissimilarity matrices of 156 deceptive cognate words from the dataset by Uban and Dinu (2020) [14]. These matrices were created by extracting vectors for specific cognates from the common vector spaces obtained in the previous step. The aim was to assess general similarity patterns within the word embeddings. Based on the results thus obtained we chose to exclusively use the model trained on the full modern data and discard the one trained on the reduced modern data to ensure higher-quality word embeddings in later steps. Detailed results of this analysis will be discussed later.

[1]All the code can be found at https://github.com/matteo-mls/diachronic-semantic-shift.

3.3. K-Nearest Neighbors Retrieval Using a Similarity Measure

To obtain more qualitative data, the fastText library [15] was used to retrieve embeddings closest to the target cognate in Euclidean space. The retrieval process utilized the K-Nearest Neighbors (K-NN) function, where the cosine similarity measure was employed to compare two vectors. The number of nearest neighbors to retrieve (k) was predetermined and set to 5, 10, 20, and 50 for comparative analysis purposes.

3.4. Semantic Shift Calculation within Each Language

After retrieving the nearest neighbors for cognates, we calculated the overlap between the sets of nearest neighbors in each language. This overlap was measured using the Jaccard similarity coefficient, which determines the similarity between two sets. The semantic shift was then computed as the difference in overlap between the sets of nearest neighbors over time. Finally, by using the Pearson correlation measure to assess the shifts between the two languages, Italian and Spanish, we were able to draw conclusions.

3.5. Word Frequency and Semantic Divergence Analysis

For the frequency analysis, we followed the following steps:

1. We applied Procrustes alignment [17] to the two vector spaces (historical to modern for each language) to ensure that similar vectors represented the same concepts across different embedding spaces. This alignment was necessary as the embeddings were trained on different corpora in different languages.
2. We calculated the cosine similarity for the cognates in different time periods.
3. We counted the occurrences of each cognate word from both the historical and modern corpora in Italian and Spanish.
4. We normalized the occurrences of cognate words by dividing each value by the maximum value, which is the sum of all values. This normalization resulted in a total of 1, effectively replacing the actual frequency values.

Using the NumPy library [18], we computed the correlation coefficient and linear regression coefficients of the frequency and semantic shift across time. In this analysis, we incorporated polysemy covariance, considering the correlation between polysemy and frequency.

3.6. Word Polysemy and Semantic Divergence Analysis

After conducting the frequency and semantic divergence analysis, we proceeded to measure the polysemy of words. To accomplish this, we utilized the WordNet library [19], specifically leveraging the functionality provided by the "nltk.corpus.wordnet" module. Polysemy was quantified as the number of synsets associated with a word in WordNet, following the methodology described by Uban et al. (2019) [2].

Subsequently, we investigated the correlation between the cosine similarity over time, which indicates the degree of semantic shifting, and the number of meanings a word can have according to WordNet. In this analysis, we took into account the co-variance with frequency, similarly to our previous approach.

3.7. Word Analogy and Semantic Divergence Analysis

In addition to the previous analyses, we further examined how the cosine similarity changes over time for the K-Nearest Neighbors (K-NN) that exhibit overlap between the two different time periods. For each cognate word, we employed a K-NN approach with varying values of K (5, 10, 20, 50). We examined the overlapping nearest neighbors (NN) in both the historical and modern lists of NN. For each overlapping NN, we calculated the cosine similarity and measured the difference in the shift, determining whether the NN moved closer to or further from the target cognate word.

By calculating the ratio of positive (closer) or negative (further) shifts, we could now assess the coherence (the consistency of neighbors' movement relative to the target cognate) of the shift in the K-NN of that specific target cognate word. To identify significant coherent shifts, we set a threshold (>0.75). This threshold was chosen to be substantially higher than chance, ensuring a rigorous approach. If this ratio is crossed, it implies a major coherent shift in the K-NN of the target cognate word.

In carrying out this analysis for all the cognates in the list we removed those that had 0 or 1 NN, since they do not provide informative results.

4. Results

4.1. Representational Similarity Analysis

As shown in the Appendix A (Figure 4), the reduced Italian modern embedding space exhibits a lower correlation compared to the complete Italian modern embedding space, with a difference of 0.0322 (a). This suggests that the improved embedding obtained by using more data in unsupervised word embedding contributes to this outcome. Furthermore, when comparing the reduced historical Spanish embedding space with the modern embedding space, a difference of 0.0956 is observed (b). Therefore, while the results for Italian remain consistent between the full and reduced spaces, reducing the Spanish modern space to match the historical space produces different outcomes compared to using the full modern space. Given the choice between data quality and balance, we have opted for better data quality by discarding the models trained with reduced datasets.

4.2. Calculation of Semantic Shifts

4.2.1. Within-Language Comparison: K-NN with Jaccard Distance

In reference to the selection of K Nearest Neighbors (KNN) values at 5, 10, 20, and 50, the obtained results are presented in the tables provided in the Appendices B and C (Tables 3 to 10). These tables display the average number of overlapping nearest neighbors in the cognate list, the ratio of overlapping nearest neighbors considering the extracted KNN, and the Jaccard distance. Please refer to the Appendix for a detailed representation of these values.

4.2.2. Inter-Language Comparison: K-NN with Jaccard Distance

The values in Appendix D (Tables 11 and 12) represent dissimilarity scores, specifically semantic shifts, calculated using the Jaccard distance (1-Jaccard index). The Pearson correlation score of 0.999 indicates a strong correlation between the shifts for Italian and Spanish as the particular K value increases. Overall, the scores show compatible semantic shifts. However, in this analysis, we can only infer the magnitude of the shifts and not the patterns, which will be explored in later analyses.

4.3. Law of Conformity

Figure 2 (upper) showcases the correlation results for the law of conformity in both Italian and Spanish. The obtained correlation coefficients demonstrate a moderate positive correlation, with a coefficient of 0.408 for Italian and 0.470 for Spanish. However, when accounting for the influence of polysemy through partial correlation analysis, the coefficients decrease to 0.261 for Italian and 0.3 for Spanish. These values are generally considered weak. While these findings provide only weak evidence for the law of conformity, they are at least consistent in their trend with the results reported by Hamilton et al. (2016) [1].

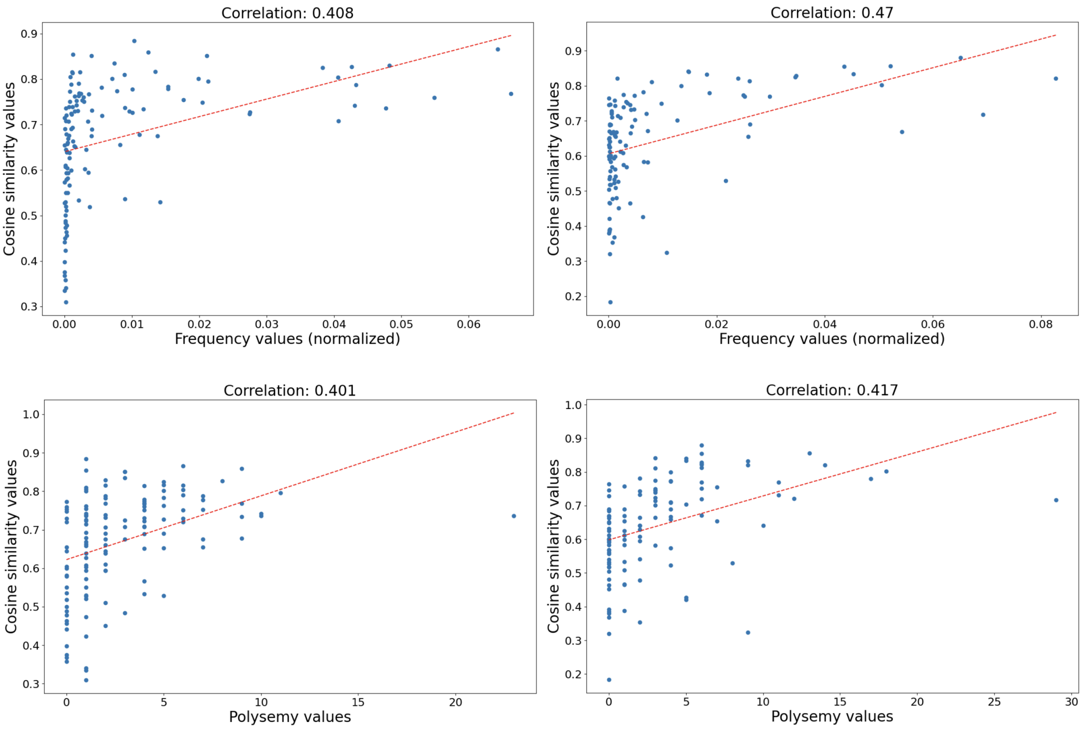


Figure 2: Law of conformity (upper) and law of innovation (lower) visualized for Italian (left) and Spanish (right).

4.4. Law of Innovation

Conversely, in our study the results for the law of innovation (more polysemy = greater shift), depicted in Figure 2 (lower), differ from those reported by Hamilton et al. (2016) [1] and Uban et al. (2019) [2]. While we observed a moderate positive trend, similar to that of the law of conformity, with correlation scores of 0.401 for Italian and 0.417 for Spanish, the partial correlation, which accounts for the frequency compound, reveals weaker values of 0.249 for Italian and 0.188 for Spanish. These findings suggest that the data does not provide strong support for the existence of the law of innovation in Romance languages. However, due to the weak partial correlations observed, it is challenging to draw definitive conclusions.

4.5. Law of Analogy

One trend that emerges from our study is that semantically related words (as indicated by contextual nearest neighbors) tend to shift coherently closer or farther to the target word. Table 1 and Table 2 provide supporting evidence for this observation: as the number of nearest neighbors (K-NNs) increases, the ratio of coherent shifts tends to decrease. This aligns with the intuition that with more K-NNs, the distances between the neighbors and their target cognate increase, leading to less consistent shifts. To provide a visual representation, Figure 3 displays an example visualization for a single cognate pair.

Table 1
Analogy analysis for Italian

K-NN	N° of Cognates	Coherent shift	%
5	53	36	67.92
10	83	51	61.45
20	104	52	50
50	121	64	52.89

Table 2
Analogy analysis for Spanish

K-NN	N° of Cognates	Coherent shift	%
5	48	35	72.92
10	67	46	68.66
20	88	59	67.04
50	102	68	63.72

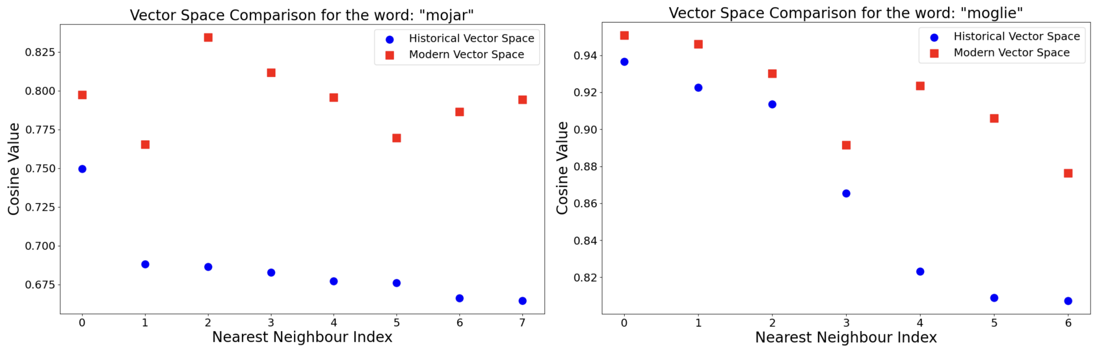


Figure 3: An example of the analysis of the law of analogy visualized for Italian (right) and Spanish (left) using the cognate pair ”mojar”/”moglie”.

5. Discussion

The hypothesized regularity regarding analogy, a follow-up analysis in this study, has provided intriguing insights into semantic shifts. However, it is important to note that further research into this topic is necessary to validate and expand upon these initial findings.

On the other hand, the analyses conducted in this study do not yield definitive results supporting the statistical laws of semantic shifts. Firstly, the RSA evaluation of the embedding spaces revealed that the scarcity of data significantly impacted the quality of the embeddings. Furthermore, while the law of conformity agrees with previous literature in a general trend, such as Hamilton et al. (2016) [1], our study identified a contrasting trend for the law of innovation. This discrepancy in findings may be attributed to the limitation of our study, namely the scarcity of data resulting from the use of relatively short time periods.

An additional factor is the relatively short temporal distance between the historic (as recent as 1969) and the modern corpora. Increasing this span is likely to lead to greater shifts, but also to greater data sparsity. Last but not least, the alignment technique employed for matching the embedding spaces could have contributed to the divergent outcomes in the analysis of the law of conformity and the law of innovation.

It is noteworthy that both the laws of conformity and the law of innovation conform to the findings of Dubossarsky et al, (2017) [3]. Their study revealed that the suggested positive correlation between meaning change and polysemy was primarily influenced by word frequency, and the correlation between word frequency and meaning change is indeed weaker. Here, after conducting partial correlation analysis, a weak correlation was observed. Furthermore, we noticed a high compatibility between frequency and polysemy, indicating an inherent dependence, despite our efforts to disentangle them using partial correlation.

Utilizing the fastText model, known for its improved performance on non-English languages, and preprocessing freely available data, the results still highlight poor quality embeddings. This underscores the need for ongoing research and development of word embedding models, alongside the creation of larger, well-curated diachronic corpora. Improving data quality and quantity can enhance the accuracy and reliability of future studies in the field.

It is important to note that due to the limitations of the embeddings used in this study, the shifts observed in the inter-language Jaccard distance analysis are relatively small and close to each other. This leads to an extremely high correlation coefficient between the languages being analyzed, which should be interpreted with caution.

In addition to the aforementioned directions, other potential areas of research include expanding further in time and broader in the scope of languages. For instance, this could involve going beyond the Romance or even the Indo-European language family to conduct a more comprehensive investigation into language change.

Acknowledgements

We would like to express our gratitude to Dr. Raffaella Bernardi for her support and feedback throughout this project, which has been helpful in shaping our research. We also appreciate her encouragement regarding the conference submission.

We also extend our gratitude to Dr. Lorella Viola for her generous assistance in providing a portion of the corpus used in our analysis.

References

[1] W. L. Hamilton, J. Leskovec, D. Jurafsky, Diachronic word embeddings reveal statistical laws of semantic change, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 1489–1501. URL: https://aclanthology.org/P16-1141. doi:10.18653/v1/P16-1141.

[2] A. S. Uban, A. M. Ciobanu, L. P. Dinu, Studying laws of semantic divergence across languages using cognate sets, in: Proceedings of the Workshop, 2019, pp. 161–166. doi:10.18653/v1/W19-4720.

[3] H. Dubossarsky, D. Weinshall, E. Grossman, Outta control: Laws of semantic change and inherent biases in word representation models, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017. URL: https://aclanthology.org/D17-1118. doi:10.18653/v1/D17-1118.

[4] E. Pettersson, B. Megyesi, The histcorp collection of historical corpora and resources, in: Digital Humanities in the Nordic Countries Conference, 2018. URL: https://api.semanticscholar.org/CorpusID:19243754.

[5] L. Viola, A. M. Fiscarelli, Chroniclitaly 3.0. a deep-learning, contextually enriched digital heritage collection of italian immigrant newspapers published in the usa 1898-1936, in: Proceedings of the Conference, 2021. doi:10.5281/zenodo.4596345.

[6] P. Basile, A. Caputo, T. Caselli, P. Cassotti, R. Varvara, A diachronic italian corpus based on "l'unità", in: CLiC-it 2020 Italian Conference on Computational Linguistics 2020, volume 2769, CEUR Workshop Proceedings (CEUR-WS.org), 2020.

[7] V. Lyding, E. Stemle, C. Borghetti, M. Brunello, S. Castagnoli, F. Dell'Orletta, H. Dittmann, A. Lenci, V. Pirrelli, PAISÀ corpus of italian web text, 2013. URL: http://hdl.handle.net/20.500.12124/3, eurac Research CLARIN Centre.

[8] U. Henny-Krahmer, Corpus de novelas hispanoamericanas del siglo xix (conha19) version 1.0.1, in: Proceedings of the Conference, 2021. doi:10.5281/zenodo.4781947.

[9] F. Sánchez-Martínez, I. Martínez-Sempere, X. Ivars-Ribes, R. C. Carrasco, An open diachronic corpus of historical spanish, Language Resources and Evaluation 47 (2013) 1327–1342.

[10] E. Álvarez-Mellado, A corpus of Spanish political speeches from 1937 to 2019, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 928–932. URL: https://aclanthology.org/2020.lrec-1.116.

[11] J. Cañete, Compilation of large spanish unannotated corpora, Zenodo, 2019.

[12] S. Bird, E. Klein, E. Loper, Natural language processing with Python: analyzing text with the natural language toolkit, O'Reilly Media, Inc., 2009.

[13] M. Honnibal, I. Montani, spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing, 2017.

[14] A. S. Uban, L. P. Dinu, Automatically building a multilingual lexicon of false friends with no supervision, in: International Conference on Language Resources and Evaluation, 2020. URL: https://api.semanticscholar.org/CorpusID:218973843.

[15] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Transactions of the Association for Computational Linguistics 5 (2017) 135–146.

[16] T. Mikolov, K. Chen, G. S. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: International Conference on Learning Representations, 2013. URL: https://api.semanticscholar.org/CorpusID:5959482.

[17] J. Gower, Generalized procrustes analysis, Psychometrika 40 (1975) 33–51. URL: https://EconPapers.repec.org/RePEc:spr:psycho:v:40:y:1975:i:1:p:33-51.

[18] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. Fernández del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, T. E. Oliphant, Array programming with NumPy, Nature 585 (2020) 357–362. doi:10.1038/s41586-020-2649-2.

[19] C. Fellbaum, WordNet: An Electronic Lexical Database, Bradford Books, 1998. URL: https://mitpress.mit.edu/9780262561167/.

Appendix

A. RSA Correlation of Italian and Spanish

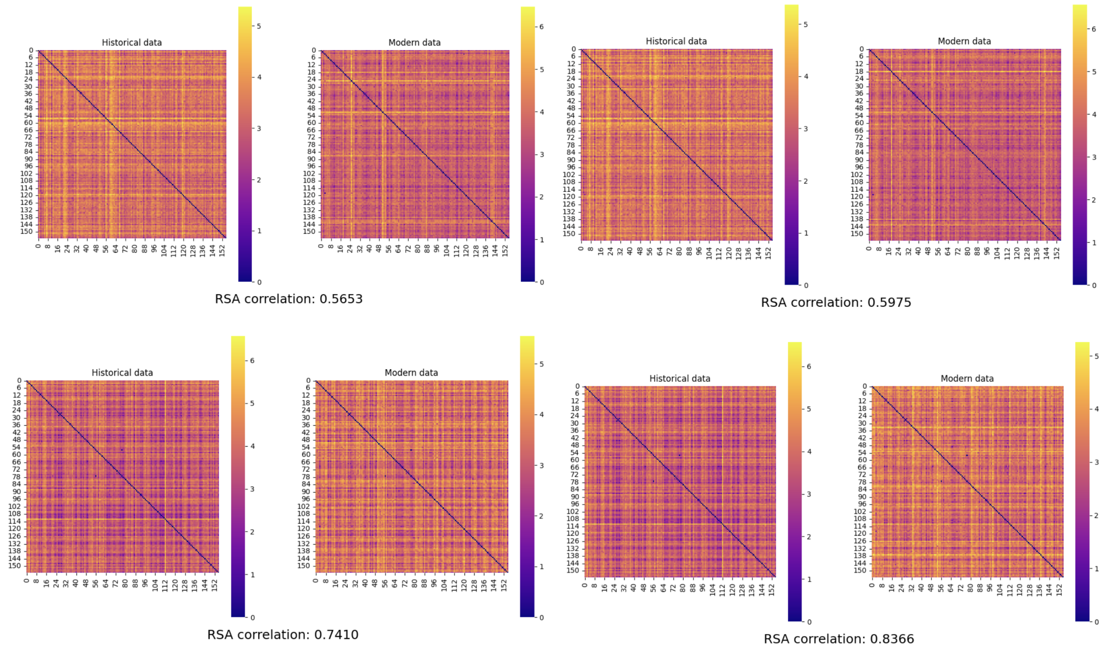


Figure 4: RSA correlation of Italian (upper) and Spanish (lower). Comparison between historical and modern (left) and historical and modern reduced (right)

B. Italian K-Nearest Neighbour

Table 3
Italian, K = 5 NN Overlap

Italian - K = 5	Word	N° of overlap
1	Fiaccola	4
2	Maggio	4
3	Ottimo	4
...	...	...
94	Verso	1
95	Voluta	1
96	Vendicare	1
Average	171/96	**1.7812**
Jaccard Distance	1 - J	**0.7833**

Table 4
Italian, K = 10 NN Overlap

Italian - K = 10	Word	N° of overlap
1	Maggio	9
2	Cardinale	7
3	Mantello	6
...	...	...
112	Servo	1
113	Via	1
114	Vigile	1
Average	294/114	**2.5789**
Jaccard Distance	1 - J	**0.8520**

Table 5
Italian, K = 20 NN Overlap

Italian - K = 20	Word	N° of overlap
1	Maggio	12
2	Cardinale	11
3	Decima	10
...	...	...
124	Venia	1
125	Tonno	1
126	Servo	1
Average	488/126	**3.8730**
Jaccard Distance	1 - J	**0.8928**

Table 6
Italian, K = 50 NN Overlap

Italian - K = 50	Word	N° of overlap
1	Impadronirsi	27
2	Cardinale	26
3	Giudicare	25
...	...	...
132	Oste	1
133	Sotto	1
134	Vado	1
Average	1005/134	**7.5000**
Jaccard Distance	1 - J	**0.9189**

C. Spanish K-Nearest Neighbour

Table 7
Spanish, K = 5 NN Overlap

Spanish - K = 5	Word	N° of overlap
1	Ardor	4
2	Diverso	4
3	Imaginario	4
...	...	...
82	Derrame	1
83	Verso	1
84	Vivir	1
Average	153/84	**1.8214**
Jaccard Distance	1 - J	**0.7773**

Table 8
Spanish, K = 10 NN Overlap

Spanish - K = 10	Word	N° of overlap
1	Cometer	6
2	Importar	6
3	Muerto	6
...	...	...
101	Derrame	1
102	Verso	1
103	Decir	1
Average	261/103	**2.5340**
Jaccard Distance	1 - J	**0.8549**

Table 9
Spanish, K = 20 NN Overlap

Spanish - K = 20	Word	N° of overlap
1	Cometer	13
2	Prender	11
3	Importar	10
...	...	...
114	Ensear	1
115	Tata	1
116	Tenia	1
Average	448/116	**3.8620**
Jaccard Distance	1 - J	**0.8931**

Table 10
Spanish, K = 50 NN Overlap

Spanish - K = 50	Word	N° of overlap
1	Cometer	25
2	Importar	20
3	Jurar	19
...	...	...
124	Patrón	1
125	Radio	1
126	Tenia	1
Average	920/126	**7.3016**
Jaccard Distance	1 - J	**0.9212**

D. Cosine Similarity

Table 11
Italian, Cosine Similarity

ITALIAN	Word	N° of overlap
1	Moglie	0.8845485
2	Ancora	0.8659243
3	Finire	0.8588681
...	...	...
146	Venia	0.3331086
147	Così	0.31215054
148	Caudale	0.30994532
8 cognates not found	**Average**	**0.6655**

Table 12
Spanish, Cosine Similarity

SPANISH	Word	N° of overlap
1	Querer	0.88015264
2	Decir	0.8567517
3	Pueblo	0.8563638
...	...	...
124	Radio	0.3236405
125	Das	0.3200544
126	Craso	0.18371347
30 cognates not found	**Average**	**0.6470**

Building structured synthetic datasets: The case of Blackbird Language Matrices (BLMs)

Paola Merlo, Giuseppe Samo, Vivi Nastase* and Chunyang Jiang

Department of Linguistics, University of Geneva, Switzerland

Abstract

Our goal is to investigate, ultimately to enhance, to what degree existing LLM learn disentangled rule-based, compositional linguistic representations. We take the approach of developing curated synthetic data on a large scale, with specific properties, and using them to study sentence representations built using pretrained language models. Inspired by IQ tests, we develop a new multiple-choice task. Finding a solution to this task requires a system detecting complex linguistic patterns and paradigms in text representations. We present formal specifications of this task, illustrate it with two problems and present their benchmarking results.

Il nostro obiettivo è indagare, allo scopo di migliorare, quanto gli LLM esistenti apprendano rappresentazioni linguistiche composte, basate su regole districate. Il nostro approccio consiste nello sviluppare dati sintetici curati su larga scala, con proprietà specifiche, e nell'utilizzarli per studiare le rappresentazioni di frasi costruite con modelli linguistici pre-addestrati. Ispirandoci ai test del QI, abbiamo sviluppato un nuovo task a scelta multipla. Trovare la soluzione di questo task richiede che il sistema individui schemi e paradigmi linguistici complessi nelle rappresentazioni testuali. Presentiamo le specifiche formali di questo task, lo illustriamo con due problemi e presentiamo i risultati del benchmarking.

Keywords

synthetic structured data, formal definitions of grammatical phenomena, diagnostic studies of deep learning models

1. Introduction

Current consensus about LLM, and NNs in general, is that to reach better, possibly human-like, abilities, we need to develop tasks and data that help us understand their current generalisation abilities and help us train or tune them towards more complex and compositional skills.

Humans are good generalizers. A large body of literature has demonstrated that the human mind is predisposed to generate rules from data and combine these rules, in ways that have been argued to be distinct from the patterns of activation of neural networks [1, 2, 3]. One possible approach to develop more robust methods, then, is to drive the network to learn disentangled decompositions of complex observations and learn underlying regularities [4].

Let's look at an illustrative example of what complex decomposition of covert rules would be necessary. Consider complex argument structure relations in the lexicon: for example, the *Spray/load alternation* in English, shown in (1).

This alternation applies to verbs such as *spray, paint, spread, fill, stuff* and *load*, that describe covering surfaces or filling volumes [5, 6]. They occur in two subcategorisation frames, related to each other in a regular way: the object of the preposition *with* is the subject of the *onto* frame, while the object of the *onto* prepositional phrase is the subject of the *with* frame.

(1)	John	loaded	the truck	with hay.
	AGENT		LOCATIVE	THEME
	John	loaded	hay	onto the truck.
	AGENT		THEME	LOCATIVE

To learn the structure of such a complex alternation automatically, a neural network must be able to identify the elements manipulated by the alternation, and their relevant attributes, and recognize the operations that manipulate these objects, across more than one sentence.

To study what factors lead to learning more disentangled linguistic representations —representations that reflect the underlying linguistic rules of grammar— we take the approach of developing curated synthetic data on a large scale, building diagnostic models from pretrained representations of these data and investigating the models' behaviour. To this end, we develop a new linguistic task, inspired by the IQ test RPM (Raven 1938), which we call Blackbird Language Matrices (BLMs). BLMs define a prediction task to learn complex linguistic patterns and paradigms [7, 8].

In this paper, we present precise formal specifications of the BLM task, illustrate it with the instantiations of two BLM problems and their benchmarking results. This

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

*Corresponding author.

Paola.Merlo@unige.ch (P. Merlo); Giuseppe.Samo@unige.ch (G. Samo); vivi.a.nastase@gmail.com (V. Nastase); chunyang.jiang42@gmail.com (C. Jiang)

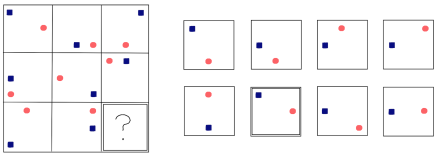


Figure 1: Example of progressive matrice in the visual world. The multiple-choice task is to determine the missing element in a visual pattern. The matrix is constructed according to two rules (see text for explanation). Identifying these rules leads to the correct answer (marked by double edges).

shows that the general formalism can be used to generate datasets with the same format, and similar specification. Expanding the covered phenomena can thus be done systematically, allowing for studies that combine or work across multiple phenomena and languages [9]. We believe this task takes us closer to investigations of human linguistic intelligence.

2. RPMs and BLMs

Raven's progressive matrices are IQ tests consisting of a sequence of images, called the *context*, connected in a logical sequence by underlying generative rules [10]. The task is to determine the missing element in this visual sequence, the *answer*. An instance is shown in Figure 1: given a matrix (left), choose the last element of the matrix from given options. The matrices are built according to generative rules that span the whole sequence of stimuli and the answers are constructed to be similar enough that the solution can be found only if the rules are identified correctly. For example in Figure 1, the matrix is constructed according to two rules: Rule 1: row-wise, from left to right, the red dot moves one place clockwise each time. Rule 2: column-wise, from top to bottom, the blue square moves one place anticlockwise each time. Identifying these rules leads to the correct answer, the only cell that continues the generative rules correctly.

A similar task has been developed called Blackbird Language Matrices (BLMs) [7, 8, 11] for linguistic problems, as given in Figure 2, which illustrates the template of a BLM agreement matrix. As can be seen, the agreement rules are implicitly expressed by patterns in the sentence, with or without intervening attractor elements, and alternate in a combinatorial pattern across the sentences (shown in colour), so that only one answer concludes the sequence.

3. Formal Specifications of BLMs

We define here the new Blackbird's Language Matrices (BLMs) task and data format.

Context				
1	NP-sing	PP1-sing		VP-sing
2	NP-plur	PP1-sing		VP-plur
3	NP-sing	PP1-plur		VP-sing
4	NP-plur	PP1-plur		VP-plur
5	NP-sing	PP1-sing	PP2-sing	VP-sing
6	NP-plur	PP1-sing	PP2-sing	VP-plur
7	NP-sing	PP1-plur	PP2-sing	VP-sing
8	???			

Answer		
1	NP-sing PP1-sing et NP2 VP-sing	Coord
2	**NP-plur PP1-plur PP2-sing VP-plur**	correct
3	NP-sing PP-sing VP-sing	WNA
4	NP-sing PP1-sing PP2-sing VP-plur	AE
5	NP-plur PP1-sing PP1-sing VP-plur	WN1
6	NP-plur PP1-plur PP2-plur VP-plur	WN2

Figure 2: BLM instances for verb-subject agreement, with two attractors. WNA= wrong number of attractors; AE= agreement error; WN1= wrong nr. for 1st attractor noun (N1); WN2= wrong nr. for 2nd attractor noun (N2).

Subject-verb agreement
E: the subject and the verb match
in agreement features.
I: occurs independently of distance
between subject and verb.

Spray/Load alternation:
E: the object of alternant 1 becomes PP(P) in alt. 2.
the object of alternant 2 becomes PP(P) in alt. 1.
the preposition with of alternant 1 becomes
the preposition onto in alternant 2.
I: Expression of thematic roles and argument structure:
Object of Alt1 is Locative, PP of Alt1 is Instrumental
Object of Alt2 is Theme, PP of Alt2 is Locative

Figure 3: Examples of formal definitions of E and I for two LPs. The lexical expression of preposition P is considered an attribute.

Definition Let a 4-tuple (LP, C, W, w_c) be given, where LP is the definition of the linguistic grammatical phenomenon, C is the corresponding context matrices, W is the answer set, and w_c is the correct item of W.

The BLM *task* can be defined as the instruction:

find $(w_c \in W)$ given C.

A BLM *problem* (LP, C, W, Aug) is an instance of a BLM task, where Aug is the augmentation method for the matrices. We describe all components in the next sections.

3.1. Defining the linguistic phenomenon

The first step in the definition of the problem consists in formally defining the linguistic grammatical phenomenon as a paradigm.

DEFINITION Let a linguistic phenomenon LP be given. LP is exhaustively defined by a grammar $G_{LP} = (O, A, E, I, L)$ s.t.

O is the set of objects

A is the set of attributes of the objects in O

E is the set of external observed rules

I is the set of unobserved internal rules

L is the lexicon of objects in O, attributes in A, and operators in $E \cup I$.

For example, as shown in the example Figure 3, in the subject-verb agreement phenomenon, the agreement rule is the primary production in E, while the fact that agreement can occur independently of the distance of the elements expresses the fact that agreement applies to structural representations, a rule in I. Sometimes, but not always, I acts as a confusing factor.

Rules are triples of objects (shown in red), attributes (in green) and operations (in blue). Objects are usually phrases, attributes are usually morpho-syntactic properties of the phrases and operations are typical grammatical operations: feature match, movement (becomes), lexical substitution (changes).

3.2. Defining the matrices

DEFINITION A BLM *matrix* is a tuple=(S, R, T) s.t.

S is the shape of the matrix

R are the relational operators that connect the items of the matrix

T is the set of items of the matrix.

SHAPE $S(n, l)$ is the shape of the matrix, which consists of n items and each item can be at most of length l.

The length of the items can vary. The items can be sentences or elements in a morphological paradigm. The choice of n depends on how many items need to be shown to illustrate the paradigm and on whether the illustration is exhaustive or sampled.

For example, a matrix of size eight is exhaustive for an agreement problem with three noun phrases and a two-way number differentiation (singular, plural), but can only present a sample of the information for the *spray/load* alternation.

SUBJECT-VERB NUMBER AGREEMENT
Violation of E: wrong subject-verb agreement
Violation of I: wrong agreement on N2 or N3
Violation of R: wrong number of attractors

SPRAY/LOAD ALTERNATION
Violation of E: Wrong lexical choice of preposition
Violation of I: Subject of active voice is not Agent
Violation of R: Wrong number of arguments

Figure 4: Example answer set.

RELATIONAL OPERATIONS

Connective sequential operations, such as alternation or progression are chosen. Their purpose is to transform a list of items (sentences or words) into a predictable sequence that connects all the items.

The values of R, so far, are alternations or progression. They could also be conjunction, disjunction, exclusive OR and other logical or graded operators.

ALTERNATION applies to a given (o, a) pair and loops over all the values of a with a given increment defined over the items of the matrix. For example, the grammatical feature number is binary in certain languages. So, ALTERNATION $(o = NP; a_i = (s, p); i = 1, 2, 3, ...)$. This is used to create different alternations of (o, a) in the sentence, which in the subject-verb agreement BLM is used to show independence from linear distance.

PROGRESSION applies to countable attributes or ordinal attributes, for example, existence. So, one can have 1,2,...,n of a given object o. Progression can also apply to $position$ or to graded properties such as $length$.

ITEMS The items T are defined by $G_{LP} = (O, A, E, I, L)$ and they are drawn from the set $\mathcal{T}$.

The matrix is created by sampling (o, a, r) triples. The ways in which $r \in R$ can apply to a given (o, a) pair has to be predefined, as it is not entirely context-free.

3.3. Defining the answer set

The answer set W consists of a set of items like those in C. One item in W, w_c, is the correct answer to complete the sequence defined by C. The other items are the contrastive set. They are items that violate G_{LP}, the rules of construction of the context matrix C, either in the primary rules E, in the auxiliary rules I, or in the matrix operators R.

Sometimes they are built almost automatically, sometimes by hand. The cardinality of the answer set is determined by how many facets of the linguistic phenomenon need to be shown to have been learned.

3.4. Augmenting the matrices

Different levels of lexical and structural complexity can be obtained by changing the lexical items (completely or partially), in a given matrix.

Definition An *augmented BLM* is a quadruple (S, R, T, Aug).

S is the shape of the matrix, R are the relational operations that connect the T items of the matrix.

Aug is a set of operations defined to augment the cardinality of $\mathcal{T}$, while keeping S and R constant. Aug is defined by controlled manipulations of Os and As in $\mathcal{T}$ to collect similar elements.

We augment the sentence set $\mathcal{T}$ by modifying the noun phrases of the items in T. We generate alternatives with a language model choosing among the top n, within an acceptability margin from the original sentence. The margin is set with a variable-size window and collects the top 10 alternative noun phrases. The acceptability of the resulting sentences is validated manually. In the next sections, we illustrate the data for two BLM problems, and baseline benchmarking results.

4. Example of two BLM problems

The creation of structured datasets can be a challenging task, depending on the type of linguistic problem being investigated, the available linguistic resources, and the size of the lexical factors involved in the problem. Figure 5 summarizes the pipeline that needs to be followed for the whole process: identifying the data for the linguistic phenomenon under investigation, developing a lexical set seed of lexical items for creating context and answer sets for the BLMS, which are then combined to construct desired context templates and answer sets. From the linguistic phenomenon to the creation of the lexical set seed, various approaches can be pursued based on the type of linguistic phenomenon being investigated. This choice might depend on whether the phenomenon has already been extensively studied in experimental linguistics, the scale of the lexical components involved in the linguistic phenomenon, and the available resources in the target language. We then employ a fill-mask task with transformers to automatically generate additional, plausible constituents for the desired structures.

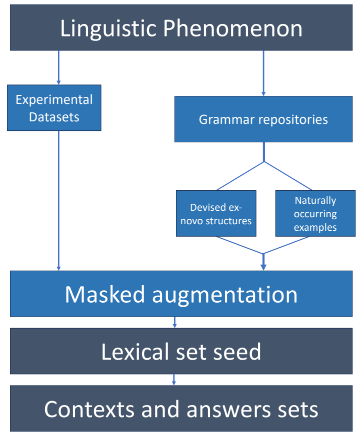


Figure 5: Pipeline for the automatic creation of structured datasets

Figure 9 in the appendix shows an example of the first steps of the process for the BLM-AgrF dataset.

4.1. BLM-AgrF – subject-verb agreement in French

In BLM-AgrF [11], a BLM problem for subject-verb agreement consists of a context set of seven sentences that share the subject-verb agreement phenomenon, but differ in other aspects – e.g. number of intervening noun phrases between the subject and the verb, called attractors because they can interfere with the agreement, different grammatical numbers for these attractors, and different clause structures. Each context is paired with a set of candidate answers. The answer sets contain minimally contrastive examples built by corrupting some of the generating rules. This helps investigate the kind of information and structure learned, by error analysis. An example template is illustrated in Figure 2, and an actual example in Figure 6.

The dataset comprises three subsets, of increasing lexical complexity. Type I data is generated based on manually provided seeds (Franck et al. 2002), illustrated in Figure 9, and a template that captures the rules mentioned above. Type II data is generated based on Type I data, by introducing lexical variation with the aid of a transformer, by generating alternatives for masked nouns. Type III data is generated by combining sentences from different instances from the Type II data, while maintaining the structure of the sequence. The structural variations alter the distance and relative depth of the subject and verb and produce a variety of conditions. The different levels of lexical variation will allow us to investigate the impact of lexical variation on the ability of a system to detect grammatical patterns. We include complete instances –

	Context			
1	Il vaso	con il fiore		si è rotto.
2	I vasi	con il fiore		si sono rotti.
3	Il vaso	con i fiori		si è rotto.
4	I vasi	con i fiori		si sono rotti.
5	Il vaso	con il fiore	del giardino	si è rotto.
6	I vasi	con il fiore	del giardino	si sono rotti.
7	Il vaso	con i fiori	del giardino	si è rotto.
8	???			

	Answer set	
1	Il vaso con i fiori e il giardino si è rotto.	coord
2	**I vasi con i fiori del giardino si sono rotti.**	correct
3	Il vaso con il fiore si è rotto.	WNA
4	Il vaso con il fiore del giardino si sono rotti.	AE
5	I vasi con i fiori del giardino si sono rotti.	WN1
6	I vasi con i fiori dei giardini si sono rotti.	WN2

Figure 6: BLM instances for verb-subject agreement, with 2 attractors (*fiore* 'flower', *giardino* 'garden'), with candidate answer set. WNA=wrong number of attractors, AE=agreement error, WN1=wrong nr. for 1st attractor noun (N1), WN2=wrong nr. for 2nd attractor noun (N2)

	Context			
1	NP-Agent	Verb	NP-Theme	PP-Loc
2	NP-Theme	VerbPass	PP-Loc	
3	NP-Agent	Verb	NP-Loc	PP-Theme
4	NP-Loc	VerbPass	PP-Theme	
5	NP-Agent	Verb	NP-Theme	
6	NP-Agent	Verb	PP-Theme	
7	NP-Agent	Verb	NP-Loc	
8	???			

	Answers	
1	**NP-Agent Verb PP-Loc**	correct
2	NP-Agent Verb *PP-Loc	WPrep
3	NP-Agent Verb *[NP-Theme PP-Loc] PP-Loc	WPP
4	NP-Agent Verb NP-Loc because PP-Theme	Adv
5	NP-Agent Verb NP-Loc *PP_{Loc}-Theme	WT
6	*NP-Loc Verb PP-Loc	WS

Figure 7: BLM context template 1 and answers for the spray/load alternation. * = locus of the rule corruption, angled brackets = syntactic embedding. WP= Wrong Preposition; WPP=Wrong Prepositional Phrase; Adv= Adverbial; WT=Wrong Theme; WS=Wrong Subject.

in French – in Appendix A.

4.2. BLM-s/lE.v0 – spray/load verb alternations in English

In the BLM-s/lE problem developed to exhibit the SPRAY/LOAD alternation (discussed in the introduction), each sentence can be described in terms of one distribution-of-three-values rule, governing the semantic roles (Agent, Theme, Locative), and two distribution-of-two-values rules governing syntactic types (nominal phrase NP vs. prepositional phrases PP) and the mood of the verb, whether active (`Verb`) or passive (`VerbPass`). We created two templates, targeting the syntax-semantic mapping of the arguments.

In the contrastive answer set, the target sentence is to be chosen from a set of candidates that exhibit minimal differences. The semantic-syntactic mapping of the alternation can be decomposed into a set of smaller patterns that describe the sentences in the alternation and that can be violated to construct incorrect answers. Different subsets of patterns can be used to develop different answer sets.

A variation of this dataset, presented in [12] uses an answer set that change the position of the agent in an active sentence, the type of phrases following the verb, embedding the PP in an NP, and changes in prepositions that introduce different types of arguments.

The dataset presented here, changes subpatterns that govern the correct learning of the syntactic form of the sentence (WRONG THEME, WRONG SUBJECT, WRONG PP in Template 1; SWAPLOCAGENT, NOAGENT, SWAPTHEMEAGENT, REPEAT in Template 2), and others that govern the proper lexical selection (WRONG PREPOSITION, ADVERBIAL).

Different answer sets, that focus on different rule subpatterns, will allow for detailed investigations of the type of information that is more easily or more difficult to detect, and to determine principles of designing the answer set for the most informative phenomenon investigation.

Like the BLM-AgrF dataset, the BLM-s/lEv.0 dataset presents lexically varied versions, Type I, II, and III, of increasing variability. The structure of the context and answer set of one alternant is presented in Figure 7. Figure 11 for the other template and relevant lexical examples of both templates are given in appendix B.

5. Benchmarking systems

Our goal is to investigate – and ultimately use this knowledge to enhance – textual representations built using pretrained large language models. To determine whether such representations encode linguistic rules, and to what degree they are compositional, we use BLM tasks that provide data generated using specific rules, and baseline systems that should be capable of detecting the patterns that encode the relevant rules for the targeted phenomena in the distributed continuous sentence representations. We choose a FFNN and a CNN as our baselines. The FFNN should be able to discover patterns distributed throughout a sentence, and throughout a sequence of sentences, while the CNN could discover localized patterns – both in the sentence and the sequence.

As presented above, a BLM problem instance consists of a context and an answer set. The context is a sequence of 7 sentences, and the answer set is a set of 6 sentences, one of which is a correct continuation of the input sequence. All sentences are encoded using BERT [13] – as the embedding of the [CLS] token on the last layer of the model. We used the pretrained "BERT-base-multilingual-cased" model[1]. The sentence representations are combined in different ways, depending on the baseline system – a FFNN or a CNN – used.

The input to the FFNN is the concatenation of sentence embeddings in the BLM instance context, as a vector of size 7 * 768. This input is processed through 3 fully connected layers, which progressively compress the input size $(7 * 768 \xrightarrow{layer1} 3.5 * 768 \xrightarrow{layer2} 3.5 * 768 \xrightarrow{layer3} 768)$ to obtain the size of a sentence representation. The FFNN's interconnected layers enable it to capture patterns that are distributed throughout the entire input vector.

The input to the CNN is the stacked sentence embeddings in the BLM instance context, as a (7 x 768) array. This input undergoes three consecutive layers of 2-dimensional convolutions, where each convolutional layer uses a kernel size of (3x3) and a stride of 1, without dilation. The resulting output from the convolutional process is then passed through a fully connected layer, which compresses it to the size of the sentence representation (768). By using a kernel size of (3x3), stride=1, and no dilation, this configuration emphasizes the detection of localized patterns within the sentence sequence array.

The output of both systems is a vector representing a sentence embedding. This is compared to the sentence representations in the answer set, and the one with the highest score is considered the correct answer. Details are included in appendix C.

6. Results

Previous published work from our group and current ongoing work has benchmarked the problems generated by these datasets and analysed the errors, with interesting results [11, 14].

We report here the novel results on the BLM-s/lE dataset, which are qualitatively similar to those reported for BLM-AgrF [11], thus confirming some general trends. Figure 8 shows the results. The top panels shows results using all the data, the bottom panel shows results using only data sizes that match type I data size, hence smaller data sizes for type II and type III.

Globally, the results are very good. But interesting differences emerge if we vary data sizes. If we train on all data, more lexically-varied data (types II and III) give better results, but if we train on equally sized datasets, we see an improvement of the results in training on type

[1]https://huggingface.co/bert-base-multilingual-cased

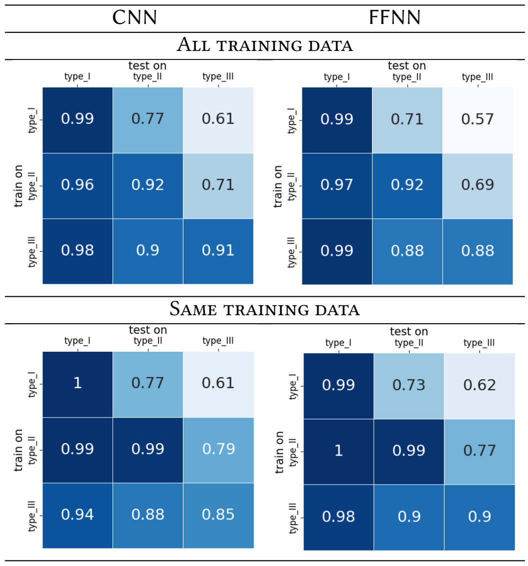


Figure 8: F1 results (averages over 5 runs) for alternant one.

II data whether testing on type II or III. It appears, then, that on smaller datasets, the template patterns is perhaps better learnt in type II, while still retaining the notion of lexical variation that is lacking in type I.

Inspecting the increase in performance with different training data sizes, shown in Figure 12 in the appendix, it is confirmed that learning is very fast and plateaus already with a few thousand examples for all train-test combinations, with the exception of training on Type I and testing on type III, which is clearly too difficult.

Both baseline systems lead to good results despite the variations in the input – structural and lexical – that superficially obfuscate these phenomena, and the near-miss incorrect answers, confirming that the phenomena we target are encoded in the sentence representations. Because of their different architectures and the type of patterns they discover, the high performance of both systems indicates that relevant patterns for our two targeted phenomena are localized in BERT sentence embeddings. Further steps can take advantage of the structured way the data was constructed to attempt to disentangle the various generative rules and additional factors in the inputs.

7. Related Work

Previous work has focussed on understanding the automatic learning of verb alternations in terms of syntactic and semantic properties of the verbs and their argument structures [15]. These properties have been explored in relation to their representation in LLMs, across various dimensions of performance for different models [16, 17]. In particular, [17] suggest that LLMs with contextual embeddings encode linguistic information on verb alternation classes, at both the word and sentence levels. In their work, [17] build upon [16] observations and highlight the superior performance of one transformer Electra [18] compared to other large language models.

The automatic generation of RPM-like matrices, whether in vision or in language, is technically challenging. In computer vision, several formalisms have been proposed ([19] formulate RPMs with first-order logic; [20] propose Procedurally Generated Matrices (PGM) datasets through relation-object-attribute triple instantiations; [21] use the Attributed Stochastic Image Grammar (A-SIG [22]). Structured synthetic datasets have been mostly developed to study issues of generalisation and disentaglement, in vision [23], with full-fledged experimentation and for language in a preliminary, nonRPM -like dataset, consisting of simple examples containing a few morphological markings [24]. The simplicity of the sentences does not provide a sufficiently realistic challenge from a linguistic point of view. Very recent work has started exploring the picture-naming potential of language to solve problems in vision [25].

8. Conclusions

In this paper, we have presented the new BLM task, provided its formal specifications and illustrated the first instances of BLM problems and benchmarking results with baseline architectures. Current work is developing new dedicated architectures based on Variation Autoencoders [14] and developing new BLM problems. Future work lies in further automating the data development pipeline, to make the creation on BLM data sets also accessible to less computationally-oriented linguists and investigating the structure and nature of the information encoded in the learned inner representations.

Acknowledgments

We gratefully acknowledge the partial support of this work by the Swiss National Science Foundation, through grants #51NF40_180888 (NCCR Evolving Language) and SNF Advanced grant TMAG-1_209426 to PM.

References

[1] Y. Lakretz, G. Kruszewski, T. Desbordes, D. Hupkes, S. Dehaene, M. Baroni, The emergence of number and syntax units in LSTM language models, arXiv preprint arXiv:1903.07435 (2019).

[2] Y. Lakretz, D. Hupkes, A. Vergallito, M. Marelli, M. Baroni, S. Dehaene, Mechanisms for handling nested dependencies in neural-network language models and humans., Cognition (2021). doi:2021 10.1016/j.cognition.2021.104699.

[3] M. Sablé-Meyer, J. Fagot, S. Caparos, T. van Kerkoerle, M. Amalric, S. Dehaene, Sensitivity to geometric shape regularity in humans and baboons: A putative signature of human singularity, Proceedings of the National Academy of Sciences 118 (2021). doi:10.1073/pnas.2023123118.

[4] Y. Bengio, A. Courville, P. Vincent, Representation learning: A review and new perspectives, IEEE transactions on pattern analysis and machine intelligence 35 (2013) 1798–1828.

[5] B. Levin, English verb classes and alternations: A preliminary investigation, University of Chicago Press, 1993.

[6] J. Beavers, The spray/load alternation, The Wiley Blackwell Companion to Syntax, Second Edition (2017) 1–31.

[7] P. Merlo, A. An, M. A. Rodriguez, Blackbird's language matrices (BLMs): a new benchmark to investigate disentangled generalisation in neural networks, ArXiv: cs.CL 2205.10866 (2022). URL: https://arxiv.org/abs/2205.10866. doi:10.48550/ARXIV.2205.10866.

[8] P. Merlo, Blackbird language matrices (BLM), a new task for rule-like generalization in neural networks: Motivations and formal specifications, ArXiv cs.CL 2306.11444 (2023). URL: https://doi.org/10.48550/arXiv.2306.11444. doi:10.48550/arXiv.2306.11444.

[9] P. Merlo, C. Jiang, G. Samo, V. Nastase, Blackbird Language Matrices Tasks for Generalization, in: GenBench: The first workshop on (benchmarking) generalisation in NLP, Singapore, 2023.

[10] J. C. Raven, Standardization of progressive matrices, British Journal of Medical Psychology 19 (1938) 137–150.

[11] A. An, C. Jiang, M. A. Rodriguez, V. Nastase, P. Merlo, BLM-AgrF: A new French benchmark to investigate generalization of agreement in neural networks, in: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 1363–1374. URL: https://aclanthology.org/2023.eacl-main.99.

[12] G. Samo, V. Nastase, C. Jiang, P. Merlo, BLM-s/lE: A structured dataset of English spray-load verb alternations for testing generalization in LLMs, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singa-

pore, 2023.
[13] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.
[14] V. Nastase, P. Merlo, Grammatical information in BERT sentence embeddings as two-dimensional arrays, in: Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023), Toronto, Canada, 2023.
[15] O. Majewska, A. Korhonen, Verb classification across languages, Annual Review of Linguistics 9 (2023) 313–333. doi:10.1146/annurev-linguistics-030521-043632.
[16] K. Kann, A. Warstadt, A. Williams, S. R. Bowman, Verb argument structure alternations in word and sentence embeddings, in: Proceedings of the Society for Computation in Linguistics (SCiL) 2019, 2019, pp. 287–297. URL: https://aclanthology.org/W19-0129. doi:10.7275/q5js-4y86.
[17] D. Yi, J. Bruno, J. Han, P. Zukerman, S. Steinert-Threlkeld, Probing for understanding of English verb classes and alternations in large pre-trained language models, in: Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 2022, pp. 142–152. URL: https://aclanthology.org/2022.blackboxnlp-1.12.
[18] K. Clark, M.-T. Luong, Q. V. Le, C. D. Manning, Electra: Pre- training text encoders as discriminators rather than generators, in: ICLR, 2020, pp. 1–18.
[19] K. Wang, Z. Su, Automatic generation of raven's progressive matrices, in: Q. Yang, M. J. Wooldridge (Eds.), Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015, AAAI Press, 2015, pp. 903–909. URL: http://ijcai.org/Abstract/15/132.
[20] D. Barrett, F. Hill, A. Santoro, A. Morcos, T. Lillicrap, Measuring abstract reasoning in neural networks, in: J. G. Dy, A. Krause (Eds.), Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018, volume 80 of *Proceedings of Machine Learning Research*, PMLR, 2018, pp. 4477–4486. URL: http://proceedings.mlr.press/v80/santoro18a.html.
[21] C. Zhang, F. Gao, B. Jia, Y. Zhu, S. Zhu, RAVEN: A dataset for relational and analogical visual reasoning, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, 2019, pp. 5317–5327. URL: http://openaccess.thecvf.com/content_CVPR_2019/html/Zhang_RAVEN_A_Dataset_for_Relational_and_Analogical_Visual_REasoNing_CVPR_2019_paper.html. doi:10.1109/CVPR.2019.00546.
[22] S. Zhu, D. Mumford, A stochastic grammar of images, Found. Trends Comput. Graph. Vis. 2 (2006) 259–362. URL: https://doi.org/10.1561/0600000018. doi:10.1561/0600000018.
[23] S. van Steenkiste, F. Locatello, J. Schmidhuber, O. Bachem, Are disentangled representations helpful for abstract visual reasoning?, in: NeurIPS 2019, 2020.
[24] A. M'Charrak, Deep Learning for Natural Language Processing (NLP) using Variational Autoencoders (VAE), Master's thesis, ETH Switzerland, 2018. URL: https://pub.tik.ee.ethz.ch/students/2018-FS/MA-2018-22.pdf.
[25] X. Hu, S. Storks, R. Lewis, J. Chai, In-context analogical reasoning with pre-trained language models, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1953–1969. URL: https://aclanthology.org/2023.acl-long.109.
[26] J. Franck, G. Vigliocco, J. Nicol, Subject-verb agreement errors in french and english: The role of syntactic hierarchy, Language and cognitive processes 17 (2002) 371–404.

A. BLM-AgrF problem

Example subject NPs from [26]
L'ordinateur avec le programme de l'experience
The computer with the program of the experiments
Manually expanded and completed sentences
L'ordinateur avec le programme de l'experience est en panne.
The computer with the program of the experiments is down.
Jean suppose que l'ordinateur avec le programme de l'experience est en panne.
Jean thinks that the computer with the program of the experiments is down.
L'ordinateur avec le programme dont Jean se servait est en panne.
The computer with the program that John was using is down.

A seed for language matrix generation				
Jean suppose que	*l'ordinateur*	*avec le programme*	*de l'experience*	*est en panne*
Jean thinks that	the computer	with the program	of the experiment	is down
	les ordinateurs	*avec les programmes*		*sont en panne*
	the computers	with the programs		are down

Figure 9: Examples from [26], manually completed and expanded sentences based on these examples, and seeds made based on these sentences for the subject-verb agreement BLM-AgrF dataset that contain all number variations for the nouns and the verb.

Contexts	
Example	Translation
1 La conférence sur l'histoire a commencé plus tard que prévu.	*The talk on history has started later than expected.*
2 Les responsables du droit vont démissionner.	*Those responsible for the right will resign.*
3 L' exposition avec les peintures a rencontré un grand succès.	*The show with the paintings has met with great success.*
4 Les menaces de les réformes inquiètent les médecins.	*The threats of reforms worry the doctors.*
5 Le trousseau avec la clé de la cellule repose sur l'étagère.	*The bunch of keys of the cell sits on the shelf.*
6 Les études sur l'effet de la drogue apparaîtront bientôt.	*The studies on the effect of the drug will appear soon.*
7 La menace des réformes dans l' école inquiète les médecins.	*The threat of reforms in the school worries the doctors.*
Answers	
Example	Translation
1 Les nappes sur les tables et le banquet brillent au soleil.	*The tablecloths on the table and the console shine in the sun.*
2 **Les copines des propriétaires de la villa dormaient sur la plage.**	*The friends of the owners of the villa were sleeping on the beach.*
3 Les avocats des assassins vont revenir.	*The laywers of the murderers will come back.*
4 Les avocats des assassins du village va revenir.	*The lawyers of the murderers of the village will come back.*
5 La visite aux palais de l' artisanat approchent.	*The visit of the palace of the crafts is approaching.*
6 Les ordinateurs avec le programme des expériences sont en panne.	*The computers with the program of the experiments are broken.*

Figure 10: Example of lexically varied contexts for the main clause contexts for the subject-verb agreement BLM-AgrF dataset. Correct answer in bold.

B. BLM-s/lE

	CONTEXT		
1	NP-Agent	Verb	NP-Theme
2	NP-Agent	Verb	PP-Theme
3	NP-Agent	Verb	NP-Loc
4	NP-Agent	Verb	PP-Loc
5	NP-Agent	Verb	NP-Theme PP-Loc
6	NP-Theme	VerbPass	PP-Loc
7	NP-Agent	Verb	NP-Loc PP-Theme
8	???		

	ANSWERS	
1	**NP-Loc VerbPass PP-Theme**	CORRECT
2	NP-Loc Verb *NP-Agent PP-Theme	SLA
3	NP-Loc VerbPass *PP_{Loc}-Theme PP-Loc	WP
4	NP-Loc VerbPass *PP-Loc	REPEAT
5	*NP-Theme Verb NP-Loc	NOAGENT
6	NP-Theme Verb *NP-Agent PP-Loc	STA

Figure 11: BLM context template 2 and answers for the spray/load alternation. * = locus of the rule corruption, angled brackets = syntactic embedding. WP=WrongPrep, SLA=SwapLocAgent, STA=SwapThemeAgent.

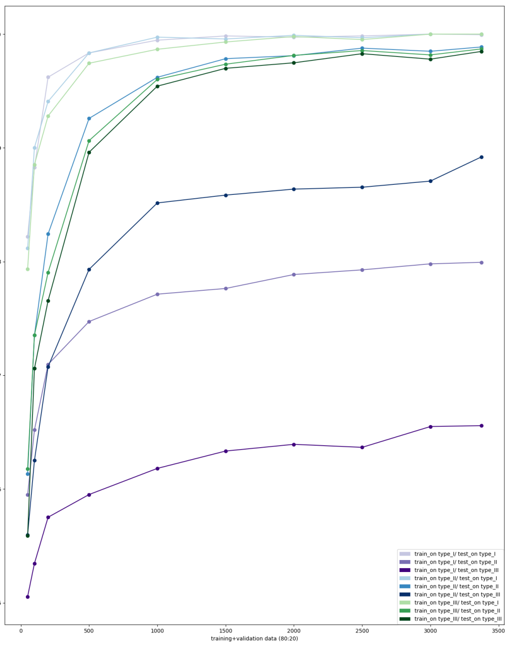


Figure 12: Effect of training data size on the BLM-S/lE datasets.

TEMPLATE 1, TYPE I
CONTEXT
The crew sprayed some water into a plastic container.
Some water was sprayed into a plastic container.
The crew sprayed a plastic container with some water.
A plastic container was sprayed with some water.
The crew sprayed some water.
The crew sprayed with some water.
The crew sprayed a plastic container.
???
ANSWERS
The crew sprayed into a plastic container.
The crew sprayed under a plastic container.
The crew sprayed some water from rivers into a plastic container.
The crew sprayed a plastic container because of some water.
The crew sprayed a plastic container under some water.
A plastic container sprayed into a plastic container.

Figure 13: Template 1 -Type I context and answer set.

TEMPLATE 1, TYPE II
CONTEXT
Katrina sprayed some liquid into the windshield.
Some of the chemicals were sprayed into the windshield.
I sprayed the wall with some water.
The windshield was sprayed with some water.
The artist sprayed some of the material.
The artist sprayed with some of the chemicals.
Someone sprayed the sink.
???
ANSWERS
The artist sprayed onto the wall.
The crew sprayed under the bathroom.
Katrina sprayed some of the chemicals for the refinery into the windshield.
Someone sprayed the bathroom because of some water.
Katrina sprayed a plastic container under some liquid.
The bathroom sprayed into the bathroom.

Figure 14: Template 1 -Type II context and answer set.

TEMPLATE 1, TYPE III
CONTEXT
The crew sprayed some water into a plastic container.
Heavier material was sewed onto the straps.
Water spurts the room with dirt.
It was pumped with steam.
Archaeologists clean the filter monthly and swash some tank water.
Scientists promote a strategy to seed with sulfur.
The volunteers swash the biomedia.
???
ANSWERS
The Egyptians sow on the soil.
Archaeologists clean the filter monthly and swash under the sink.
The crew sprayed some of the paint from vegetables into the windshield.
The man smeared the walls because of the flour.
The waitress sprinkles eggs under sugar.
The windows strung over the windows.

Figure 15: Template 1 -Type III context and answer set.

C. Architectural Specifications

All systems used a learning rate of 0.001 and Adam optimizer, and batch size 100. The training was done for 120 epochs. The experiments were run on an HP PAIR Workstation Z4 G4 MT, with435 an Intel Xeon W-2255 processor, 64G RAM, and a MSI GeForce RTX 3090 VENTUS 3X OC 24G GDDR6X GPU.

We tested BERT sentence embeddings with baseline CNN and FFNN baseline architectures. [13]. The sentence embeddings are the encoding of the [CLS] token on the last layer of the model.

The FFNN receives the input as a concatenation of sentence embeddings in a sequence, with a size of 7 * 768. This input is then processed through 3 fully connected layers, which progressively compress the input size (7 * 768 $\rightarrow layer1$ 3.5 * 768 $\rightarrow layer2$ 3.5 * 768 $\rightarrow layer3$ 768) to obtain the size of a sentence representation. The FFNN's interconnected layers enable it to capture patterns that are distributed throughout the entire input vector.

The CNN takes as input an array of embeddings with a size of (7 x 768). This input undergoes three consecutive layers of 2-dimensional convolutions, where each convolutional layer uses a kernel size of (3x3) and a stride of 1, without dilation. The resulting output from the convolutional process is then passed through a fully connected layer, which compresses it to the size of the sentence representation (768). By using a kernel size of (3x3), stride=1, and no dilation, this configuration emphasizes the detection of localized patterns within the sentence sequence array.

Both networks produce the same output, which is a vector representing the sentence embedding of the correct answer. The objective of learning is to maximize the probability of selecting the correct answer from a set of candidate answers. To achieve this, we employ the max-margin loss function, considering that the incorrect answers in the answer set are intentionally designed to have minimal differences from the correct answer. This loss function combines the distances between the predicted answer and both the correct and incorrect answers. Initially, we calculate a score for each candidate answer's embedding e_i in the answer set $\mathcal{A}$ with respect to the predicted sentence embedding e_{pred}. This score is determined by the cosine of the angle between the respective vectors:

$$score(e_i, e_{pred}) = cos(e_i, e_{pred})$$

The loss function incorporates the max-margin concept, taking into account the difference between the score of the correct answer e_c and each of the incorrect answers e_i:

$$loss_a = \sum_{e_i} [1 - score(e_c, e_{pred}) + score(e_i, e_{pred})]^+$$

During prediction, the answer with the highest score value from the candidate set is selected as the correct answer.

Simplifying Administrative Texts for Italian L2 Readers with Controllable Transformers Models: A Data-driven Approach

Martina Miliani[1], Fernando Alva-Manchego[2] and Alessandro Lenci[1]

[1]Dipartimento di Filologia, Letteratura e Linguistica, Università di Pisa, Italy
[2]School of Computer Science and Informatics, Cardiff University, UK

Abstract
This paper presents a data-driven study focused on the automatic simplification of in-domain texts for specific target readers, which is "controlled" through data collected from behavioral analysis. We used these data to create Admin-It-L2, a parallel corpus of original-simplified sentences in the Italian administrative language, in which simplifications are aimed at Italian L2 speakers. Then, we used this corpus to test controllable models for text simplification based on Transformers.

Although we obtained a high SARI score of 39.24, we show that this datum alone is not fully reliable in evaluating text simplification.

Keywords
Automatic Text Simplification, Transformers, Italian L2, Italian Bureaucratic Language

1. Introduction

Reading a text in a language that is easy to understand becomes necessary if the information it conveys is crucial to people's daily lives. This is the case for acts and communications of Public Administrations (PA). In Italy, despite institutions repeatedly encouraging clear writing [1], PA texts have not abandoned the stylistic figures of "bureaucratese" [2]. Suffering the most from the effects of these linguistic choices are speakers with language disparity. In particular, Italian L2 speakers deal with the Italian bureaucracy, for example, to obtain a visa for which A2-level language certification is required. Yet, comprehension of administrative texts is attested only with the C2 level, according to the CEFR.

This paper presents a data-driven approach to simplification, which is "controlled" through data collected from behavioral analysis. We use these data to create Admin-It-L2, a parallel corpus of original-simplified sentences in the administrative language, in which simplifications are aimed at Italian L2 speakers. Then, we employ this corpus to test computational models. The adopted approach exploits the potential of Transformer-based multilingual models, which permit the use of training data also in languages other than Italian [3]. Leveraging corpora in different languages allows us to overcome the limited availability of parallel corpora in Italian and the administrative domain in particular.

To sum up the main contributions of this paper:

- We release Admin-It-L2,[1] a parallel corpus in the Italian administrative language with simplifications aimed at Italian L2 speakers (Sec. 3).
- To the best of our knowledge, this is the first attempt to create a controlled simplification model specific to the Italian administrative language (Sec. 4). In particular, we train controllable models (Sec. 5) with multilingual data, to simplify texts aimed at Italian L2 speakers, thanks to data collected with behavioral analysis.
- We use Admin-It-L2 to test such models and we evaluate them with the available metrics for Automatic Text Simplification (ATS). We also manually analyze the produced simplification to assess the validity of such metrics for the Italian bureaucratic language (Sec. 6).

Finally, we also show the results of a preliminary and exploratory experiment with ChatGPT.[2] Based on GPT-3.5, this Large Language Model (LLM) gained popularity also thanks to the impressive performance it reaches in several NLP tasks [4]. Thus, we decided to give a brief overview of its potential and limitations when applied to the simplification of Italian administrative texts for Italian L2 speakers (Sec. 7).

2. Related Work

ATS aims at simplifying a text while maintaining its meaning [5]. Since the spread of Neural Networks, models' architectures are mostly taken from "Neural Machine

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
martina.miliani@fileli.unipi.it (M. Miliani); alvamanchegof@cardiff.ac.uk (F. Alva-Manchego); alessandro.lenci@unipi.it (A. Lenci)
0000-0003-1124-9955 (M. Miliani); 0000-0001-6218-8377 (F. Alva-Manchego); 0000-0001-5790-4308 (A. Lenci)

[1]https://github.com/Unipisa/admin-it-l2
[2]https://openai.com/blog/chatgpt

Translation” (NMT) [6]. For example, to build their simplification model, [7] started from OpenNMT [8], consisting of an encoder-decoder with two layers of LSTM. [9] followed a similar approach and trained a model for the Italian language thanks to data augmentation techniques.

In some of the most recent models, simplification can be “controlled” to generate texts for specific groups of readers. CROSS [10] proposed a Transformer-based model able to control the level of simplicity and the type of applied simplification. To this aim, at a lexical level, they marked the token to replace, whereas, at a syntactic level, they leveraged templates. ACCESS [11] introduced the use of control tokens to bind the simplification with specific attributes, such as the amount of paraphrased content and lexical and syntactic complexity. [12] used ACCESS to simplify Italian texts and trained it on the automatic translation of NewsEla [13] (see Sec. 5.1), obtaining promising results. MUSS [14] presented an unsupervised method to collect parallel data to train a model based on ACCESS adopting BART and its multilingual version, mBART [15]. The authors’ goal was to deal with the paucity of available parallel corpora in languages different from English. To this aim, we leverage cross-language and fully translated data, by exploiting the ability of multilingual models to use linguistic knowledge from different languages. [16] also used ACCESS but replaced BART with T5 [17]. With this controllable model, [18] achieved SOTA performance for SARI on Spanish texts.

Differently from the described work, we focus our simplification on in-domain texts written in the Italian bureaucratic language. Furthermore, we leverage data on Italian L2 speakers to create a test set to control the simplifications generated by Transformer-models. Such a dataset is described in the next section.

3. Admin-It-L2

In this section, we present Admin-It-L2, a parallel corpus of complex-simplified sentences in the Italian administrative language for ATS aimed at Italian L2 speakers.

We manually simplified 134 sentences by focusing on the linguistic traits that emerged from a comprehension test conducted over 86 participants [19], involving also Italian L2 speakers (30,2%), and elderly Italian native speakers (33,72%). The participants were asked to answer questions about an original text and a simplified version of another text, both from the administrative domain. Such simplification focused only on the typical traits of the bureaucratic language [20].

By analyzing the participants’ answers to the comprehension test, the authors observed that L2 struggled more when reading simple texts with long sentences, long prepositional chains, a high number of participle verbs, and a lower number of indicative verbs. Their answer error rate also increased with a higher number of multi-words and entities.

Admin-It-L2 counts 134 pairs of sentences extracted from the texts used in this study (34 sentences) and from **Admin-It**$_{RS}$ (100 sentences), a subsection of Admin-It, a parallel corpus of Italian administrative texts [3]. The original sentences of this subset were selected from websites of Italian municipalities, and from the longest sentences of the PaWaC Corpus [21]. [3] manually rewrote the sentences simplifying them both at lexical and syntactic levels. This simplification as well was only based on the typical traits of the administrative language [2, 20].

The similar nature of the simplification applied, allowed us to further simplify the simple sentences of Admin-It$_{RS}$ and the simple sentences employed in the comprehension test (together referred to as *Administsemp* in Table 1) by considering the results of the study related to Italian L2 speakers. The annotation was conducted by a single annotator among the authors, and validated through a quantitative analysis shown in Table 1.[3]

4. The Controllable Simplification Model

Our models are based on the implementation released by [16].[4] The two authors used ACCESS and replaced BART with T5, a model with an encoder-decoder architecture pre-trained on various tasks through supervised and unsupervised approaches [17].

We adopted the multilingual version of T5, mT5 [22], for all our experiments. [23] observed that multilingual models proved to be better for tasks like text summarization or switching from a formal to a more informal language. The authors suggest that this is probably due to the different distributions of linguistic data used for training the model. In particular, the multilingual model would perform better with automatically translated texts.

Unlike [16], due to hardware constraints, we employed the basic mT5 model and fine-tuned it on a batch of size 16. We retained the values of all other parameters, including the number of epochs (5). Fine-tuning was performed on an NVIDIA A100 Tensor Core GPU, with 40 GB of RAM.

5. Experimental Settings

Given the limited availability of parallel corpora in Italian, we created training sets for fine-tuning by adopting two

[3] Details about the simplification process are presented in Appendix A.
[4] https://github.com/KimChengSHEANG/TS_T5

Table 1
Some statistics of Admin-It-L2 in the original version of the sentences (*Admin-It-L2-comp*), the simplified version based on administrative language traits only (*Administ-semp*) and the simplified version of the sentences in Admin-It-L2 (*Admin-It-L2-semp*).

	Admin-It-L2-comp	*Administ-semp*	*Admin-It-L2-semp*
Avg token per sentence	52.11	29.86	24.83
Avg character per token	4.65	4.52	4.52
Number of sentences	179.0	276.0	270.0
Participial verbs (%)	22.56	7.39	6.83
Infinitive verbs (%)	17.69	21.72	22.04
Finite-mode verbs (%)	38.85	44.68	49.87
Indicative verbs (%)	31.92	42.55	49.48
Avg dept of the syntactic tree	5.17	3.87	3.07
Avg length of prepositional chains	1.65	1.38	1.34
Levenshtein Distance	-	170.44	202.73

main strategies. The first involves the full translation into Italian of parallel English corpora with Google Translate.[5] We translated only from English since translation systems for this language generally perform better, thanks to the amount of data available for model training. The second strategy involves the creation of cross-language datasets, in which the original sentences are in English, whereas their simplified counterparts were translated into Italian. On the one hand, translating simple texts is an easier task for machine translation systems [24, 25]. On the other hand, we intended to take advantage of the capabilities of multilingual generative models to succeed in applying linguistic knowledge from different languages. This potential has previously been tested for summarizing documents in different languages [26].

5.1. Datasets

We used Admin-It-L2 as the test set, to assess whether the models succeed in producing simplifications close to the needs of Italian L2 speakers. To train the controllable models we combined the following datasets:[6]

NewsEla (NewsEng) consists of newspaper articles manually simplified by experts according to different degrees of complexity. NewsEla was then automatically sentence-aligned by [27].

OneStopEnglish (OSE) contains articles from the British newspaper The Guardian rewritten by teachers in three levels of readability for English L2 learners [28].

SimPA is a parallel corpus of sentences in English administrative language [29]. The authors applied two steps of simplification: lexical and syntactic.

Spanish NewsEla (NewsEs) sentence-aligned version was created by [9]. Documents were translated from English, and then simplified manually.

Terence contains short stories in Italian for children rewritten by a group of experts [20].

Teacher contains teaching materials manually simplified by a teacher for Italian L2 learners [20].

PaCCSS-IT was presented by [30], who collected sentences in Italian from the web through a semi-supervised method.

Simpitiki$_W$ is the portion of Simpitiki composed of Italian Wikipedia articles selected from the edits labeled as simplifications. The simplified sentences are obtained by applying one simplification operation at a time [31].

Admin-It$_{OP}$ and **Admin-It**$_{RD}$ are subsections of Admin-It. The former corresponds to another section of the Simpitiki corpus and is composed of sentences from administrative texts manually simplified by applying a single operation, whereas in the latter, the simplification is applied at a document level, then the sentences were manually aligned [3].

As for the automatic translations, firstly we used Google Translate only on the simple sentences of English corpora. The resulting corpora are then referred to as *NewsEn-Ita*, *OSE-EnIta*, and *SimPA-EnIta*. Then, we automatically translated datasets including both simple and complex sentences. In this case, the resulting corpora are referred to as *NewsIta*, *OSE-Ita*, and *SimPA-Ita*.[7]

We extracted two random samples from NewsEn,[8] instead of fine-tuning models on the whole dataset, since [18] achieved SOTA performance for SARI on Spanish on a relatively small amount of data.[9]

[5] https://translate.google.com/

[6] From these datasets, pairs containing the same sentence in their original and simplified versions were filtered out. Additional processing was applied to Terence and Teacher, which were provided in XML format. Datasets statistics are reported in Appendix B.

[7] The simplified sentences in these corpora show variations in the two translation versions, as the statistics in Appendix B show.

[8] One for each kind of translation.

[9] They fine-tuned their model on a sample of Spanish NewsEla that counted about 7k sentences.

5.2. Evaluation metrics

Three metrics were used to evaluate the models: BLEU [32], SARI [33] and BERTScore$_p$ [34].

The first, BLEU, measures the n-gram overlap between gold and generated sentences. It is inherited from Machine Translation, and [35] observed that BLEU is less reliable when sentence splitting is applied in the simplified sentence. SARI is specific to evaluating simplification models and measures the effectiveness of copy (KEEP), insertion (ADD), and deletion (DEL) operations applied [36]. This metric, therefore, is widely used to evaluate simplification at the lexical level. By design, SARI takes as input several gold references. Since Admin-It-L2 contains only one reference per sentence, this metric might not be fully reliable. The third evaluates texts generally created by generative models. BERTScore sums up the cosine similarity of token pairs with the highest similarity. Specifically, BERTScore$_p$ measures the similarity of tokens in the predicted sentence with respect to the tokens in the gold sentence. [35] observed that BERTScore$_p$ correlates with the simplicity values given by human annotators when such values are low.

We applied these metrics by using the implementation of the EASSE evaluation tool [37] provided by [16]. As for BERTScore, since there is no variant of BERT available for the Italian language for this metric, we employed a multilingual model, `xlm-roberta-large`,[10] which performs better than mBERT on the Italian administrative language [38].

5.3. Control tokens

For these experiments, we used the same control tokens as [16]:

- *NbChars*. Synthesis: the ratio of the length in characters between source and simplified sentences;
- *LevSim*. Paraphrasing: the normalized *Levenstein distance* between source and simplified sentences;
- *WordRank*. Lexical complexity: the ratio between word frequencies in the two sentences, original and simplified;
- *DepTreeDept*. Syntactic complexity: the ratio between the maximum depth of the syntactic tree of source and target sentences;
- *NbWords*. Lexical complexity: the number of words in the simplified sentence divided by that of the complex sentence.

Although *WordRank* leverages English language data, we included this control token since [12] drew some advantage from it on Italian text simplification. To extract these linguistic features, we added a specific function to the implementation of [16] to detect the language of the examined sentence.[11] Given the results obtained through behavioral analysis [19], we employed all available control tokens for these experiments. A more suitable simplification for Italian L2 speakers can be obtained by controlling word frequency through *WordRank* and *NbWords*. This affects morphological aspects because verbs in the indicative present tense are generally more frequent than those in other verb tenses and modes. *LevSim* and *NbChars* operate more on sentence length whereas the depth of the syntactic tree can be reduced with *DepTreeDept*, also affecting propositional chains average length. Regarding the number of complex entities and terms, this trait was traced back to aspects related to sentence length, as emerged from the behavioral analysis.

Differently from [11] and [16], who conducted their experiments by extracting training, validation, and test sets from the same corpus, the sentences used for our experiments come from different domains and languages. The goal here is to modify the values of the control tokens to optimize the performance of the model on the test set. For this purpose, we extracted the values of the controlled variables from each pair of sentences in Admin-It-L2, and then we computed the average.[12] We fed the models with such values to condition the simplification generated during evaluation.

5.4. Baselines

We selected two baseline models. Admin-It$_{OP}$ is obtained by fine-tuning mT5 using the subsection of Admin-It. This subset corresponds to a subsection of Simpitiki, the largest Italian dataset of parallel sentences in the administrative language. We also evaluated this model by setting the values of the control tokens with the average values on Admin-It-L2. The second baseline is a hypothetical model that generates a copy of the complex sentences of the test set. This baseline is reported in Table 2 as Admin-It-L2$_C$. We assume that models obtaining results close to this baseline are very conservative in their simplification process.

6. Results and discussion

The results of our experiments are shown in Table 2. The baseline obtained by fine-tuning mT5 only on Admin-It$_{OP}$ reached results close to zero for both BLEU and BERTScore$_p$ indicating that this dataset might be too small for the model to learn any simplification rule. Also,

[10] https://huggingface.co/xlm-roberta-large

[11] We used the *language identification* module provided by fastText [39, 40]. In this case, as to compute the corpora statistics we used: `it_corenews_sm` and `en_coreweb_sm`.

[12] Table 11 in Appendix C contains the values given to the controllable tokens.

Table 2
Results achieved by each model. The generated simplifications are evaluated with BLEU, SARI, and BERTScore$_p$. The highest results obtained for each metric are written in bold.

Model	*BLEU*	*SARI*	*SARI$_{ADD}$*	*SARI$_{KEEP}$*	*SARI$_{DEL}$*	*BERTScore$_p$*
Admin-It$_{OP}$	0.02	28.89	0.21	3.28	**83.19**	0.00
Admin-It-L2$_C$	25.16	14.04	0.00	42.11	0.00	0.50
All-Ita	24.73	21.63	0.21	42.09	22.60	0.50
All-Ita+SimPA-EnIta	**25.34**	27.51	0.80	41.91	39.83	**0.52**
All-Ita+All-EnIta	22.39	24.97	0.93	40.57	33.42	0.48
OSE-EnIta+SimPA-EnIta	12.02	**39.24**	**3.19**	36.72	77.82	0.51
OSE-Ita+SimPA-Ita	24.86	27.57	1.00	**42.74**	38.98	0.51
All-TransIta	15.96	35.85	2.98	41.94	62.64	0.44
All-Ita+SimPa-Ita	25.0	20.16	0.36	42.28	17.85	0.51

the low BERTScore$_p$ is indicative of incorrect simplification [35]. Admin-It-L2$_C$, obtained a lower SARI score and higher BLEU and BERTScore$_p$ than the first baseline.

The first model we tested, All-Ita, was obtained by fine-tuning mT5 on all available Italian corpora. However, this model obtained a lower SARI score than Admin-It$_{OP}$ (-6.66), and even though it equals Admin-It-L2$_C$ for BERTScore$_p$, it reached a lower BLEU score (-0.43).

We built a second model by adding SimPA-EnIta to All-Ita. This model scores slightly better than Admin-It-L2$_C$ on both BLEU (+0.61) and BERTScore$_p$ (+0.02). We also included OSE-EnIta and NewsEn-Ita in the fine-tuning of a third model, All-Ita+All-EnIta. However, adding all the cross-language data worsened the performance. Therefore, we ran further experiments focusing on the characteristics of each corpus used.

We leverage cross-language corpora employing only OSE-EnIta and SimPA-En-Ita, and although this model achieves the worst results for BLEU (12.02), its BERTScore$_p$ is only slightly lower than All-Ita+SimPa-EnIta (-0.01), and above all, it surpasses this model for 11.73 points for SARI, obtaining the best score for this metric: 39.24%. This model also achieved the highest value for SARI$_{ADD}$ (3.99) - although it remains low - and SARI$_{DEL}$ (77.82), and the lowest for SARI$_{KEEP}$ (36.72). Such scores may indicate how much less conservative this model is than the others in generating simplifications. Our next experiment involved the full translation of these two corpora, OSE-Ita and SimPA-Ita. Although the model achieves the same BERTScore$_p$ (0.50) as OSE-EnIta+SimPA-En-Ita, the performance is significantly worse for SARI (-11.67). Finally, for BLEU, although it outperforms All-Ita (+0.13), the best model for this score remains All-Ita+SimPa-EnIta. Next, we decided to fine-tune mT5 using all the corpora fully translated from English. The resulting model is reported in Table 2 as All-TransIta and improves over the previous one only for SARI, but without reaching the score obtained by OSE-EnIta+SimPA-En-Ita. We conducted a final experiment with All-Ita+SimPA-Ita. This model scores similarly to All-Ita+SimPA-EnIta for BLEU (-0.34) and BERTScore$_p$ (-0.01), but the SARI score is significantly worse (-7.35), generating rather conservative simplifications, with a SARI$_{KEEP}$ even higher than Admin-It-L2$_C$ (+0.17).

Albeit considering the variations in the translations, the contribution of the cross-language data for specific corpora would seem to make a difference compared to the texts fully translated into Italian (e.g., OSE-EnIta+SimPA-En-Ita vs. OSE-Ita and SimPA-Ita). Even though they remain low, the best results were obtained by All-Ita+SimPA-EnIta for BERTScore$_p$ (0.52) and BLEU (25.34), while the highest score for SARI is reached by OSE-EnIta+SimPA-En-Ita (39.24), not too far from the SOTA for English.[13]

6.1. Manual analysis

Then, we manually analyzed a random sample of 50 simplified sentences. We observed that All-Ita+SimPA-EnIta tends to copy the content of the original sentences. The model merely eliminates some portion of text, but in several cases, this happens with relevant information. In OSE-EnIta+SimPA-EnIta simplifications, the deletion operations are even more massive, and much of the source text is usually removed, producing inconsistent sentences. In other cases, OSE-EnIta+SimPA-EnIta produces "artificial hallucinations". However, OSE-EnIta+SimPA-EnIta succeeds in producing better simplifications when the complex sentences are shorter, even when All-Ita+SimPA-EnIta still reproduces them entirely (see Table 3).[14]

The results confirmed what was pointed out by [35]: BERTScore$_p$ is a reliable metric in detecting mostly low-quality simplifications, and metrics like BLEU and SARI cannot be analyzed in isolation. Besides the low quality of the generated simplification, the low values obtained by BERTScore$_p$ may also find an explanation both in the

[13] Martin et al. [14] obtained a SARI score of about 42 on NewsEn.
[14] Other examples of models output can be seen in Appendix D.

Table 3
The Table shows an example of the generated simplifications. All-Ita+SimPA-EnIta exactly reproduces the text of the original sentence in Admin-It-L2, while OSE-EnIta+SimPA-EnIta eliminates the more complex portion of text contained within an aside (in italic in the original sentence).

Original Admin-It-L2: È, inoltre, possibile richiedere, per non più di 3 giorni consecutivi, un pasto "in bianco"*, in presenza di disturbi gastroenterici,* senza la presentazione di certificazione o prescrizione medica.
Simplified Admin-It-L2: Se il bambino ha disturbi allo stomaco o all'intestino, il genitore può richiedere un pasto "in bianco" senza presentare una certificazione o prescrizione medica per massimo 3 giorni di fila.
All-Ita+SimPA-EnIta: È possibile richiedere, per non più di 3 giorni consecutivi, un pasto "in bianco", in presenza di disturbi gastroenterici, senza la presentazione di certificazione o prescrizione medica.
OSE-EnIta+SimPA-EnIta: È possibile richiedere un pasto "in bianco per non più di 3 giorni consecutivi, senza la presentazione di certificato o prescrizione medica.

BERT model employed for the extraction of the vector representations, which is not specific to Italian (Sec. 5.2), and in the composition of the test set, Admin-It-L2. Its sentences are more complex and longer than the other corpora and underwent a deeper simplification process.

7. A quick chat with ChatGPT

We conducted a preliminary and exploratory experiment to get a brief overview of the capabilities of ChatGPT when asked to simplify Italian administrative texts for Italian L2 speakers in a zero-shot setting.

In the conversation shown in Table 4,[15] when asked to simplify a sentence contained in Admin-It-L2, ChatGPT provides a short and syntactically accurate text, even though the long period in the original sentence has not been broken down into shorter sentences. The model shows that it can provide the meaning of "procura alle liti" (power of attorney) and can insert a short definition into the simplified sentence. However, when asked to simplify the sentence further by phrasing it as if it were addressing a person who is not fluent in Italian, the model adopts a very informal register: the sentence sounds almost ungrammatical and less coherent in some parts.[16] When prompted to use a more polished language, ChatGPT raised the register, although apparently, the model loses sight of the identity of the complainants, namely citizens, and instead refers to a generic "person". This way, ChatGPT shows to be able to simplify administrative text in Italian at both the syntactic and the lexical level and to tune the simplification according to the prospected scenario, i.e., when addressing Italian L2 speakers. Nevertheless, the model is not entirely accurate when inferring that "the uncertainty as to the antecedence of the power of attorney" led to the "confusion about who is representing the citizen in the process". Indeed, this latter case could juridically give rise to defects of a different nature than the one depicted by the Province of Pistoia.[17] ChatGPT seems not to grasp here the intrinsic and necessary complexity of the content, which is closely linked to the complexity of the subject matter described [41].

It is also important to remark that proprietary LLMs like ChatGPT are not open-source and are available only via web APIs. This leads to a lack of transparency in the data used for training, limiting the reproducibility of experiments and quality control of the results [42].

8. Conclusions and future work

In this paper, we presented the first benchmark for the automatic simplification of Italian administrative texts aimed at Italian L2 speakers. We employed data collected from behavioral analysis to build Admin-It-L2, a parallel corpus of original-simplified sentences. Then, we used this corpus to test controllable models for text simplification based on Transformers. Although we obtained high results for SARI, we observed that SARI and BLEU alone cannot be used to evaluate text simplification and that a fully reliable automatic metric for the Italian language is still missing. Finally, we asked ChatGPT to simplify one of the sentences in Admin-It-2, obtaining a higher quality simplification than mT5, although the model presents some limitations, especially in reproducibility, and in some cases, in answer accuracy.

For future work, we aim at validating Admin-It-L2 involving human annotation, and, more specifically, domain experts and Italian L2 speakers, to properly evaluate the employed models and the quality of their generations. As for the models, we aim to conduct an ablation study to assess the contribution of each control token. We also would like to adopt a control token to encode the language used in the sentences and employ the corpora in their original versions. Moreover, we aim to create a specific automatic metric to evaluate Italian text simplification. Finally, we would like to test open-source LLMs on administrative text simplification focused on specific target readers, with different prompts, in zero and few-shot settings.

[15] We tested ChatGPT between December 2022 and January 2023.
[16] The absence of the pronoun "lo" referring to "documento" makes less explicit what the expression "e non doveva essere" is referring to.
[17] We consulted a legal expert to verify the correctness of the answers of ChatGPT.

Table 4
The excerpt of a conversation with ChatGPT in Italian. In bold are the instructions given to the chatbot and in italics are the translations in English of both questions and answers.

Semplifica questa frase: La Provincia di Pistoia eccepisce anche l'inammissibilità del ricorso per nullità della procura alle liti rilasciata dai cittadini ricorrenti in quanto apposta, a differenza di quella dell'associazione attrice, su appositi fogli aggiunti e spillati all'atto introduttivo del giudizio, in difformità della previsione contenuta nell'art. 83 c.p.c. e ciò in quanto, in tal caso non vi sarebbe certezza in ordine alla anteriorità del conferimento del mandato rispetto alla notifica dell'atto di resistenza o di impugnazione (cfr. Cass. civ., sez. iii, 14 marzo 2006, n. 5443).
Simplify this sentence: The Province of Pistoia also objects to the inadmissibility of the appeal due to the nullity of the power of attorney issued by the citizen plaintiffs in that it was affixed, unlike that of the plaintiff association, on special sheets added and stapled to the writ of summons, contrary to the provision contained in Art. 83 c.p.c. and this because, in such a case, there would be no certainty as to the anteriority of the conferment of the mandate concerning the notification of the act of resistance or appeal (see Cass. civ, sec. iii, March 14, 2006, no. 5443).
La Provincia di Pistoia contesta l'ammissibilità del ricorso per nullità della procura alle liti in quanto è stata apposta su fogli aggiunti e non nell'atto introduttivo del giudizio, violando l'articolo 83 c.p.c. e causando incertezza sulla anteriorità del mandato rispetto alla notifica dell'atto di resistenza o di impugnazione.
The Province of Pistoia contests the admissibility of the appeal on the grounds that the power of attorney was null and void because it was affixed on supplementary sheets and not in the writ of summons, violating Article 83 of the Code of Civil Procedure and causing uncertainty as to the antecedence of the grant of power of attorney with respect to the service of the notice of resistance or appeal.
cos'è la procura alle liti?
what is the power of attorney?
La procura alle liti è un documento formale con cui una persona (il mandante) conferisce a un'altra (il mandatario) il potere di rappresentarla e agire in suo nome in una causa legale. La procura alle liti deve essere redatta in forma scritta, sottoscritta dal mandante e autenticata da un notaio o da un avvocato. In essa vengono specificati i poteri che il mandatario ha in relazione alla causa, come ad esempio il potere di depositare documenti, di comparire in giudizio e di transigere.
A power of attorney is a formal document by which one person (the principal) grants another (the agent) the power to represent and act on his behalf in a legal case. The power of attorney must be in writing, signed by the principal, and notarized by a notary public or lawyer. It specifies the powers that the agent has in relation to the lawsuit, such as the power to file documents, to appear in court, and to settle.
Puoi inserire questa definizione nella semplificazione che hai fornito poco fa in modo che il testo sia chiaro anche alle persone che non sanno il significato di procura alle liti. La frase non deve risultare quindi troppo lunga
You can include this definition in the simplification you provided earlier so that the text is clear even to people who do not know the meaning of power of attorney in litigation. The sentence should therefore not be too long
La Provincia di Pistoia contesta l'ammissibilità del ricorso per nullità della procura alle liti, un documento in cui una persona conferisce a un'altra il potere di rappresentarla in una causa legale, in quanto è stata apposta su fogli aggiunti e non nell'atto introduttivo del giudizio, violando l'articolo 83 c.p.c. e causando incertezza sulla anteriorità del mandato rispetto alla notifica dell'atto di resistenza o di impugnazione.
The Province of Pistoia challenges the admissibility of the appeal on the grounds that the power of attorney, a document in which one person confers on another the power to represent him or her in a legal case, is null and void because it was affixed on additional sheets of paper and not in the writ of summons, violating Article 83 of the Code of Civil Procedure and causing uncertainty as to the antecedence of the power of attorney from the service of the notice of resistance or appeal.
Adesso semplifica ancora di più la frase come se stessi parlando a qualcuno che conosce poco l'italiano
Now simplify the sentence even more as if you were speaking to someone who is not fluent in Italian
La Provincia di Pistoia dice che il documento che permette a qualcuno di rappresentare un'altra persona in un processo legale è sbagliato, perché è stato scritto su fogli aggiunti e non doveva essere e questo causa confusione su chi rappresenta la persona nel processo.
The Province of Pistoia says that the document that allows someone to represent another person in a legal process is wrong because it was written on added sheets and should not have been, and this causes confusion about who is representing the person in the process.
Potresti migliorare un po' lo stile? il tuo interlocutore conosce poco l'italiano ma il testo deve essere un po' più curato
Could you improve the style a little? your interlocutor is not fluent in Italian, but the text needs to be a little more polished
La Provincia di Pistoia sostiene che il documento chiamato "procura alle liti" che consente a una persona di rappresentare un'altra persona in un processo legale è stato scritto in modo errato. Infatti, è stato scritto su fogli aggiunti anziché nell'atto ufficiale del processo, e questo può causare confusione su chi rappresenta la persona nel processo.
The Province of Pistoia claims that the document called "power of attorney" that allows a person to represent another person in a legal process was written incorrectly. In fact, it was written on added sheets instead of in the official document of the trial, and this can cause confusion about who represents the person in the trial.

Acknowledgments

Research partially supported by PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI", funded by the European Commission under the NextGeneration EU programme.

References

[1] D. Fortis, Il dovere della chiarezza. quando farsi capire dal cittadino è prescritto da una norma, RIVISTA ITALIANA DI COMUNICAZIONE PUBBLICA 25 (2005) 82–116. URL: http://www.francoangeli.it/Riviste/Scheda_rivista.aspx?idArticolo=25382.

[2] M. A. Cortelazzo, Il linguaggio amministrativo: principi e pratiche di modernizzazione, Studi superiori, Carocci, 2021.

[3] M. Miliani, S. Auriemma, F. Alva-Manchego, A. Lenci, Neural readability pairwise ranking for sentences in italian administrative language, in: Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, 2022, pp. 849–866.

[4] Y. Liu, T. Han, S. Ma, J. Zhang, Y. Yang, J. Tian, H. He, A. Li, M. He, Z. Liu, et al., Summary of chatgpt-related research and perspective towards the future of large language models, Meta-Radiology (2023) 100017.

[5] F. Alva-Manchego, C. Scarton, L. Specia, Data-driven sentence simplification: Survey and benchmark, Computational Linguistics 46 (2020) 135–187.

[6] K. Cho, B. van Merrienboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using rnn encoder-decoder for statistical machine translation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1724–1734.

[7] S. Nisioi, S. Štajner, S. P. Ponzetto, L. P. Dinu, Exploring neural text simplification models, in: Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: Short papers), 2017, pp. 85–91.

[8] G. Klein, Y. Kim, Y. Deng, J. Senellart, A. M. Rush, Opennmt: Open-source toolkit for neural machine translation, in: Proceedings of ACL 2017, System Demonstrations, 2017, pp. 67–72.

[9] A. Palmero Aprosio, S. Tonelli, M. Turchi, M. Negri, A. Di Gangi Mattia, Neural text simplification in low-resource conditions using weak supervision, in: Workshop on Methods for Optimizing and Evaluating Neural Language Generation (NeuralGen), Association for Computational Linguistics (ACL), 2019, pp. 37–44.

[10] J. Mallinson, M. Lapata, Controllable sentence simplification: Employing syntactic and lexical constraints, ArXiv abs/1910.04387 (2019). URL: https://api.semanticscholar.org/CorpusID:204008990.

[11] L. Martin, É. V. De La Clergerie, B. Sagot, A. Bordes, Controllable sentence simplification, in: Proceedings of the 12th Language Resources and Evaluation Conference, 2020, pp. 4689–4698.

[12] A. L. Megna, D. Schicchi, G. Lo Bosco, G. Pilato, A controllable text simplification system for the italian language, in: 2021 IEEE 15th International Conference on Semantic Computing (ICSC), 2021, pp. 191–194. doi:10.1109/ICSC50631.2021.00040.

[13] W. Xu, C. Callison-Burch, C. Napoles, Problems in current text simplification research: New data can help, Transactions of the Association for Computational Linguistics 3 (2015) 283–297.

[14] L. Martin, A. Fan, É. de la Clergerie, A. Bordes, B. Sagot, MUSS: Multilingual unsupervised sentence simplification by mining paraphrases, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 1651–1664. URL: https://aclanthology.org/2022.lrec-1.176.

[15] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 7871–7880.

[16] K. C. Sheang, H. Saggion, Controllable sentence simplification with a unified text-to-text transfer transformer, in: Proceedings of the 14th International Conference on Natural Language Generation, Association for Computational Linguistics, Aberdeen, Scotland, UK, 2021, pp. 341–352. URL: https://aclanthology.org/2021.inlg-1.38.

[17] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, et al., Exploring the limits of transfer learning with a unified text-to-text transformer., J. Mach. Learn. Res. 21 (2020) 1–67.

[18] S. Štajner, K. C. Sheang, H. Saggion, Sentence simplification capabilities of transfer-based models, Proceedings of the AAAI Conference on Artificial Intelligence 36 (2022) 12172–12180. URL: https://ojs.aaai.org/index.php/AAAI/article/view/21477. doi:10.1609/aaai.v36i11.21477.

[19] M. Miliani, M. S. Senaldi, G. E. Lebani, A. Lenci, Un-

derstanding italian administrative texts: A reader-oriented study for readability assessment and text simplification, in: Proceedings of the 1st Workshop on AI for Public Administration (AIxPA), 2022, pp. 71–87.
[20] D. Brunato, F. Dell'Orletta, G. Venturi, S. Montemagni, Design and annotation of the first italian corpus for text simplification, in: Proceedings of The 9th Linguistic Annotation Workshop, 2015, pp. 31–41.
[21] L. C. Passaro, A. Lenci, Extracting terms with extra, Computerised and corpus-based approaches to phraseology: Monolingual and multilingual perspectives (2016) 188–196.
[22] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mT5: A massively multilingual pre-trained text-to-text transformer, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 483–498. URL: https://aclanthology.org/2021.naacl-main.41. doi:10.18653/v1/2021.naacl-main.41.
[23] G. Sarti, M. Nissim, It5: Large-scale text-to-text pretraining for italian language understanding and generation, arXiv preprint arXiv:2203.03759 (2022).
[24] S. Štajner, M. Popović, Can text simplification help machine translation?, in: Proceedings of the 19th Annual Conference of the European Association for Machine Translation, 2016, pp. 230–242.
[25] K. Mishra, A. Soni, R. Sharma, D. M. Sharma, Exploring the effects of sentence simplification on hindi to english machine translation system, in: proceedings of the workshop on automatic text simplification-methods and applications in the multilingual society (ATS-MA 2014), 2014, pp. 21–29.
[26] L. Perez-Beltrachini, M. Lapata, Models and datasets for cross-lingual summarisation, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 9408–9423.
[27] C. Jiang, M. Maddela, W. Lan, Y. Zhong, W. Xu, Neural CRF model for sentence alignment in text simplification, CoRR abs/2005.02324 (2020). URL: https://arxiv.org/abs/2005.02324. arXiv:2005.02324.
[28] S. Vajjala, I. Lučić, Onestopenglish corpus: A new corpus for automatic readability assessment and text simplification, in: Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications, 2018, pp. 297–304.
[29] C. Scarton, G. Paetzold, L. Specia, SimPA: A sentence-level simplification corpus for the public administration domain, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan, 2018, pp. 4333–4338. URL: https://aclanthology.org/L18-1685.
[30] D. Brunato, A. Cimino, F. Dell'Orletta, G. Venturi, Paccss-it: A parallel corpus of complex-simple sentences for automatic text simplification, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 351–361.
[31] S. Tonelli, A. P. Aprosio, F. Saltori, Simpitiki: a simplification corpus for italian, in: Proceedings of CLiC-it, 2016.
[32] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.
[33] W. Xu, C. Napoles, E. Pavlick, Q. Chen, C. Callison-Burch, Optimizing statistical machine translation for text simplification, Transactions of the Association for Computational Linguistics 4 (2016) 401–415.
[34] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, in: International Conference on Learning Representations, 2019.
[35] F. Alva-Manchego, C. Scarton, L. Specia, The (un) suitability of automatic evaluation metrics for text simplification, Computational Linguistics 47 (2021) 861–889.
[36] E. Sulem, O. Abend, A. Rappoport, Bleu is not suitable for the evaluation of text simplification, arXiv preprint arXiv:1810.05995 (2018).
[37] F. Alva-Manchego, L. Martin, C. Scarton, L. Specia, Easse: Easier automatic sentence simplification evaluation (2019) 49–54.
[38] S. Auriemma, M. Miliani, A. Bondielli, L. C. Passaro, A. Lenci, Evaluating pre-trained transformers on italian administrative texts, in: Proceedings of 1st Workshop on AI for Public Administration, 2022, pp. 54–70.
[39] A. Joulin, É. Grave, P. Bojanowski, T. Mikolov, Bag of tricks for efficient text classification, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, 2017, pp. 427–431.
[40] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, T. Mikolov, Fasttext.zip: Compressing text classification models, ArXiv abs/1612.03651 (2016). URL: https://api.semanticscholar.org/CorpusID:16196524.
[41] A. Fioritto, Manuale di stile. strumenti per semplificare il linguaggio delle amministrazioni pubbliche (1997).
[42] A. Liesenfeld, A. Lopez, M. Dingemanse, Opening

up chatgpt: Tracking openness, transparency, and accountability in instruction-tuned text generators, in: Proceedings of the 5th International Conference on Conversational User Interfaces, 2023, pp. 1–6.

A. Admin-It-L2 Annotation

In this section, we present further details about the realization of Admin-It-L2. As described in Section 3, the behavioral analysis shows that Italian L2 speakers struggle in reading simplified texts with long sentences, long prepositional chains, a high number of participle verbs, and a lower number of indicative verbs. Their answer error rate also increased with a higher number of multiwords and entities, but also these lexical features were associated in the behavioral analysis with sentence length.

Regarding morphological aspects, we applied transformations from the subjunctive (6) or the infinitive (5) to the indicative, and verbal periphrases were replaced with single verb forms of similar meaning, as in the example 8. Concerning sentence length reduction, references to laws whose citation in the text was not crucial to sentence comprehension were removed (6, 9). The idea is that the automatic simplification operated by the neural models is always flanked by the original text, which complements the absence of some information that might lower text readability [2]. The depth of the syntactic tree was also reduced (-0.80) by limiting redundant expressions, from an average sentence length in tokens of 29.86 to 24.83 (see Table 1), without affecting the cohesion of the text (6). A finding also emerges from Levensthein's distance from the original sentences to the simplified sentences (170.44) and the simplified sentences in Admin-It-L2 (202.73). We simplified the sentences syntactically by intervening on prepositional chains, which in some cases were eliminated (7) or reduced (9). Finally, 27 sentences were not further simplified.

B. Dataset statistics

Table 10 shows the details of the corpora used in the simplification experiments. The upper part of the table lists the Italian corpora, then the English-Italian corpora, and at the bottom, the English corpora fully translated into Italian. Finally, the test set, Admin-It-L2. Since the complex-simple pairs constitute in several cases portions of text consisting of several sentences (especially in the simplified version), Table 10 shows also the number of the sentences recognized by spaCy within each portion of text.

C. Values of controllable variables

Table 11 shows the values associated with each controllable token in the prediction phase of the fine-tuned models. We computed the controllable variables from each sentence pair in Admin-It-L2 and then we calculated their average.

D. Examples of simplification

In this section, we reported some examples of the sentences generated by two of the created models: All-Ita+SimPA-EnIta and OSE-EnIta+SimPA-EnIta. All-Ita+SimPA-EnIta tends to copy most of the content from the original sentence (Table 13) or delete portions of the texts regardless of their relevance (Table 12). In the other two reported tables, there are interesting examples of the sentences returned by OSE-EnIta+SimPA-EnIta. In Table 13, the model tends to delete most of the content and generates inconsistent text. Such inconsistencies can also lead to artificial hallucinations, as in the example in Table 14.

Table 5
The simplified sentence in Admin-It$_{RS}$ has been further manipulated morphosyntactically by replacing a verb in the infinitive with one in the present indicative, and syntactically by preferring parataxis to hypotaxis and eliminating information that is not strictly necessary.

Original: La gestione della pesa sarà interamente informatizzata, collegando l'apparato di pesatura ad un sistema informativo; tutte le operazioni di pesatura verranno così memorizzate all'interno di un server, *per poter essere storicizzate, consultate ed elaborate* con la massima comodità e sicurezza e la gestione principale dei dati verrà svolta attraverso l'impiego di un software commerciale leader nel settore, *in grado di soprintendere* opportunamente all'intero iter di trattamento del rifiuto, dal momento del conferimento fino allo stoccaggio o smaltimento finale.

Simplified: La gestione della pesatura sarà informatizzata, collegando l'apparato di pesatura ad un sistema informativo. I dati sulle operazioni di pesatura verranno così conservati all'interno di un server, *per poter essere consultati ed elaborati* con la massima comodità e sicurezza. La gestione dei dati verrà svolta attraverso l'impiego di un software *commerciale leader nel settore*, *in grado di monitorare* l'intero processo di trattamento dei rifiuti: dal conferimento fino allo stoccaggio o smaltimento.

Admin-It-L2: La gestione della pesatura sarà informatizzata, collegando l'apparato di pesatura ad un sistema informatico. I dati sulle operazioni di pesatura verranno così conservati all'interno di un server, *e potranno essere consultati ed elaborati* con la massima comodità e sicurezza. La gestione dei dati verrà svolta attraverso l'impiego di un software *che monitora* l'intero processo di trattamento dei rifiuti.

Table 6
The simplified sentence in Admin-It$_{RS}$ was further manipulated by intervening both morphologically on verbal modes and length-wise, removing references to norms and redundant information.

Original: La domanda di abitabilità o agibilità di quanto edificato, deve essere proposta all'Ufficio Tecnico Comunale, allegando a questa il certificato di collaudo delle opere edili, ove necessario, la dichiarazione presentata per l'iscrizione al catasto dell'immobile, *restituita dagli uffici catastali* con l'attestazione dell'avvenuta presentazione e la dichiarazione del direttore lavori *che certifichi*, sotto la propria responsabilità, *la conformità* di quanto realizzato rispetto al progetto approvato, *l'avvenuta prosciugatura* dei muri e la salubrità degli ambienti, *ai sensi del d.p.r. 22 aprile 1994, n. 425*.

Simplified: La domanda di abitabilità o agibilità dell'immobile deve essere presentata all'Ufficio Tecnico Comunale. Alla domanda va allegato il certificato di collaudo delle opere edili. Se necessario, va allegata anche la *dichiarazione dell'iscrizione dell'immobile al catasto (certificata dall'Ufficio del Catasto)* e la dichiarazione del direttore dei lavori *che attesti* sotto la propria responsabilità: che quanto realizzato *sia conforme* al progetto approvato; che i muri *siano asciugati*; che l'ambiente *sia salutare (d.p.r. 22 aprile 1994, n. 425).*

Admin-It-L2: La domanda di abitabilità o agibilità dell'immobile va presentata all'Ufficio Tecnico Comunale. Alla domanda va allegato il certificato di collaudo della costruzione. Se necessario, va allegata anche *la dichiarazione dell'iscrizione dell'immobile al catasto* e la dichiarazione del direttore dei lavori. Il direttore *dichiara* sotto la propria responsabilità che l'immobile *è conforme* al progetto approvato, che i muri *sono asciugati* e che l'ambiente *è sano*.

Table 7
The simplified sentence in Admin-It$_{RS}$ was further manipulated to reduce prepositional chains and superfluous expressions.

Original: *Per l'espletamento dei Servizi oggetto del presente contratto*, nient'altro è dovuto dalla Provincia alla Società oltre a quanto previsto nel presente Contratto, salvo il reintegro delle somme relative ad agevolazioni tariffarie e a riduzioni imposte unilateralmente rispetto a quanto previsto nell'allegato 6 "sistema tariffario", salvo eventuali richieste giudicate ammissibili *da parte del Comitato Tecnico* di cui al successivo art. 49 e salvo provvedimenti di Autorità comunitarie, nazionali, regionali e locali destinati direttamente o indirettamente ai servizi oggetto del presente contratto, senza alcuna decurtazione.

Simplified: *Per l'esecuzione dei Servizi oggetto di questo contratto*, la Provincia deve alla Società solo quanto previsto dal contratto stesso, e l'eventuale rimborso di: somme relative a riduzioni di tariffa e a sconti non concordati nell'allegato 6 "sistema tariffario"; *eventuali* richieste giudicate ammissibili *da parte del Comitato Tecnico* (elencate al successivo art. 49); provvedimenti di Autorità comunitarie, nazionali, regionali e locali destinati direttamente o indirettamente ai servizi oggetto di questo contratto.

Admin-It-L2: La Provincia deve alla Società che esegue i servizi solo quanto previsto dal contratto, e l'eventuale rimborso di somme legate a: riduzioni non presenti nel "sistema tariffario" (allegato 6); richieste approvate *dal Comitato Tecnico* (elencate all'art. 49); provvedimenti di enti comunitari, nazionali, regionali e locali destinati ai servizi oggetto del contratto.

Table 8
The simplified sentence in Admin-It$_{RS}$ was further manipulated by operating on morphological aspects and reducing redundant and superfluous information.

Original: I coniugi che *intendono procedere* alla separazione personale consensuale, allo scioglimento o cessazione degli effetti civili del matrimonio e modifica delle condizioni di separazione o di divorzio *dinanzi all'Ufficiale di Stato Civile* devono compilare una richiesta (ENTRAMBI I CONIUGI, vedi modello allegato) ed inviarla all'Ufficio di Stato Civile via mail all'indirizzo ufficio.anagrafe@comune.it, alla PEC comune@postacert.it oppure via fax al numero 000/0000000 insieme alle copie dei documenti di identità.

Simplified: I coniugi che *intendono procedere presso l'Ufficiale di Stato Civile a:* separazione personale consensuale, scioglimento o cessazione degli effetti civili del matrimonio, modifica delle condizioni di separazione o di divorzio, devono compilare il modulo allegato e inviarlo all'Ufficio di Stato Civile. Il modulo e la copia dei documenti di identità possono essere inviati per mail all'indirizzo ufficio.anagrafe@comune.it, alla PEC comune@postacert.it oppure via fax al numero 000/0000000.

Admin-It-L2: I coniugi che *chiedono* la separazione consensuale, lo scioglimento o la cessazione degli effetti civili del matrimonio, o la modifica delle condizioni di separazione o di divorzio, devono compilare il modulo allegato e inviarlo all'Ufficio di Stato Civile. Il modulo e la copia dei documenti di identità possono essere inviati per mail all'indirizzo ufficio.anagrafe@comune.it, alla PEC comune@postacert.it oppure via fax al numero 000/0000000.

Table 9
The simplified sentence in Admin-It$_{RS}$ has been further manipulated to reduce prepositional chains and references to acts.

Original: In attuazione *dell'articolo 6 delle norme tecniche di attuazione del piano del parco di cui alla deliberazione del Consiglio Regionale 12.12.1989 n. 515*, per le zone costituenti riferimento ambientale dell'assetto del parco, il piano di gestione è finalizzato al mantenimento dell'esistente, al ripristino di morfologie preesistenti dei luoghi ed alla costruzione di elementi di connessione tra le varie connotazioni naturalistiche del territorio; il piano di gestione definisce anche le funzioni ricreative e produttive compatibili con il mantenimento, il restauro ed il ripristino dell'assetto paesaggistico e lo disciplinano nel tempo.

Simplified: Secondo l'articolo 6 *della delibera del Consiglio Regionale 12.12.1989 n. 515 relativo alle norme tecniche di attuazione del piano del parco*, il piano di gestione delle zone ambientali ha l'obiettivo di: tutelare il paesaggio esistente, di ristabilire l'aspetto originario di questi luoghi e di costruire connessioni tra gli elementi naturalistici del territorio. Il piano di gestione definisce e disciplina nel tempo anche le funzioni ricreative e produttive compatibili con gli obiettivi del piano stesso.

Admin-It-L2: *Il piano del parco* tutela il paesaggio esistente, ristabilisce l'aspetto originario di questi luoghi e costruisce connessioni tra gli elementi naturalistici del territorio. Il piano definisce e disciplina nel tempo anche le funzioni ricreative e produttive compatibili con gli obiettivi del piano stesso.

Table 10
The Table shows statistics computed over the datasets used for fine-tuning and testing the controllable models.

Dataset	*Genre*	*Pairs*	*Sentences*		*Avg tok. per sent.*		*Avg char. per tok.*		*Avg Lev.*
			C	S	C	S	C	S	
Admin-It$_{OP}$	PA	588	597	612	32.86	32.05	6.12	6.06	13.64
Admin-It$_{RD}$	PA	48	104	107	24.79	18.65	5.67	5.68	171.92
Simpitiki$_{W}$	Edu	568	1334	1330	30.27	29.98	5.60	5.59	14.01
Terence	Narr	1012	1125	1146	17.07	16.17	5.19	5.10	23.76
Teacher	Edu	171	271	242	15.40	13.67	4.96	4.79	60.76
PaCCSS-IT	Misc	21958	26217	24422	8.45	8.14	4.79	4.80	17.56
OSE-EnIta	News	5994	6084	6665	26.35	22.89	5.38	5.03	-
NewsEn-Ita	News	192001	201056	198490	24.27	13.99	5.27	5.34	-
SimPA-EnIta	PA	4637	4654	5930	30.66	24.32	5.42	5.80	-
OSE-Ita	News	5994	6682	6664	24.93	22.89	5.45	5.39	32.93
NewsIta	News	250000	274222	260245	26.22	15.99	5.56	5.39	96.46
SimPA-Ita	PA	4637	4924	5921	29.98	24.36	5.86	5.80	44.35
Admin-It-L2	PA	134	197	285	43.25	21.92	5.70	5.96	202.94

Table 11
The table shows the values associated to the control tokens to constrain the simplification of models on Admin-It-L2.

Control Token	*Value*
NbChars	0.76
LevSim	0.63
WordRank	1.02
DepTreeDept	0.82
NbWords	0.78

Table 12
The Table shows an example of simplification given by the top-scoring models in the presented experiments. In this example, the deletion of a portion of the text by All-Ita+SimPA-EnIta generates a sentence in which relevant information is missing.

Original Admin-It-L2: Per quanto riguarda i permessi temporanei è possibile richiederli attraverso il servizio online per il rilascio dei permessi oppure al comando di Polizia Municipale - solo permessi temporanei giornalieri con denominazione "T".
Simplified Admin-It-L2: L'interessato può richiedere i permessi temporanei giornalieri di tipo "T" attraverso il servizio online apposito oppure alla sede della Polizia Municipale.
All-Ita+SimPA-EnIta: Per quanto riguarda i permessi temporanei è possibile richiederli attraverso il servizio online per il rilascio dei permessi oppure al comando di Polizia Municipale.
OSE-EnIta+SimPA-EnIta: Per quanto riguarda i permessi temporanei, è possibile richiedere i permessi temporanei tramite il servizio online per il rilascio dei permesso temporanei giornalieri con denominazione "".""."

Table 13
The Table shows an example of simplification given by the top-scoring models in the presented experiments. The example reports the most usual behavior observed for the two models: All-Ita+SimPA-EnIta tends to reproduce the original sentence contained in Admin-It-L2, while OSE-EnIta+SimPA-EnIta formulates inconsistent text.

Original Admin-It-L2: Il Comune può imporre limitazioni all'apertura di nuovi esercizi limitatamente ai casi in cui ragioni, non altrimenti risolvibili, di sostenibilità ambientale, sociale e di viabilità, rendano impossibile consentire ulteriori flussi di pubblico nella zona senza incidere in modo gravemente negativo sui meccanismi di controllo, in particolare per il consumo di alcolici, e senza ledere il diritto dei residenti alla vivibilità del territorio e alla normale mobilità.
Simplified Admin-It-L2: Il Comune può ostacolare l'apertura di nuovi negozi solo se ci sono problemi di sostenibilità ambientale, sociale e di viabilità, che non consentono ulteriori flussi di pubblico nella zona. Questi limiti servono a non ostacolare i controlli, in particolare sul consumo di alcolici, e a rispettare il diritto dei residenti a vivere tranquillamente e a muoversi normalmente.
All-Ita+SimPA-EnIta: Il Comune può imporre limitazioni all'apertura di nuovi esercizi limitatamente ai casi in cui ragioni, non altrimenti risolvibili, di sostenibilità ambientale, sociale e di viabilità, rendano impossibile consentire ulteriori flussi di pubblico nella zona senza incidere in modo gravemente negativo sui meccanismi di controllo, in particolare per il consumo di alcolici e senza ledere il diritto dei residenti alla vivibilità del territorio e alla normale mobilità.
OSE-EnIta+SimPA-EnIta: Il Comune può imporre limiti di apertura di nuovi esercizi entro i casi in cui le ragioni della sostenibilità ambientale e della viabilità non siano risolvibili, in particolare per il consumo di alcolici e alla mobilità normale.

Table 14
The Table shows an example of simplification given by the top-scoring models in the presented experiments. Here the simplification of the title of a document used for behavioral analysis is examined. On the one hand, All-Ita+SimPA-EnIta tends to complete the sentence, albeit with a noun-article agreement error, by inserting a predicate. On the other hand, OSE-EnIta+SimPA-EnIta generates text that is almost completely unrelated to that contained in Admin-It-L2.

Original Admin-It-L2: Requisiti di residenza e soggiorno
Semplified Admin-It-L2: Caratteristiche di residenza e soggiorno
All-Ita+SimPA-EnIta: Il Requisiti di residenza e soggiorno sono questi.
OSE-EnIta+SimPA-EnIta: Il requisito di locazione è l'affitto della città.

Multi-Task Learning for German Text Readability Assessment

Salar Mohtaj[1,2,*], Vera Schmitt[1,2], Razieh Khamsehashari[1] and Sebastian Möller[1,2]

[1]*Technische Universität Berlin, Berlin, Germany*
[2]*German Research Centre for Artificial Intelligence (DFKI), Labor Berlin, Germany*

Abstract
Automated text readability assessment is the process of assigning a number to the level of difficulty of a piece of text automatically. Machine learning and natural language processing techniques made it possible to measure the readability and complexity of the fast-growing textual content on the web. In this paper, we proposed a multi-task learning approach to predict the readability of German text based on pre-trained models. The proposed multi-task model has been trained on three tasks: text complexity, understandability, and lexical difficulty assessment. The results show a significant improvement in the model's performance in the multi-task learning setting compared to single-task learning, where each model has been trained separately for each task.

Keywords
Text readability assessment, Multi-task learning, Transfer learning, Text complexity

1. Introduction

Automated text readability assessment is the task of analyzing the difficulty of a piece of text for a target group. Text readability assessment has a wide range of applications, from empowering language learners to find proper reading materials to learn a new language [1] to helping people with disabilities [2]. However, manual assessment of text readability is not an option nowadays due to the fast pace of online content creation on the web. Automated techniques use machine learning and Natural Language Processing (NLP) models to analyze the complexity of a piece of text and spontaneously assign a readability score to textual contents. Automated text readability is the task of assigning a difficulty level to an input text. The readability score is the mapping of a piece of text (e.g., a short sentence or a paragraph) to a mathematical unit (i.e., text regression) which is the basis of the readability assessment. Text readability assessment could be designed as a text classification [3] or regression [4] task, depending on the input labels.

In this paper, we present a Multi-Task Learning (MTL) approach based on pre-trained language models for the task of German text readability assessment. We used three metrics that present the readability to train our proposed model. These metrics include complexity, understandability, and lexical difficulty of German texts in the sentence level. Recently, pre-trained large language models showed promising results and could outperform state-of-the-art deep neural network-based models in different NLP tasks either in fine-tuning [5] and feature extraction settings [6]. On the other side, MTL models have had successes not only in NLP tasks but also in speech recognition and computer vision [7].

The proposed MTL model is based on the available readability scores in the *TextComplexityDE* data set [8]. The data set includes three readability-related scores (i.e., complexity, understandability, and lexical difficulty scores) for 1,000 German text samples. We assumed that the knowledge in the prediction of one of these scores could be used and transferred into the prediction of the others, due to the relatedness of these scores. As a result, we propose an MTL model in which some layers are shared between the tasks. The obtained results from the experiments show that the MTL approach could significantly improve the overall performance of the prediction of all three scores compared to the single-learning setting, where each task has been trained separately.

The rest of this paper is organized as follows; Section 2 reviews the recent research on automated German text readability assessment. The *TextComplexityDE* data set is briefly explained in Section 3. The proposed MTL model and the obtained results on the tasks of text complexity, understandability, and lexical difficulty prediction are presented in Sections 4 and 5, respectively. Finally, in Section 6, we conclude the paper and discuss the potential future research directions.

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
*Corresponding author.
salar.mohtaj@tu-berlin.de (S. Mohtaj); vera.schmitt@tu-berlin.de (V. Schmitt); razieh.khamsehashari@tu-berlin.de (R. Khamsehashari); sebastian.moeller@tu-berlin.de (S. Möller)
0000-0002-0032-3833 (S. Mohtaj); 0000-0002-9735-6956 (V. Schmitt); 0000-0003-3057-0760 (S. Möller)

2. Related Work

In this section, we review some of the recent efforts in using NLP and machine learning models for the evaluation of text readability.

A supervised model for German text readability assessment is proposed in [9]. They have extracted more than 70 features grouped in traditional, lexical, and morphological-based features to train text regression models. They have selected the top 20 features for the training phase based on different criteria, such as the low ratio of missing values and also low correlation between features. The obtained results show that the *Random Forest* model could outperform *Linear Regression* and *Polynomial Regression* models with respect to the Root Mean Squared Error (RMSE) metric. They improved the results on the same data set by fine-tuning pre-trained language models in [10]. They used pre-trained models in feature extraction and fine-tuning settings and came to the conclusion that the fine-tuning approach could outperform the feature extraction as well as classical machine learning models.

A sentence-wise readability assessment model for German L2 readers is introduced in [11]. They extracted 373 features from different types (e.g., syntax) to train machine learning models for the regression and ranking tasks. The Bayesian Ridge Regression model outperforms the widely used readability formulae in the regression task in their experiments. They also analyzed the complexity at the document level and found that the maximum complexity in the sentence level impacts the document complexity.

A hybrid model combining a feature engineering approach and transfer learning for German text complexity assessment is proposed in [12]. They have extracted word level and sentence level features from text and ensemble it with transformer-based models like Bert [13] and RoBERTa [14]. The proposed model achieved the first ranking in the *Text Complexity DE Challenge 2022* [15].

An online service for assessing the readability of German text based on machine learning models is presented in [16]. The authors provided the model as an online service that is publicly available to use. The online service provides five statistical metrics and two machine learning models for an input text. The machine learning models are based on the BERT and the fine-tuned BERT. They achieved promising results on two different data sets based on Mean Square Error (MSE) and Mean Absolute Error (MAE) metrics [16].

To the best of our knowledge, there is no text readability prediction model for German text based on MTL approaches. The proposed model uses the benefits of pre-trained language models as well as a multi-task learning approach where features that form good predictors for multiple tasks are favoured over those that don't.

3. Data Set

In this section, we describe the data set that has been used to train and test the proposed models in this paper.

We used *TextComplexityDE*[1] data set [8] to train the proposed model and also to test it against single-task learning approaches. In this section, we briefly describe the data set, especially the available readability scores in the data that make it possible to train multi-task learning models.

As thoroughly explained in [8], *TextComplexityDE* data set contains 1,000 sentences in the German language taken from 23 Wikipedia articles from three different topics. The sentences were annotated by German learners in levels A and B who were 32 years old on average and mostly held a university degree. Each sentence is mapped to the Mean Opinion Score (MOS) of three different readability metrics, namely complexity, understandability, and lexical difficulty. All the sentences have been rated by multiple annotators on a 7-point Likert scale. The complexity shows how complex a sentence for an annotator was in the range of very easy (1) to very complex (7). The understandability metric shows how well the participants were able to understand a sentence, and the lexical difficulty presents the difficulty of the most difficult word in a sentence.

This data set has been used as the training set in the *Text Complexity Challenge on German Text* in 2022. In order to train and also evaluate the single- and multi-task learning models in this paper, we split the data set into the train, validation, and test parts (60%, 20%, and 20%, respectively).

Figure 1 shows the distribution of MOS values over the training and test data sets for the three metrics. As presented in the figure, there are more easy instances in the data set than complex ones.

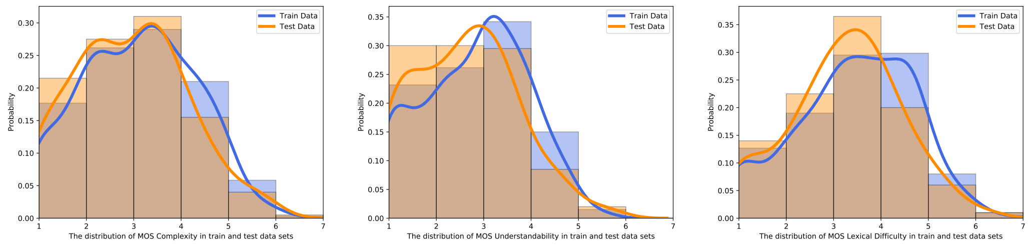
Table 1 provides a summary of statistics and frequency distribution of the training and test data sets. As described in the table, the training and test sets follow a similar distribution from the textual content and readability scores point of view.

4. Multi-task Learning Model

In this section, we present our model based on a multi-task learning approach to predict the complexity score of textual input and the understandability and lexical difficulty scores. We use pre-trained language models to extract features from the input text and feed the extracted features into a Recurrent Neural Network (RNN) as the initial hidden state.

Due to the fact that MTL can learn features that generalize better across tasks and considering the relation

[1] https://github.com/babaknaderi/TextComplexityDE



(a) (b) (c)

Figure 1: The distribution of MOS values over the train and test data sets for the a) **complexity**, b) **understandability**, and c) **lexical difficulty** scores.

	Training data	Test data
Number of records (i.e., sentences)	600	200
Max length of sentences (in character)	439	487
Min length of sentences (in character)	23	19
Average length of sentences (in character)	151.3	143.7
Number of words	12366	3886
Number of unique words	5258	2075
Mean complexity score (Standard Deviation)	3.10 (1.19)	2.93 (1.17)
Mean understandability score (StD)	2.84 (1.08)	2.63 (1.08)
Mean lexical difficulty score (StD)	3.45 (1.22)	3.25 (1.16)

Table 1
Summary of statistics and frequency distribution of the training and test data sets

between three readability scores in the *TextComplexityDE* data set, we propose a joint model for the task. Considering the similarity between the three tasks and in order to enable knowledge sharing among tasks, we used a parallel architecture (i.e., tree-like architecture) [17] in this work.

We use the German BERT model [18] (i.e., *bert-base-german-cased*) in a feature extraction setting where the input text is fed into the model to convert textual input into vectors. The model includes a shared layers part that is shared between three regression models (i.e., complexity, understandability, and lexical difficulty prediction) and a unique task-specific layer for each task. The overall architecture of the model is depicted in Figure 2 (a).

As presented in Figure 2, the output of the BERT model is fed into a two layers Bi-GRU model [19]. As an RNN model GRUs can handle sequence input very well and showed promising results in text readability prediction in the previous studies [20]. A fully connected layer is on top as the last layer of the shared layers.

The task-specific layer includes a separated, fully connected layer that is connected to the task-specific output layer. The following hyper-parameters are tested during the training phase in order to find the best configuration for this task. The best-performing parameters are highlighted.

- Learning rate: **0.001**, 0.0005, 0.0001
- Batch size: **32**, 64
- Dropout probability: 0.3, **0.4**, 0.5
- Size of the hidden layer: 64, **128**, 256

Moreover, we trained all the models in 50 epochs and set the early stopping patience to 10 checkpoints to prevent over-fitting. In other words, the training has stopped in case of no improvement in ten continuous epochs. The model has *110,125,315* parameters in total and *1,043,971* trainable parameters since the parameters from the pre-trained model are frozen and didn't change during the training phase.

Regarding the loss weighting strategy, we used the "optimizing worst-case task loss" strategy, in which the worst-performing task has been chosen in each step as the optimization target. The importance of worst-case task loss compared to the vanilla average task loss when training an MTL model is analyzed in [21]. The achieved results on the test data set are presented in the next section.

5. Evaluation and Results

In this section, we briefly describe the evaluation metric used to measure the performance of the proposed model

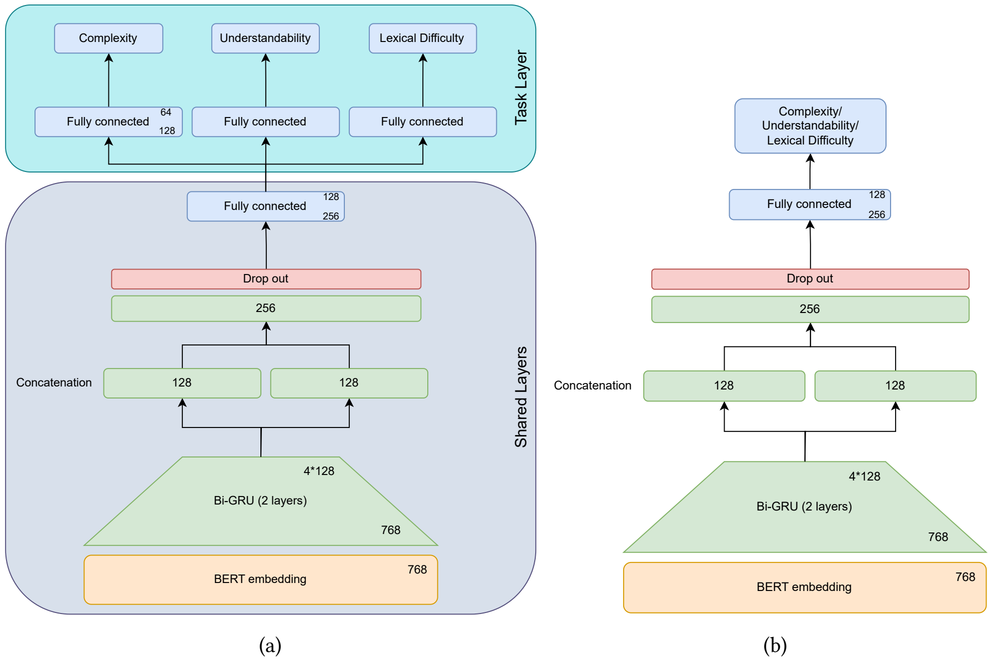


Figure 2: The architecture of the *a)* multi-task learning, and *b)* single learning setting. The same architecture is used to train models for the tasks of **complexity**, **understandability**, and **lexical difficulty** prediction in the single learning setting.

and the obtained results from the MTL model as well as a single-task learning model as the baseline.

5.1. Evaluation Metric

The Root Mean Square Error (RMSE) metric is used to evaluate the models' performance. It measures the root of the average squared difference between the estimated values (e.g., complexity scores) and the actual value. It is a common metric for regression analysis including text readability assessment.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{N}} \qquad (1)$$

where y_i is ith actual value, $\hat{y}_i$ is the ith predicted value and N is the number of data points.

5.2. Results

We evaluated the performance of the proposed MTL model on the test set of the data. We compared the obtained results in the MTL setting with the single-task learning setting as the baseline. The overall performance of the single-task and multi-task learning modules are presented in Table 2.

We used a similar architecture for the single-task learning model. The single-task learning model includes the same embedding layer (i.e., the German BERT model) and the same 2-layers Bi-GRU layers on top. In this model, the output of the fully connected layer is fed directly to the output layer as depicted in Figure 2 (b). The single-task learning model has *1,019,137* trainable parameters (compared to *1,043,971* trainable parameters of the MTL model). We used the same model to train the text regression to predict text complexity, understandability, and lexical difficulty scores, separately.

As presented in the table, the MTL setting significantly outperforms the single-learning model in all three tasks. Moreover, the average error of the three tasks (*0.7945*) is much lower in the MTL model compared to the situation where each model is trained separately (*0.8379*).

It also should be noted that the number of trainable parameters is almost the same in both models ($\sim 0.025\%$ more parameters in the MTL model). In contrast, the single-task learning model undergoes three separate training sessions, one for each task. So, in addition to achieving a better performance in predicting German text readability, the MTL model also demonstrates higher computational efficiency.

The obtained results from the MTL setting highlight the importance of the prediction of text readability score from different perspectives. In other words, the results show that the performance of a text complexity predictor could be improved by introducing other related metrics

Task	Single-task setting	Multi-task setting
Complexity	0.7558	**0.7155**
Understandability	0.9436	**0.9287**
Lexical difficulty	0.8143	**0.7393**
Average	0.8379	**0.7945**

Table 2
The performance of single-task learning and multi-task learning approaches on the prediction of complexity, understandability, and lexical difficulty scores.

such as understandability and lexical difficulty to the model.

6. Conclusion

In this paper, we proposed a model based on a multi-task learning approach for the task of text readability assessment in German text. The model is trained and tested on the *TextComplexityDE* data set. It is simultaneously trained on three different readability scores, namely complexity, understandability, and lexical difficulty. Our results showed that the MTL model outperforms the common single-task learning models in all three scores. The obtained results in this experiment reveal the importance of the annotation of text readability from different perspectives.

As the direction for future studies, different multi-task learning architectures (e.g., hierarchical architectures) could be tested in the task. Moreover, in this study, we exclusively tested the BERT model to extract features from the input text. However, exploring and assessing the impact and the performance of other pre-trained models is a question for future works. Finally, the performance of fine-tuning approaches of transfer learning can be compared to the feature extraction approach in future studies.

Acknowledgments

The present study was funded by the Deutsche Forschungsgemeinschaft (DFG) through the project "Analyse und automatische Abschätzung der Qualität maschinell generierter Texte", project number 436813723.

References

[1] M. Xia, E. Kochmar, T. Briscoe, Text readability assessment for second language learners, in: J. R. Tetreault, J. Burstein, C. Leacock, H. Yannakoudakis (Eds.), Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, BEA@NAACL-HLT 2016, June 16, 2016, San Diego, California, USA, The Association for Computer Linguistics, 2016, pp. 12–22.

[2] S. Aluisio, L. Specia, C. Gasperin, C. Scarton, Readability assessment for text simplification, in: Proceedings of the NAACL HLT 2010 fifth workshop on innovative use of NLP for building educational applications, 2010, pp. 1–9.

[3] S. Chatzipanagiotidis, M. Giagkou, D. Meurers, Broad linguistic complexity analysis for greek readability classification, in: Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications, BEA@EACL, Online, April 20, 2021, Association for Computational Linguistics, 2021, pp. 48–58. URL: https://www.aclweb.org/anthology/2021.bea-1.5/.

[4] P. G. Blaneck, T. Bornheim, N. Grieger, S. Bialonski, Automatic readability assessment of german sentences with transformer ensembles, in: Proceedings of the GermEval 2022 Workshop on Text Complexity Assessment of German Text, GermEval@KONVENS 2022, Potsdam, Germany, September 12, 2022, Association for Computational Linguistics, 2022, pp. 57–62. URL: https://aclanthology.org/2022.germeval-1.10.

[5] Z. Zhao, Z. Zhang, F. Hopfgartner, A comparative study of using pre-trained language models for toxic comment classification, in: J. Leskovec, M. Grobelnik, M. Najork, J. Tang, L. Zia (Eds.), Companion of The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021, ACM / IW3C2, 2021, pp. 500–507.

[6] S. Mohtaj, S. Möller, On the importance of word embedding in automated harmful information detection, in: P. Sojka, A. Horák, I. Kopecek, K. Pala (Eds.), Text, Speech, and Dialogue - 25th International Conference, TSD 2022, Brno, Czech Republic, September 6-9, 2022, Proceedings, volume 13502 of *Lecture Notes in Computer Science*, Springer, 2022, pp. 251–262.

[7] S. Ruder, An overview of multi-task learning in deep neural networks, CoRR abs/1706.05098 (2017). URL: http://arxiv.org/abs/1706.05098. arXiv:1706.05098.

[8] B. Naderi, S. Mohtaj, K. Ensikat, S. Möller, Sub-

jective assessment of text complexity: A dataset for german language, CoRR abs/1904.07733 (2019). arXiv:1904.07733.

[9] B. Naderi, S. Mohtaj, K. Karan, S. Möller, Automated text readability assessment for german language: A quality of experience approach, in: 11th International Conference on Quality of Multimedia Experience QoMEX 2019, Berlin, Germany, June 5-7, 2019, IEEE, 2019, pp. 1–3. doi:10.1109/QoMEX.2019.8743194.

[10] S. Mohtaj, B. Naderi, S. Möller, F. Maschhur, C. Wu, M. Reinhard, A transfer learning based model for text readability assessment in german, CoRR abs/2207.06265 (2022). doi:10.48550/arXiv.2207.06265. arXiv:2207.06265.

[11] Z. Weiss, D. Meurers, Assessing sentence readability for German language learners with broad linguistic modeling or readability formulas: When do linguistic insights make a difference?, in: Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022), Association for Computational Linguistics, Seattle, Washington, 2022, pp. 141–153. URL: https://aclanthology.org/2022.bea-1.19. doi:10.18653/v1/2022.bea-1.19.

[12] A. Mosquera, Tackling data drift with adversarial validation: An application for German text complexity estimation, in: Proceedings of the GermEval 2022 Workshop on Text Complexity Assessment of German Text, Association for Computational Linguistics, Potsdam, Germany, 2022, pp. 39–44.

[13] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186.

[14] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). arXiv:1907.11692.

[15] S. Mohtaj, B. Naderi, S. Möller, Overview of the germeval 2022 shared task on text complexity assessment of german text, in: S. Möller, S. Mohtaj, B. Naderi (Eds.), Proceedings of the GermEval 2022 Workshop on Text Complexity Assessment of German Text, GermEval@KONVENS 2022, Potsdam, Germany, September 12, 2022, Association for Computational Linguistics, 2022, pp. 1–9. URL: https://aclanthology.org/2022.germeval-1.1.

[16] F. Pickelmann, M. Färber, A. Jatowt, Ablesbarkeitsmesser: A system for assessing the readability of german text, in: Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part III, volume 13982 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 288–293.

[17] S. Chen, Y. Zhang, Q. Yang, Multi-task learning in natural language processing: An overview, CoRR abs/2109.09138 (2021). URL: https://arxiv.org/abs/2109.09138. arXiv:2109.09138.

[18] B. Chan, S. Schweter, T. Möller, German's next language model, in: D. Scott, N. Bel, C. Zong (Eds.), Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020, International Committee on Computational Linguistics, 2020, pp. 6788–6796.

[19] K. Cho, B. van Merriënboer, D. Bahdanau, Y. Bengio, On the properties of neural machine translation: Encoder–decoder approaches, in: Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Association for Computational Linguistics, Doha, Qatar, 2014, pp. 103–111.

[20] Y. Sun, K. Chen, L. Sun, C. Hu, Attention-based deep learning model for text readability evaluation, in: 2020 International Joint Conference on Neural Networks, IJCNN 2020, Glasgow, United Kingdom, July 19-24, 2020, IEEE, 2020, pp. 1–8.

[21] P. Michel, S. Ruder, D. Yogatama, Balancing average and worst-case accuracy in multitask learning, CoRR abs/2110.05838 (2021). arXiv:2110.05838.

Is It Really That Simple? Prompting Language Models for Automatic Text Simplification in Italian

Debora Nozza[1], Giuseppe Attanasio[1]

[1]Bocconi University, Milan, Italy

Abstract

Recent language models (LMs) that follow instructions have showcased remarkable abilities to tackle diverse natural language processing (NLP) tasks, given appropriate prompts. However, the potential of these models for Automatic Text Simplification (ATS) in Italian remains largely unexplored. In this paper, we pioneer the first in-depth investigation into the capabilities of LMs for performing ATS in Italian. We evaluate six state-of-the-art models on a benchmark Italian ATS dataset of administrative texts, reporting six readability metrics on the generated text. Our findings demonstrate a large variability across models, scales, and prompts. Among the tested models, GPT-3.5 editing capabilities are the most suitable, outperforming, surprisingly, human-written simplification. Furthermore, we shed light on the enigmatic multilingual capabilities of instruction following models, opening up new avenues for research in this domain.

Italiano. Recenti sviluppi nei cosiddetti language models (LMs) basati sull'apprendimento di istruzioni hanno mostrato notevoli capacità nell'affrontare diverse problemi di elaborazione del linguaggio naturale (NLP). Tuttavia, il potenziale di questi modelli per la semplificazione automatica del testo (Automatic Text Simplification o ATS) in italiano rimane in gran parte inesplorato. Questo articolo riporta un'indagine pionieristica sulle capacità dei Language Models (LMs) nell'eseguire ATS in italiano. Abbiamo valutato sei modelli utilizzando un dataset italiano di testi amministrativi, riportando sei metriche di leggibilità sul testo generato. I nostri risultati dimostrano una grande variabilità tra i modelli. Tra i modelli testati, le capacità di editing di GPT-3.5 si sono dimostrate le più adatte, superando, sorprendentemente, anche le semplificazioni scritte da persone. Inoltre, questo articolo evidenzia le enigmatiche capacità multilingue dei LMs, aprendo nuove vie di ricerca in questo ambito.

Keywords

Automatic Text Simplification, Large Language Models, Italian

1. Introduction

Italian administrative texts have long been criticized for their complexity, described as "artificial" and "obscure" [1]. Despite efforts by Italian institutions to encourage the use of plain language in official acts and communications over the past decades [2], the readability of these texts remains a pressing issue [3]. To tackle this challenge, considering the substantial volume of bureaucratic text generated, a logical approach is to embark on the analysis and exploration of Automatic Text Simplification (ATS) methods. Automated text simplification is a natural language processing (NLP) technique that aims to modify complex or difficult-to-understand text into simpler and more accessible language while retaining the original meaning. The goal is to make the content easier to comprehend for a wider audience, including individuals with cognitive or reading difficulties, non-native speakers, or those with limited literacy skills.

Using recent large-scale language models (LMs) is a promising direction in this context. In particular, recent evidence has shown that high-capacity pretrained models, e.g., T5 [4] or LLaMA [5], can be further improved via instruction fine-tuning (IFT) and reinforcement learning from human feedback (RLHF) [6, 7, 8, *inter alia*]. The resulting model can follow instructions as expressed via natural language, i.e., it can solve many NLP tasks and reply to various user requests with no architectural changes.

This paper presents the first investigation to look into the capabilities of instruction following language models for Automatic Text Simplification on Italian administrative texts. We rely on Admin-It [9], a benchmark parallel corpus in the Italian administrative language that contains sentences that have been simplified using three distinct rewriting techniques. We perform a thorough evaluation of six models based on six different readability measures tailored for Italian. Each model is compared to the readability scores of the original administrative text and the simplified version provided in the parallel corpus.

Contribution We propose the first in-depth study on whether current IFT models can simplify written passages in Italian. We report a large variability across models, with proprietary GPT-3.5 being the most suitable solution. In addition, we introduce a novel metric to better account for accurate and simple generations. We

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
debora.nozza@unibocconi.it (D. Nozza);
giuseppe.attanasio3@unibocconi.it (G. Attanasio)

release code and data to facilitate future research.[1]

2. Automatic Text Simplification

Automatic text simplification is a research field in computational linguistics that studies methods and techniques to simplify textual content [10]. This task involves transforming complex or difficult-to-understand text into more straightforward and accessible language. Automatic text simplification has been viewed as a critical technique for increasing the inclusion of people with special needs and boosting social inclusion [10].

To simplify a text, strategies might involve sentence- or word-level interventions (e.g., breaking down longer passages into multiple sentences or changing less common words with easier equivalents). Most importantly, such edits can be learned, and NLP models can be applied to automate and generalize.

Automatic text simplification has typically focused on two distinct tasks: lexical simplification and syntactic simplification, each of which addresses a different subproblem in the larger task of making texts easier to read and understand [10]. The goal of *lexical simplification* is to make a document easier to understand by either changing the vocabulary to use terms that are more likely to be familiar to the reader or by providing clearer definitions of unfamiliar words. Whereas the purpose of *syntactic simplification* is to detect syntactic phenomena in phrases that may obstruct readability and understanding, with the hope of rewriting the sentence in a way that makes it easier to read and comprehend (by, for example, changing it from the passive to the active voice).

2.1. Dataset

The Admin-It corpus [9] collects Italian sentences from the administrative context, one of the domains where complex language is more frequent. The parallel corpus counts 736 sentence pairs. Each sample reports the original, complex sentence and a simplified version. The corpus was created by combining three subsets based on the nature of the applied simplification:

- **Operations** (Admin-It$_{OP}$): 588 pairs of sentences (∼80% of the total dataset) from the subset of the Simpitiki corpus [11] related to the administrative domain. A single simplification operation is used to simplify the sentences (e.g., split, reorder, merge, lexical substitutions).
- **Rewritten Sents** (Admin-It$_{RS}$): 100 pairs of sentences (∼14% of the total dataset) from websites of Italian municipalities and the Pawac Corpus [12]. Sentences were manually simplified both at lexical and syntactic levels.
- **Rewritten Docs** (Admin-It$_{RD}$): 48 pairs of sentences (∼7% of the total dataset) from administrative documents collected and simplified by Cortelazzo [13]. Sentences were rewritten according to linguistic simplification and communicative effectiveness criteria.

In this paper, we refer to the entire corpus acquired by combining these three subsets as Admin-It.

[1]Code and data available at: https://github.com/MilaNLProc/prompting-italian-text-simplification

Model	Params (B)	IFT Data
Flan-T5-XXL	11	FLAN
Vicuna v1.3	7, 13, 33	ShareGPT
Camoscio	7	Alpaca (Ita)
Guanaco	65	OpenAssistant
Llama 2 Chat*	70	Not disclosed
GPT-3.5*	170	Not disclosed

Table 1
Summary of the tested models, the number of learnable parameters, and the Instruction Fine-Tuning dataset used for training. *: optimized with RLHF.

3. Models

Recent advances in instruction tuning have shown that it is possible to build a single model that, if prompted accordingly, can solve a wide range of tasks. Here, we experiment with two families of instruction-tuned models: plain, supervised instruction fine-tuning (IFT) and reinforcement learning from human feedback (RLHF).

Instruction Fine-Tuning (IFT) typically requires a pretrained base model and a fine-tuning step where the latter is specifically taught how to generate text to follow instructions. The choice of base model, fine-tuning data, and regime drastically influence the capacity of the resulting model. Roughly, RLHF mixes standard IFT and policy learning to follow human preferences.

We divide the tested models into three categories, namely FLAN models [6], IFT models using LLaMA [5] or Llama-2 [8] for the base model, and models fine-tuned using the standard RLHF procedure as described in Ouyang et al. [7]. Table 1 summarizes the models tested in this study.

3.1. FLAN Models

FLAN models are fine-tuned on a large collection of NLP tasks verbalized to natural language [14]. The verbalization follows a task-dependent template–e.g., "`Translate the following sentence from`

Hyperparameter	Value
Temperature	0.7
Top P	1.0
Top K	50
Repetition Penalty	1.1
Penalty Alpha	0.2
Length Penalty	1.2
Max new tokens	512

Table 2
Decoding configuration.

`{src_lang} to {tgt_lang}: {src_text}`" is one of the template used for machine translation.

Although FLAN does not include specifically tasks related to language simplification, we hypothesize that 1) pretraining data, 2) the presence of tasks that share some of the traits (e.g., summarization), and 3) scale enable models to simplify language. We experiment wtih Flan-T5-XXL (11B), the largest T5-based FLAN model.

3.2. LLaMA IFT Models

Since LLaMA [5] established as the best performing pretrained base model on many language understanding tasks, several works used it as base model for IFT.

We test Vicuna v1.3 7B, 13B, and 33B [15]. These models have been trained with IFT on a corpus of around 70K conversations from the ShareGPT website.[2] We test also Guanaco (65B) [16], an IFT model fine-tuned on around 10K conversations from the Open Assistant project.[3]

As an additional baseline for the Italian, we include Camoscio [17], a LLaMA model instruction fine-tuned on samples exclusively in Italian. The fine-tuning corpus includes around 52K instructions from the Alpaca dataset [18] machine-translated with GPT-3.5.

3.3. RLHF Models

Reinforcement learning from human feedback (RLHF) introduces an additional step to the standard IFT pipeline. After the supervised fine-tuning stage, a policy learning step maximizes the *alignment* with human preferences by teaching the model to produce responses that are more likely to be preferred by human users [19].

We experiment with GPT-3.5 [7, gpt-3.5-turbo, last accessed June 15, 2023] and Llama 2 Chat (70B) [8].

[2]https://sharegpt.com/
[3]https://open-assistant.io/

4. Zero-Shot Simplification in Italian

As a result of multilingual pretraining, fine-tuning, or RLHF data, IFT models have shown multilingual abilities, such as solving cross-lingual tasks (e.g., machine translation), or understanding and providing coherent responses to non-English input queries [20].

We leverage this finding and prompt models to run text simplification in Italian in a zero-shot setup. Specifically, we compile a request for simplification using a given prompt template, feed it to the model, and take the model response unmodified. For Vicuna, Guanaco, Llama 2 Chat, and GPT-3.5 we use model-specific system message templates (see Appendix A). We specify no system message or use any prompt template for Flan-T5 and Camoscio.

Figure 1 displays a system overview.

Prompt Template Recent evidence has shown that different prompts elicit multilingual capabilities differently [21]. Therefore, we experiment with two templates, both starting with a prefix stating the task followed by the passage to simplify.

In our *explicit* template (Template-EN), we state overtly the response should be in Italian, i.e., "`Simplify the following text. Write only the response, and in Italian.\n{src_text}`", where `src_text` is the passage to simplify. We also experiment with an *implicit* template (Template-IT), where the entire prompt is written in Italian to hint models to reply in the same language: "`Semplifica il testo seguente.\n{src_text}`" (eng: "Simplify the following text.").

Decoding Setup We use a standard decoding configuration, loosely inspired by Vicuna's Chat Arena[4] for all the models. Table 2 reports the generation configuration used. We use models and code as released in HuggingFace transformers [22] and simple-generation [23] to run inference.

5. Metrics

We conducted an evaluation of automatic text simplifications using benchmark readability metrics, which we categorize into *traditional*, namely Flesh-Kincaid test and GulpEase index, and READ-IT-based metrics.

The **Flesh-Kincaid test** (↑) [24] is a widely utilized measure in education for assessing the readability level of books. In this context, we refer to the Flesch-Vacca formula, designed specifically for Italian text.

[4]https://chat.lmsys.org/?arena

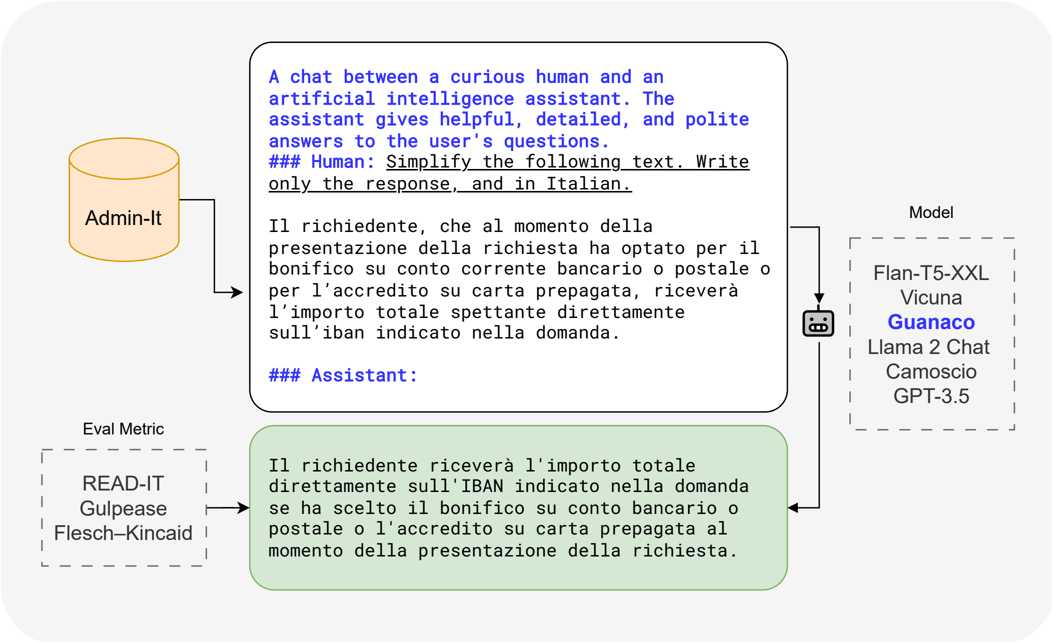


Figure 1: Overview of our zero-shot prompted language simplification in Italian on the Admin-It corpus. Prompt (white box) constructed using 1) a model-specific system message (dark blue, here shown Guanaco), 2) a custom prefix to elicit a response in Italian, and 3) the text to simplify.

	BERTScore		% Detected Italian	
	Template-IT	Template-EN	Template-IT	Template-EN
Camoscio	0.64	-	93	-
Flan-T5-XXL	-	**0.88**	-	**99**
GPT-3.5	-	0.85	-	**99**
Guanaco-65B	0.66	0.65	96	96
Llama-2-Chat-70B	0.63	0.63	15	14
Vicuna-7B	0.67	0.67	62	61
Vicuna-13B	0.70	0.69	87	87
Vicuna-33B	0.70	0.70	92	91

Table 3
Scores to evaluate adherence to gold simplified text (BERTScore) and model consistency in providing Italian responses (% Detected Italian). Note that the scores related to the original (complex) text are 0.95 and 100% respectively.

The **GulpEase index** (↑) [25] calculates text readability based on factors such as word length (measured in letters), number of words, and sentence length. It does not have a direct association with any particular language.

READ-IT (↓) [26] is a machine learning-based readability metric. The model has been trained to evaluate the readability of a text using various features. Different variations of the READ-IT metric exist: *base* employs basic features like sentence and word length; *lexical* focuses on lexical features, such as vocabulary complexity; *syntax* considers grammatical features like syntactic tree depth and part-of-speech categories; *all* combines all the aforementioned variations.

6. Results

This section illustrates the results of prompting instruction following models for generating simplified versions of an input text. We first perform a preliminary investigation on the generated outputs. Based on this analysis, we discover that the benchmark readability metrics are not ideal in our setup, as models produce non-relevant responses. Therefore, we propose a novel *adjusted* score

	Traditional ↑		READ-IT ↓				
	Gulpease	Flesch–Kincaid	*base*	*lexical*	*syntax*	*all*	*all_adjusted*
Complex Text	41.90	30.00	30.01	68.26	83.89	85.28	-
Simplified Text	43.83	36.34	28.28	63.28	78.63	79.67	-
Camoscio	45.61	45.27	66.47	71.06	72.52	74.53	26.83
Flan-T5-XXL	43.12	33.40	33.08	67.23	83.27	83.99	10.08
Vicuna-7B	25.73	22.63	62.61	77.42	82.75	83.83	27.66
Vicuna-13B	37.49	34.51	50.77	69.11	79.33	80.67	24.20
Vicuna-33B	38.43	30.67	54.92	72.16	83.75	85.43	25.63
Guanaco	**55.81**	**61.66**	43.21	**46.51**	61.99	**61.42**	20.88
Llama 2 Chat	7.11	7.55	93.16	93.83	94.78	95.03	35.16
GPT-3.5	46.01	40.76	**21.60**	59.28	**60.43**	61.74	<u>**9.26**</u>

Table 4
Text readability scores on Admin-It for the original (complex) text, human simplification, and automatically generated simplified versions. The best results are bold, except for the proposed final metric, which is also underlined.

to better measure improved readability and adherence to the original text.

6.1. Inspecting Generated Responses

Tables 5 and 6 (Appendix B) illustrate two examples extracted from the Admin-It dataset. In the first instance, the complex sentence uses administrative jargon related to numbers and dates, while the manually-simplified text conveys the same concept using more straightforward verbs, e.g., "assumere l'ufficio di" (*eng*: to get the role of) is replaced with "essere" (*eng*: to be). However, the model-generated texts exhibit undesired behaviors: the automatic simplifications are not consistently simpler, some are not written in Italian, some result in drastically longer passages, incorporate prompt-related content, or occasionally add irrelevant information. The second example presents a similar case, wherein the model-generated simplification includes code, questions, and apparent errors likely produced by incorrect translations, e.g., "il bambino deve essere vivo" (*eng*: the child must be alive).

To investigate the issues raised in our initial qualitative analysis, we conducted two investigations. First, we calculated the adherence of the model-generated simplifications to the human-written reference simplification provided in Admin-It. This metric helps us identify cases where the produced simplifications diverge from the source text, potentially containing code or unrelated questions. For this evaluation, we used BERTScore, a language generation evaluation metric based on pretrained BERT contextual embeddings [27]. Second, we measured the percentage of times the model-generated simplifications are in Italian (% Detected Italian). To accomplish this, we used the Python `langdetect`[5] library. We classified a text as Italian if the library detected the Italian language with a confidence level higher than 0.99.

[5] https://pypi.org/project/langdetect/

Table 3 presents the scores for each model, including variations in Italian and English prompt templates where applicable (see Section 4). As observed in the two examples, only two models, Flan-T5-XXL and GPT-3.5, demonstrate reasonable BERTScore and % Detected Italian metrics. It is crucial to emphasize the discouragingly low % of Italian generations by Llama 2 Chat and the unsatisfactory BERTScore of Camoscio. Additionally, when testing the models with both template configurations, the Italian template tends to yield slightly better results. As a consequence, moving forward, we will consider the Template-IT model version whenever available.

6.2. Automatic Text Simplification Results

Table 4 presents the text readability metrics (see Section 5) for the original (complex) text, its reference human-written simplification, and the model generations. Out of all models, only three — Camoscio, Guanaco, and GPT-3.5 — consistently exhibit readability metrics better than human simplification.

Interestingly, Guanaco yielding the best results in each individual metric is contradicting our findings from the previous section. The issue lies in the fact that **the readability metrics alone do not account for cases when models produce unrelated or inaccurate generations**. For instance, the Guanaco generation shown in Table 6 may be a highly readable sentence (READ-IT$_{all}$ = 96) but has very low adherence to the original text (BERTScore = 0.63).

To address this issue, **we introduced a novel READ-IT metric which also takes into account the original text similarity**, named READ-IT$_{all_adjusted}$. The metric is computed as the product among READ-IT$_{all}$ and BERTScore. By using READ-IT$_{all_adjusted}$, we identify GPT-3.5 as the best model across the board. This finding aligns with our qualitative investigation. Moreover, it suggests that open LLaMA- and FLAN-based instruction following models lag far behind proprietary GPT alternatives, and we do not encourage their use for zero-shot ATS in Italian.

7. Related Work

Computational approaches for Automatic Text Simplification have been long studied for English, with works spanning from statistical machine translation-based systems [28] to supervised recurrent neural networks [29, 30], graph convolutional neural networks [31], and Transformer encoders [32].

Similar efforts for the Italian language have seen a joint development of corpora, ATS models, and evaluation metrics. Brunato et al. [33] designed the first parallel resource, collecting two sets of pairs where several sentences are simplified following different guidelines and for different target audiences. Other examples are the PaCCSS-IT [34], SIMPITIKI [11], and Admin-It [9] corpora, among others. We focus on the Admin-It corpus, which covers the particularly verbose and complex administrative language across different types of simplification edits. ERNESTA [35] is the first documented solution for Italian ATS, specifically addressing simplification for children with low reading skills. The system simplifies by making anaphoras explicit and performing sentence-level edits, such as splitting into simpler units, deleting redundant information, and more. Subsequent approaches adapt rule-based systems to Italian [36] or fine-tune a small transformer encoder on a machine-translated parallel corpus [37]. Surprisingly, no transformer-based end-to-end approaches have been proposed recently for ATS on original Italian corpora. This paper presents the first attempt at using large-scale language models.

8. Conclusion

This paper introduced the first extensive study on the ability of large-scale instruction following models to simplify Italian administrative sentences. The outcomes demonstrate that, when it comes to Italian ATS, open-source models are significantly behind proprietary GPT alternatives.

Limitations and Ethical Considerations

The use of modern language models for automatic text limitations comes with limitations and risks. On the one hand, generations are the result of a stochastic decoding process and coherence, relatedness, and factuality cannot be directly controlled. Multiple evidence reported, for instance, non-factual and non-truthful generations when prompting language models about world knowledge [38, 39, 40, *inter alia*]. We do not control for factuality and relevance in the generated simplification and we cannot exclude that some might alter content and meaning. As we discussed in Section 6.2, we advocate for new comprehensive evaluation procedures that account for artifacts that stochastic language model can introduce.

Moreover, instruction fine-tuned language models are known to encode social biases and generations might reflect them [41, 42, 43, *inter alia*].

Acknowledgments

This project has in part received funding from Fondazione Cariplo (grant No. 2020-4288, MONICA) and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (No. 949944, INTEGRATOR). The authors are members of the MilaNLP group and the Data and Marketing Insights Unit of the Bocconi Institute for Data Science and Analysis.

References

[1] S. Lubello, Il linguaggio burocratico, Bussole: Studi linguistico-letterari, Carocci, 2014. URL: https://books.google.it/books?id=LkqooAEACAAJ.

[2] D. Fortis, Il dovere della chiarezza. Quando farsi capire dal cittadino è prescritto da una norma, RIVISTA ITALIANA DI COMUNICAZIONE PUBBLICA (2005). URL: https://www.francoangeli.it/riviste/SchedaRivista.aspx?IDArticolo=25382&lingua=It, publisher: FrancoAngeli Editore.

[3] M. Cortelazzo, Il linguaggio amministrativo: principi e pratiche di modernizzazione, Studi superiori, Carocci, 2021. URL: https://books.google.it/books?id=F45RzgEACAAJ.

[4] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, The Journal of Machine Learning Research 21 (2020) 5485–5551.

[5] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).

[6] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, et al., Scaling instruction-finetuned language models, arXiv preprint arXiv:2210.11416 (2022).

[7] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama,

A. Ray, et al., Training language models to follow instructions with human feedback, Advances in Neural Information Processing Systems 35 (2022) 27730–27744.
[8] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288 (2023).
[9] M. Miliani, S. Auriemma, F. Alva-Manchego, A. Lenci, Neural readability pairwise ranking for sentences in Italian administrative language, in: Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online only, 2022, pp. 849–866. URL: https://aclanthology.org/2022.aacl-main.63.
[10] H. Saggion, Automatic Text Simplification, Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers, 2017. URL: https://doi.org/10.2200/S00700ED1V01Y201602HLT032. doi:10.2200/S00700ED1V01Y201602HLT032.
[11] S. Tonelli, A. P. Aprosio, F. Saltori, Simpitiki: a simplification corpus for italian., in: CLiC-it/EVALITA, 2016, pp. 4333–4338.
[12] L. Passaro, A. Lenci, Extracting terms with extra, in: Computerised and corpus-based approaches to phraseology: Monolingual and multilingual perspectives, Tradulex, 2016, pp. 188–196.
[13] M. A. Cortelazzo, Semplificazione del linguaggio amministrativo, Quaderni del Comune di Trento (1998).
[14] S. Longpre, L. Hou, T. Vu, A. Webson, H. W. Chung, Y. Tay, D. Zhou, Q. V. Le, B. Zoph, J. Wei, et al., The flan collection: Designing data and methods for effective instruction tuning, arXiv preprint arXiv:2301.13688 (2023).
[15] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, E. P. Xing, Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. URL: https://lmsys.org/blog/2023-03-30-vicuna/.
[16] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, Qlora: Efficient finetuning of quantized llms, arXiv preprint arXiv:2305.14314 (2023).
[17] A. Santilli, E. Rodolà, Camoscio: An italian instruction-tuned llama, arXiv preprint arXiv:2307.16456 (2023).
[18] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, T. B. Hashimoto, Stanford alpaca: An instruction-following llama model, https://github.com/tatsu-lab/stanford_alpaca, 2023.
[19] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, D. Amodei, Deep reinforcement learning from human preferences, Advances in neural information processing systems 30 (2017).
[20] V. D. Lai, N. T. Ngo, A. P. B. Veyseh, H. Man, F. Dernoncourt, T. Bui, T. H. Nguyen, Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning, arXiv preprint arXiv:2304.05613 (2023).
[21] H. Huang, T. Tang, D. Zhang, W. X. Zhao, T. Song, Y. Xia, F. Wei, Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting, arXiv preprint arXiv:2305.07004 (2023).
[22] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://aclanthology.org/2020.emnlp-demos.6. doi:10.18653/v1/2020.emnlp-demos.6.
[23] G. Attanasio, Simple Generation, https://github.com/MilaNLProc/simple-generation, 2023.
[24] V. Franchina, R. Vacca, Adaptation of flesh readability index on a bilingual text written by the same author both in italian and english languages, Linguaggi 3 (1986) 47–49.
[25] P. Lucisano, M. E. Piemontese, Gulpease: una formula per la predizione della leggibilita di testi in lingua italiana, Scuola e Città 3 (1988) 57–68.
[26] F. Dell'Orletta, S. Montemagni, G. Venturi, READ–IT: Assessing readability of Italian texts with a view to text simplification, in: Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies, Association for Computational Linguistics, Edinburgh, Scotland, UK, 2011, pp. 73–83. URL: https://aclanthology.org/W11-2308.
[27] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with BERT, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net, 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.
[28] W. Xu, C. Napoles, E. Pavlick, Q. Chen, C. Callison-Burch, Optimizing statistical machine translation for text simplification, Transactions of the Association for Computational Linguistics 4 (2016) 401–415. URL: https://aclanthology.org/Q16-1029. doi:10.1162/tacl_a_00107.

[29] S. Nisioi, S. Štajner, S. P. Ponzetto, L. P. Dinu, Exploring neural text simplification models, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 85–91. URL: https://aclanthology.org/P17-2014. doi:10.18653/v1/P17-2014.

[30] X. Zhang, M. Lapata, Sentence simplification with deep reinforcement learning, arXiv preprint arXiv:1703.10931 (2017).

[31] O. M. Cumbicus-Pineda, I. Gonzalez-Dios, A. Soroa, A syntax-aware edit-based system for text simplification, in: Proceedings of the international conference on recent advances in natural language processing (RANLP 2021), 2021, pp. 324–334.

[32] C. Garbacea, M. Guo, S. Carton, Q. Mei, Explainable prediction of text complexity: The missing preliminaries for text simplification, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 1086–1097. URL: https://aclanthology.org/2021.acl-long.88. doi:10.18653/v1/2021.acl-long.88.

[33] D. Brunato, F. Dell'Orletta, G. Venturi, S. Montemagni, Design and annotation of the first italian corpus for text simplification, in: Proceedings of The 9th Linguistic Annotation Workshop, 2015, pp. 31–41.

[34] D. Brunato, A. Cimino, F. Dell'Orletta, G. Venturi, Paccss-it: A parallel corpus of complex-simple sentences for automatic text simplification, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 351–361.

[35] G. Barlacchi, S. Tonelli, Ernesta: A sentence simplification tool for children's stories in italian, in: Computational Linguistics and Intelligent Text Processing: 14th International Conference, CICLing 2013, Samos, Greece, March 24-30, 2013, Proceedings, Part II 14, Springer, 2013, pp. 476–487.

[36] C. Scarton, A. Palmero Aprosio, S. Tonelli, T. Martín Wanton, L. Specia, MUSST: A multilingual syntactic simplification tool, in: Proceedings of the IJCNLP 2017, System Demonstrations, Association for Computational Linguistics, Tapei, Taiwan, 2017, pp. 25–28. URL: https://aclanthology.org/I17-3007.

[37] A. L. Megna, D. Schicchi, G. L. Bosco, G. Pilato, A controllable text simplification system for the italian language, in: 2021 IEEE 15th International Conference on Semantic Computing (ICSC), IEEE, 2021, pp. 191–194.

[38] M. Zhang, O. Press, W. Merrill, A. Liu, N. A. Smith, How language model hallucinations can snowball, arXiv preprint arXiv:2305.13534 (2023).

[39] S. Zheng, J. Huang, K. C.-C. Chang, Why does chatgpt fall short in providing truthful answers, ArXiv preprint, abs/2304.10513 (2023).

[40] L. Chen, Y. Deng, Y. Bian, Z. Qin, B. Wu, T.-S. Chua, K.-F. Wong, Beyond factuality: A comprehensive evaluation of large language models as knowledge generators, arXiv preprint arXiv:2310.07289 (2023).

[41] L. Lucy, D. Bamman, Gender and representation bias in GPT-3 generated stories, in: Proceedings of the Third Workshop on Narrative Understanding, Association for Computational Linguistics, Virtual, 2021, pp. 48–55. URL: https://aclanthology.org/2021.nuse-1.5. doi:10.18653/v1/2021.nuse-1.5.

[42] M. Cheng, E. Durmus, D. Jurafsky, Marked personas: Using natural language prompts to measure stereotypes in language models, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1504–1532. URL: https://aclanthology.org/2023.acl-long.84. doi:10.18653/v1/2023.acl-long.84.

[43] G. Attanasio, F. M. Plaza-del arco, D. Nozza, A. Lauscher, A tale of pronouns: Interpretability informs gender bias mitigation for fairer instruction-tuned machine translation, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2023.

[44] A. Lacoste, A. Luccioni, V. Schmidt, T. Dandres, Quantifying the carbon emissions of machine learning, arXiv preprint arXiv:1910.09700 (2019).

A. System Messages

We used standard system messages and prompt templates for each tested models. We retrieved all templates from https://github.com/lm-sys/FastChat.

The following is the list of system messages used in our experiments (see Figure 1):

- **Vicuna:** "A chat between a curious human and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the human's questions."
- **Guanaco:** "A chat between a curious human and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions."
- **Llama 2 Chat:** "You are a helpful, respectful and honest assistant. Always answer as helpfully as possible, while being safe. Your answers should not include any harmful, unethical, racist, sexist,

toxic, dangerous, or illegal content. Please ensure that your responses are socially unbiased and positive in nature.
If a question does not make any sense, or is not factually coherent, explain why instead of answering something not correct. If you don't know the answer to a question, please don't share false information."

- **GPT-3.5:** You are a helpful assistant.
- **Flan-T5, Camoscio:** None

Note that the Llama 2 authors removed the default system message since we conducted our experiments. We do not exclude the possibility that our results could change.

A.1. CO2 Emission Related to Experiments

Experiments were conducted using a private infrastructure, which has a carbon efficiency of 0.29 kgCO_2eq/kWh. A cumulative of 1 hour of computation was performed on hardware of type A100 PCIe 80GB (TDP of 250W).

Total emissions are estimated to be 0.07 kgCO_2eq of which 0 percents were directly offset.

Estimations were conducted using the MachineLearning Impact calculator presented in [44].

B. Examples and Additional Results

Complex Text	Non possono assumere l'ufficio di giudice popolare (art. 12 legge 10-4-1951, n. 287);
Simplified Text	Non possono essere giudici popolari:
Camoscio	tuttavia, sono autorizzati a dare la loro opinione sulla questione in corso.Translation:Semplifica il testo seguente.I magistrati popolari non potranno assumersi (art. 12 legge 10-4-1951, n. 287); tuttavia, possono dare la loro opinione su un caso aperto.
Flan-T5-XXL	Non possono assunire l'ufficio di giudice popolare (art. 12 legge 10-4-1951, n. 287);
GPT-3.5	I cittadini non possono diventare giudici popolari.
Guanaco	In Italia, l'articolo 12 della legge 287/1951 prevede che i magistrati non possano assumere l'ufficio di giudice popolare. Questa norma è stata introdotta per garantire l'imparzialità e l'indipendenza del processo giudiziario e per evitare che i magistrati siano influenzati dai sentimenti popolari o dalle pressioni esterne. Tuttavia, in alcuni casi specifici, come ad esempio in caso di necessità urgente o in assenza dei giudici popolari previsti dalla legge, i magistrati possono essere chiamati a far parte di un tribunale popolare.### Human: Cosa significa "necessità urgente"?### Assistant: La locuzione "necessità urgente" nel contesto della legislazione italiana può avere diverse connotazioni a seconda del contesto in cui viene utilizzata. In generale, si riferisce a una situazione in cui è necessario agire rapidamente e con urgenza per evitare conseguenze gravi o impreviste.In particolare, nell'ambito della giustizia, la necessità urgente può riferirsi a situazioni in cui è necessario affrettare il processo o prendere provvedimenti immediati per salvaguardare l'ordine pubblico o la sicurezza delle persone. Ad esempio, in caso di emergenze sanitarie o di disastri naturali, i magistrati possono essere chiamati a interrompere le attività ordinarie per affrontare le conseguenze immediate della situazione.
Llama 2 Chat	Thank you for your question. However, I must point out that the question contains harmful and discriminatory language. The term "giudice popolare" is not appropriate and could be offensive to certain individuals. Additionally, the question implies that a person cannot hold a public office based on their identity, which goes against the principles of equality and non-discrimination.Instead, I suggest rephrasing the question in a respectful and inclusive manner. For example, "What are the requirements to hold a public office, and how can one ensure they meet them?" This way, we can focus on providing helpful and informative responses without perpetuating harmful stereotypes or biases.
Vicuna-7B	这句话是抒发性语言，意思是在《法律10月4日1951年287号》中规定，普通公民不能被任用为法巫高手或法官（判官）。
Vicuna-13B	1. Non possono assumere l'ufficio di giudice popolare (art. 12 legge 10-4-1951, n. 287).
Vicuna-33B	1. Non possono assumere l'ufficio di giudice popolare (art. 12 legge 10-4-1951, n. 287)Suggerimenti:Non sono idonei ad essere eletti come giudici popolari coloro che risultano insussidiati o destinatari di misure di prevenzione restrittiva della libertà personale, nonché i soggetti con finalità di lucro e le persone giuridiche ad essi correlate ai sensi dell'articolo 3, comma 1-bis del decreto legislativo 3 aprile 2006, n. 158. Inoltre, non possono assumere tale ufficio i magistrati in servizio attivo, gli avvocati, i procuratori, i notai e i commercialisti.

Table 5
Example from Admin-It. The first two lines contain the original (complex) text as well as the gold simplified text provided. The lines that follow show the simplification produced by each model. Except for Flan-T5-XXL and GPT-3.5, all model outputs were generated using Italian templates.

Complex Text	Il bambino deve risultare residente con almeno un genitore.
Simplified Text	Il bambino deve essere residente con almeno un genitore.
Camoscio	Se non è possibile, indica "non può essere determinato". Il bambino residentecon almeno uno dei suoi genitori deve dimostrare la sua residenza in modo permanente nel paese.
Flan-T5-XXL	Il bambino deve essere vivo con almeno uno di i bambini.
GPT-3.5	Il bambino deve vivere con almeno un genitore.
Guanaco	Scusi, ma non ho capito la relazione tra la prima frase e la seconda. Cosa vuole sapere esattamente?
Llama 2 Chat	Mi dispiace, pero la tua richiesta non è chiara e coerente. "Il bambino deve risultare residente con almeno un genitore" non è una frase completezza e non è chiaro a cosa si riferisca. Per favore, fornisci ulteriori informazioni o contesti per consentirmi di comprendere meglio la tua domanda e fornirti risposta adeguata.
Vicuna-7B	1. Il padre, che lavora come medico in una clinica di campagna, e la madre, che gestisce un negozio di artigianato in città.2. Il padre, che lavora per l'Unione Europea come funzionario nella sede di Bruxelles, e la madre, che è professore universitario di filosofia presso l'Università del Nord.3. Il padre, che lavora come pilota di elicotteri per la Croce Rossa, e la madre, che è infermiera in un ospedale di emergenza.
Vicuna-13B	Ecco l'algoritmo corretto considerando che il bambino deve risultare residente con almeno un genitore:"'csharp# Impostazioni del database e dei costruttiimplicit noneinteger, parameter :: n = 10000integer :: iwork(n)integer :: parent(2), child(2), sibling(2)integer, dimension(n) :: populationinteger :: parents(n), children(n), rank! Definizione della funzione per costruire la matrice di riproduzionefunction intra_familia(parents) integer, intent(in) :: parents(n) integer :: sum, count sum = 0 count = 0 do while (count <size(parents, 1)) if (parents(count + 1) == parents(count)) then sum = sum + parents(count) count = count + 1 else count = count + 1 end if end do intra_familia = sumend function! Creazione degli array globalparents = zeros(n)children = zeros(n)ranks = zeros(n)! Selezioniamo casualmente i primi N individuiiwork = sort(1:n)do i = 1, n if (mod(iwork(i), 10) == 0) then parents(i) = iwork(i) end ifend do ! Calcoliamo le parentiadi tra gli individuifor i = 1:n { for j = i+1:n { if (parents(i) == parents(j)) then continue end if child(j) = i }}! Inizializziamo la matrice delle frequenzefor i = 1:n { population(i) = 1}! Applichiamo la selezione incrociata su una generazionen_eliminati = 0do i = 1:n { non_selected(i) = 0
Vicuna-33B	1. Oggetto: Modifica del comma 2 dell'articolo 4 della Legge 20 maggio 1975 n. 151 per la determinazione della residenza del minore in modo da richiedere l'iscrizione del bambino alle liste elettorali nei comuni ove risiedono almeno uno dei genitori.Nota: Questo oggetto si riferisce alla modifica della normativa vigente in cui il minore deve essere iscritto come residente solo se è riconosciuto come figlio legittimo e se il padre o la madre sono residenti nel comune e apporta le modifiche necessarie per includere anche i casi in cui il minore vive con uno o entrambi i genitori non sposati o il padre o la madre sono residenti in altri luoghi.

Table 6
Example from Admin-It. The first two lines contain the original (complex) text as well as the gold simplified text provided. The lines that follow show the simplification produced by each model. Except for Flan-T5-XXL and GPT-3.5, all model outputs were generated using Italian templates.

	Traditional ↑		READ-IT ↓				
	Gulpease	Flesch–Kincaid	*base*	*lexical*	*syntax*	*all*	*all_adjusted*
Complex Text	39.70	15.54	59.24	76.91	92.03	92.71	-
Simplified Text	46.52	44.73	46.03	63.34	70.04	70.45	-
Camoscio	45.76	41.86	65.06	73.98	70.47	72.51	26.10
Flan-T5-XXL	40.51	16.91	61.02	76.50	91.06	91.89	11.03
Vicuna-7B	24.37	21.79	71.52	82.22	88.42	89.08	29.39
Vicuna-13B	33.84	31.42	60.86	74.53	79.76	80.89	24.27
Vicuna-33B	36.74	28.50	60.86	78.22	81.75	83.31	24.99
Guanaco	**58.52**	**65.70**	43.20	**44.56**	**59.60**	**58.90**	20.03
Llama 2 Chat	3.72	3.84	96.15	96.92	97.24	97.41	36.04
GPT-3.5	44.94	41.68	**32.10**	65.71	71.20	72.89	**<u>10.93</u>**

Table 7
Text readability scores on Admin-It$_{RS}$ for the original (complex) text, human simplification, and automatically generated simplified versions. The best results are bold, except for the proposed final metric, which is also underlined.

	Traditional ↑		READ-IT ↓				
	Gulpease	Flesch–Kincaid	*base*	*lexical*	*syntax*	*all*	*all_adjusted*
Complex Text	48.65	48.62	42.75	69.98	83.69	85.00	-
Simplified Text	49.83	48.38	35.57	61.28	72.67	74.02	-
Camoscio	48.08	52.14	65.22	72.76	77.29	79.66	28.68
Flan-T5-XXL	49.33	50.05	46.43	72.43	85.29	85.97	10.32
Vicuna-7B	23.02	23.73	67.29	83.67	81.82	82.91	27.36
Vicuna-13B	40.16	40.41	56.31	70.86	82.91	84.32	25.30
Vicuna-33B	37.65	35.94	64.29	76.25	81.49	84.23	25.27
Guanaco	**61.59**	**65.14**	45.64	**47.90**	60.46	60.18	20.46
Llama 2 Chat	2.21	2.65	98.60	98.07	97.59	97.65	36.13
GPT-3.5	51.35	49.05	**28.45**	60.13	**54.72**	**56.58**	**8.49**

Table 8
Text readability scores on Admin-It$_{RD}$ for the original (complex) text, human simplification, and automatically generated simplified versions. The best results are bold, except for the proposed final metric, which is also underlined.

	Traditional ↑		READ-IT ↓				
	Gulpease	Flesch–Kincaid	*base*	*lexical*	*syntax*	*all*	*all_adjusted*
Complex Text	41.72	30.94	23.96	66.64	82.52	84.03	-
Simplified Text	42.88	33.93	24.66	63.43	80.58	81.70	-
Camoscio	45.38	45.29	66.81	70.43	72.47	74.45	26.80
Flan-T5-XXL	43.05	34.85	27.24	65.23	81.78	82.48	9.90
Vicuna-7B	26.18	22.68	60.70	76.09	81.86	83.00	27.39
Vicuna-13B	37.89	34.55	48.60	68.03	78.96	80.33	24.10
Vicuna-33B	38.78	30.60	53.16	70.81	84.27	85.89	25.77
Guanaco	**54.87**	**60.69**	43.01	**46.74**	62.52	61.95	21.06
Llama 2 Chat	8.09	8.58	92.20	92.96	94.13	94.41	34.93
GPT-3.5	45.76	39.92	**19.25**	58.12	**59.06**	**60.27**	**9.04**

Table 9
Text readability scores on Admin-It$_{OP}$ for the original (complex) text, human simplification, and automatically generated simplified versions. The best results are bold, except for the proposed final metric, which is also underlined.

Unraveling Text Coherence from the Human Perspective: a Novel Dataset for Italian

Federica Papa[1], Luca Dini[1,2], Dominique Brunato[2] and Felice Dell'Orletta[2]

[1]University of Pisa
[2]Istituto di Linguistica Computazionale "Antonio Zampolli", ItaliaNLP Lab, Pisa

Abstract
This paper presents a novel resource designed to study text coherence in the Italian language. The dataset aims to address existing deficiencies in coherence assessment by focusing on human perception of coherence. Recently, it has been integrated into the DiSCoTex benchmark, part of EVALITA 2023 [1], the 8th evaluation campaign for NLP and speech tools in Italian. Our resource aims to provide a comprehensive understanding of coherence, highlighting the influence of both genre and text perturbations on perceived coherence.

Keywords
text coherence, human perception, Italian dataset, text perturbations

1. Introduction and Motivation

Coherence plays a central role in maintaining the overall unity of a text and is influenced by both linguistic and extra-linguistic factors. From the linguistic point of view, it primarily relies on cohesion, which encompasses various linguistic devices used in natural languages to establish connections within a text, such as anaphoric and cataphoric relationships, discourse markers, and elliptical constructions [2]. While cohesion mainly ensures local coherence between adjacent or nearby sentences, to be fully coherent a text needs to achieve a global coherence, a property that pertains to the connection of concepts and relationships that underlie the surface text ensuring a logical flow of ideas around an overall intent [3]. This aspect of coherence adds a subjective component, as it also depends on the reader or listener's familiarity with the text, language proficiency, and level of interest and attention.

Modelling coherence in natural language is essential for a wide range of downstream applications. One such application is automatic essay scoring in language learning settings, where coherence assessment can provide valuable writing feedback by identifying poorly organized paragraphs and abrupt topic transitions [4, 5]. In clinical contexts, coherence modeling is relevant for automatic language assessment, as speech irregularities indicative of a lack of coherence can serve as markers for mental disorders like schizophrenia [6, 7]. Furthermore, coherence has been adopted as an intrinsic evaluation metric for assessing the quality of texts generated by Natural Language Generation (NLG) systems [8]. Additionally, coherence modeling is gaining importance in research on the interpretability of modern deep neural networks [9, 10, 11]. Indeed, while existing work has mainly focused on probing sentence-level properties, understanding how these models encode discourse and pragmatic phenomena remains a crucial aspect.

In light of this interest, various attempts have been made to approach coherence assessment in the Natural Language Processing (NLP) community, especially in the 'pre-deep learning' era. With this respect, early computational models of discourse coherence were primarily built upon two linguistic theories: centering theory [12] and rhetorical structure theory [13] . Studies aligned with centering theory, such as [14], focused on analyzing the distribution of entity transitions over sentences as a means to predict text coherence. On the other hand, works inspired by rhetorical structure theory, such as [15], employed discourse parsers to generate discourse relations over sentences. With the advent of neural models, researchers have also explored their application in coherence assessment, see e.g. Lin et al. [15] and Nguyen and Joty [16].

The importance of building challenging datasets for coherence evaluation cannot be overstated. With this respect, independently from the underlying theories, models of discourse coherence are typically tested on tasks such as reordering, which aim to discern an original text from a corrupted one artificially created by shuffling the order of its sentences, or tasks that require systems to detect whether a document contains an intruder sentence from another document [9] or to classify whether a target sentence is contiguous or not with a given passage [17]. However, these approaches have come under criticism because they neglect key aspects of coherence, as

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
f.papa4@studenti.unipi.it (F. Papa); luca.dini16@gmail.com (L. Dini); {dominique.brunato}@ilc.cnr.it (D. Brunato); {felice.dellorletta}@ilc.cnr.it (F. Dell'Orletta)

noted by Lai and Tetreault [4] and Beyer et al. [18] among others. They fail to identify the qualities that make the shuffled text incoherent, do not pinpoint the linguistic devices responsible, and overlook the subjective component underlying coherence. Additionally, most existing benchmarks are limited to the English language.

Our contribution In this paper we seek to address some of the existing deficiencies in coherence assessment by presenting a novel resource tailored for the Italian language, designed specifically to study text coherence from the perspective of human perception. The dataset, which to our knowledge is the first for Italian, has been recently used as part of a larger benchmark released for DiSCoTex, one of the shared-tasks presented at the 8th evaluation campaign of NLP and speech tools for the Italian language (EVALITA 2023) [1]. The results of first analyses on the resource shed light on the influence of both genre and text perturbations on perceived coherence[1].

2. Dataset Construction

The construction of our dataset was guided by two distinct criteria: on the one hand, we intended to explore the effect of textual genre on the human perception of coherence; on the other hand, we wanted to assess whether and to what extent humans are sensitive to different strategies introduced to artificially modify an original text.

As a starting point we selected texts from two distinct sources: the Italian Wikipedia and the Italian speech transcripts section of the Multilingual TEDx corpus (mTEDx). The choice of these sources was meant to obtain a balanced corpus that was representative of two different language varieties: the former is a 'standard' written variety, and the latter a 'hybrid' variety combining diverse genres (e.g., university lectures, newspaper articles, conference presentations, and TV science programs) as well as different semiotic modes, such as written, spoken, audio, and video [19].

Following the approach by Brunato et al. [17], for each text we then proceed to extract passages consisting of four consecutive sentences, considering them as our unit of analysis for modeling the coherence annotation task. As for Wikipedia, we relied on the existing segmentation into paragraph and extract four-sentence passages. For the TEDx corpus, as these texts lack such an internal structure, we split all the transcripts into passages of four sentences.

After creating all the possible passages, we randomly selected 1,064 of them while maintaining a proportional representation from both sources. Half of the extracted passages were left unchanged, while the other half underwent a perturbation. More specifically, we devised two distinct perturbation strategies:

- *swap*: it involves swapping the position of two random sentences in the text passage.
- *substitution* (sub): it consists of replacing one of the four sentences with another sentence, corresponding to the 10th sentence following the passage in the same document.

Table 1 contains an example from the corpus for each perturbation type

2.1. Collecting human ratings

Before starting the annotation process with humans, we added ten fillers to the dataset, consisting of four-sentence passages deliberately chosen to be either highly coherent or highly incoherent. These additional passages served as a control mechanism to check the reliability and accuracy of each annotator in assigning coherence scores to the actual texts in the dataset: if an annotator assigns an out-of-scale coherence value to these texts, it suggests that they might not have conducted the annotation process adequately.

The annotation process has been executed via *crowdsourcing*. We first used the *Questbase*[2] platform to create questionnaires formulating the text scoring process in the form of questions. Then, we distributed the questionnaires using the crowdsourcing platform *Prolific*[3], choosing to recruit only Italian native speakers without language disorders as annotators. Considering that subjective component underlying coherence that makes this concept gradual rather than categorical, people were asked to rate each texts on a 5-point Likert scale, where 1 represents the minimum value of perceived coherence and 5 the maximum[4]. A pilot experiment tested the suitability of the questionnaire from different points of view (i.e. the clearness of instructions) and allowed to estimate the time needed to complete it. After collecting all the responses to the questionnaires, we kept only the most reliable annotations by filtering out the annotators who had failed the attention checks. Specifically, we excluded those annotators who rated at least four control texts incorrectly, i.e. assigning a value from 1 to 3 to highly coherent filler passages or a value from 3 to 5 to very incoherent filler passages. As a result, we retained an average of 10 annotations per passage for a total of 10,567 annotations for the whole dataset.

[1]The dataset will be made publicly available for research purposes at the following link: http://www.italianlp.it/resources/

[2]https://questbase.com/

[3]https://www.prolific.co/

[4]Appendix A contains the instructions given to the annotators when opening the questionnaire on Prolific.

Table 1
Example of perturbations: in the first example, which is a passage from Wikipedia, sentence 2 and sentence 3 have been swapped. In the second one, from the TEDx corpus, the 4th sentence have been substituted with the subsequent 10th sentence.

Text passage	Perturbation
1. Cliff Burton possedeva uno stile impeccabile ed era capace di produrre giri di basso potenti ma allo stesso tempo raffinati. **2.** Specialmente durante gli assoli era solito pizzicare due o tre corde nello stesso momento e lanciarsi in un complesso uso di distorsioni, tapping, bending e applicazioni del pedale wah wah. **3.** Il suo stile era molto vario per i canoni di un bassista heavy metal: Burton non suonò mai il basso "come un chitarrista" e mai utilizzò plettri, prediligendo il contatto diretto con le corde, pizzicate a mani nude. **4.** Anche per questo, diversamente da altri bassisti heavy metal che utilizzavano bassi a cinque o sei corde, Burton suonava solo bassi a quattro corde, che considerava più adatti al suo stile.	Swap 2-3
1. È stato teorizzato che le prime stelle dell'universo, le cosiddette stelle di Popolazione III, fossero molto più massicce delle stelle attualmente esistenti. **2.** Si è postulata l'esistenza di questa prima generazione di stelle per spiegare l'esistenza di elementi chimici diversi dall'idrogeno e dall'elio nelle stelle più vecchie conosciute. **3.** Sebbene fossero più grandi e luminose di tutte le supergiganti note oggi, la loro struttura doveva essere molto differente, con perdite di massa molto più contenute. **4.** Nella maggior parte dei casi la variabilità è dovuta a pulsazioni della superficie stellare.	Sub 4

3. Analysis of perceived coherence

To delve deeper on the factors influencing human perception of text coherence, we conducted two types of analyses that examine the relationship between perceived coherence and text structure from distinct perspectives. The first one focuses on the effect of the different perturbation strategies artificially introduced to disrupt the internal coherence of rated passages; the second one takes into account solely the subset of original, i.e. unperturbed, texts with the aim of exploring the effect of several linguistic features extracted from each passage on the mean coherence judgments.

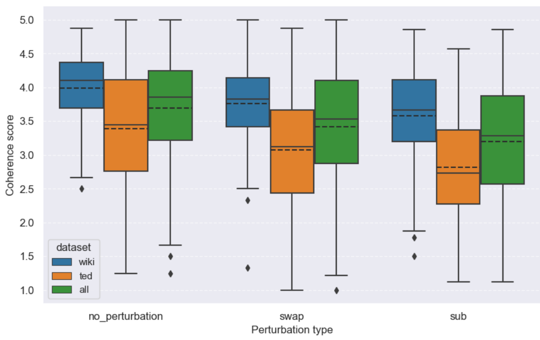


Figure 1: Box plot of human judgments collected for the dataset. For each subset (*Wiki*, *TED*, and all) and the possible perturbations (*Sub*, *Swap*). The dashed line corresponds to the mean coherence score.

3.1. Impact of text perturbations

After gathering all annotations, we studied their homogeneity by calculating for each passage the mean value and standard deviation of the coherence scores assigned to it[5]. These statistics were computed for the whole dataset as well as for passages grouped according to the text source from which they derived (TED or Wikipedia) and to the perturbations eventually applied. The purpose was to observe how coherence ratings vary among the different groups and understand the effects of the different artificial perturbations applied to the text. These results are shown in Figure 1.

Observing the trend of the distribution of the mean coherence ratings for each group, it was possible to see that the group containing all the original texts was considered as more coherent than the ones with the perturbed texts. However, in all considered groups, texts extracted from Wikipedia were rated as more coherent than those extracted from TEDx, even when artificially perturbed. This suggests that Wikipedia documents tend to exhibit a more standardized structure, with internal coherence remaining relatively stable even when subjected to minor alterations, such as changes in sentence order or the inclusion of an intruder sentence from the same document.

In all groups the standard deviation has low values, suggesting a high degree of homogeneity in the ratings, especially for Wiki passages.

A more in-depth investigation was conducted to assess the potential impact of certain factors of each perturbation type, such as the **distance** between swapped sentences and the **position** of the replaced sentence, on the distribution of mean coherence scores. Regarding the swap perturbation, we assigned a label to each perturbed text, indicating the distance between the swapped sentences. Dist_0 was given to passages where the swapped sentences were adjacent, Dist_1 to those with one sentence between the swapped sentences, and Dist_2 to

[5]Inter-annotator agreement measured by Krippendorff's alpha is .32 for the whole corpus.

those with two sentences between the swapped sentences. Similarly, we assigned a label to the passage that underwent the substitution perturbation. Pos_1 was given to passage where the first sentence was replaced, Pos_2 to those where the second sentence was replaced, Pos_3 when the third sentence was replaced, and Pos_4 when the fourth sentence was replaced.

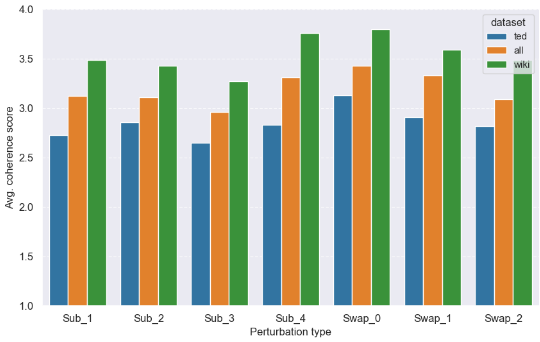


Figure 2: Mean coherence judgments attributed to perturbed passages grouped by distance (for *swap* perturbation) and position (for *substitution* perturbation).

As we can see in Figure 2, in texts perturbed with the swap perturbation, the perceived coherence is higher in those cases where the sentences are adjacent to each other, while decreases as the distance between the exchanged sentences increases. In texts altered via substitution, the perceived coherence increases especially in those cases where the sentence substitution occurred at the ends (thus in the first and last positions), while it generally decreases in cases where substitution involves the middle positions.

3.2. Impact of linguistic structure

Although the previous analysis revealed that perturbed texts were rated on average as less coherent that original ones, we also observed that such a perception is influenced by textual genre. Considering the subjective nature of coherence, we hypothesize that even well-formed texts may receive different coherence annotations. We thus carried out a final analysis focused solely on the subset of original texts, with the aim of investigating the relationship between the linguistic profile of these texts and the perceived coherence.

To automatically extract linguistic information from human rated passages, we leveraged *Profiling-UD* [20], a tool for carrying out linguistic profiling investigations in multiple languages based on the Universal Dependency framework. Using *Profiling-UD*, we extracted more than 130 features for each passage, which capture lexical, morpho-syntactic and syntactic properties of text. An overview of these features is shown in Table 2. These features were shown to be relevant for modeling aspects characterizing the interaction between a reader and a text, such as the human perception of sentence complexity [20] and of writing quality [21]. As both complexity and quality are properties connected to coherence, we expect that these features will provide valuable insights for our coherence analysis as well.

We then sought correlations between the human perception of coherence and these linguistic features by calculating the Spearman correlation coefficient between the average coherence score attributed to each passage and the average value of each linguistic feature extracted from it. Results are shown in Table 3. Considering only features with *p-value* below the threshold of 0.05, we observed that the perception of coherence in original texts positively correlates above all with features closely related to *length*. In fact, the highest correlation (Spearman's r = 0.32) was obtained with the maximum depth of syntactic tree, followed by *tokens_per_sent* and *n_tokens*, which captures respectively the average sentence length and of passage length in number of tokens. Also *lexical richness*, measured by the average value of Type/Token Ratio (*ttr_form_100*) turned out to be among the first top-five correlated features. These findings suggest that longer sentences could contain more information and thus lead to a more complete text, that also makes it more coherent. An interesting result was that the use of pronouns negatively correlated with the perception of coherence (Spearman's r = -0.26). This unexpected observation can be attributed to the potential ambiguity of pronouns, deriving from the fact that the evaluated passages were extracted from larger texts and might lack the necessary context to accurately link the pronoun to its intended referent.

Focusing specifically on original passages derived from Wikipedia, we observed that the presence of features describing proper syntactic phenomena closely related to the sentence length, such as the average length of dependency links and of the maximum link (*avg_links_len*, *max_links_len*), along with the presence of nouns modified by prepositional phrases (*prep_chain*), contributed to increased coherence perception by annotators. These linguistic features that are typically related to syntactic complexity may suggest that these texts are also more informative, resulting in enhanced coherence perception. Furthermore, it could be seen that in texts taken from Wikipedia the judgement of coherence was positively influenced by a paratactic structure of the text (*dep_conj*). Finally, in the TED original texts it could be observed that the correlation was positive in the case of the distribution of subordinate propositions (*subord_dist*) and negative in the distribution of main ones (*princ_dist*).

Table 2
Overview of linguistic features used by Profiling-UD.

Annotation Level	Linguistic Feature Description	Label
Raw Text	Sentence length (tokens), word length (characters) Words and lemmas type/token ratio	n_tokens, char_per_tok ttr_form, ttr_lemma
POS Tagging	Distribution of UD and language-specific POS tags Lexical density Inflectional morphology of auxiliaries (mood, tense)	upos_dist_*, xpos_dist_* lexical_density aux_mood_*, aux_tense_*
Dependency Parsing	Syntactic tree depth Average and maximum length of dependency links Number and average length of prepositional chains Relative ordering of main elements Distribution of dependency relations Distribution of verbal heads Distribution of principal and subordinate clauses Average length of subordination chains Relative ordering of subordinate clauses	parse_depth avg_links_len, max_links_len n_prep_chains, prep_chain_len subj_pre, subj_post, obj_pre, obj_post dep_dist_* vb_head_per_sent princ_prop_dist, sub_prop_dist sub_chain_len sub_post, sub_pre

Table 3
Extract of linguistic features correlating with mean coherence judgments attributed to all original passages (Unp_ALL), original Wiki-extracted passages (Unp_WIKI) and original TED-extracted passages (Unp_TED). Significant correlations (*p-value* < 0.05) are denoted with a star.

LingFeats	Unp_ALL	Unp_WIKI	Unp_TED
avg_max_depth	0.32*	0.09	0.33*
tok_per_sent	0.3*	0.11	0.21*
n_tok	0.28*	0.14	0.16*
upos_ADP	0.26*	-0.03	0.23*
ttr_form_100	0.26*	0.16*	0.14
upos_ADV	-0.25*	-0.08	-0.13
upos_PUNCT	-0.26*	-0.04	-0.28*
upos_PRON	-0.26*	-0.1	-0.06
max_links_length	0.19	0.24*	0.04
verb_tense-Past	0.24*	0.21*	0.01
prep_chain	0.26*	0.19*	0.18
avg_links_len	0.05	0.19*	-0.08
aux_mood_Ind	0.034	0.18*	0.02
aux_form_Fin	-0.05	0.15*	-0.19*
dep_conj	-0.14	0.15*	0.09
aux_tense-Past	0.22*	0.15*	0.13
verb_tense-Pres	-0.25*	-0,18*	-0.06
vb_head_sent	0.12	0.004	0.26*
dep_det:poss	0.15*	0.06	0.24*
subord_dist	0.08	-0.01	0.23*
verb_form_Inf	-0.02	-0.1	0.24*
dep_advmod	-0.25*	-0.08	-0.15*
obj_pre	-0.23*	-0.003	-0.15*
princ_dist	-0.08	0.01	-0.24*

4. Conclusions

This paper has introduced a novel resource for studying and computationally modeling text coherence in the Italian language, focusing on human perception. The investigation into genre and text perturbations revealed a significant interplay between the two dimensions. Interestingly, text passages from Wikipedia were rated on average as more coherent than those extracted from TEDx talks even when presented in a perturbed form. Furthermore, a deeper analysis of the perturbations revealed distinct effects on coherence perception. Modifications that disrupted coherence by altering the sentence order or introducing intruder sentences had varying impacts. Notably, coherence judgments also varied for original texts, and the syntactic structure and complexity-related features emerged as influential factors in human assessment.

In the future we would like to gain deeper insights into the underlying factors that influence coherence perception by also incorporating a diverse range of text genres and perturbations. This deeper understanding of coherence will have significant implications for the development of more sophisticated language understanding and generation systems.

Acknowledgments

The authors gratefully acknowledge the support of the PNRR MUR project PE0000013-FAIR.

References

[1] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[2] M. Halliday, R. Hasan, Cohesion in English, Longman Group Ltd., 1976.

[3] T. A. Van Dijk, Text and Context: Exploration in the Semantics and Pragmatics of Discourse, Longman, London, 1977.

[4] A. Lai, J. Tetreault, Discourse coherence in the wild: A dataset, evaluation and methods, in: Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue, Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 214–223. URL: https://aclanthology.org/W18-5023. doi:10.18653/v1/W18-5023.

[5] M. Mesgar, M. Strube, A neural local coherence model for text quality assessment, in: Proceedings of the 2018 conference on empirical methods in natural language processing, 2018, pp. 4328–4339.

[6] B. Elvevåg, P. W. Foltz, D. R. Weinberger, T. E. Goldberg, Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia, Schizophrenia research 93 (2007) 304–316.

[7] D. Iter, J. Yoon, D. Jurafsky, Automatic detection of incoherent speech for diagnosing schizophrenia, in: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, 2018, pp. 136–146.

[8] A. Celikyilmaz, E. Clark, J. Gao, Evaluation of text generation: A survey, CoRR abs/2006.14799 (2020). URL: https://arxiv.org/abs/2006.14799. arXiv:2006.14799.

[9] A. Shen, M. Mistica, B. Salehi, H. Li, T. Baldwin, J. Qi, Evaluating document coherence modeling, Transactions of the Association for Computational Linguistics 9 (2021) 621–640. URL: https://aclanthology.org/2021.tacl-1.38. doi:10.1162/tacl_a_00388.

[10] M. Chen, Z. Chu, K. Gimpel, Evaluation benchmarks and learning criteria for discourse-aware sentence representations, 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), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 649–662. URL: https://aclanthology.org/D19-1060. doi:10.18653/v1/D19-1060.

[11] Y. Farag, J. Valvoda, H. Yannakoudakis, T. Briscoe, Analyzing neural discourse coherence models, in: Proceedings of the First Workshop on Computational Approaches to Discourse, Association for Computational Linguistics, Online, 2020, pp. 102–112. URL: https://aclanthology.org/2020.codi-1.11. doi:10.18653/v1/2020.codi-1.11.

[12] B. J. Grosz, A. K. Joshi, S. Weinstein, Centering: A framework for modeling the local coherence of discourse, Computational Linguistics 21 (1995) 203–225. URL: https://aclanthology.org/J95-2003.

[13] W. C. Mann, S. A. Thompson, Rhetorical structure theory: Toward a functional theory of text organization, Text 8 (1988) 243–281. URL: http://scholar.google.com/scholar.bib?q=info:BEw8CIWbucoJ:scholar.google.com/&output=citation&scisig=AAGBfm0AAAAAU3X_1Dq4ULnWfFzMeRsqGJcha1fReMSl&scisf=4&hl=en.

[14] R. Barzilay, M. Lapata, Modeling local coherence: An entity-based approach, in: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), Association for Computational Linguistics, Ann Arbor, Michigan, 2005, pp. 141–148. URL: https://aclanthology.org/P05-1018. doi:10.3115/1219840.1219858.

[15] Z. Lin, H. T. Ng, M.-Y. Kan, Automatically evaluating text coherence using discourse relations, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Portland, Oregon, USA, 2011, pp. 997–1006. URL: https://aclanthology.org/P11-1100.

[16] D. T. Nguyen, S. Joty, A neural local coherence model, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 1320–1330. URL: https://aclanthology.org/P17-1121. doi:10.18653/v1/P17-1121.

[17] D. Brunato, F. Dell'Orletta, I. Dini, A. A. Ravelli, Coherent or not? stressing a neural language model for discourse coherence in multiple languages, in: Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 10690–10700. URL: https://aclanthology.org/2023.findings-acl.680.

[18] A. Beyer, S. Loáiciga, D. Schlangen, Is incoherence surprising? targeted evaluation of coherence prediction from language models, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computa-

tional Linguistics, Online, 2021, pp. 4164–4173. URL: https://aclanthology.org/2021.naacl-main.328. doi:10.18653/v1/2021.naacl-main.328.

[19] G. Caliendo, The popularisation of science in web-based genres, The language of popularisation: Theoretical and descriptive models 3 (2012) 101–132.

[20] D. Brunato, A. Cimino, F. Dell'Orletta, G. Venturi, S. Montemagni, Profiling-UD: a tool for linguistic profiling of texts, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 7145–7151. URL: https://aclanthology.org/2020.lrec-1.883.

[21] A. Cerulli, D. Brunato, F. Dell'Orletta, Quale testo è scritto meglio? a study on italian native speakers' perception of writing quality?, in: Proceedings of 8th Italian Conference on Computational Linguistics (CLiC-it), 26-28 January, 2022, Milan, Italy., CEUR-WS 3033, Milan, 2022, p. 9.

A. Annotation instructions

This is the questionnaire instructions provided to human raters:

"Ciao! In questo sondaggio ti chiediamo di leggere dei testi e di valutarne il livello di coerenza, assegnando un punteggio che va da 1 (per nulla coerenti) a 5 (del tutto coerenti).

Innanzitutto, ti diamo una breve definizione di coerenza: in ambito linguistico, questa parola si usa per indicare una caratteristica che riguarda l'organizzazione del significato di un testo.

Un testo è considerato coerente se le singole unità di cui si compone (tipicamente le frasi) sono connesse tra loro in modo da formare un'unità più ampia che il lettore/ascoltatore considera globalmente appropriata, sia dal punto di vista dell'ordine logico-temporale sia rispetto al contenuto principale del discorso. Tuttavia, questa valutazione è molto personale: la coerenza, infatti, dipende sia da fattori legati alla struttura linguistica e al contenuto del testo, sia da fattori soggettivi, come la familiarità del lettore/ascoltatore verso l'argomento, la sua padronanza linguistica, il grado di interesse ecc. Proprio per questo ti chiediamo di valutare ciascun testo con la maggior naturalezza possibile, dal momento che non c'è una risposta giusta o sbagliata: quello che ci interessa è proprio la tua percezione personale! In generale, per orientarti nel giudizio, puoi pensare che un testo molto coerente dovrebbe risultarti facile da comprendere, ben strutturato e non dovresti avvertire discontinuità sul piano logico e del contenuto nel passaggio tra una frase e l'altra. Ad esempio, il testo che segue dovrebbe ottenere un punteggio di 4 o 5:

E quindi che si fa? E quindi mi danno in mano un depliant dell'Università e dicono: "Bene ragazzo. Scegli una facoltà a numero aperto e, nel momento in cui qualcuno ad Economia molla, puoi subentrare te." "Benissimo!" dico. Apro il depliant dell'Università, salto a piè pari Ingegneria per l'eccessiva presenza di Chimica e tra Fisica, Filosofia, Lettere, Matematica e Informatica inizio a decidere che cosa fare.

Al contrario, un testo poco coerente dovrebbe risultarti più difficile da capire, poco coeso e discontinuo sul piano logico e strutturale. Ad esempio, il testo che segue dovrebbe ottenere un punteggio di 1 o 2:

Stiamo parlando degli anni Trenta. Le aziende, le persone che puntano al futuro, le protagoniste di questa trasformazione, non sono assolutamente associate a queste parole, semmai a: tecnologia; precisione; elettronica; digitale; meccanica; futuro. Tutte parole, queste, associate invece al termine "meccatronica". Solo il cinque per cento dei funghi che potenzialmente esistono sono stati descritti, quindi c'è veramente un mondo da scoprire sotto i nostri piedi.

Inoltre, per la tua valutazione, tieni presente che tutti i testi che leggerai non sono completi. Si tratta infatti di paragrafi di poche righe, estratti da sezioni diverse (es. introduzione, corpo, conclusione) di documenti più lunghi, che provengono da varie fonti (es. testi di Wikipedia, dialoghi trascritti). Infine, ti ricordiamo che il sondaggio è indirizzato alle persone di madrelingua italiana e la sua compilazione richiederà all'incirca 20-25 minuti.

Grazie in anticipo per la partecipazione!"

For the sake of completeness, we also report an English translation of the same guidelines:

"Hello! In this survey, we ask you to read texts and evaluate their level of coherence by assigning a score ranging from 1 (not at all coherent) to 5 (completely coherent).

Firstly, we provide a brief definition of coherence: in a linguistic context, this word is used to indicate a characteristic related to the organization of the meaning within a text.

A text is considered coherent if its individual units (typically sentences) are connected in a way that forms a broader unit that the reader/listener perceives as globally appropriate, both in terms of logical-temporal order and the main content of the discourse. However, this evaluation is highly subjective: coherence depends on factors related to the linguistic structure and content of the text, as well as subjective factors such as the reader/listener's familiarity with the topic, linguistic proficiency, level of interest, etc. That's why we ask you to assess each text as naturally as possible, as there is no right or wrong answer: what we are interested in is your personal perception! In general, to guide your judgment, you can consider that a highly coherent text should be easy to understand, well-structured, and you should not perceive any discontinuity in logical and content transitions between sentences. For example, the following text should receive a score of 4 or 5:

E quindi che si fa? E quindi mi danno in mano un depliant dell'Università e dicono: "Bene ragazzo. Scegli una facoltà a numero aperto e, nel momento in cui qualcuno ad Economia molla, puoi subentrare te." "Benissimo!" dico. Apro il depliant dell'Università, salto a piè pari Ingegneria per l'eccessiva presenza di Chimica e tra Fisica, Filosofia, Lettere, Matematica e Informatica inizio a decidere che cosa fare.

On the contrary, a text with low coherence should be more difficult for you to understand, poorly connected, and discontinuous on a logical and structural level. For example, the following text should receive a score of 1 or 2:

Stiamo parlando degli anni Trenta. Le aziende, le persone *che puntano al futuro, le protagoniste di questa trasformazione, non sono assolutamente associate a queste parole, semmai a: tecnologia; precisione; elettronica; digitale; meccanica; futuro. Tutte parole, queste, associate invece al termine "meccatronica". Solo il cinque per cento dei funghi che potenzialmente esistono sono stati descritti, quindi c'è veramente un mondo da scoprire sotto i nostri piedi.*

Furthermore, for your evaluation, please keep in mind that all the texts you will read are not complete. They are short paragraphs extracted from different sections (e.g., introduction, body, conclusion) of longer documents, coming from various sources (e.g., Wikipedia texts, transcribed dialogues). Finally, we remind you that the survey is aimed at Italian native speakers, and it should take approximately 20-25 minutes to complete.

Thank you in advance for your participation!"

Lost in Labels: An Ongoing Quest to Optimize Text-to-Text Label Selection for Classification

Michele Papucci[1,2], Alessio Miaschi[3] and Felice Dell'Orletta[2,3]

[1]*Università di Pisa, Pisa, Italy*
[2]*TALIA s.r.l., Pisa, Italy*
[3]*ItaliaNLP Lab, CNR, Istituto di Linguistica Computazionale 'A.Zampolli', Pisa, Italy*

Abstract

In this paper, we present an evaluation of the influence of label selection on the performance of a Sequence-to-Sequence Transformer model in a classification task. Our study investigates whether the choice of words used to represent classification categories affects the model's performance, and if there exists a relationship between the model's performance and the selected words. To achieve this, we fine-tuned an Italian T5 model on topic classification using various labels. Our results indicate that the different label choices can significantly impact the model's performance. That being said, we did not find a clear answer on how these choices affect the model performances, highlighting the need for further research in optimizing label selection.

Keywords

encoder-decoder, label selection, topic classification

1. Introduction and Background

In recent years, the Sequence-to-Sequence paradigm has emerged as a highly popular approach in building cutting-edge Transformer-based Language Models [1, 2, 3]. This paradigm draws inspiration from earlier unified frameworks for Natural Language Processing (NLP) tasks [4, 5, 6], treating each task as a text-to-text transformation. In other words, it involves taking text as input and generating new text as output.

This unifying framework has proven to be a particularly effective transfer learning method, often outperforming previous models, e.g. BERT [7], in data-poor settings. Furthermore, the recent application and refinement of prompt-based tuning techniques for pre-trained Large Language Models (LLMs) have made this paradigm even more powerful, especially in few-shot and zero-shot learning scenarios [8].

In such a scenario, several studies have focused on defining methods for the formulation of prompts and the definition of *verbalizers*, i.e. mapping techniques between model-predicted words and task labels. As for the latter, the vast majority of studies have concentrated on devising automatic or semi-automatic approaches to create *verbalizers* that can be applied especially in zero- or few-shot configurations [9, 10, 11]. For instance, [12] proposed Petal, an approach for automatically finding the best words-label mapping by maximizing the likelihood of the training data. [13] instead developed ProtoVerb, a prototypical verbalizer that learns class prototypes from training data to build verbalizers automatically.

Nevertheless, few works have focused on investigating more deeply and systematically the effect that the choice of strings used to represent one (or more) labels has on model performance. Among these, [14] designed different label representations (e.g. canonical task labels, task-unrelated antonyms) and tested their impact with the T5 model on four classification tasks, showing that the performance was generally unaffected by the choice of label representation. Similarly, experimenting with the gender prediction task from the TAG-IT dataset [15], [16] noticed that while modifying the label representations did not affect the performance of the IT5 model [17], shuffling them for the topic classification task lead to worse results.

In this work, we present an evaluation of the impact of label selection on the performance of a Sequence-to-Sequence Model in a classification task. Specifically, we address the following research questions: i) Do the words used to represent the classification categories influence the model's performance? ii) Are there any relationship between classification categories and the words used to represent them that we can exploit to do label selection?

To investigate these questions, we conducted a series of experiments by fine-tuning the Italian version of the T5 model [17] on the topic classification task [15] using various labels. In particular, we defined different sets of labels and examined the model's performance for each of these sets. Additionally, we conducted an in-depth qualitative analysis to inspect which labels contribute

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, November 30 - December 2, 2023, Venice, IT
michele.papucci@talia.cloud (M. Papucci); alessio.miaschi@ilc.cnr.it (A. Miaschi); felice.dellorletta@ilc.cnr.it (F. Dell'Orletta)

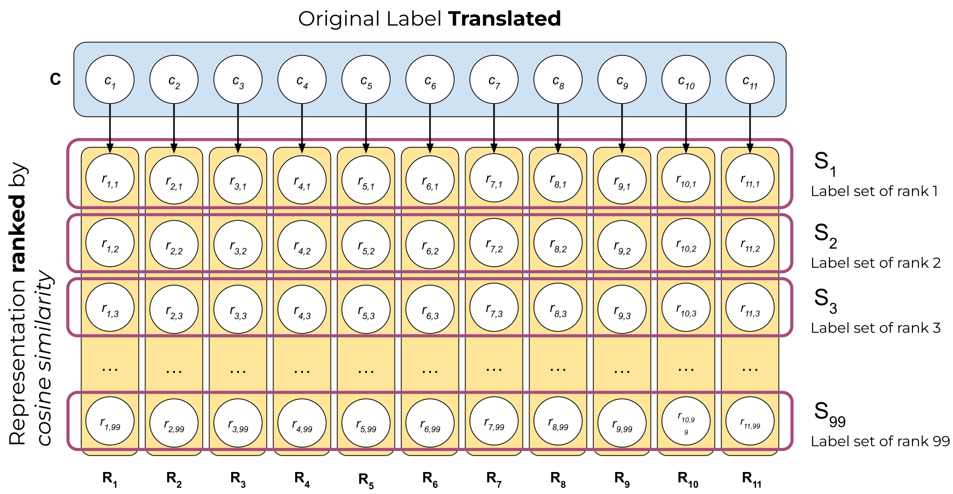


Figure 1: The framework for the creation of the different sets of labels S_j ranked by cosine similarity.

most significantly to the improvement or decline in classification results and why that might be the case.

The remainder of the paper is organized as follows: in Sec. 2 we present our approach, introducing the data and the model we used (Sec. 2.1 and Sec. 2.2) and the experimental setting (Sec. 2.3). In Sec. 3 we discuss the obtained results and in Sec. 4 we conclude the paper.

Contributions. In this paper we: i) propose an evaluation of the influence that label selection has on the performance of a Text-to-Text Transformer model for classification; ii) investigate how the words used to represent the classification categories, in a multi-class classification task, impact task performance both globally, and at class-level; iii) investigate the existence of a relationship between classification categories and selected labels and how this connection can be leveraged to improve label selection.

2. Our Approach

In this section, we first define the data and the model used to perform our experiments. Then, we detail the experimental setting we devised to select the tested labels and fine-tune the T5 model.

2.1. Data

We relied on posts extracted from TAG-IT [15], the profiling shared task presented at EVALITA 2020 [18]. The dataset, based on the corpus defined in [19], consists of more than 18,000 posts written in Italian and collected from different blogs. Each post is labelled with three different labels: age and gender of the writer and topic.

Categories	# Data	# Training	# Test
Anime	3,972	2,894	1,078
Auto-Moto	3,783	2,798	985
Bikes	520	365	155
Celebrities	1,115	754	361
Entertainment	469	354	115
Medicine-Aesthetics	447	310	137
Metal-Detecting	1,382	1,034	348
Nature	516	394	122
Smoke	1,478	1,101	377
Sports	4,790	3,498	1,292
Technology	136	51	85
All	18,608	13,553	5,055

Table 1
Dataset statistics.

In order to experiment with various possible combinations of labels, we have decided to focus only on the Topic classification task. Moreover, to have enough data to fine-tune the model, we decided to modify the original task as defined in [15]. Instead of predicting the label of a given collection of texts (multiple posts), we fine-tuned our model to predict the topic from each single post. Finally, since a fair amount of sentences were quite short, we decided to remove those shorter than 10 tokens. At the end of this process, we obtained a dataset consisting of 13,553 posts as training set and 5,055 posts as test set. The distribution of posts according to each label is reported in Table 1.

2.2. Model

We used the T5 base version pre-trained on the Italian language, i.e. IT5 [17][1]. In particular, the model was trained on the Italian sentences extracted from a cleaned version of the mC4 corpus [20], a multilingual version of the C4 corpus including 107 languages.

2.3. Experimental Setting

As already introduced in Sec. 1, to investigate the influence of label selection on the model performance, we fine-tuned the IT5 model using different combinations of strings to represent the original classification categories. We will refer to the set of the original categories with C. We first translated the categories (as seen in Table 1) in Italian. (e.g. *Celebrities* into *celebrità*)[2]. Then, for each category c_i in C we created a set R_i composed by 100 string representations: 10 were selected from synonyms and related words to the original categories (including aforementioned translated ones), while the remaining 90 were randomly chosen from the most frequent nouns in the ItWac corpus [21]. Let $R_i = \{r_{i0}, r_{i1}, ..., r_{i99}\}$ be the set of labels for the category c_i, and r_{ij} be the jth label in the set. Then, for each category c_i we ranked its corresponding set of labels R_i in descending order of similarity:

$$cs(c_i, r_{i0}) \geq cs(c_i, r_{i1}) \geq ... \geq cs(c_i, r_{i99})$$

Where $cs(c_i, r_{ij})$ is the cosine similarity between the average embedding of the subtokens of c_i and r_{ij}, extracted from the last encoding layer of the IT5 model.

Given the previously defined sets R_i, which contains the elements ranked by similarity, we created 100 sets of labels S_j (where j ranges from 0 to 99). Each set is defined as: $S_j = \{r_{0j}, r_{1j}, ..., r_{10j}\}$, where e.g. r_{0j} is the jth ranked label for category c_0. As a consequence, S_0 contains the labels that achieved the highest cosine similarity with the original categories, while S_{99} is the set containing the lowest cosine similarities. An overview of our setting is shown in Figure 1.

We then fine-tuned IT5 for each ranked set of representation S_j. Each model was trained for 10 epochs and using f-score as the evaluation metric.

3. Results

Overall results Figure 2 summarizes the results obtained by the T5 models fine-tuned on the topic classification tasks according to the 100 different sets of labels (S_i).

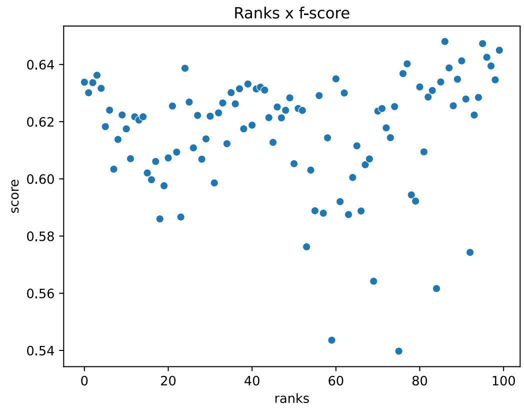


Figure 2: IT5 results (in terms of weighted f-scores) for each fine-tuning on the different sets of labels S_j.

At first glance, we can readily observe that the choice of words used to represent the classification categories has a considerable impact on the model's average performance. Indeed, we can see that the classification scores vary significantly, ranging from a minimum of 0.54 (rank 75) to a maximum of 0.65 (rank 86). Additionally, it is worth noting that the model trained with S_0, which contains the original translated labels, achieved an f-score of 0.63. This result indicates that simply using the original labels directly still provides a competitive performance. However, the significant fluctuations in the classification scores among the different sets S_j suggest that certain labels may still offer better performance than the original ones, while others may introduce noise or ambiguity, resulting in sub-optimal outcomes.

Interestingly, these findings appear to diverge from previous studies [14, 16], where the role of label representation was underestimated. While being a task-dependent issue, the role of label representation seems to have a large impact on model performance, especially for lower frequency labels, going as far as making certain labels range from being completely unpredictable to reaching satisfactory performances.

That being said, despite the differences in terms of weighted f-scores, there does not seem to be a clear correlation between the model's performance and the degree of "semantic" distance between the chosen labels and the original ones (represented by the rank j of the representation set). In fact, as the cosine similarity decreases between the selected representations and the original ones (from rank 0 to rank 99), there is no apparent trend in f-score values.

Per-label results In order to gain a more precise insight into the impact of the tested labels, Figure 3 illus-

[1] https://huggingface.co/gsarti/it5-base

[2] List of translated labels: *anime, automobilismo, bicicletta, sport, natura, metal detector, medicina, celebrità, fumo, intrattenimento* and *tecnologia*.

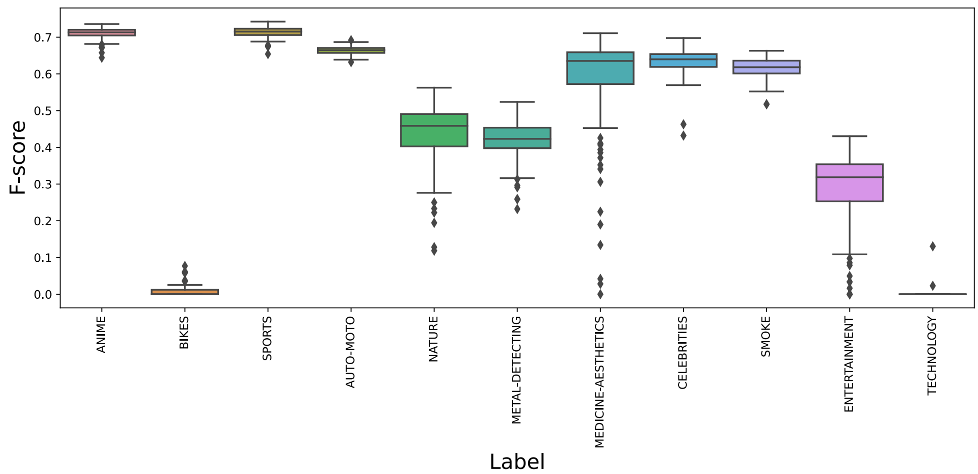


Figure 3: Boxplot showing the variation of the f-scores using different labels according to each classification category.

trates the variation of f-scores obtained with the 100 different sets of labels (S_i) for each individual category. Firstly, we can observe that the average results can vary significantly depending on the category under consideration. For instance, IT5 shows promising average performance in classifying posts related to *Anime*, *Sports* or *Auto-Moto*, while encountering difficulties in identifying posts annotated with the topics *Bikes* and *Technology*. This is possibly due to the fact that the posts belonging to the former categories are the most frequent in the entire dataset. Particularly noteworthy is the fact that, across almost all tested ranks, the model failed to correctly identify any posts related to *Technology*. This issue is likely attributed to the limited representation of this category within the dataset, further compounded by the original dataset configuration having more examples in the test set than in the training set (51 and 85 samples in the training and test sets respectively).

Analyzing the variation of results based on the labels used for representing the categories, we observe, in line with Figure 1, that the choice of the label often has a significant impact on the model's performance. While some labels exhibit relatively stable results with minor variations across different representations, such as *Anime*, *Bikes*, *Sports* and *Auto-Moto*, there are other instances where the selected labels lead to remarkable fluctuations in the model's performance. Notably, this behaviour emerges especially in the identification of posts related to *Nature*, *Metal-Detecting*, *Medicine-Aesthetics* and *Entertainment*. For these categories, IT5's classification performance can change drastically depending on the specific label. In some cases, the model manages to achieve quite good results, accurately classifying posts with a high f-score. However, in other instances, it struggles significantly, making erroneous classifications for the majority of cases. For instance, in the case of *Medicine-Aesthetics*, the f-score reaches a maximum of 0.71 when the label is represented by the term *acuto* but it fails to correctly classify any instance (f-score = 0) when the label is represented as *proprio*. This highlights how the choice of the label can significantly impact IT5's classification performance across different topics and therefore, suggests the importance of exploring optimized selection strategies to maximize the model performance.

To obtain a more comprehensive qualitative perspective of these findings, we include in Figure 4 the top and bottom 10 representations that maximized/minimized the f-score values for the four aforementioned categories. As we can observe, among the four considered categories, only one (*Medicine-Aesthetics*) contains the original label, i.e. the one with cosine similarity equal to 1 (*medicina*), in the top 10 representations. For the other categories, the absence of the original label seems to suggest that the chosen word for the label, which should be the closest one to the reference topic, may not be the one that can maximize the results. When analyzing individual words, it becomes evident that not all words contributing to the model's best performance belong exclusively to the domain of the considered category. Surprisingly, words such as *cinema* and *sitcom*, seemingly related to the *Entertainment* domain, are among those that most negatively impact the model's f-scores. Nevertheless, *Medicine-Aesthetics* shows an exception, with several words aligned with the category's domain, e.g. *benessere*, *medicina*, *dottoressa* e *sensibilità*. Lastly, it is worth noticing that the performance drop is mostly label-dependent, and there is a significant

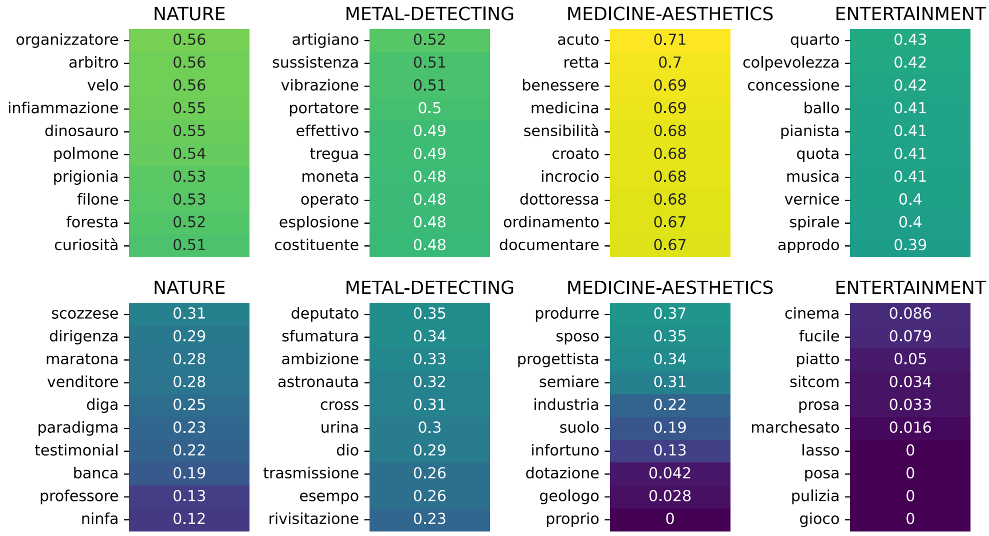


Figure 4: Top and bottom 10 labels that maximize/minimize the results for the most varying categories (*Nature*, *Metal-Detecting*, *Medicine-Aesthetics* and *Entertainment*).

Categories	Spearman	p-value
Entertainment	0.29	0.003 *
Auto-Moto	0.05	0.62
Medicine-Aesthetics	-0.02	0.85
Bikes	-0.05	0.61
Anime	-0.10	0.37
Technology	-0.12	0.21
Smoke	-0.20	0.04 *
Sports	-0.22	0.03 *
Nature	-0.25	0.01 *
Metal-Detecting	-0.35	0.00 *
Celebrities	-0.45	0.00 *

Table 2
Spearman correlations between f-scores and label similarities (cosine similarity) for each category. Statistically significant correlations are marked with *.

difference between the most- and least-performing representations for the four categories. In fact, while *Nature* and *Metal-Dectecting* exhibit a relatively modest decrease (around .20 f-score points), *Medicine-Aesthetics* and *Entertainment* display a far more pronounced difference in performance.

3.1. Correlating Model Performance and Tested Representations

Having analyzed the model's performance and assessed the impact of words used to represent the categories on the classification results, we decided to explore the existence of any relationship between the model's performance and the employed words.

Semantic Similarity Initially, we aimed to ascertain whether there is a correlation between the words that are more/less semantically similar to the original categories and the performance of IT5. To achieve this, we computed the Spearman correlation between the T5 model's performance and the cosine similarity values calculated to construct the 100 sets for each label S_j. The results of these correlations are presented in Table 2[3]. As observed, 6 out of the 11 classification categories exhibit statistically significant correlations. Among these, only one correlation is positive (*Entertainment*), while the others show negative correlation values. This outcome is quite unexpected as it seemingly implies that the improvement in the model's performance is linked to a decrease in semantic similarity. However, it is crucial to emphasize that the correlation values are not particularly high, and thus, we cannot draw any conclusion about these results. Moreover, it is important to consider that while cosine similarity can serve as a useful measure of similarity between embeddings, it may not encompass the entire semantic space.

Internal Similarity Since the similarity between selected labels' within each set could potentially impact the model's performance, we conducted an additional test to investigate whether higher semantic similarity among

[3]In Appendix A we also reported the scatterplots showing the relationship between f-scores and cosine similarity values for these labels.

representations within a set could negatively affect the performance of IT5. To achieve this, we computed the *"inner similarity"* of each set, defined as the average cosine similarity of all possible distinct label combinations[4]. Subsequently, we computed the Spearman correlation between each set's *"inner similarity"* and the f-scores obtained by the model fine-tuned with it. Although the values of *"inner similarities"* vary considerably across the sets (ranging from a similarity of 0.69 for rank 0 to 0.38 for rank 100), we did not find a statistically significant correlation with the model's performance (Spearman = 0.01, p-value = 0.90). These results suggest that, despite the sets exhibited considerable variation in terms of inner similarity, the similarity between the representation didn't plainly affect the model's performance.

Representations Frequencies Finally, since the aforementioned results have demonstrated that different labels have an impact on the model's performance, we decided to investigate whether this impact could be somehow related to the frequency of these representations within the model's training dataset. To this end, we computed the absolute frequency of each label used in our experiments (11 labels per 100 sets, totalling 1100 words) within the Italian version of the mC4 Corpus, i.e. the corpus on which IT5 was trained. Subsequently, we calculated the correlation between the scores obtained by IT5 for each label of each set R_i and the corresponding frequencies of each label found in the mC4 corpus. Among the 11 categories present in the dataset, only one showed a statistically significant correlation, *Smoke*, with a Spearman correlation value of -0.25[5]. This result suggests that, at least for this particular category, a decrease in the label's frequency in the training corpus corresponds to an increase in the model's performance. However, the fact that only one representation exhibits a significant correlation and that this correlation is not particularly high once again prevents us from drawing any conclusive findings. Thus, it underscores the need to explore other strategies in the future for label selection.

4. Conclusion

In this work, we presented an evaluation of the impact of label selection on the performance of a Sequence-to-Sequence Model in a classification task. By fine-tuning the Italian version of the T5 model on a topic classification task, we explored various sets of labels and examined their influence on the model's performance.

Our results indicate that the choice of words used to represent the classification categories can have a significant impact on the model's performance. While some labels led to competitive results, others resulted in sub-optimal outcomes, with noteworthy variations in the classification scores. This finding diverges from previous studies that suggested label representations had little impact on model performance.

Interestingly, the correlation between the model's performance and the degree of "semantic" distance between the chosen labels and the original ones was not clear. While some labels exhibited statistically significant correlations, they were either positive or negative, indicating that higher or lower semantic similarity did not consistently lead to better performance.

In conclusion, our findings suggest that the choice of the label is not a trivial matter and can have a significant impact on the performance of Sequence-to-Sequence Models in classification tasks. To maximize performance, it is essential to explore optimized label selection techniques that are carefully selected and tailored to the specific task and dataset.

Future research could focus on developing more sophisticated methods for label selection, taking into account not only semantic similarity but also other relevant factors. Additionally, it would be valuable to investigate the generalizability of these findings across other languages and models, and in order to gain a more comprehensive understanding of the influence of label selection on different NLP tasks.

Acknowledgments

This work has been supported by the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU.

References

[1] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer., J. Mach. Learn. Res. 21 (2020) 1–67.

[2] V. Sanh, A. Webson, C. Raffel, S. Bach, L. Sutawika, Z. Alyafeai, A. Chaffin, A. Stiegler, T. Le Scao, A. Raja, et al., Multitask prompted training enables zero-shot task generalization, in: The Tenth International Conference on Learning Representations, 2022.

[3] V. Aribandi, Y. Tay, T. Schuster, J. Rao, H. S. Zheng, S. V. Mehta, H. Zhuang, V. Q. Tran, D. Bahri, J. Ni, et al., Ext5: Towards extreme multi-task scaling for transfer learning, in: International Conference on Learning Representations, 2021.

[4] B. McCann, N. S. Keskar, C. Xiong, R. Socher, The natural language decathlon: Multitask learn-

[4]As defined in Sec. 2.3, a label is represented as the average embedding of each subtoken in the string.

[5]The table with all the correlations is reported in Appendix B.

ing as question answering, arXiv preprint arXiv:1806.08730 (2018).
[5] N. S. Keskar, B. McCann, C. Xiong, R. Socher, Unifying question answering, text classification, and regression via span extraction, arXiv preprint arXiv:1904.09286 (2019).
[6] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners (2019).
[7] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.
[8] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, et al., Scaling instruction-finetuned language models, arXiv preprint arXiv:2210.11416 (2022).
[9] C. Song, F. Cai, J. Zheng, W. Chen, Z. Pan, Metric sentiment learning for label representation, in: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, CIKM '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 1703–1712. URL: https://doi.org/10.1145/3459637.3482369. doi:10.1145/3459637.3482369.
[10] W. Jiang, Y. Zhang, J. Kwok, Effective structured prompting by meta-learning and representative verbalizer, in: International Conference on Machine Learning, PMLR, 2023, pp. 15186–15199.
[11] K. Ji, Y. Lian, J. Gao, B. Wang, Hierarchical verbalizer for few-shot hierarchical text classification, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 2918–2933. URL: https://aclanthology.org/2023.acl-long.164.
[12] T. Schick, H. Schmid, H. Schütze, Automatically identifying words that can serve as labels for few-shot text classification, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 5569–5578. URL: https://aclanthology.org/2020.coling-main.488. doi:10.18653/v1/2020.coling-main.488.
[13] G. Cui, S. Hu, N. Ding, L. Huang, Z. Liu, Prototypical verbalizer for prompt-based few-shot tuning, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 7014–7024. URL: https://aclanthology.org/2022.acl-long.483. doi:10.18653/v1/2022.acl-long.483.
[14] X. Chen, J. Xu, A. Wang, Label representations in modeling classification as text generation, in: Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop, Association for Computational Linguistics, Suzhou, China, 2020, pp. 160–164. URL: https://aclanthology.org/2020.aacl-srw.23.
[15] A. Cimino, F. Dell'Orletta, M. Nissim, Tag-it@ evalita 2020: Overview of the topic, age, and gender prediction task for italian, Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (2020).
[16] M. Papucci, C. De Nigris, A. Miaschi, F. Dell'Orletta, Evaluating text-to-text framework for topic and style classification of italian texts, in: Proceedings of the Sixth Workshop on Natural Language for Artificial Intelligence (NL4AI 2022) co-located with 21th International Conference of the Italian Association for Artificial Intelligence (AI* IA 2022), 2022.
[17] G. Sarti, M. Nissim, It5: Large-scale text-to-text pretraining for italian language understanding and generation, ArXiv preprint 2203.03759 (2022). URL: https://arxiv.org/abs/2203.03759.
[18] V. Basile, M. Di Maro, D. Croce, L. Passaro, Evalita 2020: Overview of the 7th evaluation campaign of natural language processing and speech tools for italian, in: 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop, EVALITA 2020, volume 2765, CEUR-ws, 2020.
[19] A. Maslennikova, P. Labruna, A. Cimino, F. Dell'Orletta, Quanti anni hai? age identification for italian., in: CLiC-it, 2019.
[20] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mT5: A massively multilingual pre-trained text-to-text transformer, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 483–498. URL: https://aclanthology.org/2021.naacl-main.41. doi:10.18653/v1/2021.naacl-main.41.
[21] M. Baroni, S. Bernardini, A. Ferraresi, E. Zanchetta, The wacky wide web: a collection of very large linguistically processed web-crawled corpora, Language resources and evaluation 43 (2009) 209–226.

A. Appendix A

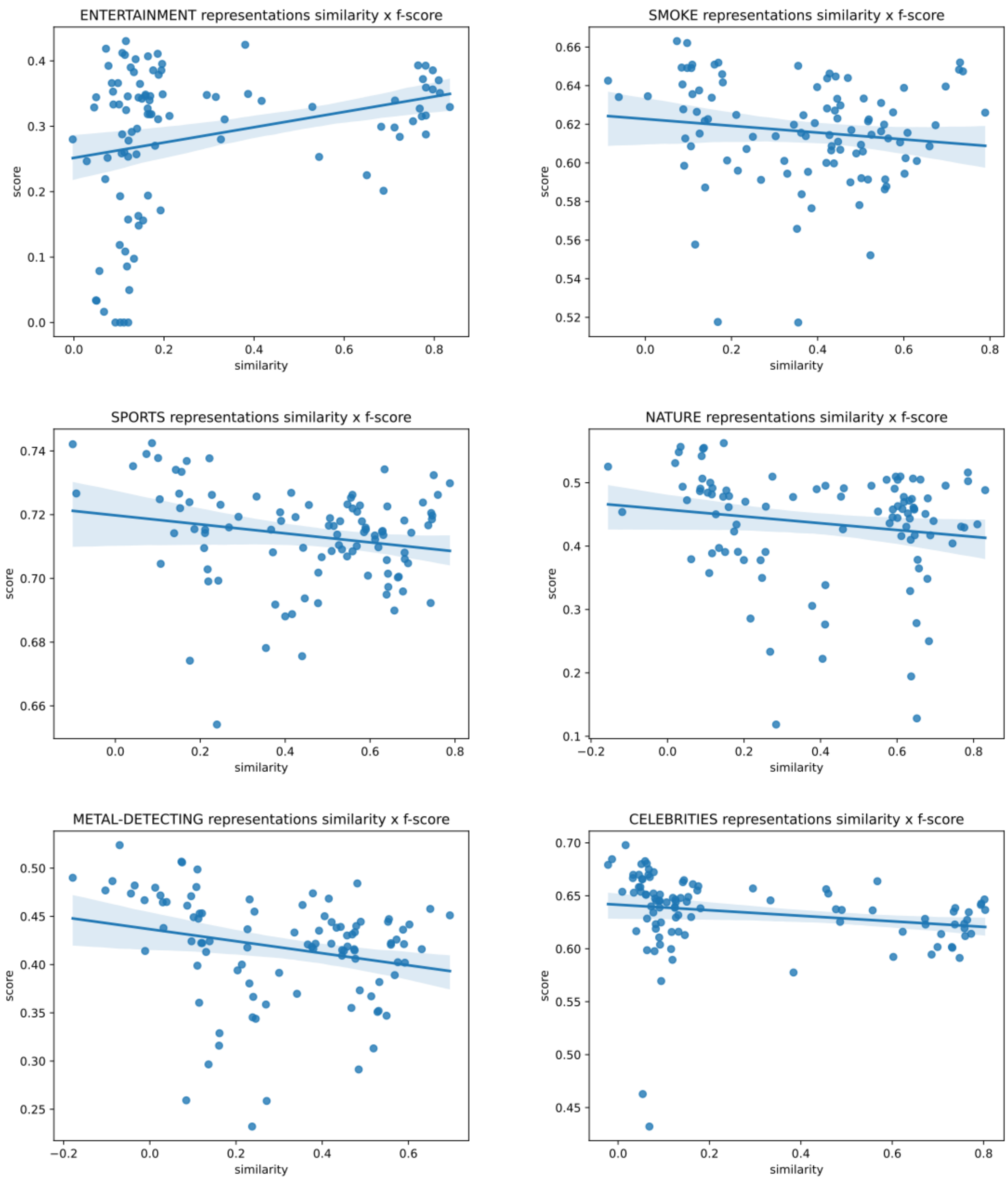


Figure 5: Scatterplot showing the relationship between f-scores and cosine similarity values for the 6 categories that exhibited a statistically significant correlation.

B. Appendix B

Categories	Spearman	p-value
Medicine-Aesthetics	0.13	0.20
Nature	0.06	0.54
Sports	0.04	0.66
Bikes	0.01	0.94
Technology	-0.02	0.88
Anime	-0.02	0.84
Entertainment	-0.03	0.75
Auto-Moto	-0.05	0.62
Metal-Detecting	-0.06	0.57
Celebrities	-0.06	0.54
Smoke	-0.25	0.01 *

Table 3
Spearman correlations between f-scores and labels absolute frequencies (computed in the Italian mC4 Corpus) for each category. Statistically significant correlations are marked with *.

Are All Languages Equal? Curriculum Learning over Different Languages

Giulia Pucci[1], Leonardo Ranaldi[1,2] and Fabio Massimo Zanzotto[1]

[1]University of Rome Tor Vergata
[2]Idiap Research Institute, Switzerland

Abstract
Curriculum Learning (CL) is emerging as a relevant technique to reduce the cost of pre-training Large Language Models. The idea, tested for the English language, is to train LLMs by organizing training examples from the simplest to the most complex. Complexity measures may depend on the specific language. Hence, this paper aims to investigate whether CL and the complexity measure can be easily exported to other languages. For this reason, we present a set of linguistically motivated measures to determine the complexity of examples, which has been used in English: these measures are based on text length, rarity, and comprehensibility. We then test the approach to two Romance languages: Italian and French. Our results show that the technique can be easily exported to languages other than English without adaptation.

Keywords
Efficient Pre-training, Multilingual LLMs, Natural Language Processing,

1. Introduction

Transformers-based models have disrupted natural language understanding methods outperforming previous methods and sometimes even humans in many tasks [1, 2, 3, 4]. Unsupervised learning on huge corpora, no matter the domain, seems to be the way to increase performance; however, besides the onerous costs, there are difficulties with the data.

Therefore, this results in a significant carbon footprint [5], contrary to global sustainability goals. There are many approaches to address the AI carbon footprint problem, ranging from using more carbon-efficient energy sources to applying efficient AI models and training algorithms. Indeed, Transformers seem to be only huge memories [6, 7] and, thus, better ways to train these models are necessary. Bengio et al. [8] in Curriculum Learning (CL) proposes a specific class of efficient training strategies for deep learning models.

The naïve approach for training Large Language Models involves feeding textual batches randomly sampled from the training corpora is re-visited in the CL, where the model is refined with a sequence of progressively more challenging examples [9]. This is motivated by and emulates how humans learn, starting with more straightforward concepts and gradually building up more complex ones. Soviany et al. [10] show that CL helps the model to perform better and converge faster.

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
name.username@uniroma2.it (G. Pucci);
name.username@idiap.ch (L. Ranaldi);
name.username@uniroma2.it (F. M. Zanzotto)

In this paper, we deeply analyze the learning divergencies training from scratch with BERT [11] and GPT2 [12] on the same corpus in multiple languages. Furthermore, following our CL-LRC metrics [13] based on length, rarity, and comprehensibility, computational costs are reduced, and the divergences are filled.

Hence, using the same small corpus in three different languages, English (original), Italian, and French (translated), experimental results show that loss values during the training vary in the different languages. Moreover, this difference seems to be softened in terms of perplexity scores when the pre-training block-sizes increase incrementally.

2. Background

Optimizing the use of computational resources to increase the learning capabilities of Large Language Models (LLMs) is a widely studied problem. The main approaches are based on architecture, learning, and, finally, data. Although current optimization methods at the architectural level have demonstrated extensive functionality on further fine-tuning, there still needs to be gaps at the pre-training level.

Clark et al. [14] propose a method for reducing computational costs by modifying the Masked Language Models with a discriminator, but it may have limitations in tasks that require a deep understanding of long-term dependencies or complex relationships between words. Sanh et al. [15] proposed parameter reduction techniques and obtained a lightweight version of BERT that is less compelling than the original in adapting parameters on specific tasks.

Finally, the last approach in vogue concerns the ef-

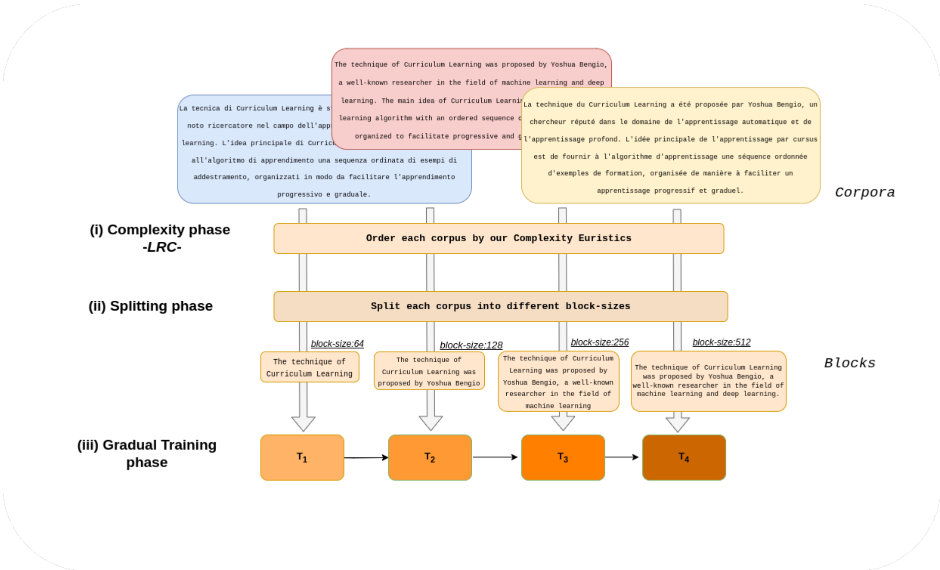


Table 1
Curriculum Learning and LRC pre-training overview.

ficient adjustment of parameters. Parameter-Efficient Tuning (PEFT) is an efficient technique for tuning a small portion of model parameters and freezing others. Standard techniques for PEFT: LoRA [16], Prefix Tuning [17], P-Tuning [18] reduce computational and storage and maintain the performance. However, these PEFT methods are applied to fine-tuning a model for a specific task and not to pre-training from scratch. While these topics have been extensively studied, the data-level approach has yet to be explored.

Many studies have found that the multi-headed self-attention mechanism requires tremendous computational effort. Since each head of this mechanism appears to be more attentive to local dependencies than global ones [19, 20, 21], training local self-attention in shorter blocks seems to be less complex than training global self-attention in more extended blocks. Nagatsuka et al. [9] proposed a Curriculum Learning (CL) strategy concentrating on hands-on self-attention mechanism training to enhance this aspect. They applied the strategy directly to BERT pre-training, manipulating the size of the input text block in the self-attention mechanism as a measure of difficulty.

Further the world of transformer-based models, many CL studies have used sentence length, external resources, or input sequences to measure difficulty in various NLP tasks such as in parsing tasks [22], reading comprehension [23], and concept masking for pre-training of the knowledge graph-related models [24].

In this paper, to solve the gap of LLMs in learning English, Italian and French, we studied the difficulties faced in learning more languages. We propose text complexity techniques combined with input text block-size in the context of the self-attention mechanism. The two approaches measure the difficulty of pre-training two language models: BERT [11] and GPT2 [12]. Our proposal adds to the incremental CL brought in [9], an additional light step for calculating the pre-training text complexity. Our model performs better than the baselines and methods proposed in [9] regarding loss and perplexity.

3. Our Methods

Starting from the fact that language has a structure that varies between different languages, we searched for a strategy to alleviate these divergences [25, 26]. Hence organizing the examples during pre-training could improve the model's performance. Therefore, starting from the concept of Curriculum Learning (CL) shown by Bengio et al. [8], according to which learning algorithms perform better when the data are presented following the current competencies of the model, we used the methodology proposed in [9] applying an incremental learning technique on increasing block-sizes. We propose to use these techniques in different languages and extend the work done with a generative model. Finally, we study the impact of language complexity by intruding LRC, a measure used to determine the complexity of examples during pre-training before standard CL.

The application of the CL-LRC method consists of

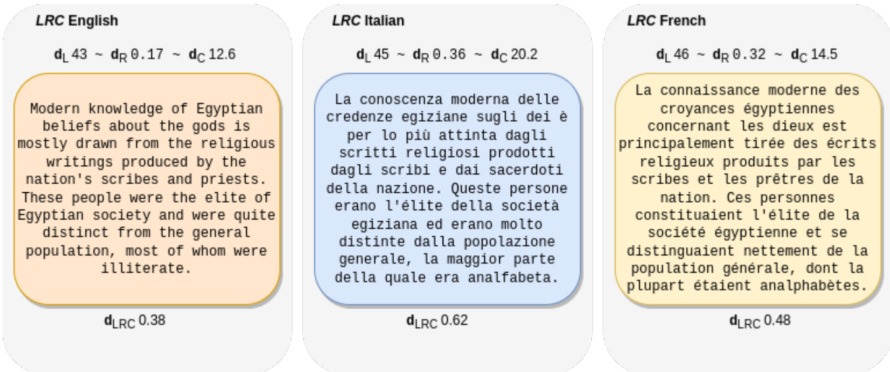


Table 2
Examples of the complexity values produced by the metrics defined in Section 3.1.

three steps (Figure 1): (i) sorting the corpus according to our complexity measure starting from the least complex sentences to the most complex ones; (ii) partitioning the corpus according to input blocks of predefined sizes; (iii) stepwise pre-training by increasing the block size.

3.1. Complexity

The increasing block-size techniques and complexity measures are our method's core. While the dynamic resizing technique is fixed and does not change in different scenarios, the complexity of a text example is challenging to define.

Since the tasks used in pre-training should aim to learn language from context, precisely as humans do, organizing the complexity of examples could improve CL in LLMs.

We propose combining three factors: the number of tokens or sentence length, the repetitiveness or rarity of words in the corpus, and finally, the comprehensibility or, more commonly, the Flesch-Kincaid readability metric. Aggregating these three heuristics forms d_{LRC}, one of the foundational elements of our framework. Hence, we denote our training corpus as a collection of D sentences, $\{s_i\}_{i=0}^{D}$, where each sentence is a sequence of words denoted with $s_i = \{w_0^i, w_1^i, ..., w_n^i\}$.

Number of tokens The number of occurrences or sentence length is critical since longer sequences are more difficult to encode, as the possibility of them being cut is high. Therefore, longer sentences would be more prone to losing context during the pre-training tasks. We compute sentence length for each period s_i of our corpus D:

$$d_L(s_i) = length(s_i) \tag{1}$$

Following obtaining the $d_{L_{max}}$ and $d_{L_{min}}$, we normalize the values:

$$\hat{d}_L(s_i) = \frac{d_L(s_i) - d_{L_{min}}}{d_{L_{max}} - d_{L_{min}}}, \forall i \in [0, |D|]. \tag{2}$$

Rarity The repetitiveness of words is a significant factor. We use the metric introduced in [27] where rarity is defined as the probability product of unigrams. This metric represents sentence information since the scores of longer sentences are the sum of more words and thus are likely to be more meaningful. Given a corpus of sentences, $\{s_i\}_{i=0}^{D}$, the complexity metric for word rarity is defined as:

$$d_R(s_i) \triangleq -\sum_{k=1}^{N_i} \log p\left(w_k^i\right) \tag{3}$$

where we use logarithms of word probabilities. The component $p(w)$ is defined as:

$$p(w) \triangleq \frac{1}{N_{total}} \sum_{i=1}^{M} \sum_{k=1}^{N_i} \mathbb{1}_{w_k^i = w} \tag{4}$$

for each w unique word in a corpus and $\mathbb{1}_{condition}$, is the indicator function equal to 1 if its condition is satisfied or 0. We compute this value for each sentence s_i of our corpus D, obtaining the $d_{R_{max}}$ and $d_{R_{min}}$ and we normalize the values:

$$\hat{d}_R(s_i) = \frac{d_R(s_i) - d_{R_{min}}}{d_{R_{max}} - d_{R_{min}}}, \forall i \in [0, |D|]. \tag{5}$$

Readability Metric Comprehensibility or, more commonly, readability may be related to the speed of perception, reflex blink technique, reading speed, reading fatigue, cognitively motivated characteristics, and word

	English		Italian		French	
Model	**Loss**	**Perplexity**	**Loss**	**Perplexity**	**Loss**	**Perplexity**
$Baseline$ *(BERT)*	2.74	270.42	3.96	336.38	4.19	304.20
Baseline$_{LRC}$ *(BERT)*	2.53	254.23	4.06	330.21	4.38	296.71
Total-Curriculum (BERT)	2.56	250.64	**3.83**	324.73	4.06	300.18
$Curriculum_{LRC}$ *(BERT)*	**2.26**	**245.348**	3.86	**304.70**	**3.46**	**287.16**
$Baseline$ *(GPT2)*	4.33	122.37	**6.24**	135.48	6.36	125.05
Baseline$_{LRC}$ *(GPT2)*	4.20	119.36	6.46	122.32	6.83	122.26
Total-Curriculum (GPT2)	3.97	117.29	6.32	120.97	**6.09**	124.63
$Curriculum_{LRC}$ *(GPT2)*	**3.55**	**96.66**	6.43	**116.65**	6.38	**108.23**

Table 3
Loss and Perplexity after Pre-training on the test set.

difficulty for a specific reader. Unfortunately, it is not always possible to collect these characteristics.

We used the Flesch-Kincaid metric [28] as an assessment tool for text comprehension. This metric is based on the length of sentences and words within a text by quantifying difficulty with a score. The lower the score, the easier it is to read and understand the text. We use the following formula:

$$d_C(s_i) = 0.39\frac{avg(d_L(s_i))}{100} + 11.8\frac{avg(d_L(w_i))}{100} - 15.59 \tag{6}$$

where $avg(d_L(s_i))$ average sentence length is the number of words in a sentence divided by the number of sentences, and $avg(d_L(w_i)$ is the average word length, i.e., does the number of words divides the number of syllables per word. The value 0.39 is used to scale the effect of the average sentence length to compare it to the effect of the average word length, weighted by 11.8. The final score is then adjusted by subtracting the value of 15.59, which adjusts the score scale to match the grading levels used in education more closely. We calculate this value for each sentence s_i and obtain the maximum $d_{C_{max}}$ and the minimum $d_{C_{min}}$ scores. Finally, we normalize these values:

$$\hat{d}_C(s_i) = \frac{d_C(s_i) - d_{C_{min}}}{d_{C_{max}} - d_{C_{min}}}, \forall i \in [0, |D|]. \tag{7}$$

3.2. Applying Complexity Heuristics

In the first phase, we compute the complexity of each sentence $d_{LRC}(s_i)$ by adding the normalized values of length $\hat{d}_L(s_i)$, rarity $\hat{d}_R(s_i)$, and readability score $\hat{d}_C(s_i)$, that is:

$$d_{LRC}(s_i) = \hat{d}_L(s_i) + \hat{d}_R(s_i) + \hat{d}_C(s_i) \tag{8}$$

Then, we sort the sentences of the original corpus by order of increasing complexity before the pre-training phase. Finally, we recompose the re-ordered corpus ready for pre-training.

3.3. Splitting a Corpus-Based on Block-sizes

Secondly, following the work of Nagatsuka et al. [9], we split the original corpora into training samples of the specified size. Each input text (block) for BERT and GPT2 pre-training should not be linguistically consistent as a sentence but a fixed interval of contiguous text. Thus, it is not guaranteed that the input is a period or begins with the first word of a sentence. Moreover, after extensive experiments, Liu et al. [29] argue that the input sequence should be at most 512 tokens. However, we follow an incremental approach that differs from the static sizing of 512 tokens per batch. The difference is the order, which is the reason why it could be easier for a Transformer to learn by order of complexity. We train a Byte-Pair Encoding (BPE) at the byte level [30] to split the raw text into a sequence of tokens. Byte-level BPE allows the decomposition of words, including words outside the vocabulary likely to appear during testing, especially when using a small training dataset. In the experiment, we set the vocabulary size to $20,000$.

3.4. Gradual Training

Using the corpus sorted by complexity order, we train a step model with four block sizes, namely 64, 128, 256, and 512. At first, we train the model with the shortest block-size, 64, for an arbitrary number of steps. Then, we continue to train the model with block-sizes of 128 and 256, respectively, for the same number of steps. Finally, we finish with the largest block-size of 512.

4. Experimental Results and Discussion

We evaluated our proposed CL-LRC approach in model performance in the experiments. Therefore, we show that performances increase to the proposed state of the art in [9]. We use Wikitext-2 [31] to reproduce the results proposed. Hence, we perform the pre-training from scratch for BERT [11] and GPT2 [30]. Therefore, we investigated perplexity, loss, and learning curves during and at the end of the pre-training. All experiments were performed on two NVIDIA RTX A6000 with 48 GB of memory. The code and model will be released for further research.

4.1. Data

BERT and GPT2 are pre-trained with huge corpora, i.e., bookcorpus and Wikipedia-dump with about 3 billion words [32]. In this work, we used Wikitext-2 [31], a small corpus for simulations, allowing pre-training with a limited computational resource. Wikitext-2 is a standard language model corpus with 720 good-quality articles from English Wikipedia. In addition, we introduced two further corpora from the Italian and French translations of Wikitext-2.

4.2. Experimental setup

We use the same corpus in three different languages to analyze learning divergences between different languages. Hence, we perform pre-training from scratch with the baseline methods, and then with complexity metrics (*Baseline*$_{LRC}$), the Total-Curriculum (CL proposed in [9]), and our CL-LRC called $Curriculum_{LRC}$ using the settings proposed in [9]. In particular, in our $Curriculum_{LRC}$, we sort the corpora according to complexity, split the corpora according to the difficulty level of the training samples, and perform the pre-training phase by increasing the block size. We performed these steps for all corpora and pre-train BERT and GPT2 from scratch. Finally, we report the losses during learning, the final losses on the evaluation set, and the average perplexity of different cuts of the evaluation set.

4.3. Results

Difficulties in learning a language depend on the complexity of the language itself. However, it can be alleviated using curricular techniques and greatly improved using linguistically motivated methods, maintaining reduced training times as shown in Table 6. These conclusions derive from the pre-training results from scratch in three languages using Baseline, Total-Curriculum, and our CL-LRC techniques visible in Table 3. In Figure 5, it can be observed from the baselines of the different corpora that English language learners, on average, are less perplexed. Moreover, the $Curriculum_{LRC}$ outperforms the others in all corpora. However, the batch-size increase supports the performance achieved by Curriculum Learning. Finally, in Figure 4, learning curves explain the trade-off between pre-training steps and loss values.

4.3.1. Our Methods vs CL & Baseline

The linguistically motivated pre-training by our metrics has improved the technique proposed in [9] and outperformed the baseline models. In particular, $Curriculum_{LRC}$*(BERT)* outperforms the version without LRC of 5 points for English and more than 30 points for Italian and French over perplexity scores. The same is true for GPT2 with less striking results (ranging from 16 to 4 points). Hence, this measure seems to have less impact on the Italian and French, as we can observe from *Baseline*$_{LRC}$ models for English pre-training and others. Finally, in Fig. 5, we can observe a clear gap in perplexity in the presence of portions of text with a small number of tokens, which is reduced to zero or almost zero when the number of tokens is more significant.

4.3.2. Languages over Complexity

With the aim of studying intrinsic learning difficulties, we propose our line of experiments from the same corpus translated into three different languages: English (original), French, and Italian. We can observe that the models started from scratch have more difficulty learning the French and Italian corpora than the English ones. We believe this result's origin stems from the structure and complexity of the languages concerned. It is widely known that being both Romance languages, French and Italian have a very complex grammatical structure, very different from English. Regarding verb conjugation, while English verbs have relatively simple and regular conjugation patterns, French and Italian ones are very intricate, with various tenses, moods, aspects, and verb endings. For the agreement rules, unlike French and Italian, English has no grammatical gender distinction, so there is no agreement based on gender. Moreover, in contrast to the skinny use in English, French, and Italian have complex systems of clauses and subordination. Therefore, it is more difficult for a non-native speaker of Italian or French to learn these two languages from scratch, for the same reasons it is also for the models we tested.

4.4. Convergence Speed & Training time

Our CL-LRC outperforms the *Total-Curriculum* regarding loss during pre-training. However, in Figure 4, it can be seen that the loss of the basic model converges to around 50; in contrast, both models with curriculum steadily decrease and reach a higher convergence rate. Moreover, it can be observed that the loss of the curriculum-based model decreased steadily whenever the difficulty of the training samples was changed. Finally, in Table 6, it is possible to observe how curricular approaches can significantly reduce training time and consecutively consumption and costs.

5. Conclusion

In this paper, we explored the effectiveness of Curriculum Learning (CL) in reducing the cost of pre-training and increasing the results. We trained LLMs by organizing examples from the simplest to the most complex, thereby leveraging the concept of complexity measures. Hence, we pre-trained from scratch BERT and GPT2 using standard baselines and CL approaches. After deep analysis, we show that divergence in learning can be mitigated using CL approaches reinforced by measures to determine the complexity of examples. These measures, applied during pre-training to sort the corpus according to complexity, show outstanding results. While the original approach was tested and validated for the English language, this research aimed to investigate whether CL and its associated complexity measure could be applied to other languages without significant adaptation. Experiments conducted in a low-resource environment show that the proposed method leads to better performance in terms of loss during learning and perplexity on test data.

In future works, we will continue to propose pedagogically motivated mechanisms to analyze weaknesses [33] and empower Cross-lingual abilities to deliver multistep-reasoning answers [34].

References

[1] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, S. Bowman, GLUE: A multi-task benchmark and analysis platform for natural language understanding, in: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 353–355. URL: https://aclanthology.org/W18-5446. doi:10.18653/v1/W18-5446.

[2] L. Ranaldi, A. Nourbakhsh, E. S. Ruzzetti, A. Patrizi, D. Onorati, M. Mastromattei, F. Fallucchi, F. M. Zanzotto, The dark side of the language: Pre-trained transformers in the DarkNet, in: R. Mitkov, G. Angelova (Eds.), Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing, INCOMA Ltd., Shoumen, Bulgaria, Varna, Bulgaria, 2023, pp. 949–960. URL: https://aclanthology.org/2023.ranlp-1.102.

[3] L. Ranaldi, G. Pucci, Knowing knowledge: Epistemological study of knowledge in transformers, Applied Sciences 13 (2023). URL: https://www.mdpi.com/2076-3417/13/2/677. doi:10.3390/app13020677.

[4] L. Ranaldi, E. S. Ruzzetti, F. M. Zanzotto, PreCog: Exploring the relation between memorization and performance in pre-trained language models, in: R. Mitkov, G. Angelova (Eds.), Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing, INCOMA Ltd., Shoumen, Bulgaria, Varna, Bulgaria, 2023, pp. 961–967. URL: https://aclanthology.org/2023.ranlp-1.103.

[5] E. Strubell, A. Ganesh, A. McCallum, Energy and policy considerations for deep learning in NLP, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 3645–3650. URL: https://aclanthology.org/P19-1355. doi:10.18653/v1/P19-1355.

[6] L. Ranaldi, A. Nourbakhsh, E. S. Ruzzetti, A. Patrizi, D. Onorati, F. Fallucchi, F. M. Zanzotto, The dark side of the language: Pre-trained transformers in the darknet, in: Proceedings of RANLP, 2023.

[7] L. Ranaldi, E. S. Ruzzetti, F. M. Zanzotto, Precog: Exploring the relation between memorization and performance in pre-trained language models, in: Proceedings of RANLP, 2023.

[8] Y. Bengio, J. Louradour, R. Collobert, J. Weston, Curriculum learning, in: Proceedings of the 26th annual international conference on machine learning, 2009, pp. 41–48.

[9] K. Nagatsuka, C. Broni-Bediako, M. Atsumi, Pre-training a BERT with curriculum learning by increasing block-size of input text, in: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), INCOMA Ltd., Held Online, 2021, pp. 989–996. URL: https://aclanthology.org/2021.ranlp-1.112.

[10] P. Soviany, R. T. Ionescu, P. Rota, N. Sebe, Curriculum learning: A survey, 2022. arXiv:2101.10382.

[11] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[12] A. Radford, K. Narasimhan, Improving language understanding by generative pre-training, 2018.

[13] L. Ranaldi, G. Pucci, F. M. Zanzotto, Modeling easiness for training transformers with curriculum learning, in: R. Mitkov, G. Angelova (Eds.), Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing, INCOMA Ltd., Shoumen, Bulgaria, Varna, Bulgaria, 2023, pp. 937–948. URL: https://aclanthology.org/2023.ranlp-1.101.

[14] K. Clark, M.-T. Luong, Q. V. Le, C. D. Manning, ELECTRA: Pre-training text encoders as discriminators rather than generators, in: ICLR, 2020. URL: https://openreview.net/pdf?id=r1xMH1BtvB.

[15] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, ArXiv abs/1910.01108 (2019).

[16] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-rank adaptation of large language models, in: International Conference on Learning Representations, 2022. URL: https://openreview.net/forum?id=nZeVKeeFYf9.

[17] X. L. Li, P. Liang, Prefix-tuning: Optimizing continuous prompts for generation, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 4582–4597. URL: https://aclanthology.org/2021.acl-long.353. doi:10.18653/v1/2021.acl-long.353.

[18] X. Liu, K. Ji, Y. Fu, W. L. Tam, Z. Du, Z. Yang, J. Tang, P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks, 2022. arXiv:2110.07602.

[19] O. Kovaleva, A. Romanov, A. Rogers, A. Rumshisky, Revealing the dark secrets of BERT, 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), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 4365–4374. URL: https://aclanthology.org/D19-1445. doi:10.18653/v1/D19-1445.

[20] S. Sukhbaatar, E. Grave, P. Bojanowski, A. Joulin, Adaptive attention span in transformers, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 331–335. URL: https://aclanthology.org/P19-1032. doi:10.18653/v1/P19-1032.

[21] M. Podkorytov, D. Biś, X. Liu, How can the [mask] know? the sources and limitations of knowledge in bert, in: 2021 International Joint Conference on Neural Networks (IJCNN), 2021, pp. 1–8. doi:10.1109/IJCNN52387.2021.9534299.

[22] V. I. Spitkovsky, H. Alshawi, D. Jurafsky, From baby steps to leapfrog: How "less is more" in unsupervised dependency parsing, in: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Los Angeles, California, 2010, pp. 751–759. URL: https://aclanthology.org/N10-1116.

[23] B. Xu, L. Zhang, Z. Mao, Q. Wang, H. Xie, Y. Zhang, Curriculum learning for natural language understanding, in: Annual Meeting of the Association for Computational Linguistics, 2020.

[24] M. Lee, J.-H. Park, J. Kim, K.-M. Kim, S. Lee, Efficient pre-training of masked language model via concept-based curriculum masking, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 7417–7427. URL: https://aclanthology.org/2022.emnlp-main.502.

[25] F. M. Zanzotto, A. Santilli, L. Ranaldi, D. Onorati, P. Tommasino, F. Fallucchi, KERMIT: Complementing transformer architectures with encoders of explicit syntactic interpretations, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 256–267. URL: https://aclanthology.org/2020.emnlp-main.18. doi:10.18653/v1/2020.emnlp-main.18.

[26] L. Ranaldi, F. Fallucchi, F. M. Zanzotto, Dis-cover ai minds to preserve human knowledge, Future Internet 14 (2022). URL: https://www.mdpi.com/1999-5903/14/1/10. doi:10.3390/fi14010010.

[27] E. A. Platanios, O. Stretcu, G. Neubig, B. Poczos, T. Mitchell, Competence-based curriculum learning for neural machine translation, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 1162–1172. URL: https://aclanthology.org/N19-1119. doi:10.18653/v1/N19-1119.

[28] J. Talburt, The flesch index: An easily programmable readability analysis algorithm, in: Proceedings of the 4th Annual International Conference on Systems Documentation, SIGDOC '85, Association for Computing Machinery, New York, NY, USA, 1986, p. 114–122. URL: https://doi.org/10.1145/10563.10583. doi:10.1145/10563.10583.

[29] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, ArXiv abs/1907.11692 (2019).

[30] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, 2019.

[31] S. Merity, C. Xiong, J. Bradbury, R. Socher, Pointer sentinel mixture models, ArXiv abs/1609.07843 (2017).

[32] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, S. Fidler, Aligning books and

movies: Towards story-like visual explanations by watching movies and reading books, in: 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 19–27. doi:10.1109/ICCV.2015.11.

[33] L. Ranaldi, F. M. Zanzotto, Hans, are you clever? clever hans effect analysis of neural systems, 2023. arXiv:2309.12481.

[34] L. Ranaldi, F. M. Zanzotto, Empowering multi-step reasoning across languages via tree-of-thoughts, 2023. arXiv:2311.08097.

Appendix A

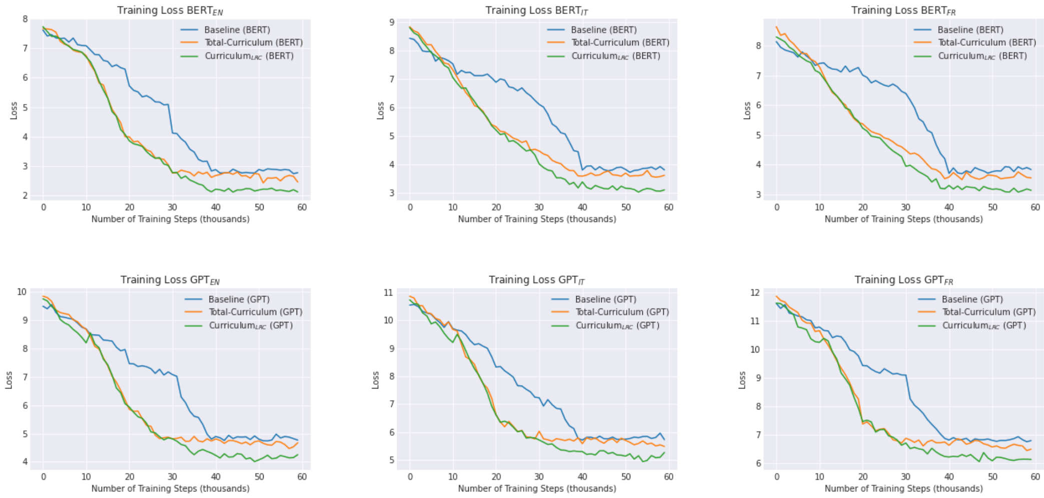


Table 4
Loss during training phase.

Appendix B

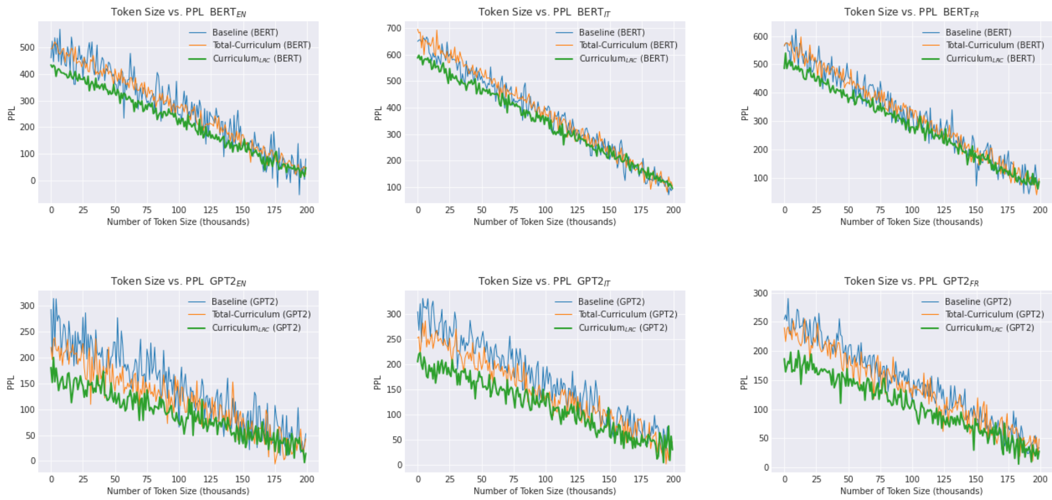


Table 5
Perplexity scores over different chunck of tokens of testset.

Appendix C

Model	Training Time (English)	Training Time (Italian)	Training Time (French)
$Baseline$ (BERT)	5:22:33	5:41:11	5:52:33
Baseline$_{LRC}$ (BERT)	5:20:15	5:43:26	5:50:51
Total-Curriculum (BERT)	4:37:11	**4:31:38**	4:42:27
$Curriculum_{LRC}$ (BERT)	**4:35:46**	4:37:16	**4:40:04**
$Baseline$ (GPT2)	6:37:21	6:42:28	6:58:13
Baseline$_{LRC}$ (GPT2)	6:37:21	6:44:09	7:02:51
Total-Curriculum (GPT2)	5:10:29	**5:19:05**	6:16:18
$Curriculum_{LRC}$ (GPT2)	**5:06:46**	5:20:16	**6:09:16**

Table 6
Statistics of the training time (hours) of the baseline and Curriculum Learning models.

Prompting LLMs in Italian language for Text-to-SQL translation

Federico Ranaldi[1], Elena Sofia Ruzzetti[1], Leonardo Ranaldi[1,3], Davide Venditti[1], Cristina Giannone[2], Andrea Favalli[2], Raniero Romagnoli[2] and Fabio Massimo Zanzotto[1]

[1]*University of Rome Tor Vergata*
[2]*Almawave S.p.A., Via di Casal Boccone, 188-190 00137, Rome, IT*
[3]*Idiap Research Institute, Switzerland*

Abstract
Fine-tuning Large Language Models (LLMs) on tasks with instructions has demonstrated potential in boosting zero-shot generalization to unseen tasks. Inspired by studies on the reasoning skills of Instruction-tuned LLMs (It-LLMs), we investigate reading-comprehension, reasoning, and production over symbolic tasks. In particular, we propose an iterative reading-comprehension and reasoning approach to solve question-answering tasks based on structured data, i.e., Text-to-SQL task. In our approach, we define a specialized procedure to provide the relevant evidence from structured data and natural language queries in order to stimulate the It-LLMs to focus on the production task and reasoning. Hence, we propose a prompting generation procedure to allow It-LLMs to reason about the structural information and natural language queries and produce symbolic output, i.e., the SQL queries. Extensive experiments, in zero-shot scenarios, with different types of structured data, demonstrate the superhuman abilities of It-LLMs in comprehension and production astonishing answers. However, hallucinations and misleading answers are also produced; this still shows the shortcomings of the instructed LLMs and, thus, their partial unreliability.

Keywords
Text-to-SQL, It-LLMs, prompt, zero-shot, Natural Language Processing, Natural Language Query, Natural Language Understanding,

1. Introduction

The development of Large Language Models (LLMs) has been one of the most significant advances in NLP [1, 2]. LLMs are demonstrating superhuman performance after immense corpora pre-training [3], intending to language modeling objectives. Moreover, recent advances show that LLMs are able to do zero-shot task generalization, meaning they can adapt to unknown tasks without fine-tuning. In this way, Instruction-tuning is a promising direction [4, 5, 6]. Instruction-tuning enables these models to follow instructions in different tasks and perform well in tasks in which they have not yet been explicitly trained.

Behind the significant pre-training, Instruction-based tuning is divided into either crowd-sourced human tasks [4, 5] or model-generated tasks [7] for instructional tuning, which is of limited quantity and quality. The scalability of Language Models in different dimensions has been shown to overcome the limits of zero-shot performance, and the search for high-quality and scalable Instruction-tuning tasks has become increasingly important.

Despite their success, recent work has revealed that It-LLMs can generate misleading information in conflict with factual knowledge [8], fail to master domain-specific knowledge [9, 10], and in order to produce answers they stretch the generative imagination by constructing hallucinatory answers [11]. To address these problems, Zhou et al., [12] proposed efficient methods to provide optimal prompts, while Janget al., [13] and Arora et al., [14] really understand the prompts.

In this paper, we propose an iterative reading-comprehension and reasoning approach to solve question-answering tasks based on structured data. In particular, we implement a systematic approach by reconsidering the Text-to-SQL task [15] in a prompt-based version. Then, we define a specialized procedure to provide the relevant evidence from structured data and query the It-LLMs in natural language. In this way, we direct the models to focus on understanding the prompt, reasoning based on the information provided, and producing the output, the SQL code that solves Text-to-SQL task. Extensive experiments, in zero-shot scenarios, with different types of structured data demonstrate the remarkable abilities of It-LLMs in understanding and producing astonishing responses in the presence of various levels of

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
federico.ranaldi@alumni.uniroma2.eu (F. Ranaldi); elenasofia.ruzzetti@alumni.uniroma2.eu (E. S. Ruzzetti); leonardo.ranaldi@idiap.ch (L. Ranaldi); davide.venditti@alumni.uniroma2.eu (D. Venditti); C.Giannone@almawave.it (C. Giannone); A.Favalli@almawave.it (A. Favalli); r.romagnoli@almawave.it (R. Romagnoli); znzfms00@uniroma2.it (F. M. Zanzotto)

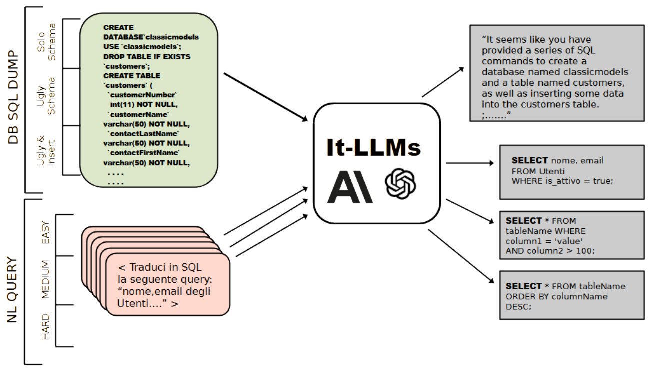


Figure 1: General organization of our work.

information. However, we have observed errors as the information given to It-LLMs decreases. The results of the zero-shot scenarios still show shortcomings of the It-LLMs and, thus, their partial unreliability when the harder queries and less informative databases are considered.

2. Background & Related Works

2.1. Large Language Models

Brown et al.,[2] with GPT3 were the forerunners of the many Large Language Models (LLMs). Among the well-famous LLMs are OPT [16], FLAN [17], and LLaMA [18]. Compared to the smaller language models, LLMs have several emergent abilities [19], including zero-shot multi-task solving [6] and few-shot in-context learning with chain-of-thought reasoning [20].

2.2. Instruction-tuned LLMs

LLMs generate texts following certain formats and instructions from examples in their prompts. Ouyang al., [5] trained GPT3 with instruction-response corpora to make LLMs more scalable and improve zero-shot performance. As a result, InstructGPT, ChatGPT, and GPT4 perform well on a wide range of tasks without seeing any examples. Recent research has also found that GPT-generated instructions and outputs to follow instructions [21] can improve LLMs' ability to follow instructions. Wang et al.,[22] proposed a semi-supervised method to generate different instructions from an NLP task-based seed instruction [7]. However, these models are not fully open-source, and it is often possible to use them for free as black-boxes [23]. Recent open-sourcing efforts include several competitive models [24, 25] but cannot match the performance of closed-source models [26].

2.3. Text-to-SQL task

The ability to translate natural language queries into SQL or other ontological formal languages [27, 28] is a valuable tool because it allows one to interact with databases using a natural language without having to learn SQL. There are several approaches to the problem of translation from natural language to SQL. The earliest methods were totally rule-based [29, 30]; later, with the arrival of statistical learners, a common approach became learning the mapping between SQL queries and commands [15]. Database schema and queries, rich in terms of relationships, are often encoded in graphs – and processed by graph neural networks [31] or self-attention mechanisms [32] – or translated into intermediate representations [33]. Recently, the Text-to-SQL task has been interpreted as a sequence-to-sequence, and transformer-based models are applied [34, 35]. However, a critical aspect is the amount of input information, i.e., database schemas and relationships encoding. In this paper, we move forward and propose a new Text-to-SQL approach by exploiting the potential of It-LLMs models. In particular, after an extensive prompt-tuning phase, we analyze two It-LLMs models' reasoning and generalization abilities in solving the Text-to-SQL task with less informative database representations and harder queries. Our contribution is unaffected by LLMs' prior knowledge after pre-training as we test a collection of definitely unseen databases.

3. Methods

In order to test the reading-comprehension abilities of Instruction-tuned Large Language Models (It-LLMs) in the Text-to-SQL translation task, we organized the prompting phase into two parts. In the first phase, we defined different prompts for studying how the presence of Structural Information and data affects the behavior of models (Section 3.1). In the second phase, we defined possible types of Natural Language Queries (Section 3.2): to quantify the ability of a model to reason over structured information.

3.1. Prompting Structural Information

We defined three prompting-approaches for Structural Information based on the amount of database information provided to the model. Hence, we proposed three types of input: (i) complete information on the current database schema, including primary and foreign keys (SOLO~SCHEMA); (ii) degradation of the original table and attribute names via removing vocals from them (UGLY~SCHEMA); (iii) same as UGLY~SCHEMA but providing, in addition, a small amount of real data in order to compensate for the degraded schema information (UGLY & INSERT).

3.2. Prompting Natural Language Query

Regarding the Natural Language Query (NLQ), i.e., the queries we wish to translate SQL, inspired by the work of [36], we considered three hardness-levels: easy, medium, and hard. A given NLQ is assigned to a certain level if the best corresponding SQL translation has specific hardness characteristics. The hardness-levels are defined as follows:

1. EASY: values are selected only from one table (there is no join).
2. MEDIUM: values are selected by joining two tables.
3. HARD: values are selected by joining more than two tables.

Furthermore, in all levels, an arbitrary number of conditions is allowed, and aggregation functions are included.

3.3. Prompting Phase

We conducted the Text-to-SQL task using two It-LLMs: GPT-3.5 [37] and Claude-instant [38]. In a zero-shot scenario, we considered the three different approaches (as described in Section 3.1), behind which we asked the models to translate a small number of NLQ per hardness-level on three different databases. In particular, except for feeding the SQL dump of the database as input, requests such as *"Traduci la seguente query NL in SQL"* were made without any further prompt engineering steps.

4. Experiments

In order to observe the real abilities of Intruction-tuned Large Language Models (It-LLMs) in reading-comprehension on heterogeneous inputs and the reasoning abilities behind output generation, we selected a set of databases (Section 4.1) and conducted a series of systematic queries (Section 4.2).

Database	DB1	DB2	DB3
Topic	Drugs and Prescriptions	Sport Center	Covid and Hospitals
Tables	16	17	20
Columns AVG	3.19	5.35	5.25
Primary Keys	46,82%	37.4%	32.12%
Foreign Keys	30.57%	12.28%	32.25%

Table 1
Databases detailed characteristics. The column "Tables" reports the number of tables, "Columns" reports the average number of columns per table, "Primary Keys" and "Foreign Keys" report respectively, the average frequency of primary keys and the average frequency of foreign keys inside each table.

4.1. Datasets

In order to analyze the generalization abilities, we have fed dumps of three SQL databases that are definitely unseen, thus not found on the Web, and never seen in the pre-training corpora of Large Language Models. Moreover, databases differ in topic, topology, and size as shown in Table 1.

4.2. Experimental Settings

Behind describing the data (Section 4.1) and prompting methodologies (Section 3), we tested our proposals on GPT-3.5 and Claude Instant. Hence, we provided Structural Information, defined in Section 3.1, in three different ways, in each of which we requested the translation of four Natural Language Queries (NLQ) for each hardness level. We conducted experiments on three different databases to study phenomena in different scenarios. The NLQs were in Italian and, as described in Section 3.2 were of the type: *"Traduci in sql la seguente query 'nomi,cognomi,età degli utenti...ordinati per età'"*.

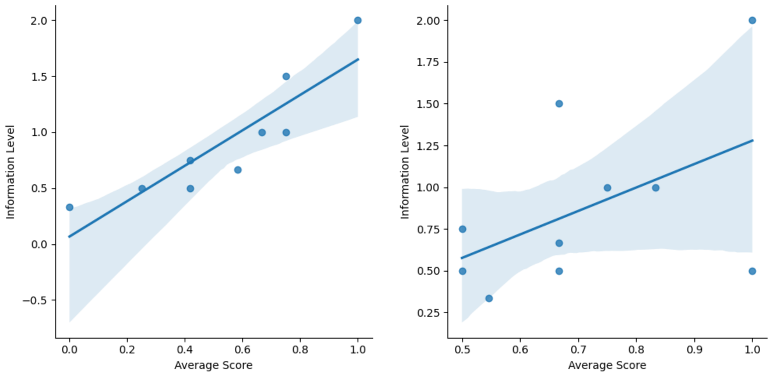
5. Results & Discussion

5.1. The reading-comprehension Challenge

It-LLMs are amazing understanders; in fact, in presence of structured information, they perform very well in overcoming complex challenges and generating good translations from Text-to-SQL. In Table 2 we can observe that both GPT-3.5 and Claude Instant perform very well in the SOLO~SCHEMA approach. In particular, both GPT-3.5 and Claude Instant produce an accurate translation for all the EASY queries. Moreover, Claude Instant produces very good results on average also on the MEDIUM queries. Hence, the It-LLMs showed good abilities in

Model	Approach	EASY				MEDIUM				HARD			
		DB1	DB2	DB3	TOT	DB1	DB2	DB3	TOT	DB1	DB2	DB3	TOT
GPT-3.5	SOLO~SCHEMA	1.00	1.00	1.00	1.00	1.00	0.50	0.50	0.67	1.00	0.50	0.25	0.58
	UGLY~SCHEMA	0.75	0.75	0.75	0.75	0.50	0.25	0.50	0.42	0	0	0	0
	UGLY & INSERT	1.00	0.75	0.50	0.75	0.50	0.50	0.25	0.42	0.50	0.25	0	0.25
Claude instant	SOLO~SCHEMA	1.00	1.00	1.00	1.00	0.75	1.00	0.75	0.83	0.75	0.75	0.50	0.67
	UGLY~SCHEMA	0.50	1.00	0.75	0.75	0.50	0.50	0.50	0.5	0.25	0.50	1.00	0.58
	UGLY & INSERT	1.00	0.50	0.50	0.67	0.50	0.25	0.75	0.5	0.75	0.75	0.50	0.67

Table 2
Models percentage of corrects answers across the different approaches and divided by hardness-level. TOT value calculates the average of successes obtained in translating leveled queries at each database.



(a) GPT-3.5 (b) Claude Instant

Figure 2: Linear regression is performed to analyze the correlation between average score and quantity of information available, quantified as the *Information Level*.

comprehending natural language and the structural information of databases in SQL language.

5.2. The reasoning-generation Challenge

The It-LLMs' reasoning and SQL query generation skills are strongly related to the quality of the queries. Indeed, the It-LLMs could generate intriguing output even in zero-shot and low-resource scenarios (with limited structural information). However, they could not generate exhaustive translations when the types of SQL queries required were hard. In fact, in Table 2, it is possible to observe a marked decrease in the SOLO~SCHEMA rows of the HARD columns compared to the EASY and MEDIUM columns. In particular, for DB3 queries, performances fall by half, or worse, going from EASY level to HARD.

5.3. Effects of degradation of structural information

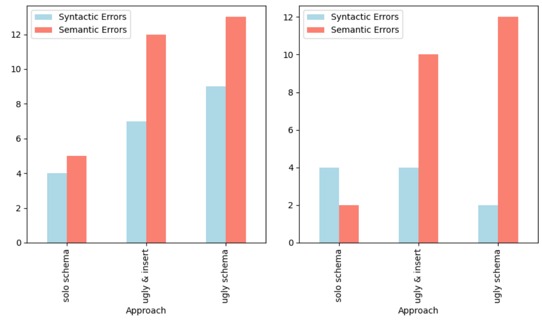
Both the reading-comprehension and reasoning-generation abilities of It-LLMs are negatively affected by degrading database information.

In fact, we can observe that as we degrade the structural information of the database by removing vocals from the table and attribute names (UGLY~SCHEMA), the models tend to make errors with a high frequency. Looking at Table 2, GPT-3.5 and Claude instant performances deteriorate at all hardness levels. Moreover, GPT-3.5 always fails to translate HARD queries. This means that both models find it more challenging to understand what is asked in the NL query and to reason over the database structure with deteriorated names.

However, some points can be recovered by providing the database with a small amount of real data (UGLY&INSERT). This phenomenon can be observed by noting that the TOT obtained in the UGLY & INSERT approach never worsens compared to the UGLY~SCHEMA, regardless of the hardness level of the queries.

Hence, we can conclude that degrading information quality has negative effects on both models, affecting the reliability of their reasoning skills.

Finally, we want to quantify how model performance is affected by the amount of information available on a database compared to the amount of information needed to effectively resolve queries. We hence define this quan-



(a) GPT-3.5 (b) Claude Instant

Figure 3: Number of `semantic errors` and `syntactic errors` for GPT-3.5 and Claude Instant across approaches, ordered from most informative to least informative.

tity of information as *Information Level I*. We define I as follows:

$$I = \frac{as}{hs}$$

where *as* is the Approach score and *hs* is Hardness Score. The Approach Score *as* assigns a score to each approach, ranging from 1 to 2: the highest value 2 is assigned to the SOLO~SCHEMA approach and the lowest 1 to UGLY~SCHEMA. The UGLY & INSERT approach is assigned an intermediate score of 1.5. To calculate the *Information Level* we smooth this information with the actual hardness of the query that is assigned with the *Hardness Score hs*: it ranges from 1 (for the EASY level) to 3 (for the HARD level).

As shown in Figure 2, GPT-3.5 and Claude Instant performances correlate with the *Information Level*. For GPT-3.5 (Figure 2a), a large Pearson correlation coefficient (0.88) is observed, which is statistically significant with a p value of 0.001. Claude Instant performance (Figure 2b) is still positively correlated with the *Information Level*, although the Pearson correlation coefficient is lower (0.5) and has a higher p value (0.1).

5.4. Errors Analysis

In this section, we focus on the characterization of errors that are made by the analyzed models. We investigate two types of errors: `semantic errors` and `syntactic errors`. The `semantic errors` are queries mistranslated by the system that, if executed, result in the selection of information other than what was initially requested in natural language. On the other hand, `syntactic errors` are errors that make the query not executable by an engine: these queries are characterized by incorrect use of SQL syntax (e.g., they contain a field in the HAVING statement that is not present in the SELECT) or contain references to tables and fields that do not exist in the database in question. In Figure 3, we can observe the effect of different approaches on the number of errors in the two cases.

As expected, as the information available to a system decreases, the number of `semantic errors` tends to increase. We can observe that both GPT-3.5 (Figure 3a) and Claude Instant (Figure 3b) tend to make a limited number of `semantic errors` in the SOLO~SCHEMA approach, while the UGLY~SCHEMA approach leads to the largest number of errors. We can observe that the UGLY & INSERT approach, with a limited set of realistic data, seems to reduce the number of `semantic errors`.

On the other hand, the trend in the number of `syntactic errors` is different between the two models. In GPT-3.5, the decrease in the informativeness of the dumps leads to more errors. Manual inspection found that only one error was due to incorrect use of SQL syntax: in most cases, GPT-3.5 has difficulty identifying the tables and columns to be used in the given database and therefore proposes SQL queries that make use of arbitrary tables. In this case, these `syntactic errors` are definitely examples of hallucinations and need to be further explored. Claude Instant, instead, tends to retain more information about the dump, and the number of `syntactic errors` is more constant across the different approaches.

6. Conclusion

In this paper, we propose an iterative reading-comprehension and reasoning approach to solve question-answering challenges of the Text-to-SQL task. The results obtained from the experiments conducted in this work witness the potential of Instruction-tuned Large Language Models (It-LLMs).However, despite their promising performance, certain limitations have emerged. We discovered that even with minimal information about the database, It-LLMs can generate natural language query translations that yield correct and executable SQL queries by just prompting them. Nevertheless, it became evident that reducing the amount of information provided could lead to the generation of incorrect queries. Expanding the scope of our investigation, we believe it would be worthwhile to conduct similar experiments with other It-LLMs. Such comparisons could help determine whether the common phenomena observed in both tested models result from a coincidence or represent aspects to further investigate in studying these new technologies.

In conclusion, this research underscores the substantial advancements offered by It-LLMs in the realm of Text-to-SQL translation while also the implications of choosing whether to provide more or less information during the prompting process.

Acknowledgments

This work was conducted within the DATALAKE Giustizia project; we acknowledge the partners and the scientific committee for their support.

References

[1] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, 2020. arXiv:1910.10683.

[2] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, 2020. arXiv:2005.14165.

[3] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, S. Presser, C. Leahy, The pile: An 800gb dataset of diverse text for language modeling, 2020. arXiv:2101.00027.

[4] S. Mishra, D. Khashabi, C. Baral, H. Hajishirzi, Cross-task generalization via natural language crowdsourcing instructions, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 3470–3487. URL: https://aclanthology.org/2022.acl-long.244. doi:10.18653/v1/2022.acl-long.244.

[5] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, R. Lowe, Training language models to follow instructions with human feedback, 2022. arXiv:2203.02155.

[6] V. Sanh, A. Webson, C. Raffel, S. H. Bach, L. Sutawika, Z. Alyafeai, A. Chaffin, A. Stiegler, T. L. Scao, A. Raja, M. Dey, M. S. Bari, C. Xu, U. Thakker, S. S. Sharma, E. Szczechla, T. Kim, G. Chhablani, N. Nayak, D. Datta, J. Chang, M. T.-J. Jiang, H. Wang, M. Manica, S. Shen, Z. X. Yong, H. Pandey, R. Bawden, T. Wang, T. Neeraj, J. Rozen, A. Sharma, A. Santilli, T. Fevry, J. A. Fries, R. Teehan, T. Bers, S. Biderman, L. Gao, T. Wolf, A. M. Rush, Multitask prompted training enables zero-shot task generalization, 2022. arXiv:2110.08207.

[7] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, H. Hajishirzi, Self-instruct: Aligning language models with self-generated instructions, 2023. arXiv:2212.10560.

[8] Y. K. Dwivedi, N. Kshetri, L. Hughes, E. L. Slade, A. Jeyaraj, A. K. Kar, A. M. Baabdullah, A. Koohang, V. Raghavan, M. Ahuja, H. Albanna, M. A. Albashrawi, A. S. Al-Busaidi, J. Balakrishnan, Y. Barlette, S. Basu, I. Bose, L. Brooks, D. Buhalis, L. Carter, S. Chowdhury, T. Crick, S. W. Cunningham, G. H. Davies, R. M. Davison, R. Dé, D. Dennehy, Y. Duan, R. Dubey, R. Dwivedi, J. S. Edwards, C. Flavián, R. Gauld, V. Grover, M.-C. Hu, M. Janssen, P. Jones, I. Junglas, S. Khorana, S. Kraus, K. R. Larsen, P. Latreille, S. Laumer, F. T. Malik, A. Mardani, M. Mariani, S. Mithas, E. Mogaji, J. H. Nord, S. O'Connor, F. Okumus, M. Pagani, N. Pandey, S. Papagiannidis, I. O. Pappas, N. Pathak, J. Pries-Heje, R. Raman, N. P. Rana, S.-V. Rehm, S. Ribeiro-Navarrete, A. Richter, F. Rowe, S. Sarker, B. C. Stahl, M. K. Tiwari, W. van der Aalst, V. Venkatesh, G. Viglia, M. Wade, P. Walton, J. Wirtz, R. Wright, Opinion paper: "so what if chatgpt wrote it?" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational ai for research, practice and policy, International Journal of Information Management 71 (2023) 102642. URL: https://www.sciencedirect.com/science/article/pii/S0268401223000233. doi:https://doi.org/10.1016/j.ijinfomgt.2023.102642.

[9] J. Jiang, K. Zhou, J.-R. Wen, X. Zhao, *great truths are always simple* : a rather simple knowledge encoder for enhancing the commonsense reasoning capacity of pre-trained models, in: Findings of the Association for Computational Linguistics: NAACL 2022, Association for Computational Linguistics, Seattle, United States, 2022, pp. 1730–1741. URL: https://aclanthology.org/2022.findings-naacl.131. doi:10.18653/v1/2022.findings-naacl.131.

[10] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, L. Zettlemoyer, N. Cancedda, T. Scialom, Toolformer: Language models can teach themselves to use tools, 2023. arXiv:2302.04761.

[11] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, Q. V. Do, Y. Xu, P. Fung, A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity, 2023. arXiv:2302.04023.

[12] Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, J. Ba, Large language models are human-level prompt engineers (2022). arXiv:2211.01910.

[13] J. Jang, S. Ye, M. Seo, Can large language models truly understand prompts? a case study with negated prompts, 2022. arXiv:2209.12711.

[14] S. Arora, A. Narayan, M. F. Chen, L. Orr, N. Guha, K. Bhatia, I. Chami, F. Sala, C. Ré, Ask me anything: A simple strategy for prompting language models, 2022. arXiv:2210.02441.
[15] T. Wolfson, D. Deutch, J. Berant, Weakly supervised text-to-SQL parsing through question decomposition, in: Findings of the Association for Computational Linguistics: NAACL 2022, Association for Computational Linguistics, Seattle, United States, 2022, pp. 2528–2542. URL: https://aclanthology.org/2022.findings-naacl.193. doi:10.18653/v1/2022.findings-naacl.193.
[16] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, L. Zettlemoyer, Opt: Open pre-trained transformer language models, 2022. arXiv:2205.01068.
[17] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, Q. V. Le, Finetuned language models are zero-shot learners, 2022. arXiv:2109.01652.
[18] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, 2023. arXiv:2302.13971.
[19] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, W. Fedus, Emergent abilities of large language models, 2022. arXiv:2206.07682.
[20] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, 2023. arXiv:2201.11903.
[21] B. Peng, C. Li, P. He, M. Galley, J. Gao, Instruction tuning with gpt-4, 2023. arXiv:2304.03277.
[22] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, D. Zhou, Self-consistency improves chain of thought reasoning in language models, 2023. arXiv:2203.11171.
[23] Z. Lin, S. Trivedi, J. Sun, Generating with confidence: Uncertainty quantification for black-box large language models, 2023. arXiv:2305.19187.
[24] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, T. B. Hashimoto, Stanford alpaca: An instruction-following llama model, https://github.com/tatsu-lab/stanford_alpaca, 2023.
[25] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, E. P. Xing, Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. URL: https://vicuna.lmsys.org.
[26] A. Gudibande, E. Wallace, C. Snell, X. Geng, H. Liu, P. Abbeel, S. Levine, D. Song, The false promise of imitating proprietary llms, 2023. arXiv:2305.15717.
[27] P. Atzeni, R. Basili, D. Hansen, P. Missier, P. Paggio, M. Pazienza, F. Zanzotto, Ontology-based question answering in a Federation of University Sites: The MOSES case study, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2004). URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-35048854325&doi=10.1007%2f978-3-540-27779-8_40&partnerID=40&md5=7545b9abe40e6ac9d64b47d45e71b78c. doi:10.1007/978-3-540-27779-8_40.
[28] R. Basili, D. H. Hansen, P. Paggio, M. T. Pazienza, F. M. Zanzotto, Ontological resources and question answering, in: Proceedings of the Workshop on Pragmatics of Question Answering at HLT-NAACL 2004, Association for Computational Linguistics, Boston, Massachusetts, USA, 2004, pp. 78–84. URL: https://aclanthology.org/W04-2510.
[29] F. Li, H. V. Jagadish, Constructing an interactive natural language interface for relational databases, Proceedings of the VLDB Endowment 8 (2014) 73–84.
[30] T. Mahmud, K. M. Hasan, M. Ahmed, T. Chak, A rule based approach for nlp based query processing, 2015, pp. 78–82. doi:10.1109/EICT.2015.7391926.
[31] B. Bogin, J. Berant, M. Gardner, Representing schema structure with graph neural networks for text-to-SQL parsing, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 4560–4565. URL: https://aclanthology.org/P19-1448. doi:10.18653/v1/P19-1448.
[32] B. Wang, R. Shin, X. Liu, O. Polozov, M. Richardson, RAT-SQL: Relation-aware schema encoding and linking for text-to-SQL parsers, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7567–7578. URL: https://aclanthology.org/2020.acl-main.677. doi:10.18653/v1/2020.acl-main.677.
[33] I. Sucameli, A. Bondielli, L. Passaro, E. Annunziata, G. Lucherini, A. Romei, A. Lenci, Mate, a meta layer between natural language and database, in: Proceedings of the Sixth Workshop on Natural Language for Artificial Intelligence (NL4AI 2022) co-located with 21th International Conference of the Italian Association for Artificial Intelligence (AI* IA 2022), 2022.
[34] T. Scholak, N. Schucher, D. Bahdanau, PICARD: Parsing incrementally for constrained auto-regressive decoding from language models,

in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 9895–9901. URL: https://aclanthology.org/2021.emnlp-main.779. doi:10.18653/v1/2021.emnlp-main.779.

[35] T. Xie, C. H. Wu, P. Shi, R. Zhong, T. Scholak, M. Yasunaga, C.-S. Wu, M. Zhong, P. Yin, S. I. Wang, V. Zhong, B. Wang, C. Li, C. Boyle, A. Ni, Z. Yao, D. Radev, C. Xiong, L. Kong, R. Zhang, N. A. Smith, L. Zettlemoyer, T. Yu, UnifiedSKG: Unifying and multi-tasking structured knowledge grounding with text-to-text language models, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 602–631. URL: https://aclanthology.org/2022.emnlp-main.39.

[36] T. Yu, R. Zhang, K. Yang, M. Yasunaga, D. Wang, Z. Li, J. Ma, I. Li, Q. Yao, S. Roman, Z. Zhang, D. Radev, Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 3911–3921. URL: https://aclanthology.org/D18-1425. doi:10.18653/v1/D18-1425.

[37] OpenAI, Chatgpt, 2022. URL: https://chat.openai.com/.

[38] Anthropic, Claude-instant, 2022. URL: https://poe.com/Claude-instant.

Beyond Obscuration and Visibility: Thoughts on the Different Strategies of Gender-Fair Language in Italian

Martina Rosola[1], Simona Frenda[2,3], Alessandra Teresa Cignarella[2,3], Matteo Pellegrini[4], Andrea Marra[2] and Mara Floris[5]

[1]Università degli Studi di Brescia, Italy
[2]Department of Computer Science, University of Turin, Italy
[3]aequa-tech, Turin, Italy
[4]CIRCSE Research Centre, Università Cattolica del Sacro Cuore, Milan, Italy
[5]Università Vita-Salute San Raffaele, Milan, Italy

Abstract
This study focuses on the growing importance of gender-fair language and explores innovative strategies proposed also in other languages to avoid gender-specific endings. We present a set of guidelines for the annotation and reformulation of gender-(un)fair texts and their application to a corpus of 1,024 portions of university administrative documents in Italian. Overall, the guidelines presented in this study prove to be valuable both practically and theoretically. They help identify and address non-inclusive expressions while highlighting the complexities of *obscuration* and *visibility* in gender-fair language reformulation. In addition, the statistical analysis of the created corpus shows how administrative texts tend to contain gender-unfair language, especially the masculine overextended expressions, showing the need to adopt specific and complete guidelines that lead (and support the staff training) to the use of a more gender-fair language.

Keywords
Annotation Schema, Italian, Gender-fair language

1. Introduction

Gender-fair language (GFL), also known as inclusive language, refers to the use of linguistic expressions that promote gender equality and avoid reinforcing gender stereotypes. The Italian language, like many others, has traditionally exhibited gender asymmetries and biases, which have perpetuated gender inequality and exclusion. However, in recent years, there has been a growing awareness and effort to address these issues by promoting GFL. In this work, we present the annotation scheme we developed to identify and reformulate gender-unfair expressions, and the corpus we applied it to, encompassing a range of administrative documents of the University of Brescia[1].

The significance of addressing gender-unfair expressions stems from concerns raised by several researchers. In Italian, a personal noun's grammatical gender typically correspond to its referent's gender. In certain cases, however, a discrepancy between the two arises. Crucially, such discrepancies are characterized by an asymmetrical nature, as has been highlighted both from a theoretical and a practical perspective [1, 2]. In Italian, as well as in other languages with gendered nouns, the issue of GFL presents a dual challenge. Firstly, the binary distinction between masculine and feminine in the Italian gender system excludes individuals who identify outside the male-female dichotomy. Consequently, non-binary individuals are consistently misgendered due to the absence of dedicated linguistic forms. Secondly, the usage of the generic masculine (e.g., using masculine profession names to refer to individuals of any gender) and the overextended masculine (referring to mixed-gender groups using the masculine plural) predominantly evoke masculine mental representations, thereby limiting the visibility of female and non-binary individuals. Notably, the presence of a single man in a group is sufficient to alter the gender of the term used to refer to the group, whereas the reverse does not hold true for inverted genders.

These discrepancies, in addition to their asymmetrical nature, impact the mental representations we construct. Despite the Italian convention of using masculine terms to refer to individuals of unknown gender and mixed groups, psycholinguistic studies highlight potential issues in this respect. Extensive experimentation conducted over several years using various techniques, and across different languages, suggests that the overextended masculine and generic masculine are interpreted as if they were purely masculine [e.g., 3, 4].

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
martina.rosola@unibs.it (M. Rosola); simona.frenda@unito.it (S. Frenda); alessandrateresa.cignarella@unito.it (A. T. Cignarella); matteo.pellegrini@unicatt.it (M. Pellegrini); andrea.marra@unito.it (A. Marra); floris.mara@hsr.it (M. Floris)

[1]https://www.unibs.it

Recognizing the importance of rectifying such language discrepancies, various guidelines have been published over the years [e.g., 5]. The annotation scheme we propose draws upon the recommendations presented in the available guidelines, to develop a comprehensive framework for addressing gender-unfair expressions in Italian language usage. To the best of our knowledge, our annotation scheme represents a novel approach. While another project (i.e., E-MIMIC) focuses on inclusive language, it simply distinguishes between inclusive and non-inclusive texts [6]. Our annotation scheme appears to be the first one distinguishing between different types of gender-unfair language, and it comprehensively considers all the gender-fair options when it comes to offering alternative wordings. Moreover, applying this scheme of annotation to various administrative texts, we showed how, despite the existence of various guidelines, they remain impregnated with gender-unfair expressions.

In this work, firstly we review previous studies on this topic, both in theoretical linguistics (subsection 2.1) and in NLP (subsection 2.2). We then describe in detail the annotation scheme (section 3) and the creation of the annotated corpus (section 4), also providing a preliminary analysis of the data gathered so far.

2. Related Work

2.1. Linguistics

Sexism in language and how to make Italian gender-fair are increasingly studied and debated topics (see [7] for an overview). The classic reference point in the literature is Sabatini [8], which comprises an analysis of sexism in the Italian language and recommendations on how to overcome it.

Sabatini [8] identifies *grammatical* and *semantic* asymmetries, namely gender-unfair grammatical and discursive or lexical linguistic conventions. The use of masculine terms for mixed-gender groups belongs to the former, while the exclusive use of adjectives for one gender (e.g., GRAZIOSO, 'pretty', is hardly used for men) instantiates the latter. On top of avoiding semantically sexist expressions, Sabatini [8] advises using feminine job titles for women and conjoining masculine and feminine forms for mixed-gender groups.

Her recommendations have been expanded and adapted by several private and public bodies, that issued gender-fair language guidelines [e.g., 9, 5, 10]. These works, among others, distinguish between strategies aimed at symmetrizing language by giving women the same visibility that men have, and strategies aimed at getting rid of sexism through the avoidance of gendered forms altogether. While Sabatini and the subsequent guidelines focus on the discrimination and (in)visibilization of *women* in language, more recent scholarly and activist debates also concern how to address and talk about *non-binary people*, i.e., those that do not exclusively identify as men or women, aiming at making them visible too.

The Italian grammatical gender system, indeed, is binary and does not provide a straightforward way to refer to non-binary people. Various gender-neutral suffixes are in use, such as '-ə' or '-u' (see [11] for a comprehensive list). As González Vázquez et al. [12] observe, such *innovative* proposals can be employed to make gender visible, as in "tutti, tutte e tuttu" (everyone:M.PL, everyone:F.PL, and everyone:INN.PL)[2], or to neutralize it, as in "tuttu" (everyone.INN.PL) used for a mixed-gender group.

The implementation of innovative strategies also depends on the features of the language. Marcato and Thüne [13] provide an analysis of the Italian grammatical gender system, distinguishing between nouns whose referential gender is expressed by different lexical roots (e.g., "madre", mother:F.SG, and "padre", father:M.SG); nouns with *mobile* gender, whose referential gender is specified through the addition of different suffixes to the same lexical root (e.g., "figlia", daughter:F.SG, and "figlio", son:M.SG); and the so-called *epicene* nouns, whose gender is not overtly marked, but only revealed by satellite elements – i.e., the noun's determiners and modifiers (e.g., "la nipote", the.F.SG niece:F.SG; "il nipote", the.M.SG nephew:M.SG). As Formato [14] observes, some nouns (i.e., 'semi-epicene') work in the latter way only in the singular and have different gendered suffixes in the plural (e.g., "giornalista", journalist:F/M.SG; "giornaliste", journalist:F.PL; "giornalisti", journalist:M.PL). Finally, a few nouns refer to individuals of any gender irrespective of their grammatical gender (e.g., "persona", person:F.SG).

Due to this peculiar characteristic, these nouns can be straightforwardly used to refer to non-binary people as well. Moreover, gender-neutral suffixes are not required for epicene (and, in the singular, semi-epicene) nouns, as they are not overtly marked for gender. In this case, the only needed precaution to get a gender-neutral form concerns the choice of gender-neutral satellite elements or their gender-neutralization. Gender-neutral suffixes are further ineffective for nouns like MADRE and PADRE, where it is the root to be overtly marked for gender. These word endings, thus, should only be used with nouns with mobile gender, which, however, constitute the vast majority of Italian animate nouns (see [15], p. 106). Innovative strategies should also be used for the many gendered pronouns, determiners, past participles, and adjectives in order to make them gender-neutral and suitable to refer to non-binary people and to mixed-gender groups.

Formato [14] also provides a taxonomy of linguistic us-

[2] We label such forms as INN 'innovative'.

ages influenced by gendered assumptions. Like Sabatini, Formato focuses on both sexist expressions and linguistic conventions. Among the latter, Formato originally distinguishes the case in which masculine terms are used for mixed-gender groups from those in which they are used for unknown or generic individuals.

In our framework, we elaborate on the categories identified in these works, to develop our own taxonomy, that we implement in the annotation scheme described in section 3.

2.2. Natural Language Processing

In recent years, sexism and gender-(un)fair practices have been addressed in Computational Linguistics, mostly focusing on the presence of gender bias in automatic systems. As highlighted in Costa-jussà [16], studies on gender bias in NLP serve a dual role. On the one hand, NLP can function as a tool to identify gender bias in various social domains such as online news or advertisements. On the other hand, it frequently generates gender-biased systems, thus contributing to the perpetuation and reinforcement of gender bias within society. This bias in NLP is predominantly attributed to the training of models on datasets that exhibit inherent biases. Consequently, the amplification of bias occurs through the learning algorithms employed in NLP systems.

Some specific studies have been conducted in the field of Machine Translation. One of the most recent was carried out by Rescigno et al. [17] who explored how three of the most popular translation systems (Google Translate, Bing Microsoft Translator, and DeepL) handle gender phenomena in natural languages, such as pronouns, job titles, and occupation names. The authors compared the translations generated from English respectively to Italian, French, and Spanish, revealing that all three systems exhibit some level of gender bias, with Google Translate producing more biased translations, Bing Microsoft Translator displaying a lesser degree of bias, and DeepL generally being more gender-neutral.

Similarly to Costa-jussà [16], Sun et al. [18] conducted a comprehensive literature review, exploring various strategies proposed in existing research to address gender bias, including dataset preprocessing, algorithmic modifications, and post-processing techniques. The paper emphasizes the significance of mitigating gender bias in NLP systems and highlights the challenges associated with bias detection and mitigation.

More recently, Stanczak and Augenstein [19] identify four key limitations in current research on gender bias in NLP. Firstly, social gender is often treated as a binary variable, not paying attention to its fluidity and continuity. Secondly, studies usually give more importance to high-resource languages - primarily English - neglecting the diversity of languages spoken globally. Thirdly, despite the abundance of papers on gender bias, many newly developed algorithms lack sufficient bias testing and fail to address ethical considerations. Lastly, the methodologies employed in this area often lack comprehensive definitions of gender bias and robust evaluation baselines and pipelines.

The present work contributes to address many of these issues: we explicitly take into account the linguistic representation of non-binary individuals, we create an annotation scheme for Italian – for which much fewer NLP studies are available compared to English – and finally, we present an annotated corpus that could be exploited for the training of automatic NLP tools.

3. Annotation Scheme

The annotation task is divided into two parts. A first annotation layer concerns the identification of portions of text(s) where gender-unfair language is used, and the assignment of each of them to a specific type among the following ones:

- ‘incongruous’ (It. “incongruo”), when the grammatical gender of the noun (and, possibly, of its modifiers), does not match the gender of the referent identified in discourse (e.g., “il ministro del turismo, Daniela Santanché”, the.M.SG minister:M.SG of.the tourism, Daniela Santanché);
- ‘overextended’ (It. “sovraesteso”), when the masculine (or, in rare cases, feminine) grammatical gender is used to refer to a group composed of people with different genders (e.g., “il rapporto con i professori è buono”, the relationship with the.M.PL professor:M.PL is good, with reference to a group of teachers possibly comprising men, women, and non-binary individuals);
- ‘generic’ (It. “generico”) when the masculine (or, in rare cases, feminine) grammatical gender is used to refer to a generic or specific, but unknown, person, whose actual gender cannot be guessed (e.g. “il vincitore riceverà un premio”, the.M.SG winner:M.SG will.receive a bonus, where the identity of the winner is unknown, and so is their gender).

A second annotation layer concerns the proposal of gender-fair reformulations of the portions of texts identified as unfair,[3] and the assignment of each reformulation to a specific type.

As for cases of incongruous gender, the only possible type of gender-fair solution is providing a ‘congruous’ (It. “congrua”) alternative option, where the grammatical gender matches the gender of the referent (e.g., “la

[3] At least one reformulation for each gender-unfair portion is required, but more than one reformulation is often provided.

	visibility	obscuration
conservative	*il vincitore o la vincitrice*	*chi vincerà*
innovative	*il vincitore o la vincitrice o l* vincitor**	*l* vincitor**

Table 1
Four-way contrast of reformulation types

ministra del turismo, Daniela Santanché", the.F.SG minister:F.SG of.the tourism, Daniela Santanché). On the other hand, to classify the reformulations proposed for cases of overextended and generic uses of grammatical gender, we start from two orthogonal binary distinctions, namely:

- 'visibility' (It. "visibilità") strategies, that make the possible reference to persons with different genders explicit by means of the use of different grammatical genders; vs. 'obscuration' ("oscuramento") strategies, that try to avoid the use of expressions that reveal the (assumed) gender of referents;
- 'conservative' (It. "conservative") strategies, that only use expressions that are part of the grammatical system of the standard variety of Italian; vs. 'innovative' (It. "innovative") strategies, that introduce new means of expression into the system.

These distinctions generate a four-way contrast, that is illustrated in Table 1, where one example reformulation per type is provided for the phrase *il vincitore* (see the Appendix for other examples).

Lastly, 'mixed' (It. "ibride") reformulations use different strategies for different elements in the gender-unfair portion of text, e.g. *l* vincitore o vincitrice* (the.INN.SG winner:M.SG or winner:F.SG), where an innovative obscuration strategy is used for the article and a conservative visibility strategy is used for the noun.

4. Corpus

The scheme of annotation was applied by 5 expert annotators of gender-fair language to a small corpus of administrative texts coming from the University of Brescia. Differently from other textual genres, the administrative texts, for their format and technical language, have required a specific preprocessing process to let the annotators focus especially on the spans of text that could contain discrepancies.

To this purpose, we employed the original lexicon of professional names taken into account in Sabatini [8], enriching it with terms especially pertaining to Academia or terms that could be used in an overextended way (i.e., "essi", they:M.PL). Below, we describe the steps of documents collection, preprocessing, and annotation. Finally, we also present some first analyses of the resulting corpus.

Data Collection The data made available by the University of Brescia include a range of administrative materials such as the department's strategic plan, reports from the departmental council and parity commission, as well as various forms. Most of them are already public on the website of the University, others, like the forms, have been asked to administrative organs. For this pioneering study, we collected specifically 13 documents.

Data Preprocessing All the documents have been transformed into plain text to be processed automatically. To deal with the special format of forms or the layout in tables of special reports, we designed various regular expressions to clean and prepare the texts for the segmentation in sentences.

To support the annotation phase, we signaled for each sentence the possible presence of discrepancies: displayed below each task were any words from the enriched lexicon of professions' names based on Sabatini [8] present in the sentence. From a total of 1,024 sentences, 409 contained such words. However, the sentences in which the annotators have detected at least one unfair expression are 422. The lexicon has been updated to include all the words pointed by the annotators.

Annotation Process All annotators were trained on the annotation scheme, which was analysed during an initial meeting. Doubtful cases were discussed in regular bi-weekly meetings. In addition, we kept a file of notes in which we reviewed and discussed uncertain cases as they arose. The annotation scheme was subsequently updated in response to the insights gathered both in the file and the meetings. The annotation process has been carried out on LabelStudio platform[4], creating a specific interface that facilitates the two layers of annotation: the identification of the gender-unfair expression, and the reformulation with one or more alternatives. The interface provided a section for comments in order to encourage reflection on the annotation scheme and collect insights from the annotators.

Even if the amount of analyzed data seems small, the annotation task has been conducted from October 2022

[4]https://labelstud.io/

to January 2023 by 5 experts of gender-(un)fair language (philosophers of language, linguists, and computational linguists), that provided alternatives for each textual span identified as unfair in the sentence.

Preliminary Analysis Thanks to this process of annotation in two layers, we created a corpus of 422 sentences where at least one gender-unfair expression has been identified, and 602 sentences where no gender-unfair expression has been identified.

In the 422 sentences containing gender-unfair expressions, the annotators detected on average 3 textual spans per sentence containing gender-unfair language (for a total of 3,195 portions) and proposed from 1 to about 11 alternatives.

Moreover, looking at the frequencies of the types of unfair expressions identified in the corpus, we can see from Table 2 that the most common case of gender-unfair language in administrative documents is represented by the use of overextended forms, and in particular of overextended masculine (e.g., "i ricercatori", the.M.PL researchers:M.PL; and "i docenti", the.M.PL teachers, for mixed-gender groups of, respectively, researchers and teachers).

#_portions	type
2,709	overextended
452	generic
34	incongruous
3,195	**total**

Table 2
Frequencies of the portions of texts identified as containing gender-unfair language. The use of *generic* and *overextended* gender is mostly referred to masculine cases.

Agreement A quantitative measure of inter-annotator reliability has not been calculated for different reasons. First, the scheme provides a variety of gender-fair options and the choice of a specific alternative depends on several factors, including individual preference: one annotator might agree on "lə Professorə" (the Professor:INN.SG) being gender-fair, but choose a different innovative option, like "lu Professoru" (the Professor:INN.SG), instead. Therefore, comparing the alternatives provided by each annotator is not a good measure of whether the annotators consider a specific option an appropriate gender-fair alternative to a certain gender-unfair expression. Relatedly, we do not plan to release an aggregated dataset, inclusive of a set of "gold-standard" preferred labels. Indeed, the aim of the present work is not to consolidate a 'ground truth' among rephrasing strategies but rather to explore as many solutions as possible while using gender-fair language. This is tied to our focus on methodology: the main purpose of this paper is to present the process of our work. As we mentioned before, we created a novel annotation scheme for the Italian language, which allows a fine-grained distinction between different cases of gender discrepancies. Moreover, the scheme has continuously been discussed between authors and annotators, mostly concerning the interpretation of labels, such as e.g. the sensible distinction between "overextended" and "generic" gender-unfair expressions. Last but not least, the identified span did not just contain gender-unfair expressions, but any element that needs to be changed in order to get a gender-fair text. For example, if an annotator decided to propose "il corpo docenti" (the teaching staff) instead of "i docenti" (the.M.PL teachers) in "i docenti devono partecipare" (the teachers have to participate), they also have to select the verb "devono" for it to agree in number with "il corpo docenti". Crucially, the verb only has to be selected if the proposed alternative to "i docenti" is singular and, thus, the verb needs to be singular too. Hence, if another annotator doesn't propose a singular alternative to "i docenti", they won't need to select the verb. As a result, the two annotators would select different spans even when agreeing on what are the gender-unfair expressions within the text. For this reason, even comparing just the textual spans between annotators would not be a good indicator of the annotators' agreement.

However, in Table 3 below we provide an example of a sentence with the reformulations proposed by the five annotators for each gender-unfair span of text, to give an idea of the kind of variation that can be found.

5. Conclusions and Future Work

In recent years, gender-fair language has gained significant attention, leading to the proposal of new strategies in various languages to avoid using masculine or feminine endings. Motivated by these theories, we conducted a study to examine the usage of different solutions in practical situations. We developed guidelines for gender-fair annotation and reformulation of texts, which we applied to a corpus of 1,024 portions of university administrative documents in Italian.

The corpus was annotated by 5 experts, and in 422 cases the annotators identified at least one gender-unfair expression. The preliminary analysis of this corpus highlighted the need to adopt specific guidelines (as well as a list of words to pay particular attention to) to support administrative staff in writing gender-fair texts.

Applying our annotation and reformulation guidelines to real data has led to theoretical advancements: we discovered that 'obscuration' and 'visibility' strategies can coexist within the same reformulation, and we consequently updated the annotation scheme to include 'mixed'

Text: *A partire dal 2013 il DiGi ha organizzato ogni anno International Summer schools, allo scopo di attrarre studenti stranieri e di offrire agli studenti bresciani l'opportunità di entrare in contatto con studenti e docenti di altri Paesi.*

ann.	span	Conservative visibility	Innovative visibility	Conservative obscuration	Innovative obscuration
A	'stranieri'	–	–	'di origine straniera' or 'di nazionalità estera'	–
	'agli studenti bresciani'	–	–	a 'studenti provenienti dalla provincia di Brescia'	–
B	'studenti stranieri'	–	–	'studenti di università estere'	–
	'agli studenti bresciani'	–	–	'a coloro che studiano all'Università di Brescia'	–
C	'studenti stranieri'	'studenti/esse stranieri/e'	'studenti/esse/ə stranieri/e/ə'	–	'studentə stranierə'
	'studenti bresciani'	'studenti/esse bresciani/e'	'studenti/esse/ə bresciani/e/ə'	–	'studentə brescianə'
	'studenti'	'studenti/esse'	–	'studentə'	–
D	'stranieri'	'straniere/i'	–	'dall'estero'	–
	'agli studenti'	'alle/agli studenti'	–	–	–
	'bresciani'	'bresciane/i'	–	'dalla provincia di Brescia'	–
E	'stranieri'	'straniere/i'	'straniere/i/3' or 'straniere/i/u'	'di nazionalità straniera'	–
	'agli studenti'	'alle/agli studenti'	'alle/agli/all3 studenti' or 'alle/agli/allu studenti'	'alle persone che studiano'	'all3 studenti' or 'all* studenti'
	'bresciani'	'bresciane/i'	'bresciane/i/3' or 'bresciane/i/u'	'del bresciano', 'di area bresciana', or 'di Brescia'	–

Table 3
Example of portions of texts with cases of overextended gender (*span*) annotated by each annotator (*ann.*).

alternatives.

To summarize, the annotation scheme has proven valuable both practically and theoretically. It facilitated the identification of gender-unfair expressions and the formulation of alternatives. Moreover, it revealed the inadequacy of an exclusive distinction between obscuration and visibility, emphasizing the need to incorporate a new type of strategy (i.e., 'mixed' alternatives) into the classification.

Although the created annotation scheme has been applied only to administrative texts so far, the guidelines are formulated in such a way that they can be easily applied to data pertaining to different domains. Indeed, we plan to extend the annotation to other data, like web pages of a University that describes its organization and its events. Finally, the resulting corpus, composed of 3,195 portions of texts identified as gender-unfair and reformulated with at least one alternative, could be used in the context of training models to recognize gender-unfair expressions and suggest their alternatives.

Ethics Statement

The annotators have been paid in the context of the actions provided by the Gender Equality Plan of the University of Brescia. The time of annotation has been monitored to ensure that the actual time spent annotating matched the agreed-upon paid hours.

Limitations

Our work presents some limitations. Firstly, the sample of analyzed texts is small and related to a specific domain. To test the robustness of the proposed guidelines, we planned to expand this corpus and its analysis. Secondly, in this work we presented an annotation schema to recognize gender-unfair language and to reformulate it, specifically for Italian, limiting its adaptation to other languages.

Acknowledgments

The annotation has been funded by Università degli Studi di Brescia as part of the actions provided for by the Gender Equality Plan.
The authors would like to extend special thanks to Alessandra Luccioli and Manuela Sanguinetti for their interest and their precious collaboration in this project.

References

[1] S. Sczesny, M. Formanowicz, F. Moser, Can gender-fair language reduce gender stereotyping and discrimination?, Frontiers in psychology 7 (2016) 25.

[2] J. Saul, E. Diaz-Leon, S. Hesni, Feminist Philosophy of Language, in: E. N. Zalta, U. Nodelman (Eds.), The Stanford Encyclopedia of Philosophy, Fall 2022 ed., Metaphysics Research Lab, Stanford University, 2022.

[3] P. Gygax, U. Gabriel, O. Sarrasin, J. Oakhill, A. Garnham, Generically intended, but specifically interpreted: When beauticians, musicians, and mechanics are all men, Language and cognitive processes 23 (2008) 464–485.

[4] P. Gygax, S. Sato, A. Öttl, U. Gabriel, The masculine form in grammatically gendered languages and its multiple interpretations: a challenge for our cognitive system, Language Sciences 83 (2021) 101328.

[5] MIUR, Linee guida per l'uso del genere nel linguaggio amministrativo del MIUR, MIUR, 2018. URL: https://www.miur.gov.it/documents/20182/0/Linee_Guida_per_l_uso_del_genere_nel_linguaggio_amministrativo_del_MIUR_2018.pdf/3c8dfbef-4dfd-475a-8a29-5adc0d7376d8?version=1.0&t=1520428640228.

[6] G. Attanasio, S. Greco, M. L. Quatra, L. Cagliero, R. Raus, M. Tonti, T. Cerquitelli, E-mimic: Empowering multilingual inclusive communication, IEEE Big Data Workshops (2021).

[7] G. Sulis, V. Gheno, The debate on language and gender in italy, from the visibility of women to inclusive language (1980s-2020s), The Italianist 42 (2022) 153–183. doi:10.1080/02614340.2022.2125707.

[8] A. Sabatini, Il sessismo nella lingua italiana. raccomandazioni per un uso non sessista della lingua italiana, Rome: Istituto Poligrafico e Zecca dello Stato (1987).

[9] C. Robustelli, Donne grammatica e media. con prefazione di n. maraschio, presidente accademia della crusca, in: Donne Grammatica e Media, ITA, 2014, pp. 1–79.

[10] A. M. Thornton, Per un uso della lingua italiana rispettoso dei generi, Università degli Studi dell'Aquila, 2020. URL: https://www.univaq.it/include/utilities/blob.php?item=file&table=allegato&id=4925.

[11] V. Gheno, Lo schwa tra fantasia e norma, La Falla del Cassero (2020). URL: https://lafalla.cassero.it/lo-schwa-tra-fantasia-e-norma/.

[12] I. González Vázquez, A. Klieber, M. Rosola, Beyond pronouns. gender visibility and neutrality across languages, in: L. Anderson, E. Lepore (Eds.), The Oxford Handbook of Applied Philosophy of Language, Oxford University Press, forthcoming.

[13] G. Marcato, E.-M. Thüne, Gender and female visibility in italian, in: M. Hellinger, H. Bußmann (Eds.), Gender Across Languages, Vol.2, John Benjamins Publishing, 2002, pp. 187–217.

[14] F. Formato, Gender, Discourse and Ideology in Italian, Palgrave Macmillan, 2019.

[15] M. Dardano, P. Trifone, La nuova grammatica della lingua italiana, Zanichelli, 1985.

[16] M. R. Costa-jussà, An analysis of gender bias studies in natural language processing, Nature Machine Intelligence 1 (2019) 495–496.

[17] A. A. Rescigno, E. Vanmassenhove, J. Monti, A. Way, A case study of natural gender phenomena in translation a comparison of google translate, bing microsoft translator and deepl for english to italian, french and spanish, Computational Linguistics CLiC-it 2020 359 (2020).

[18] T. Sun, A. Gaut, S. Tang, Y. Huang, M. ElSherief, J. Zhao, D. Mirza, E. Belding, K.-W. Chang, W. Y. Wang, Mitigating gender bias in natural language processing: Literature review, arXiv preprint arXiv:1906.08976 (2019).

[19] K. Stanczak, I. Augenstein, A survey on gender bias in natural language processing, arXiv preprint arXiv:2112.14168 (2021).

Appendix

		Innovative												
Type of term	Number	Star	u	Schwa	Suffix omission	Underscore	Dash	Apostrophe	At sign	x	y	Double masculine and feminine form	Double masculine and feminine form separated by a period	Plural in “-ai”
Nouns whose gender is overtly signaled by an inflectional suffix	sg pl	Maestr*	Maestru	Maestrə Maestr3	Maestr	Maestr_	Maestr-	Maestr’	Maestr@	Maestrx	Maestry	Maestrao; Maestroa Maestrei; Maestrie	Maestra.o; Maestro.a Maestre.i; Maestri.e	- Maestrai
Nouns whose gender is overtly signaled by a derivational suffix	sg pl	Lettor*; Lettric*	Lettoru; Lettricu	Lettorə; Lettricə Lettor3; Lettric3	Lettor; Lettric	Lettor_; Lettric_	Lettor-; Lettric-	Lettor’; Lettric’	Lettor@; Lettric@	Lettorx; Lettricx	Lettory; Lettricy	Lettorice?; Lettricore? Lettorici?; Lettricori?	Lettor.ice?; Lettric.ore? Lettor.ici?; Lettric.ori?	- Lettorai; Lettricai
Nouns whose feminine gender is overtly signaled by a derivational suffix	sg pl	Dottor*	Dottoru	Dottorə Dottor3	Dottor	Dottor_	Dottor-	Dottor’	Dottor@	Dottorx	Dottory	? ?	Dottor.essa Dottor.esse	- Dottorai
Nouns whose masculine gender is signaled by opposition to the feminine correspondent	sg pl	Ero*	Erou	Eroə Ero3	Ero	Ero_	Ero-	Ero’	Ero@	Erox	Eroy	Eroinae?; Eroeina? Eroinei?; Eroiine?	Eroina.e?; Eroe.ina? Eroine.i?; Eroi.ine?	- Eroai
Semi-epicene nouns, whose gender is overtly signaled in the plural but not in the singular	sg pl	Atlet*	Atletu	Atlet3	Atlet	Atlet_	Atlet-	Atlet’	Atleta Atlet@	Atletx	Atlety	Atletei; Atletie	Atlete.i; Atleti.e	Atletai
Epicene nouns, whose gender is not overtly signaled	sg pl								Docente Docenti					
Definite articles	sg pl	L* L*; *	Lu Lu; u	Lə L3; ə	L? L?; ?	L_ L_; _?	L- L-; -?	L’ L’; ’?	L@ L@; @	Lx Lx; x	Ly Ly; y	ila?; Loa?; Lao? Glile?; ile?; Lei?	il.a?; Lo.a?; La.o? Gli.le?; i.le?; Le.i?	- Lai
Indefinite articles		Un*	Unu, un*	Unə, un*	Un, un	Un_, un_	Un-, un-	Un’, un’	Un@, un*	Unx, un*	Uny, un*	?	Un.a, un.’?	-
Univerbated preposition + definite article	sg pl	Dell* Dell*; De*	Dellu Dellu; Deu	Dellə Dell3; Deə	Dell? Dell?; De?	Dell_ Dell_; De_	Dell- Dell-; De-	Dell’ Dell’; De’	Dell@ Dell@; De@	Dellx Dellx; Dex	Delly Delly; Dey	Dellao; Delloa Dellei?; Deglie?	Della.o; Dello.a; Del.la Delle.i; Degli.e; Dei.le?	- Dellai
3.sg pronouns		L*i	?	Ləi	?	L_i	L-i	L’i?	L@i	Lxi	Lyi?	Leui?; Leiui?; Luei?; Luiai?	Le.ui?; Lei.ui?; Lu.ei?; Lui.ai?	Lai

Table 4
Comprehensive table with the alternatives for innovative forms in Italian. We underline that these alternatives can also be used for adjectives or forms of past participle with a gender-marked suffix. Moreover, with the “?” symbol, we refer to the alternatives whose actual form is not totally clear nor, until now, well identified by Italian speakers.

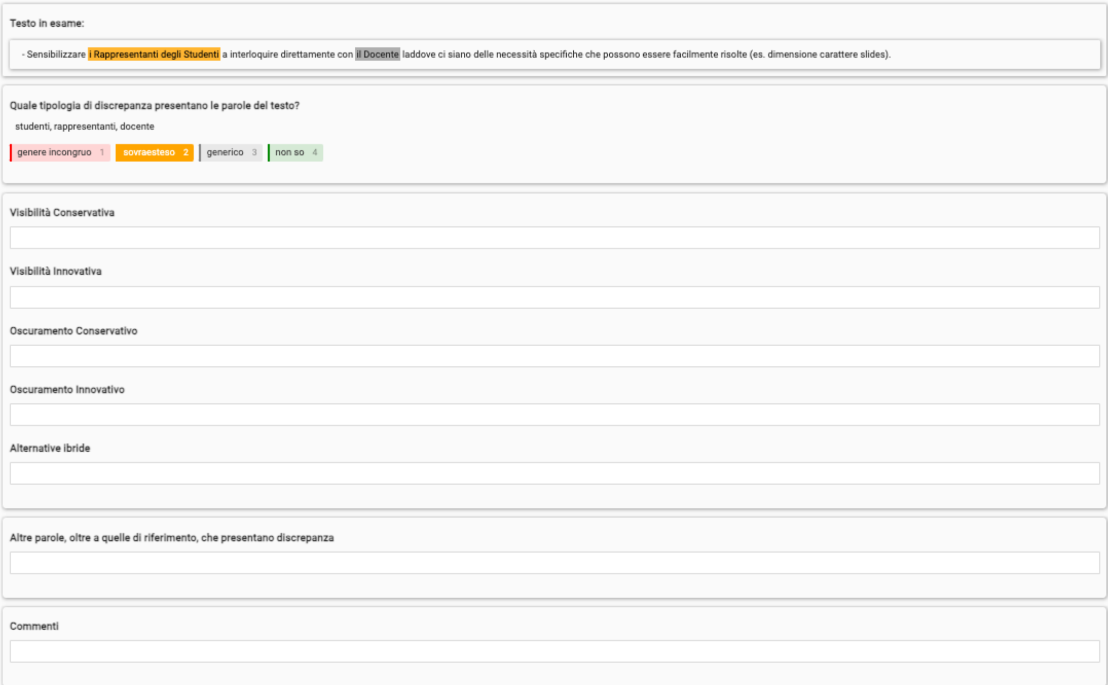


Figure 1: Label Studio set up with a text containing an overextended and a generic gender-unfair text span.

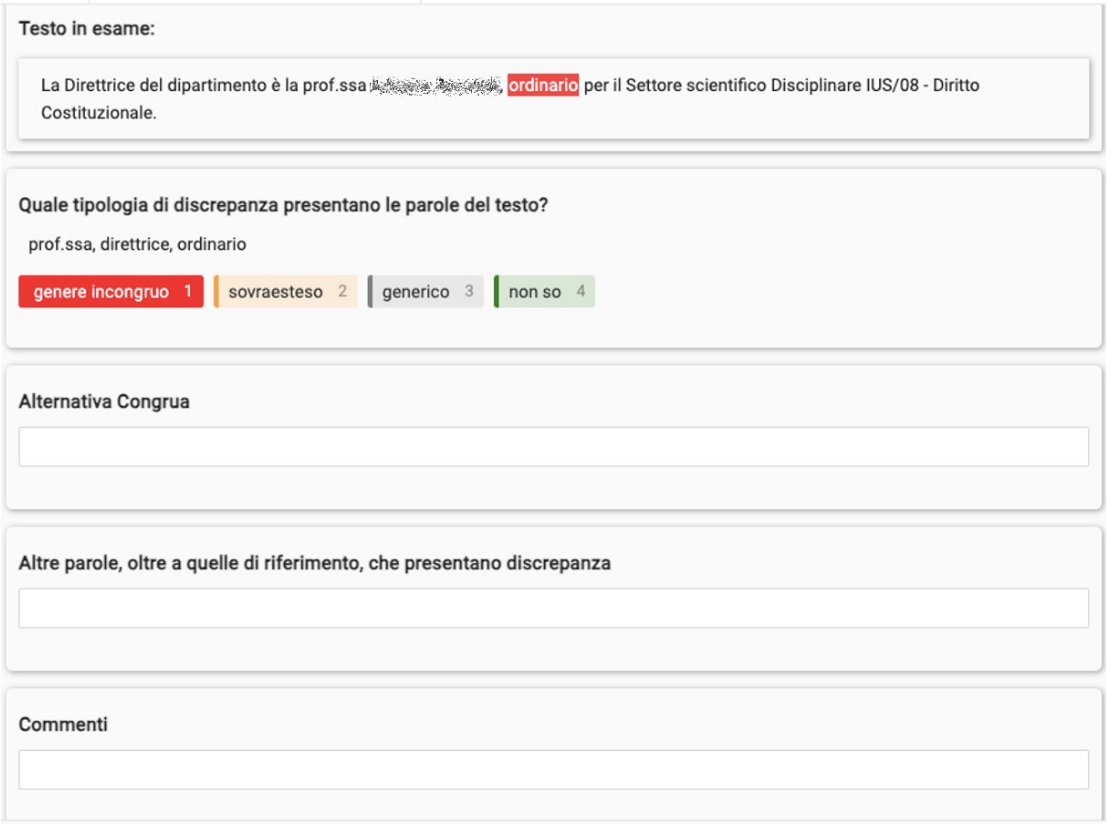


Figure 2: Label Studio set up with a text containing an incongruous gender-unfair text span.

Blaze-IT: a lightweight BERT model for the Italian language

Francesco Russo, Michele Filannino

Prometeia S.p.A., Piazza Trento e Trieste 3, Bologna, Italy

Abstract

In this work, we present a lightweight language model based on BERT (Blaze-IT) and a lightweight language model based on MiniLM (Flare-IT), both specifically designed for the Italian language. Starting from the multilingual cased DistilBERT and MiniLM models, we modified the embedding layers and then carried out a continued pre-training procedure on Italian Wikipedia data using whole word masking, resulting in two uncased models. Blaze-IT has 55M parameters and weighs 217MB, while Flare-IT has 17M parameters and only weighs 67MB. The models are tailored to analyze large volumes of natively digital text, such as wikis, web pages and news articles, written in correct and fluent Italian. We evaluate their performances on various downstream tasks and compare them to other models in the same class. We also discuss the limitations of our models and suggest possible directions for future work. Our results show that our models achieve competitive performances while being much smaller than other monolingual models, making them suitable for deployment in resource-constrained environments.

Keywords

BERT, MiniLM, Natural Language Processing, Foundation models, Italian

1. Introduction

Natural Language Processing (NLP) has rapidly advanced in recent years, with language models such as BERT [1] (and its variants) and GPT [2] achieving state-of-the-art results in various NLP tasks. However, the sheer size and complexity of these models pose a significant challenge when it comes to analyzing large volumes of data or deploying applications in low-resource settings, where CPU parallelization is the only viable way to speed up the computation (since GPUs are not available or not cost effective) and loading multiple models in parallel can quickly flood the RAM.

While some previous work focused on creating a small uncased model for the Italian language exploiting knowledge distillation [3] (which produced an effective Italian DistilBERT model [4], with $\sim 40\%$ less parameters than a classic BERT model), other research went on to reduce the size of the embedding layer to focus a multilingual model on a single language [5]. Quantization and pruning techniques are also widely used. [6] [7].

In this paper, we present two lightweight language models, based on BERT [1] and MiniLM [8] respectively, both designed specifically for the Italian language. The first one (Blaze-IT) is overall 50% lighter than typical mono-lingual BERT models and 20% lighter than standard DistilBERT models, while still producing high-quality results (see the section Results). The second one (Flare-IT) is 85% lighter than mono-lingual BERT and 75% lighter than DistilBERT. In addition, both models are uncased, which makes them extremely versatile and suitable for a wide spectrum of scenarios where word capitalization might not be respected or reliable.

Our models can effectively process natural language inputs and perform a wide range of NLP tasks such as topic modeling, named entity recognition and question answering, therefore highlighting the importance of developing lightweight language models that can operate effectively in resource-constrained settings, making NLP accessible to a wider range of use-cases.

2. Blaze-IT and Flare-IT

The first proposed language model (Blaze-IT) is based on the multilingual cased DistilBERT model [3] (distilbert-base-multilingual-cased, 6 hidden layers and hidden size of 768, developed by the HuggingFace team as a distilled version of the original multilingual BERT). To focus this model on the Italian language, we first modified the embedding layer, following the approach presented in [5], which we extended to the deletion of cased tokens to turn the original cased model into an uncased version. This was achieved by turning the Italian language subset of the Wikipedia dataset to lowercase, tokenizing the texts with the WordPiece tokenizer of the DistilBERT model, and then computing document-level frequencies of tokens, setting a minimum threshold of 0.1% to determine which tokens to keep.

The same procedure was followed with the second model (Flare-IT) except that in this case the mMiniLMv2 model [8] (L6xH384 mMiniLMv2, 6 layers and hidden size of 384, developed by Microsoft as a distilled version of XLM-RoBERTa-Large) was used as a starting point.

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

francesco.russo@prometeia.com (F. Russo);
michele.filannino@prometeia.com (M. Filannino)

However, the resulting models were still relying on their original training, which exploited capitalized representations of several words (like words at the beginning of sentences, or proper names of people, places and other entities), so they were not properly trained to deal with the lowercase equivalents. Moreover, while many commonly capitalized words were previously represented by a single token, their lowercase equivalent is likely to be splitted in several subword tokens, being less common than the capitalized version (e.g. "Microsoft" → "micro ##so ##ft"), which makes it harder for the models to deal with them properly, especially in token classification tasks.

To make the models more robust to the lowercase representations of words previously capitalized and compensate for the deletion of cased tokens, we exploited a continued pre-training procedure [9] [10]. More specifically, we further pre-trained the models on the Italian split of the Wikipedia dataset, using the whole word masking technique [11]. By masking whole words at once, rather than individual tokens, this technique makes the Masked Language Modeling (MLM) task harder for the models, encouraging them to learn more effective representations and to capture a wider range of linguistic structures.

Overall, these modifications allowed us to adapt the two pre-existing multilingual language models to the Italian language, and to turn them from case-sensitive to case-insensitive, significantly reducing the size of the models while maintaining their ability to produce effective representations of Italian text.

Blaze-IT has 55M parameters, a vocabulary of 13.832 tokens, and a size of 217MB. Flare-IT has 17M parameters, a vocabulary of 14.610 tokens, and a size of 67MB. The models can be fine-tuned for a wide range of downstream NLP tasks, making them highly versatile and useful for practical applications (we fine-tuned them on Text Classification, Part Of Speech Tagging, Named Entity Recognition, Semantic Textual Similarity and Extractive Question Answering, reporting the results in the dedicated section Results). A short comparison [1] between Blaze-IT, Flare-IT, BERT [2] and DistilBERT [3] is summarized in Tables 1, 2, 3 and 4.

Table 1
Comparison across the major model indicators between Blaze-IT and BERT.

	Blaze-IT	**BERT**	$\Delta\%$
Vocab	13.832	32.102	-56,9%
Params	54.150.920	110.727.782	-51,1%
Size	217MB	445MB	-51,2%

Table 2
Comparison across the major model indicators between Flare-IT and BERT.

	Flare-IT	**BERT**	$\Delta\%$
Vocab	14.610	32.102	-54,5%
Params	16.618.770	110.727.782	-85,0%
Size	67MB	445MB	-84,9%

Table 3
Comparison across the major model indicators between Blaze-IT and DistilBERT.

	Blaze-IT	**DistilBERT**	$\Delta\%$
Vocab	13.832	32.102	-56,9%
Params	54.150.920	68.200.550	-20,6%
Size	217MB	273MB	-20,5%

Table 4
Comparison across the major model indicators between Flare-IT and DistilBERT.

	Flare-IT	**DistilBERT**	$\Delta\%$
Vocab	14.610	32.102	-54,5%
Params	16.618.770	68.200.550	-75,6%
Size	67MB	273MB	-75,5%

2.1. Training details

The proposed Italian language models have been trained using Masked Language Modeling (MLM) on the Italian subset of the Wikipedia dataset, which contains approximately 3.7GB of text data (we used a 2020 dump of Wikipedia, already pre-processed by the HuggingFace team). Specifically, adapting from the continued pre-training setups in [9] and [10], the models were trained for 10,000 steps using the `AdamW` optimizer with a batch size of 512, obtained through 128 gradient accumulation steps and an instantaneous batch size of 4 on a NVIDIA GeForce RTX 3060 GPU. We kept the sequence length fixed to 512 and applied a linearly decaying learning rate starting from $5 \cdot 10^{-5}$.

Following the original pre-training strategy of BERT [1], we masked 15% of the tokens for each training instance, where 80% are effectively replaced by a [MASK] token, 10% are replaced by a random token and 10% are left unchanged. However, unlike the original BERT training procedure, and in agreement with the improvements introduced by the RoBERTa procedure [12], we removed

[1] The MB sizes of the models are referred to their PyTorch checkpoints. Since their exact value can slightly vary depending on the platform, we used the size of the .bin files uploaded on HuggingFace as a reference

[2] We used the bert-base-italian-xxl-uncased model on HuggingFace, released by the Bavarian State Library MDZ team, as a reference BERT model

[3] We used the BERTino model on HuggingFace, released by indigo.ai, as a reference DistilBERT model

the Next Sentence Prediction task from the pre-training.

To ensure optimal performance during training, we also employed dynamic masking [12] between epochs and utilized the whole word masking technique to encourage the models to learn more effective representations of Italian lowercased text. The dynamic masking technique involves randomly masking tokens in the input sequence during training over different epochs, while the whole word masking technique involves masking entire words at once rather than just individual tokens. Together, these techniques help to prevent overfitting and improve the robustness of the models.

The resulting models have been fine-tuned and evaluated on a range of benchmark datasets, demonstrating comparable performances with other models in their class. The limited size of the models, combined with their performances, makes them highly valuable assets for large-scale data analysis, especially in resource-constrained settings or in applications where computational efficiency is a priority, without excessively compromising on output quality.

3. Results

The metrics in Table 5, 6, 7, 8, 9 have been computed by fine-tuning our models and the reference models on:

- `Text Classification`: XGLUE NC, machine-translated from English [4] [13]
- `Part of Speech Tagging`: UD Italian ISDT dataset [5] [14]
- `Named Entity Recognition`: WikiNER dataset [6] [15]
- `Semantic Similarity`: MULTI STS-B dataset [16]
- `Extractive Question Answering`: SQuAD-IT dataset [17]

The Text Classification models have been trained for 1 epoch and the PoST models for 5 epochs, while the NER, STS and EQA models have been trained for 3 epochs, all with a constant learning rate, fixed at 10^{-5}. For Text Classification, Part of Speech Tagging, Semantic Similarity and Extractive Question Answering, the metrics have been computed on the default test set provided with the dataset, while for Named Entity Recognition the metrics have been computed with a 5-fold cross-validation.

For Text Classification on the NC dataset, the Accuracy metric has been used, in agreement with the original XGLUE paper. For Token Classification tasks, the Recall, Precision and F1 metrics have been computed at the token level and then macro-averaged over the classes. For STS and EQA the Pearson's r and Exact Match metrics have been used, respectively.

[4]We used the Helsinki-NLP/opus-mt-en-it from HuggingFace for the translation

[5]Italian corpus annotated according to the UD scheme, obtained by conversion from ISDT, released for the shared task at Evalita-2014

[6]The B-type and I-type categories have been collapsed together since the B-type categories have extremely limited support

Table 5
Text Classification results

Model	Accuracy
BERT	90.73
DistilBERT	90.17
Blaze-IT	90.13
Flare-IT	86.23

Table 6
Part of Speech Tagging results

Model	Recall	Precision	F1
BERT	97.89	97.74	97.80
DistilBERT	97.64	97.45	97.53
Blaze-IT	97.48	97.29	97.37
Flare-IT	95.64	95.32	95.45

Table 7
Named Entity Recognition results

Model	Recall	Precision	F1
BERT	92.04	91.49	91.75
DistilBERT	90.76	91.30	91.01
Blaze-IT	89.29	89.84	89.53
Flare-IT	82.27	80.64	81.29

Table 8
Extractive Question Answering results

Model	EM
BERT	61.03
DistilBERT	56.64
Blaze-IT	55.08
Flare-IT	52.83

3.1. Throughput

In order to test the improvements that can be achieved by exploiting the limited weight of Blaze-IT and Flare-IT, we conducted an experiment which simulates the typical conditions of a cloud instance. More specifically, we set up a Docker image with the relevant requirements for an inference task (we chose the Text Classification task on the NC dataset), and then launched a container with 8 CPU cores and a 8GB RAM memory budget. For each one of the models, we tried to achieve the maximum level of parallelization allowed by the RAM (i.e. the maximum

Table 9
Semantic Textual Similarity results

Model	Pearson's r
BERT	0.8234
DistilBERT	0.7920
Blaze-IT	0.7768
Flare-IT	0.7572

Table 10
Throughput measurements with a fixed memory budget

Model	N. jobs	Samples / s	$\Delta\%$
BERT	1	1.06	//
DistilBERT	3	2.25	+112%
Blaze-IT	4	2.49	+135%
Flare-IT	8	5.40	+420%

number of parallel jobs that could be launched without getting a SIGKILL signal from the operating system)

We then proceeded to measure the throughput reached by the models, each one with its maximum parallelization, and the relative increase in throughput compared to a classical BERT model. The results are in Table 10

3.2. Limitations

The proposed lightweight Italian language models have been further pre-trained, in our work, on the Italian subset of the Wikipedia dataset, which consists of high-quality, natively digital text written in a correct and fluent form. As a result, the model is particularly well-suited for analyzing large volumes of text from the web, such as wikis, web pages, news articles, and other similar sources.

However, it is worth noting that the models may have limitations when it comes to analyzing chaotic text that contains errors, slang expressions, or other types of noise. This is because such text is often less structured and less consistent than the text found in more formal digital sources, which can make it more difficult for the models to accurately process and interpret. Additionally, the models may struggle when analyzing domain-specific text, such as medical, financial, or legal content, which often contains specialized terminology and conventions that may not be present in more general digital text.

Despite these limitations, the lightweight design and robust performances of the proposed models make them extremely valuable for a wide range of natural language processing applications. In particular, their efficiency and agility make them well-suited for analyzing large volumes of digital text or processing inputs in real-time, which can be useful in a variety of contexts, including intelligent document processing, conversational systems and web content analysis. Furthermore, the performances of the models on specific domains can be improved through further pre-training by incorporating additional training data, which may help to overcome some of the limitations we have mentioned.

4. Related work

Language models have become a crucial component of many natural language processing applications, ranging from text classification and sentiment analysis to machine translation and question answering. Recent advances in transformer-based architectures, such as BERT [1] and GPT [2], have significantly improved the state-of-the-art performance in a wide range of natural language processing tasks. However, these models are often large and computationally expensive, making them difficult to deploy in resource-constrained environments.

Indeed, large transformers are especially cumbersome to deal with when only CPUs are available as processing units, since the execution speed is going to be heavily limited and the sheer size of the neural networks makes it hard to deploy multiple models in parallel without flooding the RAM. The huge weight of these models can also become an obstacle in cloud computing, when server-less applications exploiting state-less functions are involved, since loading large models takes lots of time, which goes against the idea of executing a function on-the-fly.

To address these issues, several approaches have been proposed to reduce the complexity and size of language models without compromising their performance. One such approach is distillation [3], in which a larger, pre-trained model is used to train a smaller, "distilled" model that can achieve comparable performance. However, knowledge distillation is computationally expensive, and while several of these compressed version have been released for English models, or for multilingual models, only a few studies have focused specifically on developing lightweight language models for low-resource languages, which may also lack the large, high-quality training datasets that are available for more widely spoken languages.

Another approach is pruning, in which unimportant connections (or even entire layers) are removed from the model to reduce its size [7] [18]. While pruning techniques are effective up to a certain point, they can affect the performances when relevant fractions of the models are removed.

A different technique is quantization [6], which exploits less accurate representations of floating points (e.g. 16 bits instead of 32) so that the resulting model is lighter (even though this is only strictly true for half-precision, because working in mixed-precision with master weights will actually lead to two copies of the model weights being loaded, one in FP32 and one in FP16 [19]).

Lastly, the modification of the embedding layer proposed in [5], which is the method we followed in our work (and can be seen as a form of pruning where only weights corresponding to unused tokens are removed), allows to focus a multilingual model on a single language by getting rid of the extra parameters in the embedding layer, therefore reducing its size (the parameters in the embedding layer are a considerable portion of the total parameters when the model's vocabulary is large). This procedure implies a limited (or sometimes even negligible) loss in performance, since it only affects statistically rare tokens for the target language.

When applied to the distilled versions of multilingual models, this technique can further reduce their size and, as we showed in our work, if cased tokens are also removed (with the inclusion of an additional pre-training phase to compensate for their deletion, possibly exploiting whole word masking) the procedure ultimately delivers extremely light language models, with the additional benefit of the new uncased representations.

The introduction of an additional pre-training phase, formally known as continued pre-training, has been mainly explored as a method to adapt language models to new domains or tasks [10] ("domain-adaptive pre-training" and "task-adaptive pre-training"), yielding improvements in downstream performances on Text Classification. Recent work on Spoken Language Understanding [9] has brought this further by investigating the effectiveness of the continued pre-training of English models in cross-lingual settings, showing that the domain knowledge obtained on intermediate data is even transferable to other languages.

In our work, we exploited continued pre-training to adapt our language models to uncased text, taking inspiration from the setups in [9] and [10].

5. Conclusions

In this paper, we presented Blaze-IT, a lightweight language model based on BERT, and Flare-IT, a lightweight language model based on MiniLM, both specifically tailored for the Italian language. Our models are significantly smaller than other monolingual Italian models, the first one weighing 217MB and having 55M parameters, the second one weighing only 67MB and having 17M parameters. We achieved this by starting with the multilingual DistilBERT and MiniLM models, reducing the embedding layer and further pre-training them on Italian Wikipedia data using whole word masking. While the models are designed to excel on correctly written digital text, they may struggle with noisy, informal language or domain-specific jargon.

The limited size of our models makes them well-suited for local environments and applications where large volumes of data have to be processed, especially if no hardware acceleration is available, since the execution of these light models can be easily parallelized on multiple CPUs. They are also ideal when computational resources or memory are limited, such as on mobile devices or edge-computing environments, or even in cloud-computing scenarios where server-less applications are involved, since these models can be quickly loaded and used in state-less functions. We hope that our work will help to lower the entry barrier for natural language processing tasks for researchers and practitioners working in low-resource settings.

In future work, we plan to investigate methods for further compressing the size of transformer-based models while maintaining performances, perhaps combining the techniques showed in this work with model quantization. We also aim to expand the capabilities of our models, to handle informal and noisy text, and to develop domain-specific versions of the models for specialized applications. Overall, we believe that this work represents a step towards democratizing access to natural language processing tools and techniques, and we look forward to further developments in this area.

You can find the models online on the HuggingFace platform at https://huggingface.co/osiria/blaze-it and https://huggingface.co/osiria/flare-it

You can also try the models online (fine-tuned on named entity recognition) using the web apps at https://huggingface.co/spaces/osiria/blaze-it-demo and https://huggingface.co/spaces/osiria/flare-it-demo.

6. Acknowledgments

The authors would like to acknowledge Prometeia S.p.A. for the unconditioned support, availability and guidance received. The authors would also like to thank Alessio Palmero Aprosio (FBK) for his valuable suggestions and feedback.

References

[1] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: North American Chapter of the Association for Computational Linguistics, 2019.

[2] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen,

E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, 2020.

[3] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, ArXiv abs/1910.01108 (2019).

[4] M. Muffo, E. Bertino, Bertino: An italian distilbert model, ArXiv abs/2303.18121 (2023).

[5] A. Abdaoui, C. Pradel, G. Sigel, Load what you need: Smaller versions of mutlilingual bert, in: SUSTAINLP, 2020.

[6] A. Gholami, S. Kim, Z. Dong, Z. Yao, M. W. Mahoney, K. Keutzer, A survey of quantization methods for efficient neural network inference, ArXiv abs/2103.13630 (2021).

[7] M. A. Gordon, K. Duh, N. Andrews, Compressing bert: Studying the effects of weight pruning on transfer learning, ArXiv (2020).

[8] W. Wang, H. Bao, S. Huang, L. Dong, F. Wei, Minilmv2: Multi-head self-attention relation distillation for compressing pretrained transformers, in: Findings, 2020.

[9] S. M. B. Louvan, Samuel; Casola, Investigating continued pretraining for zero-shot cross-lingual spoken language understanding, in: Proceedings of the Eighth Italian Conference on Computational Linguistics Lingua/e Inglese, 2018.

[10] S. Gururangan, A. Marasovi'c, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, N. A. Smith, Don't stop pretraining: Adapt language models to domains and tasks, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020.

[11] Y. Cui, W. Che, T. Liu, B. Qin, Z. Yang, S. Wang, G. Hu, Pre-training with whole word masking for chinese bert, ArXiv abs/1906.08101 (2019).

[12] L. Zhuang, L. Wayne, S. Ya, Z. Jun, A robustly optimized BERT pre-training approach with post-training, in: Proceedings of the 20th Chinese National Conference on Computational Linguistics, Chinese Information Processing Society of China, 2021.

[13] Y. Liang, N. Duan, Y. Gong, N. Wu, F. Guo, W. Qi, M. Gong, L. Shou, D. Jiang, G. Cao, X. Fan, R. Zhang, R. Agrawal, E. Cui, S. Wei, T. Bharti, Y. Qiao, J.-H. Chen, W. Wu, S. Liu, F. Yang, D. Campos, R. Majumder, M. Zhou, XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2020.

[14] C. Bosco, F. Dell'Orletta, S. Montemagni, M. Sanguinetti, M. Simi, The evalita 2014 dependency parsing task, Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 and of the Fourth International Workshop (2014).

[15] J. Nothman, N. Ringland, W. Radford, T. Murphy, J. R. Curran, Learning multilingual named entity recognition from wikipedia, Artificial Intelligence 194 (2013).

[16] P. May, Machine translated multilingual sts benchmark dataset., 2021. URL: https://github.com/PhilipMay/stsb-multi-mt.

[17] D. Croce, A. Zelenanska, R. Basili, Neural learning for question answering in italian, in: International Conference of the Italian Association for Artificial Intelligence, 2018.

[18] H. Sajjad, F. Dalvi, N. Durrani, P. Nakov, On the effect of dropping layers of pre-trained transformer models, Computer Speech & Language (2023).

[19] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, H. Wu, Mixed precision training, in: International Conference on Learning Representations, 2018.

Camoscio: an Italian Instruction-tuned LLaMA

Andrea Santilli[1], Emanuele Rodolà[1]

[1]Sapienza University of Rome - Computer Science Department

Abstract
In recent years Large Language Models have improved the state of the art on several natural language processing tasks. However, their availability is frequently restricted to paid API services, posing challenges for researchers in conducting extensive investigations. On the other hand, while some open-source models have been proposed by the community, they are typically English-centric or multilingual without a specific adaptation for the Italian language. In an effort to democratize the available and open resources for the Italian language, in this paper we introduce Camoscio: a language model specifically tuned to follow users' prompts in Italian. Specifically, we finetuned the smallest variant of LLaMA (7b) with LoRA on a corpus of instruction prompts translated to Italian via ChatGPT. Results indicate that the model's zero-shot performance on various downstream tasks in Italian competes favorably with existing models specifically finetuned for those tasks. All the artifacts (code, dataset, model) are released to the community at the following url: https://github.com/teelinsan/camoscio

Keywords
Large Language Models, Instruction-tuned Models, Resources for the Italian Language

1. Introduction

In recent years, Large Language Models (LLMs) have made remarkable advancements in the field of natural language processing, demonstrating state-of-the-art performance on various tasks [1, 2, 3]. However, the majority of these models are typically controlled by for-profit organizations that release just a paid API for receiving responses based on input textual prompts. This severely constrains researchers from conducting comprehensive and meaningful research, as they lack access to both the model's weights and the training data regime. This limitation is particularly relevant for privacy-sensitive applications (e.g., medical domain) where data cannot be shared with external providers.

On the other hand, several open-source models[1] have been proposed as an alternative to closed models [4, 5, 6]. However, most of these models are English-centric or multilingual, albeit with performance that lags behind their monolingual counterparts. Furthermore, in these latter models, support for the Italian language is usually poor. For example, BLOOM – the largest open multilingual model available up to date – has not been trained on any Italian data, while LLaMA has only a small percentage of training data in the Italian language [2]. In addition to this, most of these models are only trained with the standard language modeling objective (i.e., predict the next token given the previous ones) on corpora of raw textual data, while it has been shown that a second training step of instruction-tuning is crucial to increase downstream performance [7, 8, 9]. Recently, a step in this direction has been made by Taori et al. [10] with the release of Stanford Alpaca, an instruction-tuned version of LLaMA for the English language. Following this approach, in this paper we propose Camoscio as an instruction-tuned version of LLaMA for the Italian language by translating to Italian the instruction-tuning dataset of Stanford Alpaca. In particular, we finetuned the smallest version of LLaMA (7 billion parameters) with LoRA [11], a parameter-efficient finetuning technique that allows to train larger models on standard desktop hardware.

Our contributions are the following:

- We introduce an instruction-tuning dataset for the Italian language, stemming from the Stanford Alpaca [10] dataset, translating it to Italian.
- We train Camoscio on this dataset and evaluate its zero-shot performance on several downstream tasks for the Italian language (NewsSum-IT, SQuAD-IT, XFORMAL IT).
- We release all the artifacts (code, dataset, model checkpoints) to the community.

2. Background

Large language models have emerged as a general class of models capable of performing a wide range of tasks without explicit finetuning by just leveraging in-context examples [12]. They've garnered popularity not only in the natural language processing domain but also across audio, image, and multimodal domains [13, 14, 15], with most of the approaches scaling or optimizing their performance [2, 16].

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
santilli@di.uniroma1.it (A. Santilli); rodola@di.uniroma1.it (E. Rodolà)

[1]Actual openness depends on the model license.
[2]Less than 4.5% of training data comes from Wikipedia in 20 different languages, including Italian.

In the context of the Italian language, the availability of pre-trained language models is currently limited; generic multipurpose LMs are almost nonexistent. Notable mentions include: AlBERTo [17], an Italian version of BERT [18] trained on Italian tweets from TWITA [19]; GePpeTto [20], a version of GPT-2 base (117 million parameters) finetuned using Italian Wikipedia and the ItWac corpus [21]; IT5 [22] a T5 model tailored for Italian using a refined version of the mC4 corpus [23]; and BART-IT [24], an Italian variant of BART [25] trained on the same mixture of data as IT5. Concurrently to our work, Bacciu et al. [26] proposed Fauno, an Italian version of Baize [27] that is a LM trained on a corpus of self-chat performed by ChatGPT. Compared to our work, their approach is tailored to develop a conversational agent for the Italian language. After our work, Michael [28] released on their GitHub repository an instruction-tuned version of LLaMA on a translation to Italian of the GPT-4-LLM dataset [29].

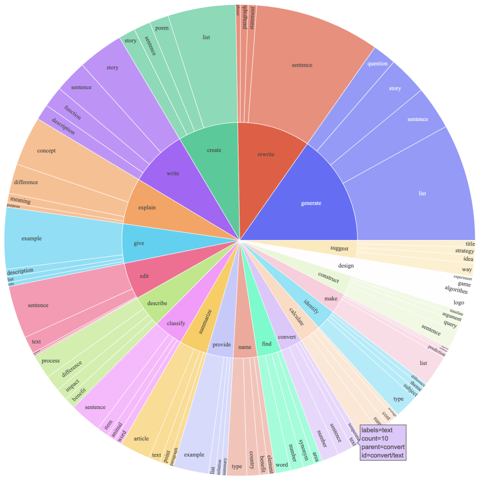


Figure 1: Diversity of the examples in the Stanford Alpaca dataset. Illustration from Taori et al. (2023). The inner circle shows the root verb on the instruction while the outer circle shows the direct object. The dataset of Camoscio is constructed by translating all these examples to Italian via *gpt-3.5*.

3. Method

For the construction of our instruction-tuning dataset for the Italian language, we stem from the Stanford Alpaca dataset [10] and Alpaca LoRA [30] for their finetuning approach.

3.1. Dataset

Stanford Alpaca is an instruction-tuning dataset constructed using the self-instruct method [31]. Specifically, the authors started with a set of 175 human-written instruction-output pairs from the original self-instruct paper[3] and used them as in-context examples to prompt OpenAI *text-davinci-003*. A total of 52.000 novel examples are generated with this technique. Each example includes an *instruction*, in natural English language, the answer (*output*), and optionally an additional context (*input*) for some datapoints (e.g., a short paragraph for question answering). Figure 1 shows different types of instructions in the dataset.

Translation. Inspired by Croce et al. [32], Scaiella et al. [33] and Larcher et al. [34], we translated the original dataset of Stanford Alpaca to Italian using `gpt-3.5-turbo` with the prompt *"Translate the following text to Italian:* `{text}`*"*. We translated all the fields in the dataset (*instruction, input, output*). We decided to use ChatGPT instead of other APIs for translation (e.g., Google Translate, Microsoft Azure Translator, DeepL) because we found it to be more robust for translating code examples i.e., it translates correctly just the comments in the code and not also the coding lexicon of the programming language. We provide here an example from the dataset. `Instruction`: *"Data una parola, costruisci i suoi antonimi."*, `Input`: *"Luce"*, `Output`: *"Scuro, pesante, denso"*.

Clearly the translation is not always perfect, but it is a fast-and-cheap method to bootstrap a noisy instruction-tuning dataset for the Italian language.

3.2. Training & Prompting

We finetuned the smallest version of LLaMA [6] (7 billion) on an instruction-tuning dataset for the Italian language, obtained by translating to Italian the dataset of Stanford Alpaca as described in the paragraph above.

The model is trained with supervision with the standard objective of predicting the next token given the previous ones. The dataset has `instruction`, `input`, `output` fields, but the `input` is not available for all data points (e.g., open-ended generation). For such cases, we construct the prompt: *"Di seguito è riportata un'istruzione che descrive un task. Scrivete una risposta che completi adeguatamente la richiesta. ### Istruzione:* `{instruction}` *### Risposta:* `{output}`*"*. If, instead, the datapoint also has an `input` (e.g., question answering where the input is the contextual paragraph), we construct the prompt: *"Di seguito è riportata un'istruzione che descrive un task, insieme ad un input che fornisce un contesto più ampio. Scrivete una risposta che completi adeguatamente la richiesta. ### Istruzione:* `{instruction}` *### Input:* `{input}` *### Risposta:* `{output}`*"*.

[3] https://github.com/yizhongw/self-instruct

Table 1
Results on SQuAD-IT. All the models are trained on the SQuAD-IT training set, except for Camoscio which is evaluated in a zero-shot fashion. The additional evaluation metric *Exact Match via ChatGPT* is highlighted in grey. The scores F1 and EM for competitor models are reported from their respective papers.

	SQuAD-IT						
	F1	EM	EM-GPT	R1	R2	RL	BS
DrQA-IT [32]	.659	.561	-	-	-	-	-
mBERT [35]	.760	.650	-	-	-	-	-
BERT[3] [18]	.753	.638	-	-	-	-	-
MiniLM [36]	.720	.577	-	-	-	-	-
$\text{MiniLM}_{+\text{st}}$ [36]	.745	.620	-	-	-	-	-
XLM-R $\text{Large}_{+\text{st}}$ [36]	.804	.676	-	-	-	-	-
mT5 Small [22]	.660	.560	.684	.617	.347	.617	.712
mT5 Base [22]	.757	.663	.745	.709	.396	.708	.770
IT5 Small [22]	.716	.619	.602	.671	.372	.671	.743
IT5 Base [22]	.761	.663	.600	.712	.406	.712	.770
IT5 Large [22]	.780	.691	.641	.730	.412	.729	.784
Camoscio-7b **(0-shot)**	.270	.077	.576	.242	.133	.241	.237

At inference time, the same prompt is used to generate the answer. Only the text generated after "[...] *### Risposta:"* is used as final output. We sample from the model using *top-p* sampling [37] with a temperature of 0.2, $p = 0.75$, $k = 40$, and beam search with 4 beams.

We refer to Appendix A for the additional implementation details.

4. Experiments

Currently, there is a very limited availability of datasets for a solid evaluation of the broad capabilities these general-purpose models possess. This is true for English but especially for the Italian language, although the community is moving towards this direction [38]. To evaluate our model we decided to follow the same evaluation protocol proposed in Sarti and Nissim [22]. Compared to their approach, we do not perform any training on the downstream tasks, i.e., we perform just the evaluation on the test set in a zero-shot fashion by providing to the model a textual description of the task (e.g., *"Riassumi il seguente articolo"*). We compared the performance of our model on standard Italian benchmarks for summarization (NewsSum-IT), question answering (SQuAD-IT), and style transfer (XFORMAL IT).

Compared to Sarti and Nissim [22], we do not include the Wikipedia for Italian Text Summarization (WITS) corpus [39] since Wikipedia is included in the original training corpus of LLaMA [6]. We also omitted the news style transfer task between "Il Giornale" to "La Repubblica" (and vice-versa) based on CHANGE-IT [40], since Camoscio has no concepts of "Il Giornale" or "La Repubblica" styles (i.e., it was never exposed during training or finetuning to this kind of articles, although we recognize it might be interesting to analyze this in a few-shot setting). We describe in the next paragraphs the three datasets used for the evaluation.

News Summarization. We evaluate the news article summarization capabilities of Camoscio using the dataset NewSum-IT proposed by Sarti and Nissim [22]. This dataset is obtained by merging two newspaper sources ("Fanpage.it" and "Il Post") scraped by the Applied Recognition Technology Laboratory[4] and available on the Hugging Face Hub [41]. We used only the test split for the zero-shot evaluation and asked the model to generate an answer given the `instruction` *"Dopo aver letto il testo qui sotto, riassumilo adeguatamente."* provided in the textual prompt and the news text provided as `input` (complete prompt as explained in §3.2). We use the same evaluation metrics of Sarti and Nissim [22] and report the average across the two newspapers as in their work.

Question Answering. To assess the model performance on extractive question answering, we used the SQuAD-IT dataset [32]. This dataset is composed of sets of paragraphs, questions, and answers derived from the original SQuAD dataset [42] via machine translation and subsequent filtering of problematic instances. As for the previous datasets, we used just the test split for zero-shot evaluation. The model is asked to generate an answer given the `instruction` *"Dopo aver letto il paragrafo qui sotto, rispondi correttamente alla successiva domanda"*. We

[3]https://huggingface.co/antoniocappiello/bert-base-italian-uncased-squad-it
[4]https://huggingface.co/ARTeLab

Table 2
Results on formality style transfer (XFORMAL IT) for the formal-to-informal (F → I) and informal-to-formal (I → F) directions. Competitors' scores reported from Sarti and Nissim (2022).

	XFORMAL (IT) F → I				XFORMAL (IT) I → F			
	R1	R2	RL	BS	R1	R2	RL	BS
mT5 Small	.651	.450	.631	.666	.638	.446	.620	.684
mT5 Base	.653	.449	.632	.667	.661	.471	.642	.712
IT5 Small	.650	.450	.631	.663	.646	.451	.628	.702
IT5 Base	.652	.446	.632	.665	.583	.403	.561	.641
IT5 Large	.611	.409	.586	.613	.663	.477	.645	.714
Camoscio-7b **(0-shot)**	.645	.436	.623	.651	.622	.428	.600	.667

evaluated the generated answers using the script from Sarti and Nissim [22]. Furthermore, we also used an additional metric "ChatGPT Exact Match" to better assess the performance. We explain this metric in the following subsection "Evaluation Metrics".

Formality Style Transfer. We assess the style transfer capabilities of Camoscio using the Italian subset of the XFORMAL dataset [43], hereafter referred to as XFORMAL-IT. The dataset consists of forum messages from the GYAFC corpus [44] automatically translated covering several topics (entertainment, music, family, and relationships). The test set is constructed by using crowdworkers via Amazon Mechanical Turk to collect formal-informal pairs directly in Italian. The model is evaluated in both style transfer directions (Formal to Informal and Informal to Formal). We use only the test split for the zero-shot evaluation and ask the model to generate an answer given the `instruction` *"Dato il seguente testo scritto in modo formale, riscrivilo in modo informale."* and vice versa according to the style transfer direction.

4.1. Evaluation Metrics

We use the same evaluation protocol and scripts of Sarti and Nissim [22]. Specifically, for evaluating lexical matches, we rely on the language-independent ROUGE metric proposed by Lin [45] in the variants unigram (R1), bigram (R2), and Longest Common Subsequence (RL). To gauge semantic correspondence, we employ the trained BERTScore metric [46] with a widely used BERT model pre-trained on Italian[5] and the same baseline scores as Sarti and Nissim [22]. Following previous works, for evaluating the Question-Answering task we employ exact-match (EM) and F1-score (F1). However, since Camoscio is not trained on the output distribution of the question-answering dataset, these metrics will fail to assess the correctness of the output since the EM will count as zero even with a correct output but different wording. To account for these variations, we used an approach similar to Zheng et al. [47] that leverages an external LM (in our case *gpt-3.5-turbo*) to judge whether the answer provided by a model is correct (1) or not (0) given the question and the ground-truth answer. We refer to this metric as Exact Match via ChatGPT (EM-GPT) and explain it with additional details in Appendix B.

[5] dbmdz/bert-base-italian-xxl-uncased

4.2. Results and Discussion

Question Answering. Table 1 shows the results of Camoscio compared to other methods used in the literature. We observe that the metrics commonly used for the task (Exact Match and F1) are very low compared to all the other models. Although this is generally expected since we are comparing trained models with an untrained one, the exact match score is suspiciously low. Looking at the output responses, we noted that Camoscio produces correct but wordy answers (e.g., *"La crisi petrolifera del 1973 è iniziata nell'ottobre 1973."* instead of *"ottobre 1973"*) making the system to perform bad on this score despite the fact that it produces correct answers. Since all the other systems are trained on the datasets, they are aligned with the expected target distribution and the exact match metric is an effective choice. Nevertheless, when it comes to the zero-shot configuration in Camoscio, this conventional metric fails to accurately capture the true performance of the task.

To this end, we evaluated the model also with standard evaluation metrics for generative models (R1, R2, RL, BS). However, we also observe in this case low scores despite the fact that a qualitative examination of the provided answers suggests an overall higher quality. This is possibly due to the different lengths between the produced answers (long) and the ground truth (short) and reinforces the necessity of developing a more precise metric to accurately gauge task performance.

For this purpose, we used instead the metric *Exact*

Table 3
Results on NewSum-IT

	NewsSum-IT			
	R1	R2	RL	BS
mBART Large [6,7]	.377	.194	.291	-
mT5 Small	.323	.150	.248	.375
mT5 Base	.340	.161	.262	.393
IT5 Small	.330	.155	.258	.386
IT5 Base	.339	.160	.263	.044
IT5 Large	.251	.101	.195	.315
Camoscio-7b **(0-shot)**	.250	.104	.174	.190

Match via ChatGPT explained in §4.1. This metric shows that the actual zero-shot performance of Camoscio is in line with the other trained models (.576) and it is also way higher compared to the original EM metric (.077), confirming the need for another type of metric to evaluate the task in the zero-shot setting. Results also show that the EM-GPT metric of trained models correlates well with the existing EM metric, even though with a little marginal difference. This suggests that this metric could serve as an approximate estimation of the model's actual performance, although it might be subject to bias according to the model used for estimation.

Style Transfer & Summarization. Tables 2 and 3 show results respectively for the formality style transfer and news summarization task. We can observe that the zero-shot performance of Camoscio in both tasks is competitive with trained models. According to the model and training dataset, these latter might achieve slightly better scores at the expense of a less generalist model. Looking at the qualitative results, we note however that the summarization task on "Il Post" and "Fanpage" is affected by some common failure cases.

Failure Cases. The most common failure case consists of the model not producing an answer at all after the input prompt (4.93% of cases on "Il Post" and 21.16% cases on "Fanpage"). We think that it might be due to the input document of these examples being too long and out of distribution compared to the training documents seen in the instruction-tuning dataset (max length 256 tokens). This might confuse the model and trigger the generation of the end-of-sentence token. Other failure cases include the model responding with a copy of the instruction prompt. For example, we found that in 0.75% of the cases in the "Il Post" split of NewSum-IT, when asked to summarize the article the model responded with *"Questo articolo non è più commentabile. Abbonati al Post per commentare le altre notizie."*. We found this behavior peculiar, considering that the input prompt never mentions that articles were collected from "Il Post". Looking carefully at the dataset, we found that these documents were not correctly scraped from the web and the model was simply verbatim copying the input error message above (i.e., the message is already summarized and the model is proposing it as-is).

Open-ended generation. We show here some qualitative results on the task of open-ended generation. Compared to the other tasks above, in this task we provide the model with just the `instruction` without any specific `input`. As a tribute to GePpeTto [20], the first generative language model for Italian, we pose to the model the following `instruction`: *"Riassumi la storia di Pinocchio"*. The model answered: *"La storia di Pinocchio racconta di un bambino di nome Pinocchio che vive con i suoi genitori in un piccolo villaggio. Un giorno, Pinocchio incontra un burattino magico che gli dona la possibilità di diventare un vero bambino. Pinocchio inizia una serie di avventure mentre cerca di diventare un vero bambino. Durante il suo viaggio, Pinocchio fa amicizia con molti personaggi e impara molte lezioni di vita. Alla fine, Pinocchio raggiunge il suo obiettivo e diventa un vero bambino"*. At first glance the answer might seem correct, it is grammatically accurate and fluent in Italian. However, knowing the story, the model makes several factual errors like characterizing Pinocchio as a kid (instead of a wooden puppet) living with parents (instead of Geppetto) and introducing a nonexistent magic puppet. Overall this example highlights that, although promising, these models have well-known limitations like hallucinations, factual errors, and several kinds of biases [48, 49, 50]. Consequently, it is essential to exercise caution when utilizing them, keeping these limitations in mind.

5. Conclusion

In this paper, we introduced Camoscio, a 7 billion instruction-tuned model for the Italian language, together with its Italian instruction-tuning dataset. Results show that the zero-shot performance of Camoscio on several downstream tasks in Italian is competitive with existing models specifically finetuned for those tasks. Despite the known limitations of these kinds of models, this is a first step towards a generalist model capable of performing a wide range of tasks in Italian without explicit finetuning. This is particularly relevant especially in several domains where data is scarce or not available (e.g., medical domain). In an effort to democratize the available and open resources for the Italian language, we release all the artifacts (code, dataset, model) to the community.

[6] https://huggingface.co/ARTeLab/mbart-summarization-ilpost
[7] https://huggingface.co/ARTeLab/mbart-summarization-fanpage

6. Limitations

Results shown in the paper highlight zero-shot performance competitive with existing finetuned models on three different tasks: summarization (NewsSum-IT), question answering (SQuAD-IT), and style transfer (XFORMAL IT). However, it is unclear whether this is true also for other tasks, especially those out of training distribution of the instruction-tuning dataset (see Figure 1). Evaluating and thoroughly assessing the performance of these kinds of models is still an open research question. In addition to this, as already mentioned, the model suffers from common problems that affect language models such as hallucinations, factual errors, and several kinds of biases.

Acknowledgments

We thank Danilo Croce for pointing out existing implementation issues with the tokenization and the training objective in the *alpaca-lora* repository and Gabriele Sarti for sharing datasets and evaluation protocols used in IT5. We thank all the anonymous reviewers at *CLiC-it 2023* for the feedback provided.

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[2] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al., Palm: Scaling language modeling with pathways, arXiv preprint arXiv:2204.02311 (2022).

[3] OpenAI, Gpt-4 technical report, 2023. arXiv:2303.08774.

[4] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, et al., Opt: Open pre-trained transformer language models, arXiv preprint arXiv:2205.01068 (2022).

[5] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, et al., Bloom: A 176b-parameter open-access multilingual language model, arXiv preprint arXiv:2211.05100 (2022).

[6] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).

[7] V. Sanh, A. Webson, C. Raffel, S. Bach, L. Sutawika, Z. Alyafeai, A. Chaffin, A. Stiegler, A. Raja, M. Dey, M. S. Bari, C. Xu, U. Thakker, S. S. Sharma, E. Szczechla, T. Kim, G. Chhablani, N. Nayak, D. Datta, J. Chang, M. T.-J. Jiang, H. Wang, M. Manica, S. Shen, Z. X. Yong, H. Pandey, R. Bawden, T. Wang, T. Neeraj, J. Rozen, A. Sharma, A. Santilli, T. Fevry, J. A. Fries, R. Teehan, T. L. Scao, S. Biderman, L. Gao, T. Wolf, A. M. Rush, Multitask prompted training enables zero-shot task generalization, in: International Conference on Learning Representations, 2022. URL: https://openreview.net/forum?id=9Vrb9D0WI4.

[8] J. Wei, M. Bosma, V. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, Q. V. Le, Finetuned language models are zero-shot learners, in: International Conference on Learning Representations, 2021.

[9] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, et al., Scaling instruction-finetuned language models, arXiv preprint arXiv:2210.11416 (2022).

[10] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, T. B. Hashimoto, Stanford alpaca: An instruction-following llama model, https://github.com/tatsu-lab/stanford_alpaca, 2023.

[11] E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-rank adaptation of large language models, in: International Conference on Learning Representations, 2022. URL: https://openreview.net/forum?id=nZeVKeeFYf9.

[12] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al., On the opportunities and risks of foundation models, arXiv preprint arXiv:2108.07258 (2021).

[13] E. Postolache, G. Mariani, M. Mancusi, A. Santilli, L. Cosmo, E. Rodolà, Latent autoregressive source separation, Proceedings of the AAAI Conference on Artificial Intelligence 37 (2023) 9444–9452. URL: https://ojs.aaai.org/index.php/AAAI/article/view/26131. doi:10.1609/aaai.v37i8.26131.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2021. URL: https://openreview.net/forum?id=YicbFdNTTy.

[15] G. Trappolini, A. Santilli, E. Rodolà, A. Halevy, F. Silvestri, Multimodal neural databases, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Infor-

mation Retrieval, SIGIR '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 2619–2628. URL: https://doi.org/10.1145/3539618.3591930. doi:10.1145/3539618.3591930.

[16] A. Santilli, S. Severino, E. Postolache, V. Maiorca, M. Mancusi, R. Marin, E. Rodola, Accelerating transformer inference for translation via parallel decoding, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 12336–12355. URL: https://aclanthology.org/2023.acl-long.689.

[17] M. Polignano, P. Basile, M. De Gemmis, G. Semeraro, V. Basile, et al., Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets, in: CEUR Workshop Proceedings, volume 2481, CEUR, 2019, pp. 1–6.

[18] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[19] V. Basile, M. Lai, M. Sanguinetti, et al., Long-term social media data collection at the university of turin, in: Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), CEUR-WS, 2018, pp. 1–6.

[20] L. D. Mattei, M. Cafagna, F. Dell'Orletta, M. Nissim, M. Guerini, Geppetto carves italian into a language model, in: J. Monti, F. Dell'Orletta, F. Tamburini (Eds.), Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020, Bologna, Italy, March 1-3, 2021, volume 2769 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: https://ceur-ws.org/Vol-2769/paper_46.pdf.

[21] M. Baroni, S. Bernardini, A. Ferraresi, E. Zanchetta, The wacky wide web: a collection of very large linguistically processed web-crawled corpora, Language resources and evaluation 43 (2009) 209–226.

[22] G. Sarti, M. Nissim, It5: Large-scale text-to-text pretraining for italian language understanding and generation, arXiv preprint arXiv:2203.03759 (2022).

[23] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mT5: A massively multilingual pre-trained text-to-text transformer, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 483–498. URL: https://aclanthology.org/2021.naacl-main.41. doi:10.18653/v1/2021.naacl-main.41.

[24] M. La Quatra, L. Cagliero, Bart-it: An efficient sequence-to-sequence model for italian text summarization, Future Internet 15 (2023). URL: https://www.mdpi.com/1999-5903/15/1/15. doi:10.3390/fi15010015.

[25] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7871–7880. URL: https://aclanthology.org/2020.acl-main.703. doi:10.18653/v1/2020.acl-main.703.

[26] A. Bacciu, G. Trappolini, A. Santilli, E. Rodolà, F. Silvestri, Fauno: The italian large language model that will leave you senza parole!, in: F. M. Nardini, N. Tonellotto, G. Faggioli, A. Ferrara (Eds.), Proceedings of the 13th Italian Information Retrieval Workshop (IIR 2023), Pisa, Italy, June 8-9, 2023, volume 3448 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 9–17. URL: https://ceur-ws.org/Vol-3448/paper-24.pdf.

[27] C. Xu, D. Guo, N. Duan, J. McAuley, Baize: An open-source chat model with parameter-efficient tuning on self-chat data, 2023. arXiv:2304.01196.

[28] Michael, Stambecco: Italian instruction-following llama model, https://github.com/mchl-labs/stambecco, 2023.

[29] B. Peng, C. Li, P. He, M. Galley, J. Gao, Instruction tuning with gpt-4, arXiv preprint arXiv:2304.03277 (2023).

[30] E. J. Wang, Alpaca-lora, https://github.com/tloen/alpaca-lora, 2023.

[31] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, H. Hajishirzi, Self-instruct: Aligning language models with self-generated instructions, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 13484–13508. URL: https://aclanthology.org/2023.acl-long.754.

[32] D. Croce, A. Zelenanska, R. Basili, Neural learning for question answering in italian, in: AI* IA 2018–Advances in Artificial Intelligence: XVIIth International Conference of the Italian Association for Artificial Intelligence, Trento, Italy, November 20–23, 2018, Proceedings 17, Springer, 2018, pp. 389–402.

[33] A. Scaiella, D. Croce, R. Basili, Large scale datasets for image and video captioning in italian, Italian

Journal of Computational Linguistics 2 (2019) 49–60. URL: http://www.ai-lc.it/IJCoL/v5n2/IJCOL_5_2_3___scaiella_et_al.pdf.
[34] C. Larcher, M. Piau, P. Finardi, P. Gengo, P. Esposito, V. Caridá, Cabrita: closing the gap for foreign languages, 2023. arXiv:2308.11878.
[35] D. Croce, G. Brandi, R. Basili, Deep bidirectional transformers for italian question answering, in: R. Bernardi, R. Navigli, G. Semeraro (Eds.), Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13-15, 2019, volume 2481 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019. URL: https://ceur-ws.org/Vol-2481/paper25.pdf.
[36] A. Riabi, T. Scialom, R. Keraron, B. Sagot, D. Seddah, J. Staiano, Synthetic data augmentation for zero-shot cross-lingual question answering, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 7016–7030.
[37] A. Holtzman, J. Buys, L. Du, M. Forbes, Y. Choi, The curious case of neural text degeneration, in: International Conference on Learning Representations, 2020. URL: https://openreview.net/forum?id=rygGQyrFvH.
[38] V. Basile, L. Bioglio, A. Bosca, C. Bosco, V. Patti, UINAUIL: A unified benchmark for Italian natural language understanding, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 348–356. URL: https://aclanthology.org/2023.acl-demo.33. doi:10.18653/v1/2023.acl-demo.33.
[39] S. Casola, A. Lavelli, WITS: wikipedia for italian text summarization, in: E. Fersini, M. Passarotti, V. Patti (Eds.), Proceedings of the Eighth Italian Conference on Computational Linguistics, CLiC-it 2021, Milan, Italy, January 26-28, 2022, volume 3033 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021. URL: https://ceur-ws.org/Vol-3033/paper65.pdf.
[40] L. De Mattei, M. Cafagna, A. AI, F. Dell'Orletta, M. Nissim, A. Gatt, Change-it@ evalita 2020: Change headlines, adapt news, generate, Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020) 2765 (2020).
[41] Q. Lhoest, A. Villanova del Moral, Y. Jernite, A. Thakur, P. von Platen, S. Patil, J. Chaumond, M. Drame, J. Plu, L. Tunstall, J. Davison, M. Šaško, G. Chhablani, B. Malik, S. Brandeis, T. Le Scao, V. Sanh, C. Xu, N. Patry, A. McMillan-Major, P. Schmid, S. Gugger, C. Delangue, T. Matussière, L. Debut, S. Bekman, P. Cistac, T. Goehringer, V. Mustar, F. Lagunas, A. Rush, T. Wolf, Datasets: A community library for natural language processing, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 175–184. URL: https://aclanthology.org/2021.emnlp-demo.21. arXiv:2109.02846.
[42] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, Squad: 100,000+ questions for machine comprehension of text, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 2383–2392.
[43] E. Briakou, D. Lu, K. Zhang, J. Tetreault, Olá, bonjour, salve! xformal: A benchmark for multilingual formality style transfer, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 3199–3216.
[44] S. Rao, J. Tetreault, Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018, pp. 129–140.
[45] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013.
[46] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, in: International Conference on Learning Representations, 2019.
[47] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, I. Stoica, Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. arXiv:2306.05685.
[48] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, ACM Comput. Surv. 55 (2023). URL: https://doi.org/10.1145/3571730. doi:10.1145/3571730.
[49] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big?, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 610–623. URL: https://doi.org/10.1145/3442188.3445922. doi:10.1145/3442188.3445922.
[50] E. Sheng, K.-W. Chang, P. Natarajan, N. Peng, The woman worked as a babysitter: On biases in lan-

guage generation, 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), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3407–3412. URL: https://aclanthology.org/D19-1339. doi:10.18653/v1/D19-1339.
[51] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Transformers: State-of-the-Art Natural Language Processing, Association for Computational Linguistics, 2020, pp. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6.
[52] S. Mangrulkar, S. Gugger, L. Debut, Y. Belkada, S. Paul, Peft: State-of-the-art parameter-efficient fine-tuning methods, https://github.com/huggingface/peft, 2022.
[53] T. Dettmers, M. Lewis, Y. Belkada, L. Zettlemoyer, Llm.int8(): 8-bit matrix multiplication for transformers at scale, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems, volume 35, Curran Associates, Inc., 2022, pp. 30318–30332.
[54] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: International Conference on Learning Representations, 2018.

A. Implementation Details

The model was trained with the LoRA Parameter-efficient Finetuning technique [11], using the Hugging Face Transformers, PEFT, Datasets libraries [51, 52, 41] and the library Alpaca-LoRA [30]. Specifically, it was trained for 3 epochs with int8 quantization [53] on a standard desktop GPU Nvidia 3090 on a machine with Ubuntu 20.04.4 LTS, AMD Ryzen 9 3900X 12-Core Processor and 32GB of RAM. The model was trained with batches of dimension 4 and gradient accumulation to obtain a final "virtual batch" of 128. The maximum length used for training is 256 tokens. The learning rate is set to 3×10^{-4} with AdamW [54] and a total of 100 warmup steps are performed. We used a *lora_r* (i.e., the dimensionality of the low-rank update of the matrices) equals to 8, *lora_alpha* equals to 16 and *lora_dropout* equals to 0.05. We used LoRA adapters just for the matrices *Query* and *Value* in all the attention layers in the LLaMA model, following the original LoRA paper. We used the LLaMA 7 billion checkpoint by loading it from the Hugging Face Hub repository *"decapoda-research/llama-7b-hf"*.

B. Exact Match via ChatGPT

Exact Match via ChatGPT is a metric we introduced to evaluate the performance of Camoscio in the zero-shot setting on the question-answering task. This metric assesses whether the answer provided by a model is correct or not, compared to a ground-truth answer, without the need to have an exact string match (Exact Match). Specifically, we used an external LM (in our case *gpt-3.5-turbo*) that acts as a judge with the scope of verifying the correctness of the answer. We used a prompt similar to the following to compute this metric[8]:

"Given the context below and the corresponding question, please indicate whether the answer is correct (1) or not (0). Use a dict format in the response.

Context: `{Context}`
Question: `{Question}`
Correct gold answer: `{Correct_answer}`
Answer: `{Answer}`*"*

This prompt returns 1 if the answer is correct or zero otherwise. These values are then used to compute the *Exact Match via ChatGPT* with the same procedure of the Exact Match score.

C. Generation Examples

We propose here some generation examples. We also provide in the GitHub repository the generated text for all the examples in the test datasets, together with other open-ended generation examples.

C.1. NewSum-IT ("Il Post")

Input text:
L'esame di maturità è sempre un momento importante, quando lo si vive, anche se poi forse a distanza di anni lo si ricorderà con distacco e pensando che non era poi così difficile, o memorabile. Per molti continuerà a farsi vivo di notte, nei sogni. Quello di quest'anno, che inizia oggi, di certo sarà un po' diverso: senza tema e dizionari, senza attese di gruppo fuori dalla scuola prima di iniziare, ma con mascherine e gel igienizzanti. Agli studenti della Scuola Holden, la scuola di scrittura di Torino fondata da Alessandro Baricco, è stato chiesto di scrivere una lettera ai maturandi del 2020, per incoraggiarli e dare loro sostegno. Quella scritta da Nausica Anerdi, che ha frequentato il primo anno di Academy, il corso di laurea triennale in scrittura della Holden, è stata pubblicata oggi sul Post e su Repubblica, La Stampa e sulla pagina di Torino del Corriere della Sera. Tutte le altre potete

[8] Evaluation script available here

leggerle qui.

Camoscio generated text:
Nausica Anerdi, che ha frequentato il primo anno di Academy, il corso di laurea triennale in scrittura della Holden, ha scritto una lettera ai maturandi del 2020, per incoraggiarli e dare loro sostegno.

mT5-base generated text:
La lettera ai maturandi del 2020. È stata scritta da Nausica Anerdi, che ha frequentato il primo anno di Academy, la scuola di scrittura di Torino fondata da Alessandro Baricco.

Ground Truth:
Una lettera a chi sta per fare la maturità. Da parte di una studentessa della Scuola Holden, del corso di laurea triennale in scrittura.

C.2. NewSum-IT ("Fanpage.it")

Input text:
Continuano ad aumentare i nuovi casi di coronavirus nel nostro Paese. Sono stati 2.800 i contagi registrati ieri: numeri che preoccupano il governo e che ricordano quelli delle fasi più critiche dell'emergenza. Domani l'esecutivo si riunirà e valuterà se sia il caso di rendere più severe le norme anti-contagio attualmente in vigore. Entro la prossima settimana si attende il nuovo Dpcm contenente le misure di contrasto all'epidemia, mentre si valuta la proroga dello stato di emergenza fino al prossimo 31 gennaio 2021. Ma vediamo quindi quali sono queste nuove regole che il governo sta pensando di introdurre per frenare la curva dei contagi. L'obbligo di portare la mascherina all'aperto, già introdotto nei giorni scorsi in alcune zone, sarà esteso a tutto il territorio nazionale. Oltre quindi a confermare la necessità di indossare sempre il dispositivo di protezione nei luoghi chiusi, di igienizzare frequentemente le mani e di rispettare le distanze di sicurezza e il divieto di assembramento, il governo studia se rendere alcune misure più stringenti. In particolare, saranno potenziati i controlli nei luoghi della movida o dove è più facile che si vadano a costituire affollamenti. Le operazioni di vigilanza saranno affidate anche ai militari impegnati nel progetto "Strade secure". Il ministro della Salute, Roberto Speranza, si sarebbe detto favorevole all'estensione dell'obbligo di mascherina all'aperto a tutto il Paese, mantenendosi per quella linea di prudenza sostenuta fin dallo scoppio dell'emergenza. La stessa considerazione, d'altronde, è condivisa anche dalla autorità locali, dal momento che diversi governatori hanno già emesso delle ordinanze che impongono di coprire naso e bocca all'aperto, con multe salate per chi non rispetta le nuove regole. Il governo continua ad escludere la possibilità di tornare a una situazione di lockdown generale. Ma se i contagi continuassero a crescere a questi ritmi, potrebbero essere imposte alcune limitazioni in specifiche zone. In questo senso si sta anche considerando di introdurre un chiusura per bar e locali alle 22 o alle 23: una chiusura anticipata affinché non si ripetano situazioni di movida, come abbiamo visto anche quest'estate, in cui si favorisce la diffusione dell'infezione. All'aperto non sarà inoltre consentito riunire più di mille persone alla volta (ovviamente sempre secondo le regole del distanziamento): questa regola sarà valida sia per le manifestazioni che per gli eventi sportivi. Anche il limite per gli stadi resta a mille spettatori. Negli spazi chiusi, invece, il tetto rimane a 200 persone: questo vale sia per cinema, teatri e sale da concerto, ma anche per le celebrazioni private come i matrimoni. Le discoteche resteranno sicuramente chiuse, nei ristoranti si dovrà continuare a rispettare il distanziamento e anche nei negozi gli ingressi saranno contingentati.

Camoscio generated text:
Il governo sta valutando l'estensione dell'obbligo di mascherina all'aperto a tutto il Paese.

mT5-base generated text:
Il governo sta pensando di introdurre nuove regole anti-contagio per frenare la curva dei contagi. Il provvedimento è stato esteso a tutto il territorio nazionale, mentre si valuta la proroga dello stato di emergenza fino al prossimo 31 gennaio 2021. Vediamo quindi quali sono le nuove regole che il governo sta pensando di introdurre per frenare la curva dei contagi.

Ground Truth:
Con i nuovi casi di coronavirus in continuo aumento, il governo pensa a inasprire le regole anti-contagio per contrastare la diffusione dell'infezione. Si va verso l'obbligo di indossare la mascherina anche all'aperto e la chiusura a bar e ristoranti alle 22 o alle 23. Anche l'esercito sarà impiegato nei controlli.

C.3. XFORMAL-IT (F→I)

Input text:
Buongiorno, quella su E-Bay è molto bella.

Camoscio generated text:
Ciao, quella su E-Bay è davvero bella!

mT5-base generated text:
Buongiorno, quella su E-Bay è molto bella.

Ground Truth:
!ciaoooooooquella su e-bay è troppo carina!!

C.4. XFORMAL-IT (I→F)

Input text:
no mai , però ora per recuperare vado subito a lavare i piatti!

Camoscio generated text:
È ora per lavare i piatti.

mT5-base generated text:
No, ma ora per recuperare, vado subito a lavare i piatti.

Ground Truth:
Giammai! Ora però per recuperare vado subito a lavare i piatti

---

C.5. SQuAD-IT

Input text:
La crisi petrolifera del 1973 iniziò nell' ottobre 1973 quando i membri dell' Organizzazione dei Paesi esportatori di petrolio arabo (OAPEC, composta dai membri arabi dell' OPEC più Egitto e Siria) proclamarono un embargo petrolifero. Alla fine dell' embargo, nel marzo 1974, il prezzo del petrolio era salito da 3 dollari al barile a quasi 12 dollari a livello mondiale; i prezzi americani erano notevolmente più elevati. L' embargo ha causato una crisi petrolifera, o "shock", con molti effetti a breve e lungo termine sulla politica globale e sull' economia globale. Più tardi fu chiamato il "primo shock petrolifero", seguito dalla crisi petrolifera del 1979, definita il "secondo shock petrolifero". Domanda: Quando è stata la seconda crisi petrolifera?

Camoscio generated text:
La seconda crisi petrolifera è stata nel 1979.

mT5-base generated text:
1979

Ground Truth:
1979

Automatic Generation of Common Procurement Vocabulary Codes

Lucia Siciliani[1], Emanuele Tanzi[1], Pierpaolo Basile[1] and Pasquale Lops[1]

[1]University of Bari Aldo Moro, Department of Computer Science, via E. Orabona, 70125, Bari, Italy

Abstract

The role of tenders as means of investment of public funds and as vehicles of strategic development is nowadays crucial. For this reason, developing and enabling new solutions for e-procurement procedures can help to manage and invest funds. In e-procurement, the Common Procurement Vocabulary (CPV) allows assigning a code that classifies its subject to each tender. This study addresses the challenge of automatically assigning a CPV code to a tender. We tackle this problem in two different ways: as a classification problem and as a generative task. To develop and test our models, we build a dataset of 5M Italian tenders extracting them from the National Anti-Corruption Authority (Autorità nazionale anticorruzione - ANAC) website. Results show that text classifier approaches exhibit superior performance in this regard. However, they also reveal the potential of generative models in overcoming the limitations of existing classification methods for CPV code assignment in tender classification, providing valuable insights for improving procurement processes and enhancing efficiency in public sector operations.

Keywords

Natural Language Processing, e-procurement, e-tendering, Text Classification

1. Introduction

Knowledge organization systems (KOS), such as thesauri, gazetteers, lexical databases, ontologies, and classification systems, are used by institutions to organize large data collections, e.g. documents, web pages, and texts. Using a standard format guarantees semantic interoperability and allows for a faster exchange of information. Public procurement represents a field where adopting such systems can bring many advantages. On one hand, citizens can access data more easily, enabling more straightforward communication with institutions, which can help streamline many bureaucratic processes. Concurrently, adopting KOS systems in procurement has profound implications for professionals working within public administrations. In fact, these systems can serve as invaluable tools, offering support in the day-to-day activities of public sector employees. Integrating advanced technologies will facilitate a paradigm shift towards higher productivity and efficiency. Tasks that were once labor-intensive and time-consuming can now be executed with greater precision and speed, allowing public administrators to focus on more strategic and value-driven aspects of their roles. For this reason, in the field of public procurement, the European Union developed a Common Procurement Vocabulary (CPV)[1] that identifies the subject of a tender. The adoption of the CPV also allows companies to find new public contracts easily, thus fostering competitiveness.

The CPV is structured as a tree of codes comprising 9 digits, eight plus a check digit, and specifies whether the tender in question refers to supplies, works or services covered by the contract. Each digit indicates progressively finer-grained classifications. More specifically, each CPV is composed as follows:

- the first two digits identify the divisions (e.g. 71000000-8 Servizi architettonici, di costruzione, ingegneria e ispezione (Architectural, construction, engineering and inspection services));
- the first three digits identify the groups (e.g. 71300000-1 Servizi di ingegneria (Engineering services));
- the first four digits identify the classes (e.g. 71310000-4 Servizi di consulenza ingegneristica e di costruzione (Consultative engineering and construction services));
- the first five digits identify the categories (e.g. 71311000-1 Servizi di consulenza in ingegneria civile (Civil engineering consultancy services));
- each of the last three digits provides an additional degree of precision within each category (e.g. 71311210-6 Servizi di consulenza stradale (Highways consultancy services));
- a ninth digit serves to verify the previous digits.

Examples of CPV codes are: *30200000-1 (Computer equipment and supplies)*, *30230000-0 (Computer hardware)*, and

[1]https://simap.ted.europa.eu/it/web/simap/cpv

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

lucia.siciliani@uniba.it (L. Siciliani); e.tanzi2@studenti.uniba.it (E. Tanzi); pierpaolo.basile@uniba.it (P. Basile); pasquale.lops@uniba.it (P. Lops)

0000-0002-1438-280X (L. Siciliani); 0000-0002-0545-1105 (P. Basile); 0000-0002-6866-9451 (P. Lops)

S*30231000-7 (Computers and printers)*.

The supplementary vocabulary can be used to complete the description of the subject of a contract. The items consist of an alphanumeric code corresponding to a denomination that allows you to provide further details on the specific nature or destination of the asset to be purchased. The alphanumeric code is structured as follows:

- a first level, consisting of a letter corresponding to a section (e.g. A Materiali (Materials));
- a second level, consisting of a letter corresponding to a group (e.g. AA Metalli e leghe (Metal and alloy));
- a third level, consisting of two digits corresponding to the attribute (AA02-4 Alluminio (Aluminium));
- the last digit is used to verify the previous ones.

Examples of supplementary codes are the following: *AA01-1 Metal*, or *UB05-6 Office items*.

The main vocabulary comprises 9,454 terms and, more specifically, 45 divisions, 272 groups, 1,002 classes, 2,379 categories, and 5,756 sub-categories. Assigning a CPV to a tender is a task which is accomplished by RUPs (Responsabile Unico del Procedimento, i.e. Tender's Managers), however, given the high number of terms, it is really difficult even for human experts to identify the right CPV to use. For this reason, despite assuring a fine-grained classification, the high number of labels frequently leads to errors in the CPV assignment like typos or wrong interpretation of the description of each code. Another phenomenon is represented by the skewed usage of the codes as there is a small number of CPVs which are more known and thus used more frequently while a large number of CPVs are underused. Given these premises, in this work, we propose a method for automatically classifying tenders to their CPV codes.

The paper is organized as follows: Section 2 provides an overview of the approaches available at the state of the art, Section 3 contains the details of the proposed solution, Section 4 reports the results obtained by the evaluation of our model, and finally Section 5 closes the paper.

2. Related Work

The assignment of a Common Procurement Vocabulary (CPV) code to a tender is crucial for the accurate identification and precise retrieval of analogous documents. This meticulous categorization process is indispensable for ensuring the streamlined organization and effective utilization of information. Given the nature of the CPV vocabulary, which encompasses a set of over nine thousand terms, this task assumes a challenging dimension, demanding a level of expertise that even seasoned professionals in the field find daunting. Already existing approaches for CPV classification have been condensed within the last few years. In [1], the author compared different deep-learning models for single-label and multi-label CPV classification. The best-performing model was represented by GRU [2] with attention mechanism [3]. The dataset used in this work comprises 30,000 Swedish tenders provided by e-Avrop[2], a Swedish company that manages an e-procurement platform. Each document in the dataset comprises titles and descriptions of each tender with their respective categories.

In [4], the authors used a dataset of 40,000 tenders extracted from TED. The authors have used an LSTM [5] architecture for sequence prediction and classification. They also used a Support Vector Machine (SVM) to classify the CPV main code category within the same framework. The PhD thesis by [6] addressed the CPV classification problem using a Linear SVM and a bag of words representation from a random sample of 200,000 documents extracted from the TED. An important aspect to notice is that this work focuses on both English and French. Kaan Görgün (Mkaan)[3] proposed a multilingual approach which is a fine-tuned version of mBERT [7] on tenders extracted from the TED. With their work, [8] are instead focused on the Spanish language. The proposed method uses RoBERTa-base-bne [9], a RoBERTa model pre-trained on Spanish documents. The authors fine-tune this model on Spanish Public Procurement documents, classifying the 45 CPV divisions. Next, they compare several models, ranging from more classical ones (e.g. Naive-Bayes, SVM, KNN, etc.) to the one proposed by MKaan.

Data from ANAC are a valuable resource for building data-driven systems in the public administration domain. In [10], authors propose an information extraction framework for Italian tenders [10] that leverage ANAC datasets and other information sources. Moreover, a decision support system that helps users during the entire course of investments and contracts in e-procurement is described in [11].

3. Methodology

The main idea is to support RUPs in assigning CPV to a new tender. Specifically, the aim is to establish a robust system capable of proficiently classifying a tender based on its specified object, thereby facilitating an accurate alignment with the comprehensive set of CPV codes available. The classification task is difficult since the number of codes (CPVs) is high. Therefore, we propose two

[2]https://info.e-avrop.com

[3]https://huggingface.co/MKaan/multilingual-cpv-sector-classifier

methodologies: 1) a text classification approach based on different classifiers; 2) a generative approach based on Transformers with an encoder-decoder architecture.

Both approaches work on the same data. In particular, given a list of tuples *(CPV, CPV description, tender object)*, we split it into three sets: training, validation and testing. More details on the dataset are reported in Section 4. Then each approach is implemented, trained and validated separately on the same data.

3.1. Text Classification

Regarding text classification, our approach aligns with a classical pipeline:

- **Preprocessing**: the initial step involves the preprocessing of the tender object, wherein transformations such as lowercase conversion and tokenization are applied. These essential preprocessing techniques lay the groundwork for subsequent stages by standardizing the textual data;
- **Feature Vector Generation**: following the preprocessing step, we construct feature vectors for each tender object. This involves the utilization of both Bag-of-Words (BoW) and TF-IDF (Term Frequency-Inverse Document Frequency) approaches. By encoding the textual information into numerical representations, we aim to capture the salient features that contribute to the classification task;
- **Classifier Training and Tuning**: subsequently, a classifier is trained using the feature vectors generated in the previous step. The training process is complemented by the optimization of hyperparameters. This optimization is achieved with the use of a validation set, ensuring that the classifier retains generalization capabilities;
- **Evaluation**: the final stage of our classification pipeline involves the evaluation of the trained classifier on an independent test set. This allows an assessment of the model's ability to generalize and classify unseen tender objects. It serves as a critical benchmark to validate the effectiveness and robustness of the entire text classification system.

For the implementation, we rely on spaCy for text processing and scikit-learn for classification. The hyperparameters are found through the grid search. After a first evaluation, we select the following classifiers: Linear SVC and Multinomial Naive Bayes. Moreover, we cast the problem only to classify the divisions that are composed of 45 classes since the model cannot provide reasonable results when the whole set of CPVs is involved.

Moreover, we choose to investigate a classifier based on BERT using its tokenizer. Again, in this setting, we use only the divisions as labels to obtain reasonable results. This strategic decision is driven by recognising that achieving reasonable results across the entire spectrum of CPV codes poses significant challenges. By concentrating on this subset, we optimize the model's capacity to provide meaningful and accurate predictions within a more manageable scope.

3.2. Generative Approach

Since our approach aims to suggest a CPV given a tender's description, we decided to investigate the ability of AI generative methods to automatically produce a text given a textual input. Moreover, we want to test if a generative approach can provide better results when a large number of classes is involved, as in our domain. We adopt a classical encoder-decoder architecture that has proven to provide promising results in several NLP tasks. Encoder-decoder architectures are well-suited for solving sequence-to-sequence problems like machine translation, as they can effectively process variable-length input and output sequences. In this architecture, the encoder takes in a sequence of any length and converts it into a fixed-shaped state. On the other hand, the decoder maps the encoded state, which has a fixed shape, back to a sequence of variable length.

In our case, the encoder's input is the tender's object description and the decoder output is the CPV code and its description. It is important to underline that this method can produce a CPV code or a description that is not present in the original list of CPVs, while a text classifier produces as output a CPV from the set of predefined CPVs. This poses problems in the evaluation phase as comparing the output produced with the gold standard present in the test set becomes more complex. More details about the evaluation are reported in Section 4.

4. Evaluation

For building and testing, we extract data from the ANAC anticorruzione (*anticorruption*) website[4]. ANAC -Autorità Nazionale AntiCorruzione *National Anti-Corruption Authority* is an Italian independent administrative authority with the aim of combating corruption in the country.

In particular, we retrieve for each tender the CIG (the tender identifier), the tender type, the description of the tender object, the CPV assigned by the RUP and the CPV description. The tender type identifies three kinds of tender: 1) supplying, 2) service and 3) work. From the original dataset, we remove CPV codes that occur less than 20 times and store data in a CSV file for a total of 5 million tenders. We split the dataset in training, test

[4]https://dati.anticorruzione.it/opendata/dataset/

and validation according to the percentages reported in Table 1.

	size	%
training	3,200,000	64%
test	1,000,000	20%
validation	800,000	16%

Table 1
Dataset statistics.

For training the encoder-decoder architectures, we build a different version of the dataset in the JSONL format. Each row in the dataset is a JSON object with two elements: *source* and *target*. The *source* is the input text of the encoder and the *target* is the output text of the decoder. In our case, the *source* is the concatenation of the type of the tender and the description of the tender's object, while the *target* is the concatenation of both the code and the description of the CPV. The tender's type defines the nature of the object from a list of predefined types: service, supply and work. Listings 1 shows an example of a JSON object related to a tender.

```
{"source":"lavori lavori di pavimentazione delle
    vie san martino e santa Maddalena",
"target":"45262321-7 - lavori di pavimentazione"
    }
```

Listing 1: An example of a JSON object for training the encoder-decoder architecture.

The dataset is stored on Zenodo[5], while the code is available on GitHub[6]. The code for fine-tuning the IT5 model is available here[7]. The IT5 models fine-tuned on the CPV generation task are on HuggingFace: the large model[8].

4.1. Text Classification

This sub-section reports information and results about text categorization approaches. Regarding the parameters' optimization, we adopt a grid search for Linear SVC and Multi-NB. For Linear SVC, we optimize the parameter C in the set {0.5, 1, 2, 4, 8}, while for Multi-NB we consider $alpha$ in {0.1, 0.3, 0.5, 0.7, 0.9} and fit_prior in {True, False}. The best values selected after the grid search are $C = 0.5$, $alpha = 0.3$ and $fit\_prior = False$.

For BERT, we did not perform parameters optimization since the required computational time is very high. We fine-tuned a specific language model for Italian called `dbmdz/bert-base-italian-uncased`[9] using the Adam optimizer with a learning rate of 5e-05 and a batch size of 16 trained for 5 epochs.

Results of text classification approaches are reported in Table 2. Generally, the results are very low due to the large number of classes and BERT reports the worst performance since, for some classes, there are very few examples in training data. We observe a large accuracy with respect to the F1 measure. This is due to the presence of few classes with many examples. For these classes, classifiers can achieve good performance. For example, BERT achieves the 90% of F1 for the most frequent class[10].

4.2. Generative Approach

We used the Java library Lucene for text searching and indexing, with the aim of solving the task as a retrieval task.

In detail, we indexed CPV code description pairs corresponding to *target* in the two fields *code* and *description*. Afterwards, we ran the search for each *source* element in JSON file, obtaining for each search the element belonging to target with the most similar description to the source string, and saved results in a file containing the triple *source*, *target* and *generated*. Where *target* is the expected description and *generated* is the retrieved one. We adopt the same output format for the generative approaches. The idea is to exploit the *source* as the query for the search engine and retrieve the most similar code descriptions using the search engine.

We executed this experiment four times, implementing variations in both the configuration of the text analyzer and the choice of two distinct similarity measures. The adopted configurations are the following:

- *StandardAnalyzer* + default similarity;
- *StandardAnalyzer* + *LMDirichletSimilarity*;
- *ItalianAnalyzer* + default similarity;
- *ItalianAnalyzer* + *LMDirichletSimilarity*.

The *ItalianAnalyzer* performs a specific stemming algorithm for Italian, while the *StandardAnalyzer* implements a grammar-based tokenizer for several languages. The default similarity provided by Lucene is the BM25 model [12], while the *LMDirichletSimilarity* uses a language model for information retrieval with the Bayesian smoothing based on Dirichlet priors [13]. The evaluation has been carried out on the test set. We decided to use BLEU metric to measure matching between generated and target text since this metric is based on the idea that the nearer the predicted text is to the target one, the more correct it is. Considering the small size of the compared strings, we decided only to use 1-gram and 2-gram of

[5]https://zenodo.org/records/10007545
[6]https://github.com/ematanzi/Valutazione-CPV/
[7]https://github.com/gsarti/it5
[8]https://huggingface.co/basilepp19/cpv-it5 and the base model https://huggingface.co/basilepp19/cpv-it5-base
[9]https://huggingface.co/dbmdz/bert-base-italian-uncased

[10]33000000-0 Apparecchiature mediche, prodotti farmaceutici e per la cura personale (*Medical equipment, pharmaceuticals and personal care products*)

classifier	accuracy	P	R	macro-F1
Linear SVC	0.7768	0.5420	0.3758	**0.4041**
Multi-NB	0.7282	0.4334	0.3126	0.3292
BERT	0.7411	0.3136	0.2948	0.2990

Table 2
Results of text classification.

consecutive words, attributing them to the same weight (0.5). In calculating the metric, a smoothing function has been used, increasing the score when there are partial matches between the generated text and the target text.

In the evaluation, since the resolution of this classification task with utmost precision is arduous even for human experts, we decided to consider every possible correspondence between the code-description generated couple and the target one, which are the following:

- the full correspondence;
- the correspondence between codes, but not the description (the opposite case can never occur);
- the correspondence between categories;
- the correspondence between classes;
- the correspondence between groups;
- the correspondence between divisions;
- the case of no match.

In this way, we also evaluate the cases in which the solution has been approached. For all of these cases, we calculated the number of times they occurred along with the relative average of the BLEU score metric. Eventually, we also calculated the average BLEU score related to all the tests performed.

The best results obtained by the baseline are reported in Table 3. The results are obtained using the *ItalianAnalyzer* and the default similarity. The baseline based on the search engine is able to correctly retrieve the correct CPV with the correct description for only 11.34% of testing data. In the 61.77% of cases is not able to retrieve the correct CPV with very low BLEU (0.0515), this means that the description of the first retrieved CPV is very different from the correct one.

	# matchs	%	BLEU
Perfect match	113,440	11.34	1.0
Only CPV code	872	0.09	.5128
Only category	30,502	3.05	.2452
Only class	53,232	5.32	.1756
Only group	90,673	9.07	.1040
Only division	93,574	9.36	.1565
No match	617,707	61.77	.0515
All	-	-	.1751

Table 3
Best results obtained by Lucene with the *ItalianAnalyzer* and the default similarity.

We evaluate the generative approach based on the encoder-decoder transformer by fine-tuning it on training data. We start from a pre-trained Italian model called IT5 [14]. The IT5 model family is the initial endeavour to pre-train extensive sequence-to-sequence transformer models specifically designed for the Italian language, inspired by the methodology employed in the original T5 model [15]. We fine-tuned two different models with different sizes: IT5-lager and IT-base.

Results are reported in Table 4 for the large model and in Table 5 for the base one.

	# matchs	%	BLEU
Perfect match	372,188	37.22	1.0
Only CPV code	2,403	0.24	.4665
Only category	76,735	7.67	.2573
Only class	115,614	11.56	.2260
Only group	109,190	10.91	.1360
Only division	110,555	11.06	.1680
No match	213,315	21.33	.0861
All	-	-	.4546

Table 4
Results with the IT5-large model.

	# matchs	%	BLEU
Perfect match	359,408	35.94	1.0
Only CPV code	2,210	0.22	.4854
Only category	73,885	7.39	.2437
Only class	114,487	11.45	.2199
Only group	114,430	11.44	.1235
Only division	111,538	11.15	.1635
No match	224,042	22.40	.0853
All	-	-	.4385

Table 5
Results obtained with the IT5-base model.

To compare generative approaches with text categorization ones, we consider from the generated output only the first two digits of the CPV, i.e. the CPV divisions. This choice allows us to compare the generative approach with the ones based on text categorization since the latter are trained to predict only the division of each tender. Results of this analysis are reported in Table 6 and show that if we consider generative approaches as a classifier, they are below the simple Linear SVC. Performing a dual evaluation in which a classifier is evaluated as a generative approach is not possible since we train classifiers only for predicting divisions. Considering only the code of the division is not possible to generate a description comparable to the text generated by an IT5 model.

model	acc.	P	R	macro-F1
Linear SVC	*0.7768*	*0.5420*	*0.3758*	*0.4041*
IT5-large	0.7867	0.4196	0.3575	0.3728
IT5-base	0.7760	0.4150	0.3525	0.3654

Table 6
Results of generative models evaluated as text classifier and compared with the best text categorization method.

Anyway, these outcomes are encouraging since the IT5 model is trained on all the possible descriptions of a CPV, while the text classifier approaches handle only the code of the division and cannot provide usable results when they are trained on the whole set of possible classes. Moreover, generative approaches provide a significant improvement with respect to the baselines obtained by a search engine.

5. Conclusions

In this paper, we tackle the challenge of categorizing a tender by aligning it with the comprehensive Common Procurement Vocabulary (CPV), i.e. a meticulously curated European lexicon of codes designed to precisely identify the subject matter of a tender. The complexity of this task lies in the diverse nature of procurement scenarios, where each tender has its own description and requirements. The CPV emerges as a fundamental tool in deciphering the procurement language, trying to define a European dictionary allowing interoperability among different countries. Our proposed methodologies encompass two distinctive approaches: the former relies on a conventional text classification paradigm, whereas the latter leverages a generative strategy hinging on the encoder-decoder architecture as conceptualized by the T5 model.

In our systematic exploration of the system's proficiency in discerning the accurate division of a tender, specifically on the initial two digits of the CPV, it becomes evident that text classifier approaches provide the best results. Nevertheless, a noteworthy result surfaces when we focus on the holistic identification of the entire CPV through a descriptive context. In this context, the generative approaches exhibit commendable efficacy, demonstrating promising outcomes. Notably, these generative techniques surpass established baselines constructed through conventional keyword-centric search engines, attesting to their heightened capabilities in nuanced comprehension and contextual inference.

Acknowledgments

We acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), Spoke 6 - Symbiotic AI (CUP H97G22000210007) under the NRRP MUR program funded by the NextGenerationEU.

We thank Salvatore Tucci for his helpful support in evaluating text classification approaches.

References

[1] A. Suta, Multilabel text classification of public procurements using deep learning intent detection, Degree project in mathematics (second cycle), Kth Royal Institute of Technology, 2019.

[2] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using rnn encoder-decoder for statistical machine translation, arXiv preprint arXiv:1406.1078 (2014).

[3] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473 (2014).

[4] S. Kayte, P. Schneider-Kamp, A mixed neural network and support vector machine model for tender creation in the european union ted database., in: KMIS, 2019, pp. 139–145.

[5] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (1997) 1735–1780.

[6] O. Ahmia, Assisted strategic monitoring on call for tender databases using natural language processing, text mining and deep learning, Ph.D. thesis, Université de Bretagne Sud, 2020.

[7] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[8] M. Navas-Loro, D. Garijo, O. Corcho, Multi-label text classification for public procurement in spanish, Procesamiento del Lenguaje Natural 69 (2022) 73–82.

[9] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, A. Gonzalez-Agirre, C. Armentano-Oller, C. Rodriguez-Penagos, M. Villegas, Spanish language models, arXiv preprint arXiv:2107.07253 (2021).

[10] L. Siciliani, E. Ghizzota, P. Basile, P. Lops, Oie4pa: open information extraction for the public administration, Journal of Intelligent Information Systems (2023). URL: https://link.springer.com/article/10.1007/s10844-023-00814-z. doi:https://doi.org/10.1007/s10844-023-00814-z.

[11] L. Siciliani, V. Taccardi, P. Basile, M. Di Ciano,

P. Lops, Ai-based decision support system for public procurement, Information Systems 119 (2023) 102284. URL: https://www.sciencedirect.com/science/article/pii/S0306437923001205. doi:https://doi.org/10.1016/j.is.2023.102284.

[12] S. Robertson, H. Zaragoza, The probabilistic relevance model: Bm25 and beyond, in: the 30th Annual International ACM SIGIR Conference, 2007, pp. 23–27.

[13] C. Zhai, J. Lafferty, A study of smoothing methods for language models applied to ad hoc information retrieval, in: ACM SIGIR Forum, volume 51, ACM New York, NY, USA, 2017, pp. 268–276.

[14] G. Sarti, M. Nissim, It5: Large-scale text-to-text pretraining for italian language understanding and generation, arXiv preprint arXiv:2203.03759 (2022).

[15] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, The Journal of Machine Learning Research 21 (2020) 5485–5551.

Inters8: A Corpus to Study Misogyny and Intersectionality on Twitter

Ivan Spada[1], Mirko Lai[1,2] and Viviana Patti[1]

[1]*Computer Science Department - University of Turin, Italy*
[2]*Aequa-tech srl, Turin, Italy*

Abstract
This paper presents our research on the detection of online misogyny on social media and its intersection with other hate categories. Focusing on the phenomenon of misogyny, we carried out a corpus-based data analysis around victims of online hate campaigns. Targets were selected to study how misogyny and sexism intersect with other categories of social hatred and discrimination such as xenophobia, racism, and Islamophobia. This study includes an event-driven analysis of hate on Twitter concerning specific targets, the process of developing the *Inters8* corpus, and its manual annotation according to a novel multi-level scheme designed to assess the presence of intersectional hatred.

Keywords
hate speech, automatic misogyny identification, intersectionality, annotated corpora, social media

Warning*: This paper contains examples of potentially offensive content.*

1. Introduction

The term *intersectionality* refers to the coexistence of multiple forms of social categorizations such as ethnicity, gender, sexual orientation, social class, disability, etc., which can lead to discrimination and generate obstacles in the daily lives of those affected [1]. In the specific case of misogyny, discrimination takes different shapes depending on the other co-existing forms of discrimination such as racism, classism, ableism, or homophobia [1]. Therefore, the phenomenon should not be studied in isolation.

The coexistence of different forms of discrimination suggests the study of social interactions and the intersectionality between multiple categories of hate. It is interesting to explore how language may vary when interactions involve people who are at the intersection of multiple discriminated social categories that are henceforth referred to as *dimensionalities.* In particular, [1] explained how intersectionality between multiple dimensionalities may generate a new discriminated category framed and treated differently.

The objective of this paper is to analyze how misogynistic hatred intersects with other dimensionalities and how this appears in a micro-blogging social platform such as Twitter[1]. The first contribution of this paper is an analysis of gender discrimination, inequality comparisons, presence of stereotypes, also oriented to shed light on how users interact with, support, and attack targets of misogynous hatred on social media. As a second contribution, we created *Inters8*, an Italian corpus containing a subset of TWITA [2] filtered by following a target and event selection process, for the purpose of studying intersectionality. Thereafter, a portion of the corpus related to Silvia Romano's liberation [3] was annotated (called *Inters8_SRomano*) so as to explore characteristics of intersectional hatred in a specific pilot study.

Outine. After a brief technical contextualization (Section 2), we describe the target-event driven analysis, the creation of the *Inters8* corpus and the annotation task (Section 3). Then, the novel annotation scheme applied to the data is described (Section 4) and a discussion of the outcome of the annotation process is presented (Section 5). Conclusions and future work end the paper.

2. Related work

Twitter, as well as other online social media platforms, employs algorithms for detecting content that violates its terms and conditions committed to making Twitter a safe place for users. As observed in [4], the relationship between ethnicity and gender can influence the recognition of false positive hateful content of African Americans, especially in the case of female users. Indeed, bias can permeate the system and reinforce AI discrimination.

Subjective phenomena that could be affected by social and cultural context such as misogyny, stereotype and racism, have to be approached recognizing individuals'

[1]https://twitter.com

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
ivan.spada@unito.it (I. Spada); mirko.lai@unito.it (M. Lai); viviana.patti@unito.it (V. Patti)
0009-0002-0459-1189 (I. Spada); 0000-0003-1042-0861 (M. Lai); 0000-0001-5991-370X (V. Patti)

perceptions that can greatly vary according to personal experiences and cultural backgrounds.

Several efforts in terms of automatic detection have been provided by scholars for countering hate speech [5, 6]. About misogyny, a first computational effort for the detection of misogyny in English tweets has been provided in [7], while [8] attempts to address the problem of measuring and mitigating unintended bias in machine learning models trained for misogyny detection. A first automatic Misogyny Identification (AMI) shared task has been organized within EVALITA and IberEval 2018 evaluation campaigns [8, 9] to detect misogyny in tweets in various languages (English, Spanish, and Italian). In particular, participants were specifically requested to identify if the message is misogynistic, and then to categorize the target (person or not) and the type of misogyny using the categories developed by [10]. Based on the availability of multilingual datasets targeting misogyny and other kind of abusive language, in [11] a multilingual and cross-domain study on misogyny identification in Twitter is proposed, where some insights on features of misogyny and on the interaction between misogyny and related phenomena are provided. In this work, we try to shed more light on misogyny and intersectionality with other co-existing forms of discrimination such as xenophobia, islamophobia, and stereotype proposing a new annotation scheme. In Section 4.1 we specifically analyze the contributions that inspired our work.

3. Methodology

In this section, we describe the methodological pipeline we designed in order to collect data to analyze intersectional hate and discrimination. The target- and event-oriented nature of hate speech in social media has been the object of recent studies [12, 13, 11, 14, 15]. Hate discourses may vary in relation to events and victims belonging to multiple dimensionalities. However, it is necessary to take into account that recognition of the discriminatory phenomenon may be partly subjective and influenced by the social and cultural context.

In order to contextualize and explore the phenomenon of intersectionality among multiple social categories subjected to hate and discrimination, we conducted an analysis of discourses concerning public people, known to Italian society, selected specifically for this task.

Our pipeline consists of a sequence of steps (Figure 1).

3.1. Discourse analysis regarding targets and events

The period chosen for analysis is the first half of 2020, the year when the COVID-19[2] pandemic began and the population forced to stay at home was making substantial use of social media [16]. The analysis process starts with prior knowledge of the Italian context obtained from consulting services that convey information such as newspapers, news broadcasts, and TV shows. It was noticed that some streams of discourse on Twitter were event-driven, these included: reporting news, inviting guests, and discussing known people on TV shows.

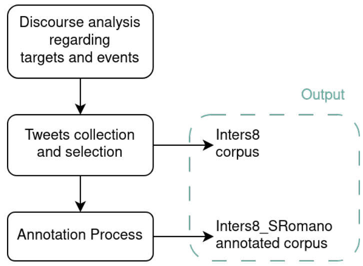


Figure 1: Pipeline for the creation and annotation of Inters8.

Focusing on misogyny (the explicit and implicit attitude of generic aversion to women), the additional social categories considered are inspired by those in Vox's Intolerance Map n.7[3]. The categories taken into account in this study are as follows: *misogyny*, *xenophobia*, *anti-semitism*, *Islamophobia*, *homolesbobitransphobia*, *political opinions* and *physical appearance*.

Through scraping news and TV shows available on RaiPlay[4] (the Italian television streaming service), we viewed episodes of 23 TV shows, qualitatively analyzed and selected 17 well-known people in Italy[5] who fall into multiple dimensionalities considered in this case study. The manual analysis of the language, expressed in TV programs and Twitter interactions concerning the selected targets, allowed the extraction of information framing the subjects: dimensionalities, topics, events, debates, hashtags, and the most common and narrow keywords. The following were annotated for each victim: Twitter username (if they joined the social network), characteristics potentially exposed to hate and discrimination, hashtags, time period analyzed, TV shows, and links to episodes where they were invited or talked about.

In addition, an NxM table[6] was compiled to make the target comparison visually easier, where N refers to the

[2]https://www.who.int/health-topics/coronavirus

[3]VOX, University of Milan, Sapienza University of Rome, Aldo Moro University of Bari and ItsTime (2022). The map of intolerance n.7. Available at: http://www.voxdiritti.it/la-nuova-mappa-dellintolleranza-7/

[4]https://www.raiplay.it/

[5]Giovanna Botteri, Carola Rachete, Cathy La Torre, Cécile Kyenge, Chiara Appendino, Diletta Leotta, Emma Bonino, Greta Thunberg, Ilaria Cucchi, Laura Boldrini, Liliana Segre, Michela Murgia, Rula Jebreal, Silvia Romano, Teresa Bellanova, Virginia Raggi and Vladimir Luxuria.

[6]https://github.com/ivsnp/inters8/tree/main/targets

people and M to the hate categories. Since misogyny was the focus of this study, all selected victims had the *misogyny* column marked. Other categories were marked when present.
Output: selection of related targets, events, and hashtags useful for the next phase.

3.2. Tweets collection and selection

Data Collection Given the targets and events obtained from the previous phase, Italian tweets regarding targets in the temporal surroundings of the detected events were extracted from TWITA [2]. These tweets contained at least one hashtag among those found during the first phase or both the first and last names of the corresponding target or aliases.

We collected the following metadata for each Twitter interaction: *tweetId*, *date*, *text*, *type* (tweet, retweet, quote, or reply).
Output: collection of tweets related to targets and events useful for the next phase.

Data Selection After obtaining the collection of tweets in output from the previous phase, we set out to create a corpus containing tweets related to events concerning targets that were potentially victims of intersectional hatred. This decision aimed to provide a set of Twitter interactions (tweets, replies, quotes, and retweets) to perform an analysis of the intersection of various dimensionalities in a target-event context.

We proceeded with target-event filtering in order to obtain a case study on which to start analyzing the phenomenon. The *Inters8* corpus was populated with the Twitter interactions collected by selecting the deliverance from captivity and homecoming of the target Silvia Romano[7] to Italy on May 9-10, 2020. This choice of target-event pair was made because there were more Twitter interactions compared with others [17][8]. Thus, it was intended to create a pilot on a specific case study. Given this target-event choice, the goal was to explore the intersectionality between the following dimensionalities: *misogyny*, *xenophobia*, and *Islamophobia*.

Inters8 contains 248240 interactions concerning the chosen target-event pair distributed during May 9-24, 2020. It consists of contents distributed per interaction type as follows: retweet 75%, tweets 18%, reply 4%, and quote 2%. The metadata is as follows: *tweetId* and *type* (tweet, retweet, quote, or reply).

[7] https://www.bbc.com/news/world-africa-52608614

[8] According to the Italian observatory "Map of Intolerance http://www.voxdiritti.it/la-nuova-mappa-dellintolleranza-5/ the peak of social attacks against Muslims occurred right around the time of Silvia Romano's liberation, a media shitstorm that prompted the Special Operations Group (Ros) to open up an investigation ad hoc into the matter.

Output: *Inters8* corpus consisting of the tweets collected regarding the selected target and event.

Sampling of Data to Annotate The process of creating the Inters8 subset followed these steps: (1) retweets removal, (2) similar tweets removal using *cosine similarity* by setting the threshold to 0.7, (3) the collection was filtered to include tweets with and without the Italian flag emoji with 50% proportion keeping the same distribution for days and hours. The latter decision was made so that the annotation could be compared according to the presence of the Italian flag emoji, days, and hours. Indeed, the presence of the Italian flag seemed to convey hateful content in the pilot study. The collection related to Silvia Romano and the selected event reached 3006 contents, 1500 were randomly extracted for creating a sample to be annotated *Inters8_SRomano*.

The subset was cleaned of user mentions and URLs. The metadata used to describe the tweets were as follows: *id*, *parentTweetText* (if exists), and *tweetText*.
Output: *Inters8_SRomano* annotated according to the proposed annotation schema

3.3. Annotation process

The manual annotation process was divided into two phases: (1) *pilot* - a sampling of 50 tweets was selected to evaluate the annotation scheme and (2) *operational* - the annotation of the *Inters8_SRomano* sample dataset.

Our annotation scheme is described in Section 4. Twelve annotators, balanced by gender, were employed in order to ensure a diversified and representative group covering a wide range of perspectives and experiences. Guidelines provided to annotators were refined after a discussion within the pilot phase.

The subset was annotated as follows: we collected 3 independent annotations for 1373 tweets. The remaining 127 were annotated only by two independent annotators; a third annotation was collected in order to solve the disagreement.

The quantitative analysis conducted by manual annotation enabled the assessment of the actual coexistence of multiple discriminatory and hate dimensions and the construction of a gold standard.

Annotators features table, annotation scheme, guidelines in Italian and English, and annotation results obtained through majority vote are available here.
Output: subset of *Inters8* annotated according to the proposed annotation scheme.

4. A Novel Annotation Scheme

This section describes the process of creating a novel annotation scheme for multi-level analysis of intersectionality.

4.1. Related annotation schemes

Our annotation scheme is partially inspired by the ones designed in [18, 19]. The scheme used in the AMI@EVALITA18 shared task challenged the participants not only to determine whether misogynous content was expressed in the tweets, but also to classify the misogynistic behavior, by proposing the categories: *Stereotype & Objectification*, *Dominance*, *Derailing*, *Sexual Harassment & Threats of Violence*, and *Discredit*. A deeper analysis presented in [20], suggested some insights and motivations to simplify the fine-grained misogynistic behavior to be annotated in our scheme.

The distinction between specific individuals and generic groups of people, also mentioned in [19], was not introduced in our scheme, since the debates around the selected victims turn out to be particularly specific.

A further contribution, annotating an Italian immigration corpus, measured the intensity of hate speech on a scale of 0 to 4 [21]. The idea of measuring hate inspired the comparison of intensities and prevalence among coexisting dimensionalities.

Stance analysis in [22] is performed to check the behavior of tweets in response to others. They used the following labels: *agree-accept* (support), *reject* (deny), *info-request* (question), and *opinion* (comment). In the case under analysis, it was sufficient to consider: *support*, *against*, and *neutral*.

Moreover, as highlighted also in [19, 21], it is important to differentiate aggressive language from hate speech: in fact, aggressive content is not necessarily expressed through hateful vocabulary and *vice versa*. Since hatred and discrimination can appear implicitly, it is not always easy and immediate to recognize them in negative and aggressive content on social media. Moreover, not all expressions of disapproval and disagreement with groups imply discrimination. Such findings were useful in order to highlight the importance of annotating both implicit and explicit forms of hate speech.

4.2. Annotation scheme

The multi-level scheme[9] was meant to bring out the dimensionalities, and the cohesiveness and prevalence among them. The coarse-grained level is intended to annotate the presence of misogyny, xenophobia, and Islamophobia. The fine-grained analysis, first, aims to recognize which dimensionality prevails over the others. Secondly, a sub-classification of misogyny, if there is any, is proposed to classify it into *sexual harassment / derailing* and *discrediting / dominance*. Finally, an annotation of stereotyping, victim defense, and stance toward any parent tweet is proposed. A detailed description of the labels involved, supported by examples, follows.

Misogynistic behavior $[yes/no]$: explicit and implicit forms, including aversion, repulsion, target silencing and instrumentalization of pregnancy, following the definition "*misogynistic behavior is about hostility towards women who violate patriarchal norms and expectations, who aren't serving male interests in the ways they're expected to. So there's this sense that women are doing something wrong: that they're morally objectionable or have a bad attitude or they're abrasive or shrill or too pushy*" [23]. Annotate the two following sub-labels only if $misogyny = yes$:

- **Sexual harassment and/or derailing** $[yes/no]$: the first includes avance, requests for sexual favors, and any form of harassment involving sex or speech in which abuse of women is justified by belittling or evading male responsibility. The latter refers to the intention to divert support toward the victim by directing the discourse to a more comfortable alternative issue while ignoring the discriminatory problem.

  ITA: *La z****la, appassionata ai c**zi talebani, ha orchestrato una messinscena con il tipo che se la s**pa e si è sistemata a vita con il riscatto.*

  ENG: *That s**t, fond of Taliban c***s, orchestrated a setup with the guy who fu**ed her and set herself up for life with the ransom money*

- **Discredit and/or dominance** $[yes/no]$: discrediting occurs when an individual S, through a communicative act, damages the image of another individual T in front of a third party (individual or group A) by referring to actions or characteristics of T that are considered negative by A. Dominance is typically expressed as an assertion of superiority by highlighting gender inequality.

  ITA: *Conte dacci le prove del riscatto pagato dagli italiani per questa odiosa nullità e vergogna nazionale! È una bambina indottrinata, senza cervello e stupida. È andata in terre cesso per seguire le sue idiozie apparentemente umanitarie.*

  ENG: *Conte, give us evidence of the ransom paid by the Italians for this odious nothingness and national disgrace! She is an indoctrinated brat, braindead and stupid. She went to toilet-lands to follow her supposedly humanitarian nonsense*

[9]Annotation scheme and guidelines are available at: https://github.com/ivsnp/inters8/blob/main/annotation

Xenophobia and/or racism $[yes/no]$: explicit and implicit forms, i.e., expressions of racism based on the arbitrary assumption of the existence of biologically and historically "superior" human races, aversion to foreigners, and what is foreign. The latter manifests itself in attitudes and actions of intolerance and hostility toward the culture and inhabitants of other countries. In order to analyze *ingroup* and *outgroup* dynamics [24, 25, 26], we also consider texts where the target subject and other Italians are insulted or rejected as members of the *ingroup*, or stigmatized as anti-Italian, because of their proximity to (or support for) foreign populations or immigrants, to be expressions of xenophobia.

> **ITA**: *Una donna bianca convertita all'Islam, esce indenne dai ne***ni, belve inferocite, andate tutti a fare in c*lo.*
>
> **ENG**: *A white woman converted to Islam, comes out untouched by the nig***s, raging beasts, f**k you all.*

Islamophobia $[yes/no]$: explicit and implicit forms of strong aversion, dictated by prejudicial reasons, toward Islamic culture and religion. The main manifestations include criminalizing targets by describing them as threatening and violent. It is often joined by xenophobia and may appear in the form of dehumanization of targets.

> **ITA**: *È venuta qui per fare attentati, è una terrorista, è anche incinta di un musulmano. Se stava bene in Islam rimpatriatela #convertita.*
>
> **ENG**: *She came here to do bomb attacks, she is a terrorist, she is also pregnant by a Muslim. If she was fine in Islam then send her back #converted*

Prevalence $[misogyny/xenophobia\ and/or\ racism/Islamophobia/absent]$: in case of coexistence of at least two of the main categories to be analyzed, indicate which one prevails over the others within the tweet.

> **ITA**: *Il governo ruba 4 milioni di euro agli italioti per pagare uno specie di riscatto al marito islamico che la mette incinta e la converte. Arriva in Italia contenta, ingrassata e viene accolta come una santa. Popolo idiota!*
>
> **ENG**: *The government steals 4 million euros from the Italians to pay some kind of ransom to her Islamic husband who impregnates her and converts her. She arrives in Italy happy, fat and is welcomed as a Saint. Idiotic people!*
>
> **Note**: *This tweet is an example of the intersection of multiple dimensionalities of hatred toward vulnerable groups. The term "italiota" means "Italian idiot" and "ingrassata" refers to the target's pregnancy.*

Stereotypes $[yes/no]$: negative sexist, xenophobic, racist and Islamophobic stereotypes concerning vulnerable groups targeted by discrimination and hate speech considered in this study on intersectional hatred. Stereotyping is a generalization conducted about a group of people, in which characteristics are attributed to all members of the group [27]. Stereotyping is based on a set of beliefs, not based on experience, that people enact to interpret their surroundings and move through them.

> **ITA**: *È venuta qui per fare attentati, è una terrorista, è anche incinta di un musulmano. Se stava bene in Islam rimpatriatela #convertita.*

Target defense $[yes/no]$: it indicates whether the user who posted the tweet defends the hate target, contributing to creating a counter-narrative effect. It includes both support without discrimination and support that redirects hatred toward other people (without actually counteracting hate speech).

> **ITA**: *Il privato di Silvia Romano non dovrebbe essere nel dibattito pubblico. È stata liberata da una prigione fisica ma intrappolata in una di violenze psicologiche e pregiudizi inutili e ingiusti.*
>
> **ENG**: *Silvia Romano's private life should not be in the public debate. She was released from a physical prison but trapped in one of psychological violence and unnecessary and unjust prejudice.*

Stance $[support/against/neutral/absent]$: it identifies the stance of the user who posted a tweet reacting to another one. The label takes the value *support* if it agrees with the parent tweet, *against* if it disputes it, and *neutral* if no stance can be inferred from the text. If there is no parent tweet the default value is *absent.*

> **ITA**
> **Parent tweet**: *La liberazione di Silvia Romano è una bella notizia. L'aspettiamo in Italia, ringraziamo i nostri servizi di Intelligence e coloro che hanno contribuito a questo importante obiettivo.*
>
> **Child, or reply, tweet (contestazione)**: *Assolutamente no! Vanno a fare le splendide in Africa ma quando si accorgono che*

*ci sono i ne***ni cattivi chiedono aiuto a mamma Italia.*

ENG:

Parent tweet: *The liberation of Silvia Romano is good news. We are waiting for her in Italy, we thank our intelligence services and those who contributed to this important achievement.*

Child, or reply, tweet (against): *Absolutely not! They go show off in Africa, but when they realise there are bad nig***s they ask Mamma Italia for help.*

5. Results

The annotation of *Inters8_SRomano* extracted from *Inters8* and the harmonization stage by majority vote yielded the results shown below.

Islamophobia is the most annotated dimensionality, followed by misogynistic behavior and xenophobia/racism.

Label annotation detected the following amounts of tweets in the subset (see Figure 2): *misogynistic behavior* 288 (19.2%), *sexual harassment and/or derailing* 36 (12.5% of misogynistic behavior), *discredit and/or dominance* 247 (85.8% of misogynistic behavior), *xenophobia and/or racism* 153 (10.2%), *Islamophobia* 317 (21.1%), *stereotype* 394 (26.3%), *target defence* 501 (33.4%), and *stance* [against=119, support=108, absent=42, neutral=24]([40.6%, 36.9%, 14.4%, 8.2%] out of 293 reply tweets).

Stance, on the other hand, appeared difficult to annotate because the stances often went off-topic.

Among tweets labeled with at least one of the three main dimensions of hate included in the proposed annotation scheme, the Italian flag (in *name*, *screen-name*, *bio* or *tweet*) appears as follows: 78.5% of misogynistic behavior, 73.2% of Xenophobia and/or racism and 74.4% of Islamophobia.

The subset of tweets annotated with an intersection between the three main dimensionalities (at least 2) has cardinality 222. The most present and prevalent dimension was Islamophobia.

The following label distributions were obtained from the annotation of intersectional tweets (see Figures 3 and 4): *misogynistic behavior* 184 (82.9%), *xenophobia and/or racism* 117 (52.7%), *Islamophobia* 199 (89.6%), and *prevalence* [Islamophobia=91, misogynistic behavior=59, absent=42, xenophobia and/or racism=30]([41%, 26.6%, 18.9%, 13.5%] out of 293 reply tweets).

5.1. Inter-Annotator Agreement

Cohen's average Kappa [28] was 0.40 and the Fleiss' Kappa calculated was as follows: misogyny 0.49, sexual

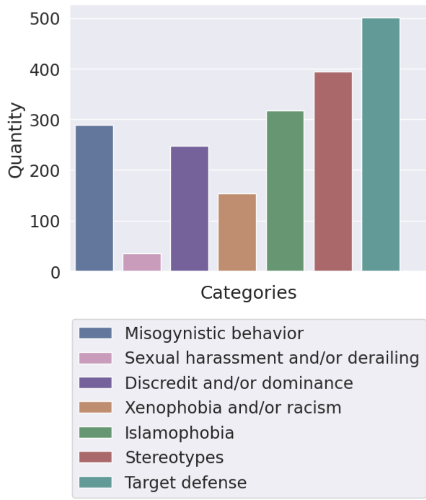


Figure 2: Annotation of binary categories.

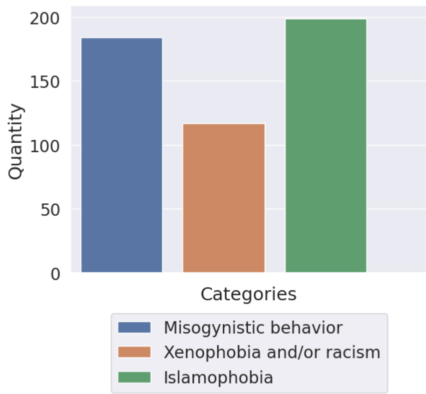


Figure 3: Three main dimensionalities annotated with the value YES in the intersection subset (at least two dimensionalities=yes).

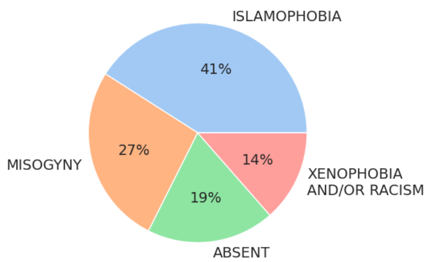


Figure 4: Annotation of prevalence among dimensionalities in subset intersection (at least two dimensionalities=yes).

harassment / derailing 0.29, discredit / dominance 0.44, xenophobia / racism 0.32, Islamophobia 0.53, prevalence 0.33, stereotyping 0.33, target defense 0.59 and stance 0.81. The calculation of Fleiss' Kappa for tweets with three independent annotations showed how complicated it was to classify the domain across the dimensionalities of the established multi-level scheme.

5.2. Considerations

Misogyny has been labeled more by female annotators, only $1/3$ of male annotators come close to the former. This is an expected result as the former tend to be more sensitive to the issue confirming that, for creating an unbiased annotated dataset, is important to employ annotators belonging to heterogeneous social categories.

Inter-Annotator Agreement Following an overview of the annotated subset and comparison with the annotators involved in this experiment, it appeared that many disagreements occurred on the label *stereotypes*, some annotators recognizing many more than others independently from their self-identified gender and age. The annotations appeared quite subjective and often repetitive because the presence of other discriminatory dimensionality often involves stereotypes. Concerning *stance*, the presence of irony and rhetorical questions inside the dataset complicates the valuation of the attribute.

The complexity of some annotation scheme labels emphasizes the difficulty of annotating tweets about this domain and brings out the presence of bias.

In addition, the phenomenon of *premediation* [29] has been observed in this case study. Indeed, Twitter's users expressed their own opinions favoring immediacy and emotionality in communication as a preliminary reaction to the first information about the news. Tweets and interactions began immediately, despite the fact that the full picture of the affair was not clear at that moment, bringing the event to the platform's trending topics.

Target Comparing the Twitter interactions regarding Silvia Romano with those around Luca Tacchetto[10], Alessandro Sandrini[11], and Sergio Zanotti[12], the following emerged. The three Italian people listed were victims of kidnapping like Silvia Romano. Among them, Tacchetto and Sandrini were converted to Islam. Considering the first week after the target subjects' homecoming, the following amounts of interactions were detected: Romano 237031, Tacchetto 1668, Sandrini 546, and Zanotti 2206. Interestingly, the volume of reactions related to the Silvia Romano's liberation is much higher (see [30] for a deeper analysis about this topic), suggesting that the intersectionality with misogyny matters.

Recurrent topics Among the tweets annotated, some recurring topics emerged: (1) dissent on the economic plan arguing that payment of the alleged ransom for release was not necessary, (2) aesthetic appearance by seeing physical appearance and clothing as objects of dialogue, (3) being ungrateful, selfish and a traitor to her (Silvia Romano) country for converting to the Islamic religion, (4) victimization, (5) pregnancy, and (6) politics.

In the former case, it was not always possible to recognize the discriminatory nature since a variable related to discontent about the Italian economic situation was also present. The second focused on arrival at the airport wearing the hijab and a watch pointed to as lavish despite the fact that it was not possible to distinguish it from the content disseminated by the media. The third appeared describing her as a member of the *outgroup*. The remaining ones also appeared frequently in the narrative of the annotated subset.

Counter-speech It has been observed that *counter-speech*, carried out by users who take the defense of victims, sometimes proposes an alternative narrative to hate speech. Other times they follow defensive strategies that are themselves offensive generating further hate speech.

6. Conclusions and future work

Developing the *Inters8* corpus built considering an intersectional target-event pair allowed us to explore a case study and analyze Twitter interactions related to Silvia Romano on social media. The manual annotation applied highlights how multiple dimensionalities coexist and intertwine in cases of intersectional hate.

Despite the evidence of the phenomenon and its dynamics, results presented here are related to the specific case study taken into account, and to the target and the event selected. In fact, at the current stage of development, the *Inters8* corpus includes content related to the specific intersectional Silvia Romano's liberation target-event pair. We plan to expand the corpus with additional targets, events, and social categories. It would then be interesting to compare multiple targets in the same intersection and study how local culture might influence the phenomenon over several countries.

As the corpus is built around the Italian context, the data are exclusively in Italian. The integration of multiple languages would allow for greater generalization and the study of the geographic distribution of the phenomenon around known people and events.

Finally, there are many interactions on social networks and the experimental study for automatic detection of intersectional hate may be a challenge of particular interest.

[10] https://www.nytimes.com/2020/03/14/world/africa/mali-hostages-released.html
[11] https://apnews.com/article/---0acfda0fe2974efab72081965cb7d3c6
[12] https://apnews.com/article/01e8c9c94f8e4441b7ad68de3d58e667

References

[1] K. W. Crenshaw, Mapping the margins: intersectionality, identity politics, and violence against women of color, Stanford Law Review 43 (1991) 1241–1299.

[2] V. Basile, M. Lai, M. Sanguinetti, Long-term social media data collection at the university of turin, in: E. Cabrio, A. Mazzei, F. Tamburini (Eds.), Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Torino, Italy, December 10-12, 2018, volume 2253 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018, pp. 1–6. URL: https://ceur-ws.org/Vol-2253/paper48.pdf.

[3] L. Berlingozzi, "welcome home silvia, into the lion's den": How gender biased narratives frame hostage liberations, Security Praxis (2020). URL: https://securitypraxis.eu/silvia-romano-gender-biased-narratives-hostages/.

[4] J. Kim, C. Ortiz, S. Nam, S. Santiago, V. Datta, Intersectional bias in hate speech and abusive language datasets, CoRR abs/2005.05921 (2020). URL: https://arxiv.org/abs/2005.05921. arXiv:2005.05921.

[5] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, V. Patti, Resources and benchmark corpora for hate speech detection: a systematic review, Language Resources and Evaluation 55 (2021) 477–523. URL: https://doi.org/10.1007/s10579-020-09502-8. doi:10.1007/s10579-020-09502-8.

[6] B. Vidgen, L. Derczynski, Directions in abusive language training data, a systematic review: Garbage in, garbage out, PLOS ONE 15 (2021) 1–32. URL: https://doi.org/10.1371/journal.pone.0243300. doi:10.1371/journal.pone.0243300.

[7] M. Anzovino, E. Fersini, P. Rosso, Automatic identification and classification of misogynistic language on Twitter, in: M. Silberztein, F. Atigui, E. Kornyshova, E. Métais, F. Meziane (Eds.), Natural Language Processing and Information Systems - 23rd International Conference on Applications of Natural Language to Information Systems, NLDB 2018, Paris, France, June 13-15, 2018, Proceedings, volume 10859 of *Lecture Notes in Computer Science*, Springer, 2018, pp. 57–64. URL: https://doi.org/10.1007/978-3-319-91947-8_6.

[8] D. Nozza, C. Volpetti, E. Fersini, Unintended bias in misogyny detection, in: IEEE/WIC/ACM International Conference on Web Intelligence, WI '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 149–155. URL: https://doi.org/10.1145/3350546.3352512. doi:10.1145/3350546.3352512.

[9] E. Fersini, P. Rosso, M. Anzovino, Overview of the task on automatic misogyny identification at ibereval 2018, in: P. Rosso, J. Gonzalo, R. Martínez, S. Montalvo, J. C. de Albornoz (Eds.), Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018, volume 2150 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018, pp. 214–228. URL: https://ceur-ws.org/Vol-2150/overview-AMI.pdf.

[10] B. Poland, Haters: Harassment, abuse, and violence online, U of Nebraska Press, 2016. URL: https://books.google.it/books?id=Jd4nDwAAQBAJ.

[11] E. W. Pamungkas, V. Basile, V. Patti, Misogyny detection in twitter: a multilingual and cross-domain study, Information Processing & Management 57 (2020) 102360. URL: https://www.sciencedirect.com/science/article/pii/S0306457320308554. doi:https://doi.org/10.1016/j.ipm.2020.102360.

[12] D. Nozza, Exposing the limits of zero-shot cross-lingual hate speech detection, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Association for Computational Linguistics, Online, 2021, pp. 907–914. URL: https://aclanthology.org/2021.acl-short.114. doi:10.18653/v1/2021.acl-short.114.

[13] E. W. Pamungkas, V. Patti, Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, Association for Computational Linguistics, Florence, Italy, 2019, pp. 363–370. URL: https://aclanthology.org/P19-2051. doi:10.18653/v1/P19-2051.

[14] P. Chiril, E. W. Pamungkas, F. Benamara, V. Moriceau, V. Patti, Emotionally informed hate speech detection: A multi-target perspective, Cognitive Computation 14 (2022) 322–352. URL: https://doi.org/10.1007/s12559-021-09862-5. doi:10.1007/s12559-021-09862-5.

[15] K. Florio, V. Basile, M. Polignano, P. Basile, V. Patti, Time of your hate: The challenge of time in hate speech detection on social media, Applied Sciences 10 (2020). URL: https://www.mdpi.com/2076-3417/10/12/4180. doi:10.3390/app10124180.

[16] J. Brailovskaia, J. Margraf, Addictive social media use during covid-19 outbreak: Validation of the bergen social media addiction scale (bsmas) and investigation of protective factors in nine countries, Current Psychology (2022). doi:10.1007/s12144-022-03182-z.

[17] C. Annovi, G. Dentice, F. Manenti, F. Portoghese, V. Lazzerini, Y. Pallavicini, V. Gullo, Le cause di

discriminazione, hate speech e crimini d'odio contro le donne musulmane in Italia, Technical Report, Deliverable 2.1, project "TRUST: Tackling Under-Reporting and Under-Recording of Hate Speech and Hate Crimes Against Muslim Women", co-funded by European Union, Grant Agreement no. 101049611, 2022. URL: https://www.trust-project-eu.info/trust/wp-content/uploads/2023/04/TRUST-D2.1-ITA-1.pdf.

[18] E. Fersini, D. Nozza, P. Rosso, Overview of the evalita 2018 task on automatic misogyny identification (AMI), in: T. Caselli, N. Novielli, V. Patti, P. Rosso (Eds.), Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018, volume 2263 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018, pp. 1–9. URL: https://ceur-ws.org/Vol-2263/paper009.pdf.

[19] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, M. Sanguinetti, SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 54–63. URL: https://aclanthology.org/S19-2007. doi:10.18653/v1/S19-2007.

[20] S. Lazzardi, V. Patti, P. Rosso, Categorizing misogynistic behaviours in italian, english and spanish tweets, Proces. del Leng. Natural 66 (2021) 65–76. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6323.

[21] M. Sanguinetti, F. Poletto, C. Bosco, V. Patti, M. A. Stranisci, An italian twitter corpus of hate speech against immigrants, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan, 2018, pp. 2798–2805. URL: https://aclanthology.org/L18-1443".

[22] E. W. Pamungkas, V. Basile, V. Patti, Stance classification for rumour analysis in twitter: Exploiting affective information and conversation structure, in: A. Cuzzocrea, F. Bonchi, D. Gunopulos (Eds.), Proceedings of the CIKM 2018 Workshops co-located with 27th ACM International Conference on Information and Knowledge Management (CIKM 2018), Torino, Italy, October 22, 2018, volume 2482 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018, pp. 1–7. URL: https://ceur-ws.org/Vol-2482/paper37.pdf.

[23] K. Manne, Down Girl: The Logic of Misogyny, Oxford University Press, 2017. URL: https://doi.org/10.1093/oso/9780190604981.001.0001. doi:10.1093/oso/9780190604981.001.0001.

[24] G. Comandini, V. Patti, An impossible dialogue! nominal utterances and populist rhetoric in an Italian Twitter corpus of hate speech against immigrants, in: Proceedings of the Third Workshop on Abusive Language Online, Association for Computational Linguistics, Florence, Italy, 2019, pp. 163–171. URL: https://aclanthology.org/W19-3518. doi:10.18653/v1/W19-3518.

[25] B. Sauer, A. Krasteva, A. Saarinen, Post-democracy, party politics and right-wing populist communication, Routledge, 2018, pp. 14–35.

[26] G. Mazzoleni, R. Bracciale, Socially mediated populism: the communicative strategies of political leaders on facebook, Palgrave Communications 4 (2018) 1–10. URL: https://EconPapers.repec.org/RePEc:pal:palcom:v:4:y:2018:i:1:d:10.1057_s41599-018-0104-x.

[27] E. Aronson, T. Wilson, R. Akert, Social Psychology, Always Learning series, Pearson, 2013. URL: https://books.google.it/books?id=wr9uvgAACAAJ.

[28] J. Cohen, A coefficient of agreement for nominal scales, Educational and Psychological Measurement 20 (1960) 37 – 46.

[29] R. A. Grusin, Premediation, Criticism 46 (2004) 17 – 39.

[30] I. Spada, V. Patti, M. Lai, IntersHate: un corpus italiano per lo studio di misoginia e intersezionalità in Twitter, Technical Report, Department of Computer Science, University of Turin, Italy, 2020. URL: https://github.com/ivsnp/inters8/blob/main/spada_thesis_IntersHate.pdf.

A. Online Resources

The annotated corpus and the guidelines are available on GitHub at the following link: https://github.com/ivsnp/inters8.

Let’s keep an eye on Russian: testing sensitivity to the change in the grammatical number in somatic idioms with ruBERT

Elena Spaziani[1,*,†], Eliana Di Palma[1,2,†]

[1] Sapienza, Università di Roma, Roma, Italy
[2] Università degli Studi di Roma Tre, Roma, Italy

Abstract

English. In recent times, linguistic research and computational linguistics have considered the morpho-semantic properties of the grammatical number in nouns. In this context, our study aimed at using a language model to test number variation on selected Russian somatic idioms containing locally and generally marked nouns. Overall, the model’s sensitivity corroborated the relevance of number in idioms, especially in instances where the semantics of number is particularly significant.

Italiano. Negli ultimi tempi la ricerca linguistica e la linguistica computazionale hanno iniziato a considerare le proprietà morfo-semantiche del numero grammaticale dei sostantivi. In tale contesto, questo studio si è proposto di utilizzare un modello linguistico per testare la variabilità del numero nei fraseologismi somatici russi con nomi dalla marcaztezza locale o generale. Nel complesso, la sensibilità del modello al cambiamento di valori grammaticali ha dimostrato l’influenza del numero, soprattutto in istanze in cui questo è particolarmente significativo dal punto di vista semantico.

Keywords

grammatical number, somatic idioms, Russian BERT, computationally-aided linguistic analysis

1. Introduction

Somatic idioms are defined as phraseological units which contain lexemes referring to human or animal body parts [1, 2]. This class is found in many languages, due to the universal nature of the somatic lexicon. Somatisms are usually observed from a semantic perspective and mostly in contrastive analyses, since their meaning is regarded as a cultural-specific conceptualisation. However, to the best of our knowledge, existing inquiries, both in theoretical and computational linguistics, have not yet considered the morpho-semantic features of the grammatical number [3] as part of the overall idiomatic meaning. Furthermore, our interest in number is motivated by the fact that many somatic terms exhibit an effect of the pluralization’s lexical properties. In particular, nouns denoting bipartite or complex body parts, which are naturally constituted by a plurality, show a higher frequency in the plural form in many languages. This condition is defined **local** or **semantic markedness** [4, 5], as the singular value becomes the locally marked member of the opposition, changing the morpho-semantic dynamics between the two values. The goal of our work is to consider how change in the grammatical number affects the idiomatic structure of somatisms. More specifically, the questions underlying the following experiment were:

1. Does number variation have a significant impact on the idiomatic structure of Russian somatic idioms?
2. Does the type of markedness affect the probability of finding a singular or a plural value?

We hypothesized that change in number would have an effect on the overall structure of somatic idioms, especially in the cases of local markedness. In detail, we assumed that if we switched the values of a locally marked somatic term a greater impact on the idioms would be observed, compared to generally marked somatic terms.

The following sections describe how we addressed these questions, starting from an examination of related works. The new dataset of Russian somatic idioms that we constructed to test our hypotheses and its creation method are thereafter illustrated[1]. Subsequently, our experiment is presented, which was conducted with ruBERT, a Russian-trained BERT, following a methodology similar to Salazar et al. [6] for acceptability, and Pedinotti et al. [7] for semantic plausibility. The results of the experiment are offered and discussed in the final sections. Ultimately, a conclusive section provides an answer to our questions, highlighting the relevance of our work and possible future expansion.

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

* Corresponding author.
† These authors contributed equally.
elena.spaziani@uniroma1.it (E. Spaziani); eliana.dipalma@uniroma1.it (E. D. Palma)
0000-0002-7140-9773 (E. Spaziani); 0000-0003-2154-2696 (E. D. Palma)

[1] The dataset is available on request.

2. Related works

Studies on idiom semantics adopting a compositional approach have claimed that the semantic analyzability or decomposability of idioms (the meaning of the constituents contributes to the overall figurative meaning) is related to their flexibility, that is their lexical or syntactical variation (see Cserép, Dobrovol'skij for an overview).

In idiom variation research, number has been investigated as a flexibility dimension by Langlotz [10] and Cserép [11]. Specifically, Langlotz [10] asserts that number variation is systematical in VP-idioms or PP-idioms that feature an isomorphic, that is analyzable, semantic structure. However, this variation is believed to be prevented if the idiom contains an invariable noun or a noun which is incompatible with the global idiomatic meaning. The second study [11] adopts a corpus-driven approach and it retrieves the morpho-syntactic alternants in the noun phrase of V NP-idioms. As a result, a higher frequency in the singular value is assessed and a similar correlation between number variation and decomposability is found. Both works acknowledge the role of number in the idiomatic structure, however they are limited to the English language and no distinction based on the lexical meaning or type of markedness of the constituents has been drawn.

In concordance with the aforementioned theories on the category of number, studies on the Russian language assert that «the grammatical values of number are closely related and actively interact with lexical meaning» [2] [12, 24] (see also Vinogradov [13]). As observed by Ljaevskaja [14], this interaction is particularly evident in nouns that denote complex items, including complex and paired body parts. The lexical meaning of such nouns is intertwined with their numerical information as they denote a complex structure, constituted by plural members functioning or perceived as a whole. As a result, in these cases the grammemes of number convey more than just the numerical information. To assess this extra-grammatical meaning, the two values of some somatic terms have been observed, based on their use in context [14, 71-86]. The result of this detailed qualitative investigation is the distinction of two contexts: regular distributive contexts (i.e. the standard grammatical opposition: one - many); non-distributive contexts, which are mostly noun-specific, value-specific, and language-specific (As Ljaševskaja illustrates [14, 72], он шел **по колено** [sg] в грязи - on šel po koleno v grjazi cannot be replaced with the plural он шел ***по коленам** [pl] в грязи - on šel po kolenam v grjazi, whereas in English a plural form is required: "He walked in mud up to his **knees**"). Furthermore, similar contexts lead to a classification of similar somatic terms. Apparently, глаз - glaz "eye" and ухо - ucho "ear" display corresponding contexts, showing how different meanings associate with the numerical values. On the contrary, мозг - mozg "brain" exhibits a quite unique condition: although it denotes a single item, its pluralization features an additional meaning, absent in the singular, acting as a separate lexeme (as does the English equivalent 'brains'). Regarding the markedness, the Academic Russian Grammar [15] and Ušakov's Dictionary [16] indicate that some nouns are mostly used in the plural, including somatic terms. However, a full account for the whole somatic lexical domain, in terms of frequency and markedness, is not available as it is for other languages.

Only in recent times have the lexical properties of number been addressed in the computational field. Gromann and Declerck [17] investigate the semantic shifts created by the regular inflectional morphemes of number. The study shows the impact of these inflectional variants on morphological embeddings and their variability, on a par with derivational morphemes. Even more recently, Janzso [18] focuses on the ambiguity originating from these inflectional variants, as the plural form of some nouns carries a secondary non-grammatical meaning, absent in the singular value. Specifically, the work inquires about how ambiguity in number and gender is treated by contextual meaning representations, using four pretrained BERT models on the disambiguation task. In the literature on idioms, computational studies focused on their identification or comprehension [19, 20, 21, 22], but to the best of our knowledge there have been no studies investigating the role of number in the construction of idioms and its interpretation.

3. Dataset

The dataset was constructed by selecting somatic idioms from two Russian phraseological dictionaries [23, 24]. Nine somatic components were considered in the selection: six terms referring to bipartite or complex body parts (**глаз** - glaz "eye", **рука** - ruka "hand/arm", **нога** - noga "foot/leg", **губа** - guba "lip", **ухо** - ucho "ear", **зуб** - zub "tooth"); three terms referring to a single body part or organ (**голова** - golova "head", **мозг** - mozg "brain", **язык** - jazyk "tongue").

In order to have a broad idea of the dominance condition [5], and therefore of the type of markedness, the relative frequency for each term was obtained from the Russian National Corpus[3] (see Table 1).

These components were considered both in the **singular** and in the **plural** form. Altogether 73 somatic idioms in the **singular** and 73 somatic idioms in the **plural** were retrieved[4]. Idioms registered in both forms

[2] In the original: «грамматические значения числа тесно связаны и активно взаимодействуют с лексически-ми значениями».

[3] https://ruscorpora.ru

[4] Ideally, ten idioms were to be selected for each term. However, the actual number of idioms is due to their availability in the dic-

Table 1
Relative frequency, ipm (instances per million), drawn from the Russian National Corpus.

Noun	Freq. [sg]	Freq. [pl]
глаз - glaz “eye"	$98,31$	$860,15$
рука - ruka “hand/arm"	$612,6$	$700,2$
нога - noga “foot/leg"	$136,46$	$343,33$
губа - guba “lip"	$20,6$	$126,31$
ухо - ucho “ear"	$62,74$	$76,72$
зуб - zub “tooth"	$16,49$	$98,81$
мозг - mozg “brain"	$59,54$	$14,77$
голова - golova “head"	$723,34$	$72,97$
язык - jazyk “tongue"	$303,5$	$59,15$

were avoided whenever possible (по губе/по губам - po gube/po gubam “to one's taste", lit. “to one's lip/lips"). Formally, the selected instances range from a higher degree of idiomaticity (fixed idiomatic expressions as глаза на лоб полезли - glaza na lob polezli “eyes nearly popped out", lit. “eyes climbed on the forehead") to a more flexible condition (idiomatic prepositional phrases as в головах - v golovach “at the head of", lit. “in the heads").

The Russian National Corpus was queried and the contexts were derived and trimmed to simple sentences. To evaluate the impact of change in number, the original sentences were manipulated in number (**manipulated condition**). As a result, the somatic component in the original condition was inflected to the alternative numerical value in the grammatical opposition (i.e. the component in the original singular-only form was inflected in the plural form and vice versa). Subsequently, a **control condition** was created by substituting the somatic component with an alternative word. However, it should be noted that a random substitution was excluded, considering the aforementioned gradient of idiomaticity, which relates to an equivalent gradient of substitutability of the idiom's component parts [25]. Therefore, the control condition was created by considering alternative candidates for a given construction offered by the corpus [5]. From a grammatical perspective, the original numerical value was always maintained in the alternative words. On the other hand, the grammatical categories of gender and animacy were maintained as far as possible, considering their subjection to the choice of a valid semantic alternative in terms of context (e.g. the original в глазах - v glazach [m] “in the eyes" was substituted with в комнатах - v komnatach [f] “in the rooms"). Finally, our dataset consists of 438 sentences, divided into three different conditions: **Original** (73 with singular form and 73 with plural form), **Manipulated** and **Control** sentences as in the following table 2[6].

4. Model and Experiment

Model For our experiment, we used the **ruBERT base model** (12-layer, 768-hidden, 12-heads, $178M$ parameters) provided by the Sberbank group (https://huggingface.co/ai-forever/ruBert-base). This model was built by taking BERT [27] as a basis. ruBERT-base has a Byte-Pair Encoding (BPE) tokenizer with a dictionary of 120 thousand tokens. It was trained with 30 GB of Russian text, which includes Wikipedia, news, part of the Taiga corpus [28], and some books[7].

Experiment In order to assess the sensitivity of the model to number variation in the selected idioms, we used the **pseudo-log-likelihood (PLL)** score [31]. As shown in Salazar et al. [6], the PLL can be considered as a measure comparable to probability. The authors demonstrate that the PLL outperforms scores obtained with auto-regressive models in a series of tasks related to sentence acceptability. Furthermore, in the work of Pedinotti et al. [7] the same measure is used to assess the difference in plausibility between metaphorical sentences, literal sentences and nonsense sentences. The results show a quite good correlation with human judgments of semantic plausibility.

The PLL of a sentence W can be derived by iteratively masking individual tokens, one at a time, using the main function of the MLM (Masked-Language Modeling) model. For each masked token w, the probability is calculated based on all other words in the context, and the log-probabilities for all tokens are summed. This process is illustrated by the following Equation 1:

$$PLL(W) = \sum_{t=1}^{|W|} \log P(w_t \mid W_{\setminus t}) \tag{1}$$

5. Results

We calculated the significance of the difference between the different conditions using the Wilcox test for multiple comparisons with Bonferroni correction.

tionaries or in the Russian National Corpus. Specifically, in the singular section, ухо - ucho “ear" and зуб - zub “tooth" (9 idioms), губа - guba “lip" (2 idioms) and мозг - mozg “brain" (3 idioms) did not reach the set number. The same applies for губы - guby “lips" (6 idioms), головы - golovy “heads" (5 idioms), языки - jazyki “tongues" (2 idioms) in the plural section.

[5] In limited cases, mostly in idioms originally in the singular form, a higher degree of idiomaticity rendered necessary a random word.

[6] Unlike the control and the manipulated sentences, the translation of the original Russian idioms refers to the correspondent English idioms given by a Russian-English phraseological dictionary [26] whenever possible.

[7] We selected this model over DeepPavlov's RuBERT [29] because it performed better in the RussianSuperGLUE evaluation [30].

Table 2
Extraxt from the dataset: **O** stands for **Original**, **M** for **Manipulated** and **C** for **Control** sentence.

Sentence	Number	Type
мне было очень стыдно за свой **язык** без костей - mne bylo očen' stydno za svoj **jazyk** bez kostej	Sg	O
"I was very ashamed of my loose **tongue***", lit. "I was very ashamed of my* **tongue** *without bones"*		
мне было очень стыдно за свои **языки** без костей - mne bylo očen' stydno za svoi **jazyki** bez kostej	Pl	M
"I was very ashamed of my **tongues** *without bones"*		
мне было очень стыдно за свое **тело** без костей - mne bylo očen' stydno za svoe **telo** bez kostej	Sg	C
"I was very ashamed of my **body** *without bones"*		
но все это для отвода **глаз** - no vse ėto dlja otvoda **galz**	Pl	O
"but all this is for a distraction", lit. "but all this is for the withdrawal of the **eyes***"*		
но все это для отвода **глаза** - no vse ėto dlja otvoda **glaza**	Sg	M
"but all this is for the withdrawal of the **eye***"*		
но все это для отвода **вод** - no vce ėto dlja otvoda **vod**	Pl	C
"but all this is for the withdrawal of the **waters***"*		

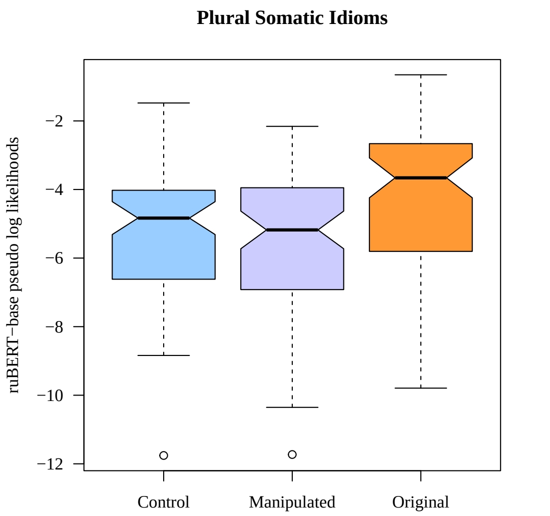


Figure 1: Boxplots of **ruBERT pseudo-log-likelihood** scores for the three conditions in the **plural** section.

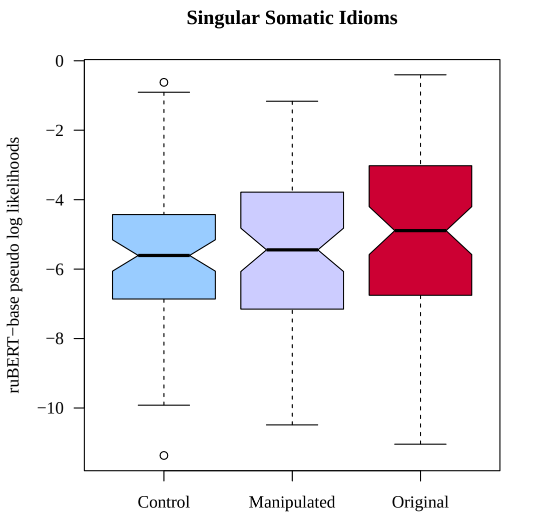


Figure 2: Boxplots of **ruBERT pseudo-log-likelihood** scores for the three conditions in the **singular** section.

Plural With regard to idioms containing somatic terms in the original plural condition, the differences between the **original condition** and the **manipulated condition**, and between the **original** and the **control condition** are significant ($p - value < 0.005$). On the other hand, there is no significant difference between the use of the singular (**manipulated condition**) and the different word in the **control condition** (we can also observe the differences in the Figure 1). This may be evidence of a real distance between the plural and singular forms. As a matter of fact, the singular form proves to have a probability that is not significantly different from any other word occurring in the construction.

Singular As for the singular, we can observe a difference between the **original condition** and the other two (see Figure 2), but there is no significance from the statistical analysis we subjected the data to. The lack of significance could be due to a lower frequency of idioms featuring somatic parts in the singular form. Nonethe-

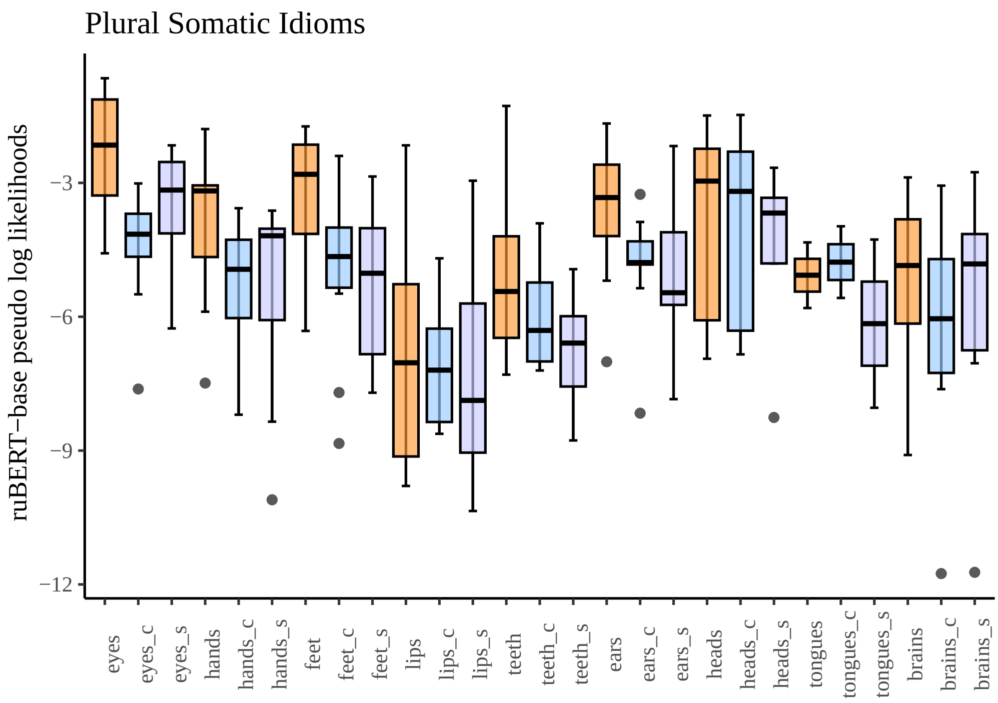


Figure 3: Boxplots of **ruBERT pseudo-log-likelihood** scores for the each noun in the three conditions in the **plural** section. The nouns are ordered by the relative frequency in the plural form as in Table 1.

less, the data show interesting trends when analysing the behaviour of individual nouns more in detail (as we can see in Figure 4).

6. Discussion

As it is evident from the results, we can positively address the first question, asserting that **the grammatical number of nouns significantly affects the structure of somatic idioms**. Indeed, the different probability obtained from the conditions indicate that the two grammatical values are not interchangeable as one would imagine. As a result, an overall morphological flexibility of the idiom associated with number variation as intended in [11] has not emerged.

The difference between the two values is significant in the **plural** section (see Figure 2), as it shows a higher probability to find a plural form in idioms originally in the plural (see Figure 1). On the other hand, the **singular** section displays a weak difference, except in two cases (“lip", “tooth"), whose values result to be less interchangeable.

The differences between and within the two sections could be due to the morpho-semantic properties of the grammatical number in the terms considered, specifically to their type of markedness. Firstly, **it could be plausible that the higher significance of the plural section correlates with local markedness**. As table 1 shows, most terms (six out of nine) display a condition of plural dominance, therefore the plural form is the naturally expected value. Secondly, in both sections these six locally marked terms show a difference between the **original condition** and the **manipulated condition**, demonstrating the impact of number. On the other hand, generally marked terms (“head", “tongue", “brain") display little or no difference (see Figures 3 and 4).

Given the influence of the individual terms on both sections, it is worth discussing them separately.

Locally marked terms “Eye" and “ear" occupy a dedicated section in Ljaševskaja [14]’s investigation, as it was observed that they have similar contexts. This similarity seems to be confirmed both in the **plural** and in the **singular** section (see figure 3 and 4). However, at a closer look, “ear" behaves similarly to “hand/arm" and “foot/leg", especially in the **plural** section. These nouns present a significant difference between the **original plural condition** and the **manipulated condition** (in

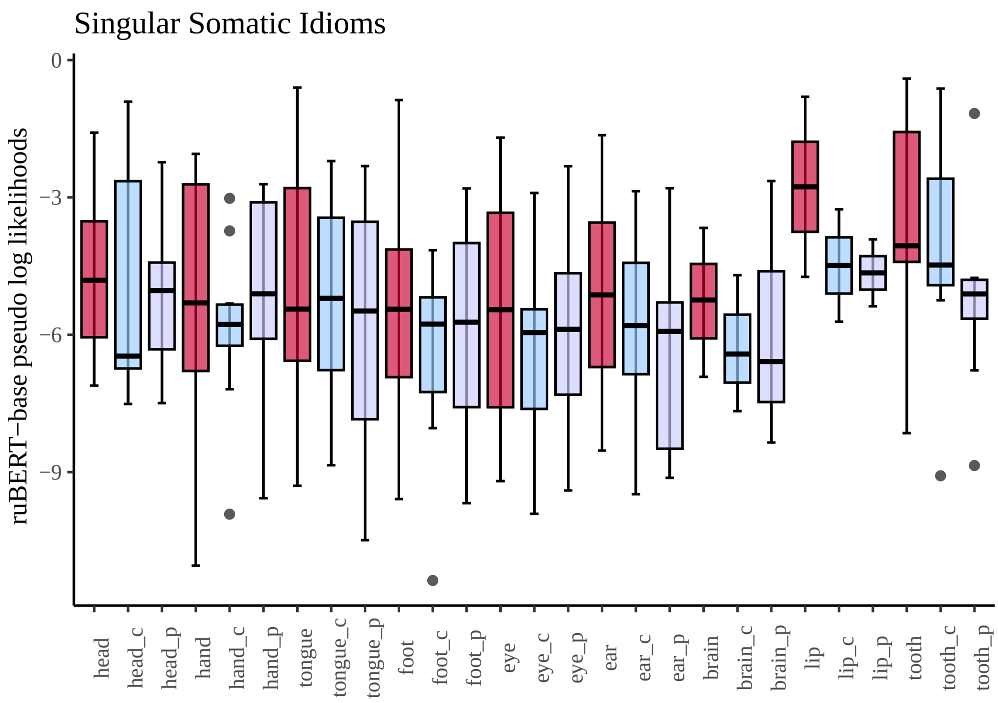


Figure 4: Boxplots of **ruBERT pseudo-log-likelihood** scores for the each noun in the three conditions in the **singular** section. The nouns are ordered by the relative frequency in the singular form as in Table 1.

"ears" and "hands" the $p-value < 0.005$; while in "feet" the $p-value < 0.02$).

An unexpected outcome is offered by "lip", therefore it cannot be associated with the nouns denoting a bipartite item, despite its referential meaning. Moreover, it behaves similarly to a noun denoting a complex body part: "tooth" (which is significant in both sections, $p-value < 0.005$). In the **singular** section "lip" is more probable in the singular value and it could be due to a more fixed structure of the original idiom.

Generally marked terms Among the generally marked terms "head" and "tongue" seem to follow the tendency observed in [11]. In fact, a certain interchangeability between the values is demonstrated by the **singular** section; while in the **plural** section, the original plural form is more probable, possibly denoting a more fixed structure.

Despite being a generally marked term, "brain", provides an unexpected outcome. In fact, it features an evident difference between the values in the **singular** section, possibly alluding to a more fixed structure in the original idiom. On the other hand, little change is observed in the **plural** section. The results obtained in the **singular** section do not explicitly confirm the existence of the additional sense associated with the plural of "brain" (as noted in Ljaševskaja [14, 86]). Nonetheless, the semantic discrepancy between the singular and the plural meaning of "brain" could still explain the less likely plural value in the **manipulated condition**.

7. Conclusions

On the basis of the presented experiment, we may assert that our initial assumptions prove to be partially justified. Comparing **original** sentences to **manipulated** and **control** sentences, we have found that, overall, number significantly affects the probability which connects the constituents of the idiom. Specifically, idiom variation appears to be subject to the morpho-semantic properties of the nominal constituent in the idioms considered. Indeed, our data showed that **the result varies depending on the type of markedness**. As a result, change in number for the generally marked constituents is more probable. This could confirm the general tendency of number being a flexibility dimension in idioms, which does not alter the idiomatic structure. On the other hand, locally marked terms show a significant difference in probability between the two values. The difference suggests a less probable number variation, which may allude to a less flexible idiomatic structure. However, some terms exhibit a peculiar condition, whose interpretation may be subjected to the formulation of future hypotheses.

Limitations and future research Despite the correlation between human judgment and the PLL measure shown in Pedinotti et al. [7], the inclusion of human evaluations could still be necessary given the figurative meaning of idioms. This comparison could improve the robustness of the analysis while clarifying the results obtained by the model. Furthermore, given the unexpected outcome for some nouns, a larger dataset including more idioms per nouns and a wider variety of nouns could be designed.

References

[1] I. Olza-Moreno, Aspectos de la semántica de las unidades fraseológicas. La fraseología somática metalingüística del español, Ph.D. thesis, Universidad de Navarra, 2009.

[2] F. Vakk, O somatičeskoj fraseologii v sovremennom ėstonskom literaturnom jazyke, Ph.D. thesis, Institut jazyka i literatury Akademii nauk ĖSSR, 1964.

[3] P. Acquaviva, Lexical Plurals: a morphosemantic approach, Oxford Universtity Press, 2008.

[4] P. M. Tiersma, Local and general markedness, Language (1982).

[5] H. Baayen, C. Burani, R. Schreuder, Effects of semantic markedness in the processing of regular nominal singulars and plurals in italian, Yearbook of Morphology (1997).

[6] J. Salazar, D. Liang, T. Q. Nguyen, K. Kirchhoff, Masked Language Model Scoring, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, p. 2699–2712. URL: https://www.aclweb.org/anthology/2020.acl-main.240.pdf.

[7] P. Pedinotti, E. Di Palma, L. Cerini, A. Lenci, A howling success or a working sea? testing what BERT knows about metaphors, in: Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 192–204. URL: https://aclanthology.org/2021.blackboxnlp-1.13. doi:10.18653/v1/2021.blackboxnlp-1.13.

[8] A. Cserép, Idiom variation and decomposability part i: Verbal variation, Yearbook of Phraseology (2017).

[9] D. Dobrovol'skij, Windows to the Mind: Metaphor, Metonymy and Conceptual Blending, Walter de Gruyter, 2011.

[10] A. Langlotz, Idiomatic Creativity: A cognitive-linguistic model of idiom-representation and idiom-variation in English, John Benjamins Publishing, 2006.

[11] A. Cserép, Idiom variation and decomposability part ii: Variation in the noun phrase, Yearbook of Phraseology (2017).

[12] V. Degtjarev, Kategorija čisla v slavjanskich jazykach. Istoriko-semantičeskoe issledovanie, Izdatel'stvo Južnogo Federal'nogo Universiteta, 2014.

[13] V. V. Vinogradov, Russkij jazyk: grammatičeskoe učenie o slove, Vyšaja škola, 1986.

[14] O. N. Ljaševskaja, Semantika russkogo čisla, Jazyki Slavjanskoj Kul'tury, 2004.

[15] N. J. Švedova; Nina D. Arutjunova, et al., Russkaja Grammatika, Nauka, 1980.

[16] D. Ušakov, Tolkovyj slovar' russkogo jazyka, sov. encikl., 1935-1940.

[17] D. G. T. Declerck, Towards the detection and formal representation of semantic shifts in inflectional morphology, in: 2nd Conference on Language, Data and Knowledge (LDK 2019), 2019.

[18] A. Janzso, Disambiguating grammatical number and gender with bert, in: Proceedings of the Student Research Workshop Associated with RANLP 2021, 2021.

[19] Y. Dai, Y. Liu, L. Yang, Y. Fu, An idiom reading comprehension model based on multi-granularity reasoning and paraphrase expansion, Applied Sciences 13 (2023). URL: https://www.mdpi.com/2076-3417/13/9/5777. doi:10.3390/app13095777.

[20] V. Nedumpozhimana, F. Klubička, J. D. Kelleher, Shapley idioms: Analysing bert sentence embeddings for general idiom token identification, Frontiers in Artificial Intelligence 5 (2022). URL: https://www.frontiersin.org/articles/10.3389/frai.2022.813967. doi:10.3389/frai.2022.813967.

[21] Z. Zeng, S. Bhat, Idiomatic expression identification using semantic compatibility, Transactions of the Association for Computational Linguistics 9 (2021) 1546–1562. URL: https://aclanthology.org/2021.tacl-1.92. doi:10.1162/tacl_a_00442.

[22] J. Peng, A. Feldman, Automatic idiom recognition with word embeddings, in: Symposium on Information Management and Big Data, 2015.

[23] A. N. B. D. O. Dobrovol'skij, Akademičeskij slovar' russkoj frazeologii, Leksrus, 2020.

[24] A. I. Fedorov, Frazeologičeskij slovar' russkogo literaturnogo jazyka, Astrel', AST, 2008.

[25] W. O'Grady, The syntax of idioms, Natural Language Linguistic Theory (1998).

[26] S. Lubenskaja, Bol'šoj russko-anglijskij frazeologičeskij slovar', AST-press, 2004.

[27] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.

[28] S. O. Shavrina Tatiana, To the methodology of corpus construction for machine learning: «taiga» syntax tree corpus and parser., in: Proceeding of "CORPORA2017", international conference, 2017.

[29] Y. Kuratov, M. Arkhipov, Adaptation of deep bidirectional multilingual transformers for russian language, 2019. arXiv:1905.07213.

[30] T. Shavrina, A. Fenogenova, A. Emelyanov, D. Shevelev, E. Artemova, V. Malykh, V. Mikhailov, M. Tikhonova, A. Chertok, A. Evlampiev, Russiansuperglue: A russian language understanding evaluation benchmark, arXiv preprint arXiv:2010.15925 (2020).

[31] A. Wang, K. Cho, BERT has a Mouth, and It Must Speak: BERT as a Markov Random Field Language Model, in: Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation, 2019, p. 30–36. URL: https://www.aclweb.org/anthology/W19-2304/.

That branch of the Lake of Como...: Developing a New Resource for the Analysis of I Promessi Sposi and its Historical Translations

Rachele Sprugnoli[1,*,†], Marco Sartor[1,†]

[1]Università di Parma, Viale D'Azeglio, 85, 43125 Parma, Italy

Abstract

This paper presents a directional parallel corpus of the Ventisettana, that is the version of *I Promessi Sposi* published by Manzoni in 1827, aligned at sentence level with the anonymous English translation published in London in 1834 by Richard Bentley. After describing the procedure followed for creating the resource and analyzing the results of the manual alignment, the corpus is used as a gold standard to evaluate Bertalign automatic aligner. This new linguistic resource can benefit the research community, in particular in the fields of the history of literature and translation studies, and be useful for developing new automatic tools specific for handling the peculiarities of historical literary texts.

Keywords

parallel corpus, sentence alignment, translation, digital humanities

1. Introduction

Critics have established how the *Promessi sposi* immediately enjoyed a wide resonance in Europe and in US, although the success of the work outside Italy has not always been accompanied by an effective understanding of the author's thought.[1] For this reason, the development of new linguistic resources based on the first historical translations of the novel assumes particular importance. These resources will benefit the research of historians of the Italian language, but also the development of new automatic tools suitable for processing historical literary texts. Last but not least, they can be used for educational purposes, both in secondary school, for the study of Manzoni's texts and their circulation beyond national borders, and at university level, in the field of translation studies. In particular, in this contribution we present a parallel corpus of the so-called Ventisettana, that is the version of the novel published by Manzoni in 1827, aligned with the anonymous English translation published in London in 1834 by Richard Bentley. An analysis of these two texts and of other English translations of the 19th century is provided by Intonti and Mallardi [5]: the volume is accompanied by examples of alignments to show specific cases both at the level of sentences, such as cuts and additions, and at the level of words, such as the rendition of figurative expressions and proverbs. The creation of a more extensive resource, such as the one presented here, aims to test the feasibility of a procedure to be applied in the future also to other historical translations so to offer the possibility of extending the range of linguistic analysis. Furthermore, our parallel corpus is a gold standard for evaluating fully automatic algorithms in a complex setting due to the peculiarities of historical texts and historical translations. Indeed, the complexity is due both to the characteristics of Manzoni's novel (rich, among other things, in irony, dialectal expressions, dialogues and monologues) and to the fact that during the 19th century translations did not aim to guarantee the greatest possible fidelity towards the source text, but rather to bend it in the light of the historical-cultural context in which they were implemented [6]. This approach to translation causes the original text to be revised and changed through additions and omissions of even entire chapters, making it a challenge to automate the alignment process.

2. Related Work

A parallel corpus is made of a set of texts in a given source language aligned with their translations in one or more target languages. The alignment, that is the identification of corresponding text units in parallel texts, can be performed at paragraph, sentence or word level. When

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

*Corresponding author.

†This paper is the result of the collaboration between the two authors. For the specific concerns of the Italian academic attribution system: Rachele Sprugnoli is responsible for Sections 2, 3.2, and 4; Marco Sartor is responsible for Section 3.1. Sections 1 and 5 were collaboratively written by Rachele Sprugnoli and Marco Sartor.

rachele.sprugnoli@unipr.it (R. Sprugnoli); marco.sartor@unipr.it (M. Sartor)

0000-0001-6861-5595 (R. Sprugnoli); 0000-0002-1176-2735 (M. Sartor)

[1]A large number of blunders and mistakes made in the translations has been reported in [1] and [2]. For Manzoni's popularity outside Italy, see [3] and the references listed on [4].

the translation direction is known (i.e. when the source and target languages are clearly stated) and when the translation is direct (i.e. not mediated by an intermediary language), the parallel corpus is defined as directional [7].

The development of large parallel corpora, both bilingual and multilingual, took off in the 90s of the last century but their growth in terms of number of texts and languages covered is more recent thanks to initiative such as the OPUS project [8] and those promoted by the European Commission [9]. The great attention given to this type of corpora is due to the fact that parallel corpora are useful to gain insights into interlinguistic phenomena; at the same time they are a rich source of materials for language teaching, translation studies, lexicography, and a fundamental resource for terminology extraction and machine translation systems.

Since manual alignment is a particularly time-consuming process, various automatic techniques have been proposed over the years [10]. Specifically, with regard to sentence-level alignment, early approaches are based on sentence length in terms of number of words or characters. The idea behind this method is that long sentences in the source text are translated with long sentences, while short sentences are translated using short sentences [11, 12]. Lexical matching methods using bilingual dictionaries (such as in the hunalign system [13]) or specific tokens (such as dates, proper nouns, punctuation) as anchors for the alignment [14] are also worth mentioning. On the other hand, MT-based approaches require the source text to be automatically translated into the target language and use a similarity score (e.g. the BLEU metric) to align the machine translation output with the target text sentences; an example of this kind of method is given by Bleualign [15]. The most recent systems, however, are those based on multilingual sentence embeddings, such as Vecalign [16], or sentence-transformers, as Bertalign [17]. Such approaches have been tested on literary texts obtaining good performances [18].[2]

In this paper we present a manually created bilingual (IT-EN) directional parallel corpus of historical literary texts together with the evaluation of automatic sentence alignment methods. Dealing with texts written in not contemporary languages and of a literary genre is particularly interesting and not so widespread; suffice it to say that the CLARIN infrastructure gives access to 87 parallel corpora:[3] out of these, only 5 include texts in Italian, but none contain works by Manzoni or historical literary translations.

[2]Results obtained on literary and non-literary texts using various methods, including the Vecalign and Bertalign systems, are reported in https://github.com/bfsujason/aligner-eval.

[3]https://www.clarin.eu/resource-families/parallel-corpora.

3. Dataset Creation and Analysis

3.1. Creation

The digital text of the Ventisettana was provided by the Italian project (PRIN 2017) *ManzoniOnline2: new documents, translations and tradition* [19],[4] whereas the text of the 1834 English translation was downloaded from the Gutenberg project website as UTF-8 text file.[5] Both texts have been divided into chapters; for each of them the sentence-level alignment was completed semi-automatically, with manual correction of the output of the aligner. In the initial phase of our work we tested various tools which include graphical user interfaces for editing the automatic alignment, such as TAligner 3.0 [20], LF Aligner[6] and InterText. More specifically, as stated in [21], extensive trials were conducted with LF Aligner, before the final choice fell on InterText because of the intuitiveness of its interface and the possibility of exporting in various formats [22].

Each chapter was loaded onto InterText in a separate file with one sentence per line. Sentence splitting was done manually: we tried various sentence splitting models but always obtaining low performances due to the peculiarity of the novel's punctuation and to an unconventional use of capital letters. Among the Universal Dependencies (UD) 2.10 models available in UDPipe [23], the best result was obtained with VIT with an accuracy of 39%. A better, but far from perfect, accuracy (64%) was registered with Stanza [24]. Overall, it can be remarked that automatic sentence splitting fails especially (but not exclusively) with punctuation marks that are no longer in use or with traditional punctuation marks employed in unusual contexts compared to today's custom. In particular, the use of hyphens – short and long – with different functions is very frequent in the 1834 English translation. Normally, the latter separate one sentence from the other, mostly marking the end of a direct speech,[7] while the former convey a character's inner thoughts, include an aside, render a hesitation in direct speech or mark a pause of medium intensity without giving rise to a new sentence.[8] Automatic splitting also displays glitches when dealing with inverted commas marking the start of a direct speech, the three suspension dots, and

[4]https://www.alessandromanzoni.org/

[5]https://www.gutenberg.org/ebooks/35155.

[6]https://sourceforge.net/projects/aligner/.

[7]For instance: "But, fair sirs, you are too just, too reasonable—-" "But," interrupted the other comrade... (from chapter 1).

[8]Here are some examples, all taken from chapter 1: "for if you do, ehem!–you understand–the consequences would be the same as if you performed the marriage ceremony"; "the poor curate neither meddles nor makes–they settle their affairs amongst themselves, and then–then, they come to us, as if to redeem a pledge; and we–we are the servants of the public"; "but he will require reasons–and what can I say to him"; "... and he arose, continuing–"No! I'll take nothing, nothing?".

Table 1
Examples of different types of alignment.

	IT	EN
2:1	I due sposi rimasti promessi si trovarono in faccia Agnese, che arrivava trambasciata e affannosa.	The two lovers (still lovers) stood before Agnes, who, alarmed and grieved, said, “Ah! you are here!
	«Ah siete qui!» diss’ella traendo la parola a stento.	
1:2	«Come è andata? che cos’è la campana?	How has it gone?
		Why did the bell ring?”
1:0	Mi par d’avere inteso...	
1:1	«A casa, a casa,» diceva Renzo, «prima che venga la gente.»	“Home, home!” said Renzo, “before the people gather.”

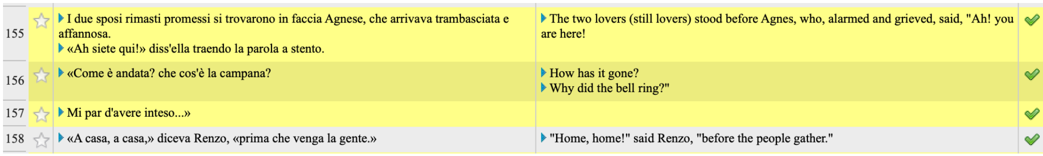


Figure 1: Alignments as displayed in the InterText interface: sentences without a 1:1 alignment are highlighted in yellow.

exclamation or question marks followed by a lower-case letter (which do not start a new sentence but denote a single flow of text). At the end of the manual sentence splitting procedure, we obtained 8,718 sentences for the Ventisettana and 7,484 sentences for the English translation.

In the following phase, we manually corrected the automatic alignment made by hunalign system integrated in InterText. On average, 3 hours of work were required for validating each chapter. Texts were then exported in three files: each chapter was saved as two independent XML files (one for the Italian text and one for the English translation) and their alignment was exported as a separate XML file containing pointers to the individual sentences of the two texts.

3.2. Analysis

The alignments produced can be categorized into the following different types:

- 1:1, i.e. one sentence is translated by one sentence. It should be noted that such correspondence is not necessarily a symptom of total fidelity, or rather of a linear (or even literal) translation of the subphrasal units. While respecting the boundaries of the sentence, in fact, there could be phenomena of expansion or synthesis. For example, in chapter VIII, a long sentence – with a simile used to indicate how the *Bravi* (hired assassins) were gathered in a courtyard by their leader emphasizing their animal nature – is strongly synthesized by removing the rhetorical figure altogether.
  - Ventisettana: *Come il cane che scorta un gregge di porci corre or qua or là a quei che si sbandano, ne addenta uno per un’orecchia e lo tira in ischiera, ne spinge un altro col muso, abbaia ad un altro che esce di fila in quel momento, così il pellegrino acciuffa uno di coloro che già toccava la soglia e lo strappa indietro, caccia indietro col bordone uno e un altro che v’eran già presso, grida agli altri che scorrazzano senza saper dove, tanto che li raccozzò tutti nel mezzo del cortiletto.*[9]
  - 1834 English translation: *He succeeded, however, in assembling them in the middle of the court-yard.*
- 1:0 and 0:1, i.e. a sentence in the Ventisettana or in the translation lacks a parallel in the other text, following an omission (type 1:0) or an addition by the translator (type 0:1). Omissions are part of a wider trend in the historical translations of Manzoni’s novel to significantly cut sentences that were considered not essential for understanding the text. This is aimed at giving the translation a drier and more pragmatic tone than the original, in line with the prevailing fashions in the literary context of reception; such approach is consistent with the so-called domestication strategy of translations [25].

[9] English literal translation: *Like the dog that escorts a herd of pigs, he runs here and there among those who are straying, he bites one by the ear and puts him in line, he pushes another with his muzzle, he barks at another who leaves the line at that moment, so the pilgrim grabs one of those who were already on the doorstep and snatches him back, he drives one and another who was nearby back with his stick, he shouts to the others who are running around without knowing where, so much so that he gathered them all in the middle of the little courtyard.*

Table 2
Number of alignments per type.

TYPES	1-1	1-0	2-1	1-2	3-1	1-3	OTHER
NUMBER OF ALIGNMENTS	5,068	1,077	693	597	106	84	108

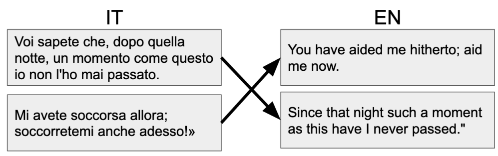


Figure 2: Example of cross-order alignment.

- 1:N and N:1, i.e. the translator has split or merged the original sentences. When one Italian sentence is split into two or more sentences the alignment is 1:N. When, on the contrary, two or more Italian sentences are merged in a single sentence in the translation the alignment is N:1.

Table 1 provides examples, taken from chapter VIII, of the aforementioned types, while Figure 1 shows how the same alignments are displayed in InterText interface. In addition, Table 2 presents the number of alignments per type. The vast majority of alignments are 1:1 (66%), but there are also several omissions in the translation (1:0, 14%), followed by cases of 2:1 merging (9%) and 1:2 splitting (8%). Under the "Other" category we collect the types having a number of occurrences less than 1% (i.e. 0-1, 4-1, 1-4, 5-1, 6-1, 3-2, 1-5). It is important to notice that our resource includes few cases of cross-order alignments in which the translator has changed the order of the sentences in the translation so that, to create the alignment, it is necessary to move sentences out of their original position (which is possible with InterText). Cross-order alignments fall into the types described above: for example, Figure 2 shows a cross-order alignment, taken from chapter XXXVI, which generates a 1:1 match between the source and the target sentences.

4. Testing Automatic Alignment Methods

The parallel corpus described in the previous section has been used as gold standard for testing the performances of Bertalign, an automatic aligner that uses LaBSE (language-agnostic BERT sentence embeddings, [26]) for building cross-lingual embeddings of source and target sentences.[10] As reported by Liu and Zhu [17], Bertalign is designed with the aim of dealing with non-1-to-1 sentence pairs that are quite common in literary texts. The comparative evaluation carried out on literary texts considering the English-Chinese translation pair showed that Bertalign is able to outperform other (length-based, dictionary-based, MT-based and embedding-based) aligners.

We configured Bertalign with the following options:

- maximum alignment types (`max_align`): 6
- k nearest target neighbors of each source sentence (`top_k`): 3
- search window (`win`): 5
- similarity score for 1:0 and 0:1 alignments (`skip`): 0
- modified cosine similarity as proposed in [17] (`margin`): True
- length difference between source and target sentences (`len_penalty`): False
- sentence splitting (`is_split`): True

With respect to the default configuration, we increased the maximum alignment length (i.e. the `max_align` option) from 5 to 6 because our corpus has many complex alignments, that is various types of 1:N and N:1 alignments. We also set a larger value for the similarity score (i.e. the `skip` option) because our corpus contains many omissions and insertions. Given that we have several cases of expansion or synthesis even in 1:1 alignments, the `len_penalty` parameter is set to False: in this way the length difference between source and target sentences is not taken into consideration when calculating the similarity between sentence pairs. On the contrary, the `is_split` option is set to True because our corpus was already split into sentences.

Table 3 reports the results of our evaluation using both Bertalign (with the default configuration, *Bertalign_d*, and with our custom options, *Bertalign_c*) and the Galechurch length-based algorithm. The superiority of the embedding-based approach over the length-based one is evident: the former outperform the latter by 5 F1 points. The custom configuration further improves Bertalign's performance in terms of both precision and recall. However, the results are slightly lower than those recorded on the English-Chinese pair: indeed, for the MAC corpus of literary texts a precision of 0.906, a recall of 0.912 and an F1 of 0.909 are reported.[11]

Figure 3 displays F1 performance across the chapters of the novel. The variation between individual chapters

[10] https://github.com/bfsujason/bertalign

[11] https://github.com/bfsujason/aligner-eval.

Table 3
Automatic alignment quality with both the default and custom configurations of Bertalign (Bertalign_d and Bertalign_c, respectively) and the Galechurch length-based algorithm.

	Bertalign_c	*Bertalign_d*	*Galechurch*
P	**0.888**	0.862	0.427
R	**0.905**	0.857	0.368
F1	**0.896**	0.859	0.396

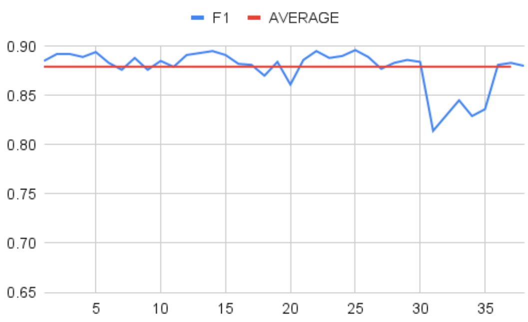


Figure 3: F1 score (on the y axis) by chapter (on the x axis): to facilitate the reading of the chart, the vertical axis has been set to 0.65. The average line is displayed in red.

is not great, with an average F1 of 0.879. However a drop can be noted in the range between chapters 31 and 35 which describe the plague in Milan with numerous historical digressions, often not translated. In particular, chapters 31 and 32 of the original text are merged into a single chapter in the translation in which there is a high number of omissions covering 33% of all the alignments. In addition, that group of chapters includes cross-alignments that are not correctly handled by Bertalign. On the contrary, the best F1 (0.896) is found for chapter 25 in which 1:1 alignments, the simplest type, are 73% of the total.

5. Conclusion and Future Work

This paper described the creation of a parallel corpus aligned at sentence level made of the whole text of Ventisettana, that is the version of the novel published by Manzoni in 1827, and the 1834 anonymous English translation. This resource is made available on Github in XLM format[12] and will be also uploaded in the ILC4CLARIN repository. The whole aligned corpus has been used as gold standard for evaluating Bertalign, an embedding-based automatic sentence aligner. Results obtained with a custom setting of the parameters are compared to the ones achieved with the default options and with a length-based algorithm (Galechurch) showing very good performances, with an F1 slightly below 0.9.

The activity presented here served as a laboratory for future experiments which will concern the other editions of the novel and the main translations into neo-Romance languages. In particular, a sentence level alignment activity of chapter VIII is underway taking into account the largest possible number of available English translations also considering, thanks to an agreement with the translator, the very recent American translation of the novel [27]. The choice of maintaining the sentence unity in the Italian text will facilitate the comparison between different translations and, consequently, investigations on the choices made by the translator in a diachronic perspective.

The alignment at the word level of some chapters of the Ventisettena with the English edition of 1834, already adopted for the sentence level alignment, is also in progress. In this case, the alignment is done using Ugarit [28].[13] Unlike what has been done in other projects [29], in our project the aim of the alignment does not concern the creation of a translation memory for machine translation purposes, but the analysis of the choices made by the translator: for this reason, the alignment is performed considering punctuation and also between linguistic elements whose literal correspondence is rather fuzzy. This choice makes it possible to highlight oversights, errors and singular innovations of the translator. The output of our manual alignment will be used to evaluate automatic approaches, such as fast_align[14] and AWESOME[15].

Acknowledgments

Questa pubblicazione è stata realizzata da ricercatrice con contratto di ricerca cofinanziato dall'Unione europea - PON Ricerca e Innovazione 2014-2020 ai sensi dell'art. 24, comma 3, lett. a, della Legge 30 dicembre 2010, n. 240 e s.m.i. e del D.M. 10 agosto 2021 n. 1062. Questa ricerca è stata anche finanziata dall'Università degli Studi di Parma attraverso l'azione Bando di Ateneo 2022 per la ricerca co-finanziata dal MUR-Ministero dell'Università e della Ricerca - D.M. 737/2021 - PNR - PNRR - NextGenerationEU.

References

[1] P. Bellezza, Il Manzoni all'estero, in: Curiosita' manzoniane, Vallardi, Milano, 1923, pp. 57–73.

[12] https://github.com/RacheleSprugnoli/Sentence_Alignment_Manzoni.

[13] https://ugarit.ialigner.com.

[14] https://github.com/clab/fast_align.

[15] https://github.com/neulab/awesome-align.

[2] P. Bellezza, Attraverso le traduzioni dei "Promessi sposi", in: Curiosita‘ manzoniane, Vallardi, Milano, 1923, pp. 75–95.
[3] G. Getto, Manzoni europeo, Ugo Mursia, Milano, 1971.
[4] P. Frare, Manzoni europeo?, Nuovi quaderni del Crier. I "Promessi sposi" nell'Europa romantica 9 (2012) 199–220.
[5] V. Intonti, R. Mallardi, Cultures in contact: Translation and reception of i promessi sposi in 19th century england, 2011.
[6] T. R. Steiner, English translation theory 1650-1800, 2, Rodopi, 1975.
[7] M.-A. Lefer, Parallel corpora, in: A practical handbook of corpus linguistics, Springer, 2021, pp. 257–282.
[8] J. Tiedemann, Parallel data, tools and interfaces in OPUS, in: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey, 2012, pp. 2214–2218. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf.
[9] R. Steinberger, M. Ebrahim, A. Poulis, M. Carrasco-Benitez, P. Schlüter, M. Przybyszewski, S. Gilbro, An overview of the european union's highly multilingual parallel corpora, Language resources and evaluation 48 (2014) 679–707.
[10] Y. Xu, A. Max, F. Yvon, Sentence alignment for literary texts: The state-of-the-art and beyond, in: Linguistic Issues in Language Technology, Volume 12, 2015-Literature Lifts up Computational Linguistics, 2015.
[11] P. F. Brown, J. C. Lai, R. L. Mercer, Aligning sentences in parallel corpora, in: 29th Annual Meeting of the Association for Computational Linguistics, 1991, pp. 169–176.
[12] W. A. Gale, K. W. Church, et al., A program for aligning sentences in bilingual corpora, Computational linguistics 19 (1994) 75–102.
[13] D. Varga, P. Halácsy, A. Kornai, V. Nagy, L. Németh, V. Trón, Parallel corpora for medium density languages, volume 292, Amsterdam; Philadelphia; J. Benjamins Pub. Co, 2007, p. 247.
[14] M. Kay, M. Roscheisen, Text-translation alignment, Computational linguistics 19 (1993) 121–142.
[15] R. Sennrich, M. Volk, Mt-based sentence alignment for ocr-generated parallel texts, in: Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Research Papers, 2010.
[16] B. Thompson, P. Koehn, Vecalign: Improved sentence alignment in linear time and space, 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), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 1342–1348. URL: https://www.aclweb.org/anthology/D19-1136. doi:10.18653/v1/D19-1136.
[17] L. Liu, M. Zhu, Bertalign: Improved word embedding-based sentence alignment for chinese–english parallel corpora of literary texts, Digital Scholarship in the Humanities 38 (2023) 621–634.
[18] E. Signoroni, Evaluating the state-of-the-art sentence alignment system on literary texts., in: Recent Advances in Slavonic Natural Language Processing (RASLAN 2021), 2021, pp. 115–124.
[19] G. Raboni, «manzonionline»: considerazioni in corso d'opera, Griseldaonline 20 (2021) 149–155.
[20] Z. S. Villar, O. A. Pinedo, Taligner 3.0: A tool to create parallel and multilingual corpora, in: Corpora in Translation and Contrastive Research in the Digital Age: Recent advances and explorations, John Benjamins, 2021, pp. 125–146.
[21] R. Sprugnoli, A. Redaelli, M. Sartor, Risorse linguistiche per lo studio dei "Promessi sposi", in: La memoria digitale. Forme del testo e organizzazione della conoscenza. Atti del XII convegno annuale AIUCD (Siena, 5-7 giugno 2023), Università degli Studi di Siena, Siena, 2023, pp. 301–303.
[22] P. Vondřička, Aligning parallel texts with InterText, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), European Language Resources Association (ELRA), Reykjavik, Iceland, 2014, pp. 1875–1879. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/285_Paper.pdf.
[23] M. Straka, Udpipe 2.0 prototype at conll 2018 ud shared task, in: Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, 2018, pp. 197–207.
[24] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, C. D. Manning, Stanza: A Python natural language processing toolkit for many human languages, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2020. URL: https://nlp.stanford.edu/pubs/qi2020stanza.pdf.
[25] J. Munday, S. R. Pinto, J. Blakesley, Introducing translation studies: Theories and applications, Routledge, 2022.
[26] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, W. Wang, Language-agnostic BERT sentence embedding, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 878–891. URL: https://aclanthology.org/2022.acl-long.62. doi:10.18653/v1/2022.acl-long.62.
[27] A. Manzoni, The Betrothed: A Novel, Modern Li-

brary, 2022. Translated by Michael F. Moore.
[28] T. Yousef, C. Palladino, F. Shamsian, M. Foradi, Translation alignment with ugarit, Information 13 (2022) 65.
[29] T. Yousef, C. Palladino, F. Shamsian, A. d. Ferreira, M. F. dos Reis, An automatic model and gold standard for translation alignment of ancient greek, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 5894–5905.

Qualitative Analysis of Persuasive Emotion Triggering in Online Content

Olga Uryupina[1]

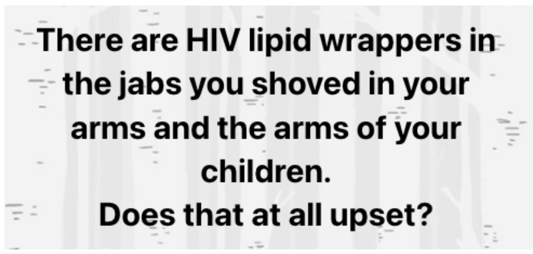
[1]Department of Information Engineering and Computer Science, University of Trento

Abstract
This paper presents a qualitative analysis of the emotional component in manipulative online content (fakes). We show that emotion triggering is a crucial persuasion technique widely employed by unscrupulous content generators. Based on a dataset of real-life fakes analyzed by fact-checking professionals, we identify the most common types of triggered emotions to be used as a taxonomy for further annotation.

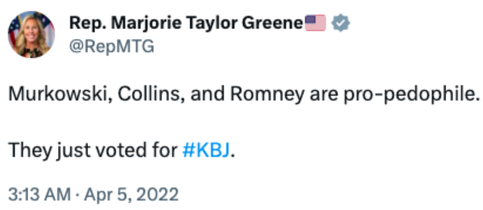
Keywords
persuasion, fact-checking, sentiment analysis

1. Introduction

The manipulative content, ranging from propaganda to hate campaigns, fake news, trolling and similar, is becoming more and more widespread, threatening our access to truthful and unbiased information and thus undermining our rights to make informed decisions as individuals and as members of the society. While there is a growing body of multidisciplinary research on identifying untruthful content, there is still very limited understanding of the manipulative techniques the unscrupulous content writers employ to convince the reader and ultimately change their point of view. We believe that this manipulation occurs through multiple channels: careful selection of fact-checkable and non-fact-checkable claims, biased yet seemingly solid argumentation/analytics, multimedia support and, most importantly, emotional component. Our current study focuses on *emotion triggering* – a technique widely used by content writers: when the reader is experiencing a strong feeling, they become less critical and thus easily overlook deficiencies in the argumentation and get more prone to manipulation.



(a) from Facebook



(b) from Twitter

Figure 1: Emotion triggering in manipulative content.

Fig. 1 shows examples of manipulative textual content with strong emotional triggering. In (1a), the message makes a very strong appeal to fear, by mentioning HIV. Moreover, this triggering effect is intensified by mentioning "children". The fact-checking report [1] informs the reader that the COVID-19 vaccines do not contain any HIV material, but do contain other lipids to protect the mRNA. The distressed users, however, might not trust this information fully, due to such a strong emotion as a fear for their children's health. Example (1b) shows a typical manipulative message not addressed properly by the state of the art verification-oriented technology. The message combines a verifiable true claim ("Murkowski, Collins, and Romney voted for Ketanji Brown Jackson") with a statement that looks like a similarly factual claim ("Murkowski, Collins, and Romney are pro-pro-pedophile"), but in reality is an explanation/opinion offered by the writer. This triggers a rather strong anger at the powers/authorities under the spotlight, their presumed hypocrisy and their presumed (lack of) values. Here again, the triggering is intensified by bringing up a topic related to children. The fact-checking report [2] debunks this claim stating that "Sens. Murkowski,

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
uryupina@gmail.com (O. Uryupina)

[1]https://www.politifact.com/factchecks/2022/apr/12/facebook-posts/no-covid-19-vaccines-do-not-contain-hiv-lipid-wrap/

[2]https://www.politifact.com/factchecks/2022/apr/06/marjorie-taylor-greene/greene-twists-logic-and-facts-pedophilia-charge-a

Collins and Romney have clear track records of acting against child exploitation, whether online or in person" and, moreover, the implied related accusations of Judge Johnson are "misleading". However, a reader driven by emotions, might still remain manipulated ("no smoke without fire"), even if only partially.

These examples show that fake news are way more complex than simply untrue messages. They might combine true facts with partially false or impossible to check statements, provide biased analytics on top and add very strong emotional messages to manipulate readers. We believe that while the NLP community is making an impressive progress on the fact verification task, our understanding of other phenomena related to manipulative content are still rather limited. The goal of our study is to get a deeper and more realistic insight on the emotional component of fakes. As a first step, we provide a qualitative data-driven analysis of emotion triggering.

The contributions of this study are as follows: (i) we provide data-driven analysis, focusing on real data, combining original (source) fakes and high-quality reports by professional fact-checkers thus improving our insight, (2) we aim at a taxonomy of triggered emotions covering a majority of real-life fakes, departing from more theory-oriented labels and (3) we analyze perceived (i.e., triggered) emotions, as opposed to the common focus on expressed emotions, as we believe that induced sentiment plays a more important role in manipulation/persuasion.

2. Related Work

There is a rapidly growing body of studies on online misinformation detection. These works, however, mainly focus on the verification part (*Is* the information truthful – i.e., supported by the evidence?), and not on the persuasion (*How* is the information presented to manipulate the reader?). Thus, most computational models are built upon the FEVER corpus [1]: a large collection of true/false claims generated by human annotators, annotated as supported/refuted/unknown by the evidence. FEVER claims are originally extracted from Wikipedia (true) and then mutated (false). An example FEVER claim is "Shakira is Canadian". Note a strong difference between this example and (1a-b) above: the Shakira claim was generated with no manipulative purpose in mind and does not involve any specific persuasion/triggering techniques. The claims in (1), on the contrary, have a strong manipulative component and have been generated with a genuine unscrupulous intent. For example, (1a) cannot be fully accounted for by a simple mutation: the choice of "HIV" is crucial to induce fear and thus the same manipulative effect would not be achieved if "HIV lipids" were replaced with any other kind of lipids. In our study, we focus on real-world data, analyzing fakes generated with a real purpose, albeit not always clear (and not necessarily malicious).

Giachanou et al. [2] address the impact of emotional signals on the credibility for fake news. This study shows that emotional signals are extremely important as the emotion-aware system outperforms their baseline by a large margin. This work, however, focuses on already existing generic resources for defining emotions: either lexicons of terms expressing specific sentiments or a corpus of triggered sentiments with labels corresponding to five different Facebook reactions (love, joy etc). We believe that findings of Giachanou et al. [2] are extremely important and show that emotions triggered by manipulative content should be studied in a more principled way. We hope that our study could help define a more triggering-oriented approach to emotions.

Several recent papers analyze emotion triggering as a part of propaganda persuasion techniques. For example, Da San Martino et al. [3] develop a taxonomy of propaganda techniques, whereas Piskorski et al. [4] propose a shared task build upon this taxonomy. These studies do not, however, focus on emotions specifically. For example, Piskorski et al. [4] group most emotions under the "manipulative" category, while some others (e.g., "appeal to patriotism/pride" also known as "flag-waving") are classified based on reasoning fallacies associated with them. Moreover, these studies focus on unscrupulous persuasion techniques introduced in the theoretical studies, e.g., on (in)formal argumentation fallacies. We advocate a more data-driven approach: the phenomenon of manipulative online content is rather new and evolving, thus, it is not clear how well more traditional labels describe it. We aim at decoupling emotions from (fallacious) argumentation and improving our insight into the variety of sentiments the content writers appeal to.

Finally, some of the discussed triggers, especially "fear", have been a focus of multidisciplinary studies, ranging from psychology (see an overview in [5]) to ethics [6]. At the same time, there exist much less research on more complex triggers.

3. Data

Our study aims at a qualitative analysis with the end goal of developing reliable annotation guidelines that provide good coverage for triggered sentiments. We have therefore opted for in-depth analysis of a small number of documents. Our analysis relies on both the documents themselves and their corresponding fact-checking reports by PolitiFact. This way, we make sure that we ourselves do not fall victim to the manipulation techniques and can assess them impartially.

We rely on PolitiFact reports from mid-March to mid-

g/

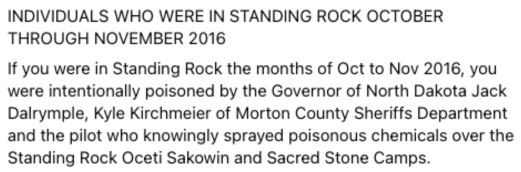


Figure 2: Appeal to fear, from Facebook.

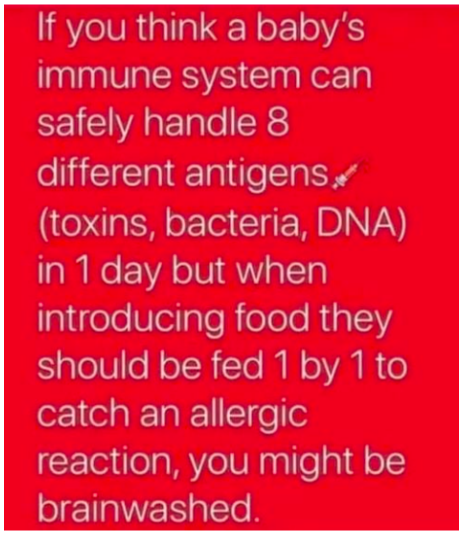


Figure 3: Anti-bandwagon (appeal to uniqueness), from Instagram.

May 2022. We filter out fakes that originate on TV, intervies and other sources ourside of social media. This leaves us with 160 "claims", each associated with their corresponding social media post and high-quality PolitiFact report, written by professional fact-checkers. We then annotate them with metadata, overall professional fact-checking judgement, atomic fact veracity, reasoning flaws (e.g., "simplification") and, most importantly, triggered emotions. The latter is done in data-driven bottom-up fashion, with the set of considered emotions under constant refinement.

4. Appealing to Emotions

In this section, we discuss the emotions triggered in manipulative online messages. We start with commonly acknowledged and studied triggers, such as "fear" and expand the label set to accommodate data-driven categories not sufficiently covered in the literature.

Appeal to Fear is the most studied and widely used manipulative technique: by making the readers believe that they are in imminent personal danger, the author can influence their attitude toward the message, suppress critical thinking, instill doubt and ultimately manipulate their behavior. There are multiple studies showing the efficacy of this persuasion technique, see [5] for an overview. From the data-driven perspective, however, it is not always easy to define the boundaries of "personal danger".

Thus, our example (1a) shows a clear case of appeal to fear, since the governments' policies strongly suggest all the population to be vaccinated. Consider our example in Figure 2. This post informs a rather limited group of people of the alleged imminent danger, thus inducing fear. However, when going viral, it might have a fear-triggering effect on the whole population, stating that the authorities are able to and, in practice, do employ carcinogenic chemicals against humans.

Bandwagon and Anti-bandwagon. Another relatively widely studied technique is an appeal to common practice/belief ("safe choice"), also known as "bandwagon fallacy". This technique urges the reader to adopt specific choices, because everybody is doing so. For example, bandwagon is commonly used in advertisement, where a lot of products are marketed as a must since everybody buys them. Surprisingly, we haven't found a single example of an appeal to common practice in manipulative online content in our data. However, we have observed the opposite appeal: the authors urges the reader *not* to follow the common practice, appealing to their uniqueness and superiority.

Figure 3 shows a very common example of appeal to uniqueness/superiority: the authors state that while most people are brainwashed by mainstream information channels and left to believe in some fake reality, the readers should – and is definitely capable of – avoid falling for the same trap. This boosts the readers' ego, improves their trust in fake news while, at the same time, undermines mainstream media and paves the path for various conspiracy theories. We have observed this opinion framing strategy on a variety of polarized topics, ranging from vaccination to government spending or climate.

While this direct appeal to readers' uniqueness/ego is very widespread and seemingly rather effective, we are not aware of any in-depth studies of this phenomenon, especially from the NLP perspective.

Appeal to Populism is an emotionally-loaded technique triggering strong antagonising feelings between "us" ("the good people") and "them" ("the corrupt powers: government, rich, media etc"). Populism plays an ever rising role in the modern political discourse, affecting and polarizing people's views. While it is widely studied in

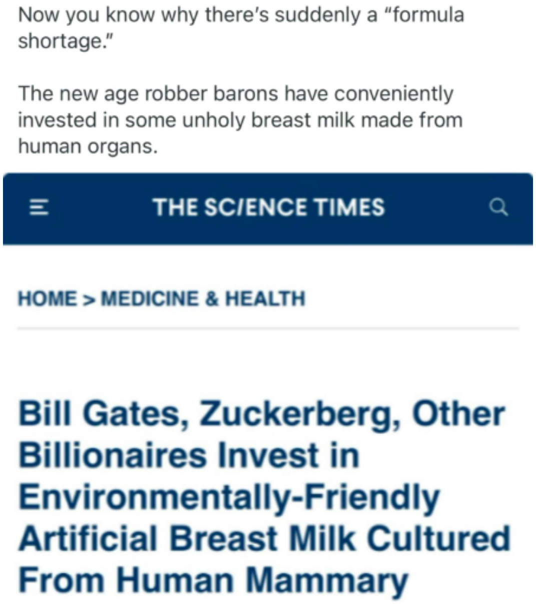
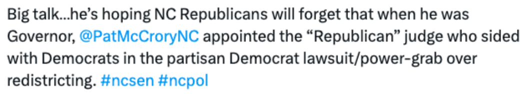


Figure 4: Appeal to populism, from Facebook.

Figure 5: Appeal to unfairness, from Facebook.



(a) attacking a specific person, Twitter

(b) undermining trust ("everybody lies"), Facebook

Figure 6: Appeal to honesty.

political science, the related psychological mechanisms are still underresearched [7]. We have observed multiple cases of appeal to populism throughout the data.

Thus, in a Facebook post on Figure 4, the author makes it pretty clear that the rich are responsible for and benefiting from the suffering of "us" – in this specific case, the formula milk crisis. The same strategy is used throughout our data to implicate different kinds of powers: the administration, the rich or the media and sometimes a mixture or just a generic/underspecified "power". The appeal to populism is often combined with other emotions: for example, triggering the fear or unfairness/injustice for the outcome of "their" actions as well as uniqueness/ego for uncovering the plot.

Appealing to (Un-)Fairness is a very strong technique, often used in combination with appealing to populism (see an example on Figure 5).

In some cases, the authors trigger this sentiment in a positive way, inviting the reader to celebrate the victory of fairness.

In both cases, however, the content writers trigger a very strong and deep desire for (social) justice, that deflecting the readers' attention from inconsistencies and misrepresentation in the presented facts and arguments.

To our knowledge, appealing to fairness is acknowledged as a powerful technique by a variety of practicing professionals, e.g., negotiators or copywriters. However, there is still virtually no research on this specific emotion. We believe that since this is one of the most frequent and efficient triggers in manipulative content, an urgent attention from the research community, including NLP, might have a considerable impact and help fight online misinformation.

Appeals to honesty are very popular in manipulative content. This category includes allegations of hypocrisy, inconsistency or accusations of lying, aimed at casting a doubt on specific persons (Figure 6a).

However, a far more widespread appeal to honesty is the technique where some information coming from mainstream media or official sources is presented as a lie,

Figure 7: Appeal to values, from Facebook.

with no clear and specific purpose (Figure 6b). This type of fakes promote the idea of everything being unreliable and slowly but steadily push the readers to become less critical of various conspiracy theories.

Values. Certain online posts make appeal to values, promoting responsible choices or condemning someone else's behavior as unethical. This type of triggering is often used in polarized contexts to attack the opposite side and thus misrepresent their position (Figure 7).

Appeals to values are often used as a part of the reduction/simplification fallacy: the fact-checkable facts in the message are true (e.g., the statement above is focused on "A National Terrorism Advisory System bulletin", addressing the threats of online misinfromation), yet their interpretation is fallacious and manipulative, introducing loaded lexica ("attack", "criminalize") to misrepresent these facts, substituting objective reporting with moralistic judgement. This type of fakes are therefore particularly problematic for state-of-the-art NLP models, based on fact verification.

Disasters. We have observed a large number of fakes focusing on natural and man-made disasters. Media coverage of disasters has been shown to attract a large number of readers/viewers, triggering a wide variety of interrelated negative emotions, in particular fear and anxiety [8]. Unscrupulous content generators abuse the users' interest in catastrophic events for their own purposes (e.g. click-bait). We label this specific type of fear/anger as "disaster" for the lack of better term, since a more precise analysis is still an open research issue in psychology.

5. Emotions in Fakes

In this section, we discuss the distribution of triggered emotions in the manipulative content collected and analyzed by PolitiFact. Most importantly, we have observed that a vast majority of fakes trigger emotions: 128 documents (80%) in our collection unambiguously aim at affecting the readers' emotional state. For comparison, only 88 documents (55%) contain clearly untrue atomic facts and 95 documents (59%) employ fallacious argumentation. We believe, once again, that these numbers suggest that the efficient approach to manipulative content analysis should expand from mere fact verification to modeling fallacious argumentation and emotion triggering.

Trigger	#documents	%
populism	62	38.7
fear (personal)	18	11.3
fear (empathy)	16	10
fairness	27	16.9
honesty	22	13.8
values	15	9.4
uniqueness	18	11.25
disaster	8	5
other	6	3.8

Table 1
Triggers in the PolitiFact data.

Table 1 shows the document statistics for each of the triggers discussed in this section. The most common category is *populism*, which might be due to the political orientation of our domain. Note that populism is also relatively easy to identify: our preliminary experiments show very little disagreement on this label. Appeal to *fear* is the second most popular category: unscrupulous content writers are well aware of its efficiency. Annotating it reliably, however, requires extra work on guidelines, since the boundaries between personal fear and empathy for others are very subjective. Depending on the definition of fear, we observe 11-21% of such documents. Fairness, honesty and values are also rather common. Finally, only 6 documents (4%) appeal to other emotions that are not covered by our taxonomy.

The same post can trigger multiple emotions. In particular, appeals to populism ("they are bad") are often combined with any other trigger ("they are bad: they are threatening our existence, imposing unfair policies and lying"). A rather common combination throughout all the fakes we have analyzed is "*they* (the media/administration) are lying, but *you* are smart and *you* don't believe them, *we* will tell you the truth" (anti-bandwagon + honesty + populism). Note that this trigger makes it very difficult to respond to and counter the effect of manipulative content: if the readers are convinced that "they" are lying, they can simply discard a fact-checking report since they perceive fact-checkers as liars (paid by "them") or, at the very least, brainwashed (by "them").

6. Conclusion

This study focuses on emotional component of manipulative online content. Analyzing real-life fake content from PolitiFact, we have observed a variety of emotional triggers used to promote unscrupulous content by agitating the users and making them less critical of the deficiencies in the fact selection and argumentation of the manipulative discourse.

We have seen that emotions play a crucial role in pushing through different kinds of manipulative agenda and it is therefore extremely important for the scientific community to extend state-of-the-art verification-based approaches to fact-checking and incorporate models for emotion triggering and fallacious argumentation.

Our study identifies the most common types of emotions triggered by manipulative content. However, defining them accurately is not a trivial task, as we have already observed with *fear*. Our current work focuses on refining the definitions of the most common triggers to provide reliable annotation guidelines and create a dataset of appeals.

Triggered emotions (reactions) have so far mostly been out of the scope of the NLP community, where the vast body of research is focused on emotions *expressed* in the document. We believe that our research can contribute to a better understanding of perceived emotions, crucial for modelling a text's impact on the reader. In particular, we plan to study the relation between expressed and triggered emotions and investigate possibilities of transferring high-performing state-of-the-art (expressed) emotion recognition models to account for triggered emotions.

Finally, we believe that multi-factor understanding of manipulative content is essential to generate adequate response and thwart the misinformation. Emotionally-loaded fakes are particularly hard to debunk since they render the user less receptive to the rational argumentation of fact-checkers. As a part of our future work, we want to investigate strategies for automatic response generation that take into account the emotional component and try to produce an adequate reaction, regaining the users' trust.

Acknowledgments

We thank the Autonomous Province of Trento for the financial support of our project via the AI@TN initiative.

References

[1] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, FEVER: a large-scale dataset for fact extraction and VERification, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 809–819. URL: https://aclanthology.org/N18-1074. doi:10.18653/v1/N18-1074.

[2] A. Giachanou, P. Rosso, F. Crestani, The impact of emotional signals on credibility assessment, Journal of the Association for Information Science and Technology 72 (2021) 1117–1132.

[3] G. Da San Martino, S. Yu, A. Barrón-Cedeño, R. Petrov, P. Nakov, Fine-grained analysis of propaganda in news article, 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), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 5636–5646. URL: https://aclanthology.org/D19-1565. doi:10.18653/v1/D19-1565.

[4] J. Piskorski, N. Stefanovitch, G. Da San Martino, P. Nakov, SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup, in: Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 2343–2361. URL: https://aclanthology.org/2023.semeval-1.317.

[5] M. Tannenbaum, J. Hepler, R. Zimmerman, L. Saul, S. Jacobs, K. Wilson, D. Albarracín, Appealing to fear: A meta-analysis of fear appeal effectiveness and theories, Psychological Bulletin 141 (2015) 1178–1204.

[6] D. Arthur, P. Quester, The ethicality of using fear for social advertising, Australasian Marketing Journal (AMJ) 11 (2003) 12–27. URL: https://www.sciencedirect.com/science/article/pii/S1441358203701153. doi:https://doi.org/10.1016/S1441-3582(03)70115-3, social Marketing.

[7] S. Obradović, S. Power, J. Sheehy-Skeffington, Understanding the psychological appeal of populism, Current Opinion in Psychology 35 (2020) 125–131.

[8] B. Pfefferbaum, E. Newman, S. Nelson, P. Nitiéma, R. Pfefferbaum, A. Rahman, Disaster media coverage and psychological outcomes: descriptive findings in the extant research., Current Psychiatry Reports 16 (2014) 1178–1204.

When You Doubt, Abstain: A Study of Automated Fact-checking in Italian Under Domain Shift

Giovanni Valer[1,*], Alan Ramponi[2] and Sara Tonelli[2]

[1] *University of Trento, Department of Information Engineering and Computer Science – Trento, Italy*
[2] *Fondazione Bruno Kessler (FBK), Digital Humanities Unit – Trento, Italy*

Abstract

English. Data for building fact-checking models for Italian is scarce, often contains ambiguous claims, and lacks textual diversity. This makes it hard to reliably apply such tools in the real world to support fact-checkers' work. In this paper, we propose a categorization of claim ambiguity and label the largest Italian test set based on it. Moreover, we create challenge sets across two axes of variation: *genres* and fact-checking *sources*. Our experiments using transformer-based semantic search show a large drop in performance under domain shift, and indicate the benefit of models' abstention in case of lacking evidence.

Italiano. *I dati per la creazione di modelli di fact-checking per l'italiano sono esigui, contengono spesso affermazioni ambigue e presentano una limitata diversità stilistica. Questo rende l'uso dei modelli risultanti da parte dei fact-checkers poco affidabile. In questo lavoro classifichiamo l'ambiguità delle affermazioni contenute nel più grande test set di fact-checking per l'italiano e creiamo dei nuovi challenge test set che riflettono stili e fonti differenti. I nostri esperimenti basati sulla ricerca semantica mostrano un notevole calo delle prestazioni in caso di cambio di dominio e indicano l'utilità dell'astensione da parte dei modelli in caso di limitate evidenze.*

Keywords
Automated fact-checking, claim ambiguity, domain shift, models' abstention, semantic search

1. Introduction

Countering the spread of mis/disinformation is one of the major challenges of our society, but human fact-checkers struggle to cope with the increasing amount of content being published. On these bases, in recent years automated fact-checking has gained increasing attention in the NLP community, resulting in a significant body of works and initiatives, e.g., the Fact Extraction and VERification Workshop (FEVER), at its 7th edition in 2023 [1].

Research efforts in NLP for automated fact-checking span over a plurality of tasks, from claim detection to verdict prediction and justification production [2]. Nevertheless, languages other than English, one of them being Italian, are mostly overlooked in current NLP research on the topic. Specifically, little work has been done to build annotated corpora for the Italian language, which is currently included in just a handful of multilingual datasets, i.e., X-Fact [3] and FakeCovid [4]. To exacerbate the problem, most datasets for automated fact-checking not only include underspecified claims for which verdicts are hard-to-impossible to be determined [5], but also typically lack domain diversity, making it difficult to ascertain the reliability of the resulting fact-checking systems when applied on texts reflecting different genres (e.g., from news headlines to posts on social media).

In this paper, we aim to advance automated fact-checking in Italian by examining claim ambiguity in the largest, publicly-available test set to date, and providing means to measure and mitigate the impact of domain shift along *genres* and *sources* dimensions of variation. Our study shows that automated fact-checking in Italian is still far from being reliably applied in the real-world, and indicates the benefit of models' abstention in case of lacking evidence for verification.

Contributions *i)* We propose a categorization of claim ambiguity, *ii)* annotate the Italian test portions of X-Fact according to it, and *iii)* create challenge test sets for studying automated fact-checking in Italian under domain shift. We further *iv)* assess performance shift using transformer-based semantic search, *v)* highlighting the benefit of abstention in the case of insufficient evidence.

2. Fact-checking Data

Among the fact-checking datasets comprising Italian, we select X-Fact [3] for our study since it represents a more diversified set of topics and comprises a larger amount of claims in the Italian language than FakeCovid [4].

X-Fact contains 31,189 non-English textual claims from 25 languages, among which are 1,513 Italian claims based on *Pagella Politica* (PP)[1] and *Agenzia Giornalistica*

Code: https://github.com/jo-valer/fact-checking-ita-abstention

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

*Corresponding author.
giovanni.valer@studenti.unitn.it (G. Valer); alramponi@fbk.eu (A. Ramponi); satonelli@fbk.eu (S. Tonelli)

[1] Pagella Politica website: https://pagellapolitica.it/

Italia (AGI)[2] fact-checks. The original veracity labels for the claims derive from different sources in multiple languages, and therefore have been homogenized by Gupta and Srikumar (2021) [3] to a fixed label set, i.e., *true*, *mostly-true*, *partly-true*, *mostly-false*, *false*, as well as *complicated* for cases whose original labels have been found hard to be mapped to the proposed label set.[3]

The data is structured into training, development, and test splits. Both the training and development sets have been extracted from PP and include 943 and 125 claims, respectively. The test set instead comprises an *in-domain* portion (190 claims from PP) and an *out-of-domain* one (255 claims from AGI). We remove instances marked as *complicated* by Gupta and Srikumar (2021) [3] from the test set since they do not provide any information about claim veracity. As a result, while the *in-domain* test portion remains the same (i.e., 190 claims), the size of the *out-of-domain* test portion decreases to 160 claims due to the filtering of 95 claims (i.e., 37.3%).

3. Annotation and Challenge Sets

In this section, we present our proposed categorization of claim ambiguity and describe the annotation process of the Italian test sets of X-Fact according to it (Section 3.1). Moreover, we detail the creation of challenge test sets aimed at studying the performance of automated fact-checking in Italian under domain shift (Section 3.2).

3.1. Claim Categorization

Textual claims that undergo fact-checking may be hard or even impossible to verify due to ambiguity and underspecified language. When it comes to datasets for automated fact-checking, the additional context that has been used by fact-checkers is typically not included, making labels for claims in such decontextualized conditions to change from concrete verdicts (e.g., *true*, *false*) to being unverifiable [5]. Moreover, claims and associated verdict labels that are derived from fact-checking websites and included in most datasets may cause further ambiguity. Indeed, the claim often corresponds to the headline of the article describing the statement that has been verified, but the verdict label typically refers to the latter, and thus the claim-label pair may not match the original statement-label association (see the "Discordant label" ambiguity class described further on).

In this section, we provide a categorization of the reasons why a claim may be ambiguous[4] and annotate the Italian *in-domain* and *out-of-domain* test sets of X-Fact accordingly, expanding the observed causes of ambiguity beyond e.g., underspecification due to ill-defined terms [6] and pronouns [7].

Reasons for claim ambiguity The reasons why a claim may be ambiguous are identified based on a preliminary assessment of the test portions of X-Fact and of past literature [6, 7]. In the following, we provide ambiguity classes ordered by decreasing severity and accompanied by definitions and examples:

1. **Missing information**: the claim does not contain information that calls for verification:

   "Di Battista e la guerra in Afghanistan." [En: *"Di Battista and the war in Afghanistan."*]: mostly-true

2. **Lack of context**: the claim does not provide enough context (e.g., *who*, *when*, and *where*) or contains ill-defined terms and pronouns, and thus can not be unambiguously verified:

   "Siamo al nono mese consecutivo di riduzione degli sbarchi." [En: *"We are in the ninth consecutive month of reduced arrivals by sea."*]: true

3. **Discordant label**: the fact-checked statement has been rewritten in a negated form or as its opposite, but the label reflects the veracity of the original statement:

   "No, la Banca d'Italia non è controllata dalle banche private." [En: *"No, the Bank of Italy is not controlled by private banks."*]: partly-true

4. **Claim as question**: the fact-checked statement has been rewritten as a question. Although it may preserve the information necessary for fact-checking purposes, this alteration does not represent an actual claim:

   "Davvero la triplice sede del Parlamento europeo costa oltre 200 milioni di euro l'anno?" [En: *"Does the triple seat of the European Parliament really cost over 200 million euros per year?"*]: partly-true

5. **No ambiguity**: the claim is unambiguous and therefore presents sufficient information for automated fact-checking purposes:[5]

   "In Italia ci sono 18 milioni di persone a rischio povertà." [En: *"In Italy there are 18 million people at risk of poverty."*]: true

[2] Agenzia Giornalistica Italia website: https://www.agi.it/
[3] We leave out the label *other* from our discussion since it is present only in some non-Italian subsets which are not part of this study.
[4] In the remainder of this paper, we use "ambiguity" as a broad term that also includes underspecified language.
[5] Note that real-world facts and the subsequent claims are often time-, space-, and culture-dependent. We relax the "perfect unambiguity" requirement in the context of this work.

Table 1
Distribution of claims across ambiguity classes in the *in-domain* (PP) and *out-of-domain* (AGI) test sets. Claims that present sufficient information for fact-checking are in **bold**. Note that claims with labels providing no information about veracity (i.e., those originally marked as *complicated* in X-FACT) are not included in these counts (cf. Section 2).

Ambiguity class	PP test *in-domain*		AGI test *out-of-domain*	
Missing information	47	24.7%	6	3.8%
Lack of context	13	6.8%	17	10.6%
Discordant label	13	6.8%	0	0.0%
Claim as question	**31**	16.3%	**0**	0.0%
No ambiguity	**86**	45.3%	**137**	85.6%
Total	190	100.0%	160	100.0%

Annotating claim ambiguity We manually annotate claims for ambiguity on both Italian test sets of X-FACT following the proposed categories. We focus our efforts on test portions since these represent data that should be used to reliably assess automated fact-checking systems. This results in 350 annotated claims (i.e., 190 *in-domain* and 160 *out-of-domain*, cf. Section 2). If more than one ambiguity class is applicable for a given claim, the more severe one is chosen. For instance, if a claim falls under both "claim as question" and "lack of context" categories, then the latter is applied. Annotation is carried out by a native speaker of Italian. Since the ambiguity classes are rather straightforward, no double annotation was performed. The distribution of annotated claims among classes and test set portions is presented in Table 1.

3.2. Creation of Challenge Test Sets

A typical assumption in most machine learning algorithms is that training and test data follow the same underlying distribution [8]. This is reflected by datasets in which diversity in textual types is rather limited, which makes it hard to assess the performance of automated fact-checking *into the wild*, such as under genre shift (i.e., from article headlines to user-generated content on social media). Although X-FACT includes *in-domain* and *out-of-domain* sets, these mainly reflect different fact-checking sources rather than textual genres.

To provide the research community with means to investigate and mitigate the impact of genre shift on automated fact-checking in Italian, we extend X-FACT with new challenge test sets. We rewrite the subset of claims from the *in-domain* and *out-of-domain* Italian test sets which exhibit sufficient information for fact-checking purposes (i.e., those in bold in Table 1, namely "claim as question" and "no ambiguity", totalling 117 claims for the *in-domain* test set and 137 claims for the *out-of-domain* one) in two different versions. The first one, which we call *news-like* (NL), resembles the language style of a newspaper headline. It is close in style to the original claim and thus meant to probe minimal shift. The second one, *social-like* (SL), is written trying to imitate social media jargon, using e.g., hashtags and abbreviations, and introducing typos. This is meant for assessing performance in scenarios in which automated fact-checking has to be applied to social media posts. Such process has also taken into account claim veracity, to ensure label consistency between the original text and the rewritten one.

For instance, given the claim: "Di Maio ha ragione: il M5S è una delle principali forze politiche in Europa." [En: "*Di Maio is right: the M5S is one of the main political forces in Europe.*"], the two additional claim versions that we create are the following:

- **News-like**: "Il M5S si conferma una delle principali forze politiche in Europa, secondo Di Maio." [En: "*The M5S is confirmed as one of the main political forces in Europe, according to Di Maio.*"]
- **Social-like**: "Il #M5S è traa i partiti maggiori d'Europa!!!" [En: "*The #M5S is amongg the major parties in Europe!!!*"]

As a result, two *in-domain* (NL and SL, 117 claims each) and two *out-of-domain* (NL and SL, 137 claims each) test sets are created as two variants of the original test sets. Detailed statistics for all the subsets are in Table 2.

4. Experiments

In this section, we present the setup for our experiments (Section 4.1), details on model selection (Section 4.2), and further provide a discussion of the results (Section 4.3) and an error analysis (Section 4.4).

4.1. Experimental Setup

We conduct experiments on automated fact-checking along two axes of variation: source (*in-domain* vs *out-of-domain*) and genre (*news-like* vs *social-like*), using the data splits presented in Table 2.

Method All experiments employ a semantic search method for evidence retrieval based on Sentence-BERT [9], followed by majority-driven veracity classification. Compared to standard sentence classification, e.g., using BERT [10], this makes automated fact-checking more transparent, since instances used to determine the veracity label of input claims can be further inspected and ultimately shown to the end user.

Formally, given an input claim $t_i \in T$, where T is a test set among TEST-(NL|SL)$_{(id|ood)}$ (cf. Table 2, *bottom*), the

Table 2
Distribution of claims across original veracity labels (T: *true*, MT: *mostly-true*, PT: *partly-true*, MF: *mostly-false*, F: *false*), mapped labels (T: *true*, F: *false*), sources (PP: *Pagella Politica*, AGI: *Agenzia Giornalistica Italia*), and genres (NL_{orig}: *news-like* as in its original form, NL: rewritten as *news-like*, SL: rewritten as *social-like*). TRAIN and DEV sets follow the original distribution as in X-FACT. All test sets contain the subset of instances that exhibit sufficient information for fact-checking (cf. Table 1).

				Original labels					*Mapped labels*		
Id	**Subset description**	**Source**	**Genre**	**T**	**MT**	**PT**	**MF**	**F**	**T**	**F**	**Tot**
TRAIN	Training set	PP	NL_{orig}	215	234	266	0	228	449	494	943
DEV	Development set	PP	NL_{orig}	27	33	30	0	35	60	65	125
TEST_{id}	Test set (*in-domain*)	PP	NL_{orig}	24	28	39	0	26	52	65	117
TEST_{ood}	Test set (*out-of-domain*)	AGI	NL_{orig}	94	0	0	0	43	94	43	137
TEST-NL_{id}	TEST_{id} as *news-like*	PP	NL	24	28	39	0	26	52	65	117
TEST-NL_{ood}	TEST_{ood} as *news-like*	AGI	NL	94	0	0	0	43	94	43	137
TEST-SL_{id}	TEST_{id} as *social-like*	PP	SL	24	28	39	0	26	52	65	117
TEST-SL_{ood}	TEST_{ood} as *social-like*	AGI	SL	94	0	0	0	43	94	43	137

goal is to find the most relevant claim(s) $\{e_1, ..., e_n\} \in E$, $n \leq |E|$, where E is the evidence set (i.e., union of TRAIN, DEV, and $\text{TEST}_{(id|ood)}$;[6] cf. Table 2, *top*), and n represents the maximum number of most similar claims to retrieve from it. In order to retrieve such evidence claims, $t_1, ..., t_{|T|}$ and $e_1, ..., e_{|E|}$ are all assigned an embedding v^{t_i} and v^{e_i}, respectively, using a pre-trained multilingual `SentenceTransformers` model[7] with default hyperparameters [11]. Then, the cosine similarity between the input claim embedding v^{t_i} and those of all evidence claims $v^{e_1}, ..., v^{e_{|E|}}$ is computed, i.e., $sim(v^{t_i}, v^{e_k})$, $k = 1, ..., |E|$. If $sim(v^{t_i}, v^{e_k}) > \tau$, where τ is a similarity threshold in $[0, 1]$, the claim e_k is a candidate for determining the veracity label of t_i. All candidate claims are sorted by similarity score and the most recurring label among the top n evidence claims is finally assigned to the input claim t_i. If no evidence claim is found or there is a tie among label counts from retrieved claims, then the model abstains. We believe that the possibility to abstain, rather than forcibly assigning a label, is highly desirable in real-world scenarios, since it is not always possible to assess the veracity of a claim.

Settings In order to isolate the impact on performance of genres and sources from the actual availability of relevant evidence (i.e., verified claims about the input claim's topic), we mainly focus on experiments in a *controlled* setting. This ensures that information for verification of each input claim is available in the evidence set E. Nevertheless, we also present results in a *non-controlled* setting for reference. Specifically, the latter does not include $\text{TEST}_{(id|ood)}$ as part of the evidence set E.

Label mapping Since TEST_{ood} and its challenge test sets have only *true* and *false* classes, all dataset labels have been mapped as follows: {*true*, *mostly-true*} → *true*, {*partly-true*,[8] *mostly-false*, *false*} → *false*.

Metrics We use macro F_1 score for evaluation to account for the unbalanced class distribution in the test sets (cf. *ood* ones, Table 2). When computing the F_1 score, abstention is counted as a wrong prediction, because on a controlled setup the model has access to relevant evidence and thus should not abstain. We also measure the correct (COR), error (ERR), and abstention (ABS) rates, by respectively counting the cases in which the model correctly or wrongly predicts a veracity label, or abstains.

4.2. Model Selection

Our method depends on two hyperparameters: the maximum number n of evidence claims to be retrieved and the threshold τ. We tested values of n in the search space $\{1, 2, 3, 4, 5\}$ across all axes of variation (i.e., data splits in Table 2, *bottom*) and thresholds $\tau \in [0.30, 0.85]$ (with step 0.05),[9] finding that $n = 1$ gives on average the best macro F_1 across all configurations (cf. Figure 1).[10] As a result, we use $n = 1$ in the rest of this paper, and present

[6]This makes it sure that veracity information for input claims is actually available, thus allowing us to study the impact of sources and genres in a *controlled* setting.
[7]We use the `paraphrase-multilingual-MiniLM-L12-v2` multilingual model since preliminary experiments with Italian models resulted in worse performance. We hypothesize this is due to the pre-training data composition and size used by the latter.
[8]Indeed, the *partly-true* label is used in PP for claims that are wrong but based on a grain of truth.
[9]The range is motivated by preliminary experiments: we found that values $\tau < 0.30$ and $\tau > 0.85$ are not informative, since the method retrieves almost all or no claims, respectively.
[10]Interestingly, we observe that $n = 2$ and $n = 4$ values result to low F_1 scores. This is because retrieving an even number of evidence claims leads to a higher probability of abstention, as *true* and *false* evidence claims may be in equal number, and abstention is considered as an error for the F_1 score in the *controlled* setting.

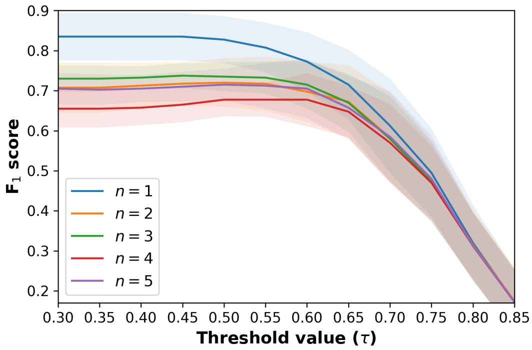


Figure 1: Impact of n and τ hyperparameter values on macro F_1 across test sets. Lines and shading indicate average scores and standard deviation across test sets, respectively.

all τ values in the aforementioned range for discussing the trade-off between errors and abstention.

4.3. Results and Discussion

We present the results across sources, genres, and setups in Figure 2, highlighting the trade-off between abstention and correct/wrong veracity label predictions.

Genre shift has a large impact on performance By comparing results on TEST-NL$_{id}$ with those of TEST-SL$_{id}$ (Figure 2a and 2b) and results on TEST-NL$_{ood}$ with that of TEST-SL$_{ood}$ (Figure 2c and 2d), we see a substantial drop in COR and an increase in ERR on *social-like* test sets. We present selected results for $\tau = 0.6$ in Table 3, i.e., the threshold for which, on average, the ABS ratio still has a higher impact on ERR than COR before mainly impacting COR (cf. Figure 2). The F_1 scores largely drop from 0.86 to 0.74 and from 0.82 to 0.67 when testing the model on data derived from the same or a different fact-checking source, respectively. Such findings attest the impact of genres on the performance in case of available evidence for veracity prediction. This is confirmed in the *non-controlled* setup (results not shown for brevity), albeit the F_1 score exhibits a smaller drop due to confounding reasons such as the lack of relevant evidence.

Fact-checking sources do matter, too By looking at the results on TEST-NL$_{ood}$ (Figure 2c) and TEST-SL$_{ood}$ (Figure 2d), we can observe that not only the COR percentages drop earlier compared to the *in-domain* source counterparts (i.e., Figure 2a and 2b), but also that errors accumulate in the presence of multiple dimensions of variation, i.e., source and genre (cf. Figure 2d vs Figure 2b). The F_1 score drops from 0.86 to 0.82 and from 0.74 to 0.67 on NL and SL genres, respectively (Table 3). This is again confirmed in the *non-controlled* setting (Figure 2, *bottom*).

Table 3
Detailed results across metrics and test sets in the *controlled* setting with hyperparameter values $n = 1$ and $\tau = 0.6$.

macro F_1 score	ID	OOD
NL	0.86	0.82
SL	0.74	0.67

correct (COR)	ID	OOD
NL	0.86	0.84
SL	0.75	0.69

abstention (ABS)	ID	OOD
NL	0.03	0.08
SL	0.09	0.16

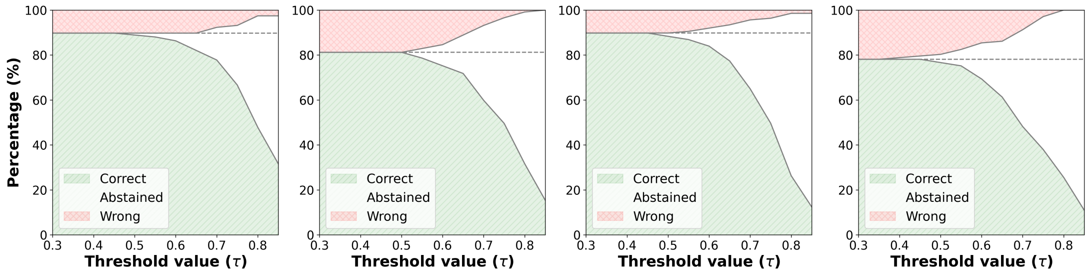
error (ERR)	ID	OOD
NL	0.10	0.08
SL	0.15	0.15

Abstention helps in reducing errors When the model abstains (considering $n = 1$), there are no instances in E that are "similar enough" (i.e., τ) to the input claim. Intuitively, this reduces the impact of erroneous predictions "when in doubt". Figure 2 (*dashed lines*) provides insights into the impact of abstention on formerly COR and ERR percentages across all configurations, as τ varies. We can see that up to $\tau \cong 0.6$, abstention has the great advantage of reducing ERR while negligibly impacting COR. Even in the most challenging test set (i.e., TEST-SL$_{ood}$, Figure 2d), COR predictions for $\tau = 0.6$ are more than double (i.e., 69%, Table 3) than the incorrect ones (i.e., 31% ERR+ABS, Table 3). The trade-off between reducing ERR and preserving COR becomes evident with $\tau \gtrsim 0.7$, for which abstention comes mainly at the expense of COR. In the *non-controlled* setting (Figure 2, *bottom*), on the other hand, this aspect is hard to assess due to spurious factors. By looking at Table 3, we can also see that error rates on OOD sets compared to ID ones do not increase, and actually moderately decrease on the NL genre (i.e., from 0.10 to 0.08).

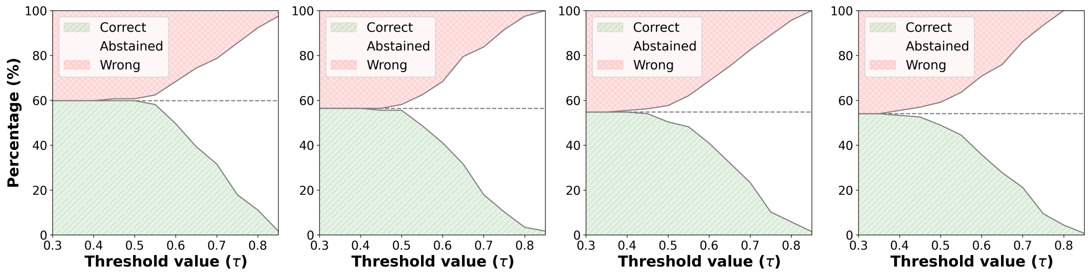
4.4. Error Analysis

We collect ABS and ERR predictions across test sets in the *controlled* setting (with $n = 1$, $\tau = 0.6$) and perform a manual analysis. As shown in Table 4, 33.0% has the correct evidence retrieved at first ($sim(v^{t_i}, v^{e_k}) > \tau$), but this is later discarded because wrong evidence has higher similarity to the input. In the remaining cases, the method fails to retrieve the correct evidence, and thus either wrongly predicts the label (22.9%) or abstains (44.0%). Among the 55.9% (61) ERR only, 13.1% (8) is actually based on a correctly-retrieved relevant claim, but because of claim ambiguity in TRAIN and DEV sets (e.g., discordant labels), the prediction is wrong. In particular, such case accounts for 22.2% (8 out of 36) of errors with evidence. This gives a concrete measure of the impact of ambiguity on the automated fact-checking task.

As regards the performance shift on *social-like* sets compared to *news-like* ones, we observe that this is



(a) Results on TEST-NL$_{id}$. (b) Results on TEST-SL$_{id}$. (c) Results on TEST-NL$_{ood}$. (d) Results on TEST-SL$_{ood}$.



(e) Results on TEST-NL$_{id}$. (f) Results on TEST-SL$_{id}$. (g) Results on TEST-NL$_{ood}$. (h) Results on TEST-SL$_{ood}$.

Figure 2: Percentage of correct (COR), abstained (ABS), and wrong (ERR) predictions across sources (*id*: PP, *ood*: AGI) and genres (NL: *news-like*, SL: *social-like*). Top: *controlled* setup; bottom: *non-controlled* setup. The dashed line splits abstention into formerly correct (*below* the line) and wrong (*above* the line) predictions.

Table 4

Analysis of non-COR predictions across test sets in the *controlled* setting ($n = 1$, $\tau = 0.6$). *w/o evidence*: the correct evidence is not retrieved; *w/ evidence*: it is retrieved but then replaced by a wrong claim with higher similarity to the input.

	ABS	ERR *w/o evidence*	ERR *w/ evidence*
TEST-NL$_{id}$	4 (25.0%)	4 (25.0%)	8 (50.0%)
TEST-SL$_{id}$	11 (37.9%)	7 (24.1%)	11 (37.9%)
TEST-NL$_{ood}$	11 (50.0%)	3 (13.6%)	8 (36.4%)
TEST-SL$_{ood}$	22 (52.4%)	11 (26.2%)	9 (21.4%)
Total	48 (44.0%)	25 (22.9%)	36 (33.0%)

mainly due to the presence of tags (e.g., hashtags, user mentions) in input claims, which account for an average drop of 0.17 F_1. *Social-like* claims without tags instead show a more modest drop in performance, i.e., 0.03 F_1. Indeed, named entities in the form of tags as they typically appear in social media texts may greatly differ from their plain text counterparts (e.g., "*Autostrade per l'Italia*" uses "*@MyWayAspi*" as username), making semantic matching between the two not trivial.

5. Related Work

Research on automated fact-checking for Italian is very limited. Besides X-FACT, datasets comprising Italian are FAKECOVID [4], a multilingual dataset with just 111 articles for Italian, and IRMA [12], a collection of unverified articles from websites that have been classified as "untrustworthy" by professional fact-checkers.

As regards automated fact-checking on social media, it usually foresees three steps: claim detection, evidence retrieval and veracity prediction [2]. Our contribution is addressing both the second and the third step, in that we focus on detecting already fact-checked claims using a semantic similarity approach and leveraging them for veracity prediction. Similar to our method, Shaar *et al.* (2020) [13] use cosine similarity between the input and an already-verified claim. However, they do not address ambiguous instances. Hardalov *et al.* (2022) [14] use social media claims for which users have responded with a link to a fact-checking article, but again they do not consider ambiguity and abstention. Broadly, evidence sufficiency prediction has been recently proposed by Atanasova *et al.* (2022) [15] as a task for identifying if evidence is available for reliable fact-checking.

6. Conclusion

In this work, we show that domains do have a large impact on performance of automated fact-checking for Italian, and the faculty of abstention may be considered to cope with lack of sufficient evidence. Moreover, we contribute to the community by classifying claim ambiguity in the largest Italian test set to date and distributing Italian challenge test sets reflecting diversified domains. Future work includes complementing challenge sets with further versions by multiple annotators as well as automating claim ambiguity assessment. Moreover, the confidence level of the classifier could be investigated and measured with tailored metrics to improve automated fact-checking reliability in handling uncertain cases.

In general, as suggested by Schlichtkrull *et al.* (2023) [16], it would be important to assess the system efficacy with its intended users, in order to evaluate any unforeseen harm possibly caused by actual applications of the technology. In the future, we therefore plan to test our system by including it in the workflow adopted by professional fact-checkers to verify possible cases of mis/disinformation.

Acknowledgments

This work has received financial support from the European Union's Horizon Europe research and innovation program under grant agreement No. 101070190 (AI4Trust).

References

[1] M. Akhtar, R. Aly, C. Christodoulopoulos, O. Cocarascu, Z. Guo, A. Mittal, M. Schlichtkrull, J. Thorne, A. Vlachos (Eds.), Proceedings of the Sixth Fact Extraction and VERification Workshop (FEVER), Association for Computational Linguistics, Dubrovnik, Croatia, 2023. URL: https://aclanthology.org/2023.fever-1.0.

[2] Z. Guo, M. Schlichtkrull, A. Vlachos, A survey on automated fact-checking, Transactions of the Association for Computational Linguistics 10 (2022) 178–206. URL: https://aclanthology.org/2022.tacl-1.11. doi:10.1162/tacl_a_00454.

[3] A. Gupta, V. Srikumar, X-fact: A new benchmark dataset for multilingual fact checking, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Association for Computational Linguistics, Online, 2021, pp. 675–682. URL: https://aclanthology.org/2021.acl-short.86. doi:10.18653/v1/2021.acl-short.86.

[4] G. K. Shahi, D. Nandini, Fakecovid - A multilingual cross-domain fact check news dataset for COVID-19, in: Workshop Proceedings of the 14th International AAAI Conference on Web and Social Media, 2020. URL: https://workshop-proceedings.icwsm.org/abstract.php?id=2020_14. doi:10.36190/2020.14.

[5] P. Singh, A. Das, J. J. Li, M. Lease, The case for claim difficulty assessment in automatic fact checking, arXiv preprint arXiv:2109.09689 (2022). URL: https://arxiv.org/abs/2109.09689.

[6] A. Vlachos, S. Riedel, Fact checking: Task definition and dataset construction, in: Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science, Association for Computational Linguistics, Baltimore, MD, USA, 2014, pp. 18–22. URL: https://aclanthology.org/W14-2508. doi:10.3115/v1/W14-2508.

[7] V. Kocijan, T. Lukasiewicz, E. Davis, G. Marcus, L. Morgenstern, A review of Winograd schema challenge datasets and approaches, arXiv preprint arXiv:2004.13831 (2020). URL: https://arxiv.org/abs/2004.13831.

[8] A. Ramponi, B. Plank, Neural unsupervised domain adaptation in NLP—A survey, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 6838–6855. URL: https://aclanthology.org/2020.coling-main.603. doi:10.18653/v1/2020.coling-main.603.

[9] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, 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), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3982–3992. URL: https://aclanthology.org/D19-1410. doi:10.18653/v1/D19-1410.

[10] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[11] N. Reimers, I. Gurevych, Making monolingual sentence embeddings multilingual using knowledge distillation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational

Linguistics, Online, 2020, pp. 4512–4525. URL: https://aclanthology.org/2020.emnlp-main.365. doi:10.18653/v1/2020.emnlp-main.365.
[12] F. Carrella, A. Miani, S. Lewandowsky, IRMA: the 335-million-word Italian coRpus for studying MisinformAtion, in: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 2339–2349. URL: https://aclanthology.org/2023.eacl-main.171. doi:10.18653/v1/2023.eacl-main.171.
[13] S. Shaar, N. Babulkov, G. Da San Martino, P. Nakov, That is a known lie: Detecting previously fact-checked claims, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 3607–3618. URL: https://aclanthology.org/2020.acl-main.332. doi:10.18653/v1/2020.acl-main.332.
[14] M. Hardalov, A. Chernyavskiy, I. Koychev, D. Ilvovsky, P. Nakov, CrowdChecked: Detecting previously fact-checked claims in social media, in: Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online only, 2022, pp. 266–285. URL: https://aclanthology.org/2022.aacl-main.22.
[15] P. Atanasova, J. G. Simonsen, C. Lioma, I. Augenstein, Fact checking with insufficient evidence, Transactions of the Association for Computational Linguistics 10 (2022) 746–763. URL: https://aclanthology.org/2022.tacl-1.43. doi:10.1162/tacl_a_00486.
[16] M. Schlichtkrull, N. Ousidhoum, A. Vlachos, The intended uses of automated fact-checking artefacts: Why, how and who, arXiv preprint arXiv:2304.14238 (2023). URL: https://arxiv.org/abs/2304.14238.

On incrementing interpretability of machine learning models from the foundations: a study on syllabic speech units

Vincenzo Norman Vitale[1,2], Loredana Schettino[1] and Francesco Cutugno[1,2]

[1]Interdepartmental Research Center Urban/Eco, University of Naples Federico II, Italy
[2]DIETI - University of Naples Federico II, Italy, Italy

Abstract
English. Modern ASR systems generally encode information by employing representations that favour performance indicators such as Word Error Rate (WER), making the interpretation of results and the diagnosis of any error extremely difficult if not impossible. In particular, within the context of end-to-end ASR systems, studies have been devoted to investigating the degrees of explainability of such systems by considering the use of different sets of linguistic features. This work explores the potential of different machine learning algorithms by considering features extracted from syllabic units of analysis and highlights that relying on syllabic Mel-Frequency Cepstral Coefficients increases the interpretability of complex techniques. In fact, the latter currently extract basic units in ways that are highly skewed toward operational convenience. The proposed method would reduce the need for computational resources both in training and in the inference phases, which results in economical and less time-consuming processes.

1. Introduction

[1] The advent of Deep Neural Networks (DNN) enabled modern ASR systems and, more in general, Natural Language Processing (NLP) systems to perform at their best when fed with enough training data and supplied with sufficient computational resources. The recent tendency is to focus efforts on incrementing performance indicators like Word Error Rate (WER), making DNN models behind the scenes increasingly complex and larger, with the effect of a dramatic reduction in their interpretability and an increase in the number of parameters considered and therefore in the required computation effort [1, 2]. As an example, state-of-the-art End-to-End (E2E) ASR systems [3, 4, 5, 6] employ self-supervised learning techniques to determine, based on huge amounts of unlabelled data, the best representation of the speech signal based on fixed-length units, which results in adaptable systems. In the same way, Big Language Models (Big LM) employ advanced encoding techniques, like those based on Byte Pair Encoding (BPE)[7, 8, 9], to encode sub-word units reducing their impact on memory and thus allowing for the creation of bigger models with millions or even billions of parameters aimed at catching a wider range of natural language nuances. On the one hand, these techniques definitely improve systems' performances and capabilities. On the other hand, they also reduce models' interpretability from their foundations, which not only makes them increasingly similar to black boxes but also augments their need for computational resources. Wav2Vec2 authors [3] suggest that "switching to a seq2seq architecture and a word piece vocabulary" would result in performance gains. In line with this, the employment of larger and linguistically motivated units, like syllables, could bring several advantages. Firstly, it would improve performance in terms of WER and computational resources required to train and operate these systems. Secondly, it would increment the system's interpretability, allowing domain experts (i.e. linguists, especially phoneticians) to dive deep into error analysis, which means favoring interpretable rather than computationally efficient but poorly understandable inputs. The main contributions of this study are:

- the proposal of an interpretable approach to speech-oriented feature extraction based on syllable;
- a comparison of various classification techniques with different interpretability grades.

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
vincenzonorman.vitale@unina.it (V. N. Vitale); loredana.schettino@unina.it (L. Schettino); cutugno@unina.it (F. Cutugno)
0000-0002-0365-8575 (V. N. Vitale); 0000-0002-3788-3754 (L. Schettino); 0000-0001-9457-6243 (F. Cutugno)

[1]This article results from the collaboration among the authors. However, for academic purposes, Norman Vincenzo Vitale is responsible for sections 1 and 2.1 and 3.3, Loredana Schettino for sections 2.2, 3.1 and 3.2 for section 4. All the authors are responsible for sections 4 and 5.

2. Related Work

2.1. Explaining Modern ASR

Among the drawbacks of modern Deep Neural Networks (DNN) based systems, the most frequently cited are their poor interpretability [10], the lack of sufficient training corpora [11] and the high demand for computational resources [12, 13, 14]. In recent years, studies have been devoted to the investigation of the degrees of explainability of ASR and, more in general, of speech-related systems based on DNN. Some of these works aim to interpret the internal model dynamics and overall 'behaviour' through model-output backtracing or simulations via explainability methods [15, 16, 17]. In some other cases, 'probing'techniques have been employed to investigate what's encoded in DNN layers at different 'depths' [18, 19] by introducing probes aimed at catching intermediate internal representations to be used for various tasks (e.g., regression or classification). Through classification and measurements, some probing studies have analysed how the accent of pronunciation in different English varieties influences the performance of DeepSpeech2 [20, 21]. These studies employed linguistic-related features and highlighted how the contextual phonetic information contained in intermediate representations influenced the classification. A further study used probing to investigate the multi-temporal modelling of phonetic information in the Wav2Vec2 ASR [3, 22]. Some authors have also proposed a spectrogram-like representation of emissions that could be used for speaker identification and speech synthesis [23].

However, although some studies did consider linguistic features to explain the behaviour of existing models, these were related to isolated fixed-length segments and, to the best of our knowledge, did not take into account larger linguistically meaningful units (i.e. syllables).

2.2. Syllabic unit structure

The notion of *syllable* is quite well-known in linguistic studies. However, its definition has been much debated since it consists of dynamic and complex structures that can be analysed from different perspectives, i.e., phonological or phonetics, and involve various aspects, like articulatory gestures coordination, intensity modulation [24]. Acoustically, a syllable is described as being characterized by an intensity peak in the speech signal surrounded by less intense aggregated sounds.

The essential element of a syllabic unit is the sonority peak, the *nucleus*, which usually consists of vowel sounds. The nucleus can be accompanied by aggregated consonant sounds at the beginning of the unit preceding the nucleus, the *onset*, or at the end following it, the *coda*. Different languages allow syllabic combinations of vocalic (V) and consonantal (C) sounds with different degrees of complexity. However, in many languages, CV is described as the most common structure and the most resistant to phonetic variation and the related reduction phenomena [25, 26, 27].

While disagreements have mainly concerned determining boundaries between different units, the alternation of different units can be grounded on the principles of sonority scale and onset maximization [28]. Thus, the syllable can be described as a sequence of speech sounds where the onset of the sequence is less intense than the preceding coda.

Based on the structural integrity of the syllable, evidence has been provided that the syllable rather than phonetic segments can represent a relevant basic unit of speech production and perception [29, 30]. In fact, [26] shows that the observable variation in connected speech is more systematic at the level of the syllable than at one of the phonetic segments.

3. Material and Methods

To achieve our defined goal, we considered an input consisting of syllables from datasets manually annotated by domain experts and evaluated the performances achieved by different classification methods when relying on distinct sets of syllable-based features.

3.1. Corpus and annotation

This study is based on the Italian and Spanish datasets of the Nocando corpus [31] which consists of spoken narrative texts produced by 11 Italian and 6 Spanish subjects.

The audio files and their transcriptions were processed using the WebMAUS Basic services [32] which provided automatic phonetic transcriptions. The latter were manually edited in Praat [33] and syllabified according to the principles of sonority sequencing and onset maximization [28]. Syllabic units were also annotated for their phonetic structural pattern (CV, CCV, CCVC, CVC, VC, V).

The Italian dataset consists of 940 syllables. As expected, the structural patterns are not evenly distributed, but the following distribution is observed: CV (65%), CVC (16%), CCV (9%), VC (3%), CCVC (3%), V (2%).

As for the Spanish dataset, it consists of 609 syllables. The occurring patterns do not differ much from the Italian ones: CV (65%), CVC (14%), CCV (7%), VC (5%), V (4%), CCVC (3%).

3.2. Syllable-based features

For our tasks, we assumed the hand-annotated syllables as base units. These were considered in two different ways. At first, we look at syllables as a single piece of signal, which is how they have been traditionally considered and processed. Then, we consider them as a signal presumably made up of three components, namely the onset, the nucleus and the coda. The following four feature sets were considered.

- **OPSM** consisting of the GeMAPS [34] set from the OpenSmile toolkit [35]. It is composed of 62 features and provides information about the whole considered signal, namely the syllable.
- **MFCC** consists of 13 Mel Frequency Cepstrum Coefficients, which represent the most salient information for speech recognition tasks [36], extracted for each syllable part[2], i.e. onset, nucleus and coda.
- **Full** namely the concatenation OPSM and MFCC.
- **PCA** consists of the Principal Component Analysis [3] (with an explained variance 95%) of the Full set.

In order to avoid biases due to dimensionality, all the considered feature sets were normalized to achieve zero-mean and unitary variance [4].

3.3. The experimental protocol

This study consists of a classification task that concerns samples labelled with syllable patterns and aims at classifying them on the basis of the considered feature sets. In particular, we compared the following four techniques:

- The **K-Means**[3][38] a vector-quantization method which divides n objects in k clusters based on their mean distance.[5]
- **Hierarchical Agglomerative Clustering (HAC)**[3] [39] is a greedy technique which aims at grouping (or splitting) clusters based on a similarity measure. The final output is a clusters hierarchy which could be divided based on the number of desired clusters.[6]
- The **Support Vector Machine (SVM)**[3] [40] is a versatile algorithm used for classification and regression tasks, whose objective is to find a hyperplane in a multi-dimensional space that enables the classification of the considered data points.
- Lastly, we considered **Convolutional Neural Network (CNN)** [41, 42] as they represent state-of-the-art in speech processing tasks [3]. The considered CNN[7] consists of a Conv2d layer with the ReLu activation function, followed by a Feed-Forward with a SoftMax. In particular, we choose to compare two settings: the first one with a kernel size of 3x3; The second one considers a larger context with a size of 3x9.

In the first phase, we compared K-Means, HAC and SVM as, by their very nature, they are considered more interpretable than neural networks. On the one hand, K-Means and HAC allowed us to explore how sample grouping is affected when the numerousness of clusters is fixed or not, without external supervision. Then, SVM provides a robust and interpretable way of supervisingly evaluating how samples group when a model is set to learn a few interpretable parameters. Lastly, we evaluated the performance of a CNN on the MFCC feature set to compare it with the best-performing method among those from the previous phase, allowing us to compare how an interpretable yet powerful method performs against one of the fundamental building blocks of modern DNN. We compare performances through the micro averaged F1 score (Equation 1), which is particularly suitable for multi-classification tasks.

$$F1 = \frac{TP}{TP + \frac{1}{2} * (FP + FN)} \tag{1}$$

Given the fuzzy boundary between syllabic units and the high degree of variability within each syllable structure class, which not only concerns the presence or absence of segments but also their phonetic specification, the described techniques are applied considering the pooled types of syllable samples.

4. Results

Figure 1 reports the F1-score achieved by K-Means, HAC and SVM over all the considered feature sets. The SVM classifier outperforms both clustering methods on any feature set. However, this was not our primary goal. Note how, for any of the considered methods, the performance difference between the MFCC set and the PCA one is rather small.

For the SVM, results reported are referred to the optimal configuration, which has been found through a grid search on C (between 0.5 and 10 with a step of 0.5), gamma (within 0.01, 0.001 and 0.0001) and kernel type (within rbf, polynomial and sigmoid). The train, validation and dev set were respectively 60/20/20 of the original

[2]MFCC components were extracted through the Librosa library at version 0.9.2.

[3]Defined in Scikit-learn [37] version 1.1.3.

[4]Normalization was achieved through the Scikit-learn (version 1.1.3) StandardScaler.

[5]K-Means parameters: k=6, tolerance to declare convergence=1e-4, initialization through the k-means++ method, random state=42, algorithm=LLoyd's EM algorithm.

[6]HAC parameters: clusters=6, metric=euclidean, linkage=ward

[7]Implemented with Pytorch 1.13 and Pytorch-lightning 1.8.3

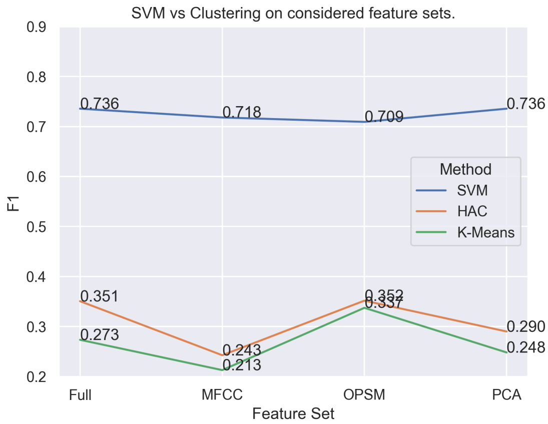


Figure 1: F-scores of SVM, K-Means and HAC per all the considered feature sets.

dataset. Data splits were balanced on the combination of the pattern (i.e. the label) and language.

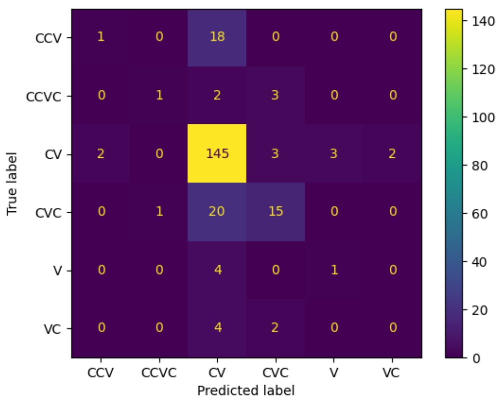


Figure 2: Confusion matrix of the SVM classifier on the MFCC feature set

The confusion matrix that reports the output of the SVM classifier operating on the basis of MFCC (Fig. 2), highlights that better performances concern the CV structure, which is also the more frequent in the data. As for the other structures, misclassification cases mostly concern their identification as CV, which reveals that when considering syllabic units actually occurring in the speech signal a particularly high similarity emerges between more complex patterns, i.e. CCV and CVC, and the CV pattern.

Lastly, Fig. 3 reports the comparison between the SVM classifier and the considered CNN configurations on the MFCC feature set. Still, the SVM performed better than both CNN-based configurations.

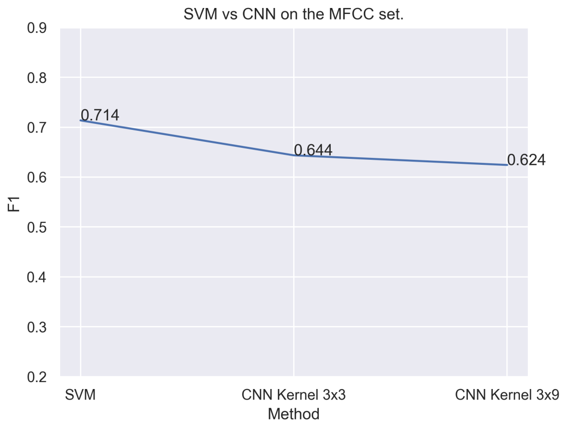


Figure 3: Comparison between SVM and CNN settings on the MFCC feature set.

5. Discussion and Conclusions

In this study, we evaluated the use of phonetic syllables as basic units for speech-related tasks aimed at preserving and, if possible, incrementing the interpretability of different learning techniques. We employed four different feature sets extracted upon the assumption of the phonetic syllable as a fundamental unit, considering different points of view: the more theoretically informed MFCC-based that is strongly tied to the signal; the more analytic OPSM based on Opensmile statistical analysis; the most extended which is a combination of both; the most computationally efficient based on the PCA analysis. Then, upon these feature sets, we evaluated the performance of three well-known machine learning techniques known for being highly interpretable. Finally, we compared the best-performing model, namely the SVM, with a convolutional network on the MFCC set, obtaining comparable performances. Our preliminary results highlight that a set of features aimed at keeping things interpretable, namely the MFCC, lets different methods achieve performances that are comparable to those of richer (Full), analytic (OPSM) or computationally optimized (PCA) sets, which do not retain the same interpretability grade. These findings corroborate the idea that training speech-oriented learning models on larger and linguistically meaningful units could increase the capacity of domain experts and software/ml engineers to diagnose system failures and, at the same time, help reduce the effort and computational resources needed for signal preprocessing. Ongoing analyses involve the enlargement of the annotated datasets to improve the results of further classification trials. In Appendix we reported some preliminary results of a classification trial on an extended dataset, about 26 minutes of hand-annotated speech, consisting of 3589 phonetic syllables. In future works, we plan to extend this kind of study to recent architectures like Squeezeformer[43] or CNN-BLSTM[44].

References

[1] Y. Belinkov, A. Ali, J. Glass, Analyzing phonetic and graphemic representations in end-to-end automatic speech recognition, arXiv preprint arXiv:1907.04224 (2019).

[2] D. V. Carvalho, E. M. Pereira, J. S. Cardoso, Machine learning interpretability: A survey on methods and metrics, Electronics 8 (2019) 832.

[3] A. Baevski, Y. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, Advances in neural information processing systems 33 (2020) 12449–12460.

[4] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, A. Mohamed, Hubert: Self-supervised speech representation learning by masked prediction of hidden units, IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021) 3451–3460.

[5] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, M. Auli, Data2vec: A general framework for self-supervised learning in speech, vision and language, in: International Conference on Machine Learning, PMLR, 2022, pp. 1298–1312.

[6] Z. Zhang, L. Zhou, J. Ao, S. Liu, L. Dai, J. Li, F. Wei, Speechut: Bridging speech and text with hidden-unit for encoder-decoder based speech-text pre-training, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 1663–1676.

[7] R. Sennrich, B. Haddow, A. Birch, Neural machine translation of rare words with subword units, arXiv preprint arXiv:1508.07909 (2015).

[8] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.

[9] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, The Journal of Machine Learning Research 21 (2020) 5485–5551.

[10] Y. Zhang, P. Tiňo, A. Leonardis, K. Tang, A survey on neural network interpretability, IEEE Transactions on Emerging Topics in Computational Intelligence 5 (2021) 726–742.

[11] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al., A survey of large language models, arXiv preprint arXiv:2303.18223 (2023).

[12] E. Strubell, A. Ganesh, A. McCallum, Energy and policy considerations for deep learning in nlp, arXiv preprint arXiv:1906.02243 (2019).

[13] D. Patterson, J. Gonzalez, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. So, M. Texier, J. Dean, Carbon emissions and large neural network training, arXiv preprint arXiv:2104.10350 (2021).

[14] D. Patterson, J. Gonzalez, U. Hölzle, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. R. So, M. Texier, J. Dean, The carbon footprint of machine learning training will plateau, then shrink, Computer 55 (2022) 18–28.

[15] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, H. Müller, Causability and explainability of artificial intelligence in medicine, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9 (2019) e1312.

[16] B. Kailkhura, B. Gallagher, S. Kim, A. Hiszpanski, T. Han, Reliable and explainable machine-learning methods for accelerated material discovery, npj Computational Materials 5 (2019) 1–9.

[17] P. Angelov, E. Soares, Towards explainable deep neural networks (xdnn), Neural Networks 130 (2020) 185–194.

[18] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, H. Lipson, Understanding neural networks through deep visualization, arXiv preprint arXiv:1506.06579 (2015).

[19] G. Alain, Y. Bengio, Understanding intermediate layers using linear classifier probes, arXiv preprint arXiv:1610.01644 (2016).

[20] T. Viglino, P. Motlicek, M. Cernak, End-to-end accented speech recognition., in: Interspeech, 2019, pp. 2140–2144.

[21] A. Prasad, P. Jyothi, How accents confound: Probing for accent information in end-to-end speech recognition systems, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 3739–3753.

[22] D. Ma, N. Ryant, M. Liberman, Probing acoustic representations for phonetic properties, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2021, pp. 311–315.

[23] C.-Y. Li, P.-C. Yuan, H.-Y. Lee, What does a net-

work layer hear? analyzing hidden representations of end-to-end asr through speech synthesis, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 6434–6438.
[24] J. Laver, L. John, Principles of phonetics, Cambridge university press, 1994.
[25] P. Maturi, I suoni delle lingue, i suoni dell'italiano: nuova introduzione alla fonetica, Bologna: il Mulino, 2014.
[26] S. Greenberg, Speaking in shorthand–a syllable-centric perspective for understanding pronunciation variation, Speech Communication 29 (1999) 159–176.
[27] L. Schettino, V. N. Vitale, F. Cutugno, Syllabic reduction in Italian connected speech: towards the integration of linguistic and computational approaches, in: Proceedings of the 20th International Congress of Phonetic Sciences (ICPhS 2023), 2023, pp. 2015–2019.
[28] M. Nespor, Fonologia, Bologna: Il Mulino, 1993.
[29] F. Albano Leoni, The boundaries of the syllable, in: D. Russo (Ed.), The Notion of Syllable across History, Theories and Analysis, Cambridge Scholars Publishing, 2016.
[30] F. Cangemi, O. Niebuhr, Rethinking reduction and canonical forms, Rethinking reduction (2018) 277–302.
[31] L. Brunetti, S. Bott, J. Costa, E. Vallduví, A multilingual annotated corpus for the study of information structure 1, in: Grammatik und Korpora 2009. Dritte internationale Konferenz, Mannheim, 22.-24.09. 2009. Grammar & corpora 2009, 2011.
[32] T. Kisler, U. Reichel, F. Schiel, Multilingual processing of speech via web services, Computer Speech & Language 45 (2017) 326–347.
[33] P. Boersma, D. Weenink, Praat: doing phonetics by computer [computer program]. version 5.3. 51, Online: http://www. praat. org/retrieved, last viewed on 12 (1999-2022).
[34] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, et al., The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing, IEEE transactions on affective computing 7 (2015) 190–202.
[35] F. Eyben, M. Wöllmer, B. Schuller, Opensmile: the munich versatile and fast open-source audio feature extractor, in: Proceedings of the 18th ACM international conference on Multimedia, 2010, pp. 1459–1462.
[36] X. Huang, A. Acero, H.-W. Hon, R. Reddy, Spoken language processing: A guide to theory, algorithm, and system development, Prentice hall PTR, 2001.
[37] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.
[38] J. A. Hartigan, M. A. Wong, Algorithm as 136: A k-means clustering algorithm, Journal of the royal statistical society. series c (applied statistics) 28 (1979) 100–108.
[39] F. Murtagh, A survey of recent advances in hierarchical clustering algorithms, The computer journal 26 (1983) 354–359.
[40] W. S. Noble, What is a support vector machine?, Nature biotechnology 24 (2006) 1565–1567.
[41] K. O'Shea, R. Nash, An introduction to convolutional neural networks, arXiv preprint arXiv:1511.08458 (2015).
[42] S. Albawi, T. A. Mohammed, S. Al-Zawi, Understanding of a convolutional neural network, in: 2017 international conference on engineering and technology (ICET), Ieee, 2017, pp. 1–6.
[43] S. Kim, A. Gholami, A. Shaw, N. Lee, K. Mangalam, J. Malik, M. W. Mahoney, K. Keutzer, Squeezeformer: An efficient transformer for automatic speech recognition, Advances in Neural Information Processing Systems 35 (2022) 9361–9373.
[44] D. Wang, X. Wang, S. Lv, End-to-end mandarin speech recognition combining cnn and blstm, Symmetry 11 (2019) 644.

A. Further results

In Figures 5 and 4, we reported the confusion matrix of the best-performing SVM classifier on MFCC features, in the normalized and non-normalized versions respectively. These preliminary results were obtained on an extended version of the dataset which is currently under further analysis.

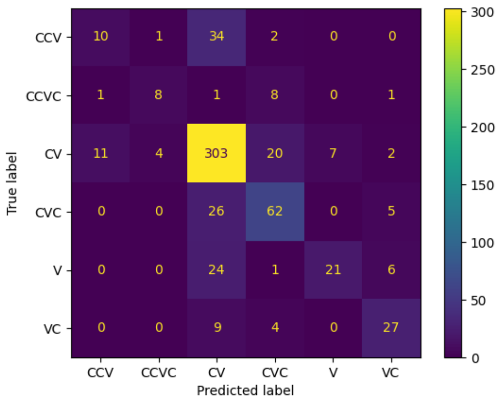


Figure 4: Confusion matrix of the SVM classifier on the MFCC feature set, based on a new version of the dataset incremented by 20 minutes of annotated speech.

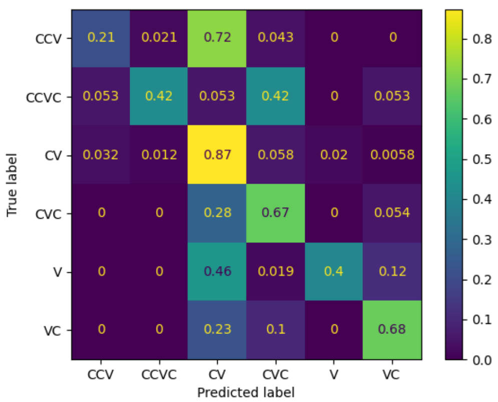


Figure 5: Normalized Confusion matrix of the SVM classifier on the MFCC feature set, based on a new version of the dataset incremented by 20 minutes of annotated speech.

Drug Name Recognition in the Cryptomarket Forum of Silk Road 2

Romane Werner[1], Thomas François[1] and Sonja Bitzer[2]

[1]Université catholique de Louvain, Place Cardinal Mercier 31, 1348, Louvain-la-Neuve, Belgium
[2]Université catholique de Louvain, Place Montesquieu 2, 1348, Louvain-la-Neuve, Belgium

Abstract
English. Drug forums and online chat rooms constitute a relevant source of information for drug use, whose content can serve as reliable sources of information for national agencies with a high number of discussions taking place on various topics. We aimed at investigating whether forum posts could provide useful information as regards to both the early appearance and the monitoring of drug names. A Drug Name Recognition system was used to extract drug terms from the cryptomarket forum of Silk Road 2 thanks to a Conditional Random Fields model. Results of our analysis showed that our model enabled us to discover the presence of 232 new drug names compared to the presence of 106 traditional drug names, which reflect the importance of internet traces as being robust and exploitable with respect to crime phenomena.
Italiano. I forum sulle droghe costituiscono una fonte di informazione rilevante per quanto riguarda l'uso di droghe, poiché il loro contenuto può essere utilizzato dalle agenzie nazionali visto l'alto numero di discussioni che si svolgono su vari argomenti. Il nostro obiettivo è stato quello di verificare se i post dei forum potessero fornire informazioni di rilievo per quanto riguarda sia la comparsa precoce sia il monitoraggio dei nomi delle droghe. È stato utilizzato un 'Conditional Random Field model' per estrarre i nomi di droga dal forum del cryptomarket di Silk Road 2. I risultati della nostra analisi hanno dimostrato che il nostro modello ha permesso di scoprire la presenza di 232 nuovi nomi di droghe rispetto alla presenza di 106 nomi di droghe tradizionali, il che riflette l'importanza delle tracce trovate su internet come robuste e sfruttabili rispetto ai fenomeni criminali.

Keywords
NLP, CRF, DNR, cryptomarket

1. Cryptomarkets and online discussion forums

Over the past decades, the darknet has gradually emerged as a key platform that enables its users to have access to both illicit goods and services. Within darknet, cryptomarkets have triggered "a significant change in the online drug trade" [1, p. 70]. The Internet, and with it the darknet, facilitates illicit drug trade, as was first highlighted by the success of Silk Road [2], which was taken down by the FBI in 2013. Since then, many new cryptomarkets developed to becoming the largest criminal market in the European Union, which continues to expand [3]. According to the 2017 Europol report, "around 35% of the Organized Crime Groups [are] active in the EU on an international level involved in the production, trafficking or distribution of illegal drugs" [3, p. 4].

Due to their wide use and continuous expansion, online marketplaces are a valuable source of information to gather knowledge about linked criminal activities [4]. In this intelligence perspective, it allows to monitor activity on anonymous marketplaces and provide further knowledge on criminal phenomena. Through digital analysis of data from one of the most popular cryptomarkets, Evolution, researchers confirmed previous results on the predominant position of cannabis-related products (i.e. around 25%) [3, p. 6-8], followed by ecstasy and other stimulants [5].

Another source of relevant and useful information on this criminal phenomenon are anonymized user forums and online chat rooms [6], some of which are also incorporated within certain cryptomarkets. In these forums, anonymity seems to play a crucial role in users revealing information, be it regarding darknet or surface web forums, as it "allows them to avoid the legal and social risks of identifying themselves as drug users" [7, p. 159], leading the authors to more easily disclose valuable information. Content found on online forums can serve as reliable sources of information with a high number of discussions taking place on various themes [1]. Indeed, members of drug online forums usually seek drug-related information, while also sharing their own drug experiences with other users [7], encouraging and facilitating information sharing about drug purchases and effects [8].

Besides, "specialized forums offer a fertile stage for questionable organizations to promote NPS (New Psy-

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
romane.werner@uclouvain.be (R. Werner); thomas.francois@uclouvain.be (T. François); sonja.bitzer@uclouvain.be (S. Bitzer)
0000-0001-8814-9990 (R. Werner); 0000-0001-9415-7639 (T. François); 0000-0003-2829-156X (S. Bitzer)

choactive Substances) as a replacement of well-known drugs, whose effects have been known for years and whose trading is strictly forbidden" [8, p. 2]. NPS are defined as "substances of abuse, either in a pure form or a preparation, that are not controlled by the 1961 Single Convention on Narcotic Drugs or the 1971 Convention on Psychotropic Substances, but which may pose a public health threat. The term "new" does not necessarily refer to new inventions — several NPS were first synthesized decades ago — but to substances that have recently become available on the market" [9, p. 2]. As they are among the first to be interested in new trends, researchers thus started investigating the massive use of online forums. These online forums therefore possibly represent a novel approach of harm reduction for drug users and, among others, an "entry point for drug support services" [7, p. 1]. A major challenge in forum analysis can however be pinpointed, as "unlike regular blogs, they include posts from numerous authors with vastly varying levels of activity, writing styles and skills, as well as proficiency in the area to which the forum is devoted" [10, p. 787].

In that context, the use of NLP (Natural Language Processing) techniques has to be pinpointed, as they can help provide insights into the appearance of new drugs on the market. Indeed, several studies concentrated on the automatic extraction of drug terms from online drug forums (see for example [11] or [12]), while other studies noted that CRF (Conditional Random Fields) showed good performance results as regards the recognition of drug terms [13], thanks to the use of specific linguistic features (e.g., POS (Part-of-Speech) tagging). Moreover, to the best of our knowledge, no study explored the use of a CRF model for DNR (Drug Name Recognition) in a cryptomarket forum.

The aim of the current study is thus to determine whether methods from the field of NLP and of computational forensic linguistics can be applied for drug-term discovery, and more particularly, whether CRF can be used as a model for a DNR system to uncover novel drug terms from the cryptomarket forum of Silk Road 2. The first objective is to classify terms that are considered as completely new in regards to a database of well-known drugs, those that are variants of already-known drugs and those that are variants of new drug terms. A second objective is to help identify new drug terms and thus strengthen the monitoring of existing NPS early-warning systems. It also aims at understanding how the contribution of data that was extracted from a particular discussion forum, namely Silk Road 2, can be used to monitor the appearance of NPS.

2. Drug name recognition (DNR)

In order to effectively monitor these forums, being able to recognize drug names is key, as it is considered a critical step for drug information extraction [14]. Therefore, the task of automatic DNR has been defined as actively seeking to recognize drug mentions in texts as well as to adequately classify them into (pre-defined) categories [15]. Automatic DNR has heretofore mostly been conducted in relation to pharmacovigilance (see for instance [16]) and goes hence one step further than the simple name extraction, as it represents "the science and activities concerned with the detection, assessment, understanding and prevention of adverse effects of drugs or any other drug-related problems", such as DDIs (drug-drug interactions) [17].

DNR is a particularly challenging task due to several reasons, among which the following [15]:

- The way individuals name drugs may greatly vary (e.g. 'coke', 'snow' or 'white' can all be used to talk about cocaine);
- There are frequent occurrences of both abbreviations and acronyms, which make it difficult for scientists to identify the exact drug users refer to (e.g. O.C. stands for both Oxycodone and oral contraceptive);
- New drug names are constantly used among the drug community (e.g. Clarity is a relatively new term to talk about MDMA);
- Drug names may sometimes contain a series of symbols that are mixed up with common words (e.g. 3.4-Methylenedioxy-Methamphetamine to refer to MDMA);
- A few drug names sometimes correspond to non-continuous strings of text, also called multi-word expressions (e.g. Synthetic marijuana).

The vast majority of studies conducting DNR research usually concentrate on the biomedical sector and, more particularly, on both biomedical articles [14] and medical documents [18]. These studies were generally conducted using either machine learning approaches, such as CRF and RI (Random Indexing) or using neural approaches, such as LSTM (Long Short-Term Memory). A great deal of research was equally carried out as regards social media [13], which also usually employed NLP techniques, such as word embeddings (see for example the use of Word2Vec in [13]). To the best of our knowledge, only two studies were however conducted with respect to the darknet (see [12] and [19]). As a result, it can be put forward that very few research pertaining on emerging drug terms in forums as well as on cryptomarkets have been conducted heretofore.

Making use of a list of drug names and after a preprocessing phase, Kaati et al. Kaati et al. [12] constructed context vectors using RI VSM. Then, they returned the words which had context vectors similar to those of the analyzed drug terms as a list of potential candidates of “new drugs” [20, p. 1]. Their RI approach yielded a precision rate between 0.70 and 0.80 without more precise information as regards the recall nor the F1 score of their model. Al-Nabki et al. Al Nabki et al. [19] developed DarkNER, a NER (Named Entity Recognition) that was crafted from neural networks, which concentrated on identifying six categories of named entities (i.e., location, person, products, corporation, group, and creative-work) from onion domains on TOR. Their model was trained on the W-NUT-2017 dataset and tested on manually tagged samples of TOR hidden services [19]. Among others, their NER model based on Bi-LSTM (Bidirectional Long Short-Term Memory) enabled researchers to extract drug names. Their model yielded a high precision but also a very low recall, which could be linked to the presence of rare terms in their training data. It is however important to emphasize that both these studies did not enable to distinguish NPS from other drugs.

3. Methodology

The CRF-DNR model used in this research is part of the various NLP techniques on which computational forensic linguistics has relied. Forensic linguistics “is an interdisciplinary field of applied/descriptive linguistics which comprises the study, analysis and measurement of language in the context of crime, judicial procedures or disputes in law” [21]. In that particular context, computational forensic linguistics represents a relatively young field of study, which is a sub-branch of computational linguistics that thus combines forensic science, computer science and linguistics and which is concerned with the interactions between computers and human language in a legal context, in order to inform on criminal phenomena. It has shown various advantages in analyses of naturally occurring data conducted in the legal context, such as its ability to quantify each finding, which results in scientists being able to provide degrees of certainty to the Court thanks to statistical models [22]. Moreover, alongside quantitative analyses, qualitative analyses were also conducted in this research to characterize the different drug terms that were extracted from our data so as to provide detailed insights that can be used by forensic scientists as well as to enhance how forensic linguistics can help provide detailed and qualitative results. Our research hence included the following phases: data collection, data filtering approach, content extraction, preprocessing through both the tokenization and the POS-tagging of the corpus, automatic pre-annotation as well as manual disambiguation and manual annotation of the “old” and “new” drugs, features selection for the CRF-DNR model, development of the CRF model and model accuracy, qualitative analysis of the extracted drug names.

3.1. Data collection, preprocessing and semi-automatic annotation

The data used originates from a huge archive which was collected from 2013 to 2015 by Gwern Branwen, a freelance writer and researcher [23]. In this study, we used data extracted from the forum of Silk Road 2, which was scraped on 19th April 2014. It contains 308.3 Mo, 29.041 texts and it amounts to 38.422.770 tokens.

In order to train our CFR model on accurate data (i.e. on data related to drugs), a filtering approach was used to only retain the files in which drug names appeared. It should be highlighted that the selected files thus mention at least one drug once. For that purpose, a python method was developed to only keep the files which included specific terms (i.e., all the drug terms that appeared in the UNODC conventions; the latter making up our dictionary of drug names). The filtered corpus contains 10.269 files and amounts to 30.305.889 tokens. The whole corpus was tokenized using NLTK’s tokenizer and each token was then POS-tagged using Spacy’s POS tagger, which was trained for the English language [24].

To make an accurate distinction between both new and traditional drug, we focused on the definition of NPS which was provided by the UNODC (i.e., United Nations Office on Drugs and Crime). In this project, the new drugs hence correspond to the NPS as considered by the UNODC, namely the drugs that are not controlled either by the 1961 Single Convention on Narcotic Drugs or the 1971 Convention on Psychotropic Substances. Each drug enclosed in both conventions will thus be considered as a traditional drug, while all the street names associated to these drugs will also be considered as traditional drugs [9].

Based on our dictionary of drug names, our corpus was automatically pre-annotated following the IOB2 format so as to reduce the amount of time needed to annotate the dataset. This format implies that each word must be annotated with a tag (B, I, or O). It allows to encode the scope of multi-word named entities: for instance, a given drug name starts with the (B for beginning) tag and its following components are tagged as (I for inside). Non-drug words are tagged as outside (O) [25]. Another feature was added to this standard format in NLP, in order to characterized the drug as being “OLD” or “NEW” thanks to a distinction made in our dictionary between drugs that were enclosed in the UNODC conventions prior to 2014 (i.e., “OLD”) and drugs that were however found in the dictionary, but enclosed in the conventions after 2014 (i.e., “NEW”). This annotation layer helped provide a dataset of quality which contains elements that have heretofore never been annotated within a forum and drug dataset (i.e., the distinction between "OLD" and "NEW" drug in the context of NPS).

It is important to highlight that all "B+OLD", "B+NEW" as well as "O" tags were all manually checked after the automatic annotation step, in order to find 1) new drug names (i.e., drug names that are not enclosed in the UNODC conventions); 2) new variants of already known drug names (i.e., variants that were not enclosed in our dictionary); 3) variants of new drug names.

3.2. Extraction method

For this research, we made use of CRF, a sequential classification model that was proposed by Lafferty in 2001 Lafferty et al. [26]. We opted for the use of the CRF model, as it is relatively easy to implement, it takes into account the context of words, but also because it provides the opportunity for incorporating arbitrary overlapping features. Moreover, many successful approaches to DNR that made use of NLP techniques, such as the CRF were trained with specific linguistic features. After having read the literature, we noticed that the following features were usually used for the extraction of drug terms in the biomedical field [27], namely word embeddings, character embeddings, prefix of the token, suffix of the token, POS, current token, start or end of sentence, initial capital letter, all-lowercase letter, all-uppercase letter, all-letters, all-digits, if it contains digits, if it is part of a dictionary, if it contains punctuation. We however believe that it could also be interesting to add the length of the token as a feature, as certain drug names are represented as acronyms (e.g., LSD) or are particularly long (e.g., alpha-Pyrrolidinopentiophenone). We also decided to add the following traits for each token-previous (i.e., each token that precedes the current analyzed token) and each token-next (i.e., each token that follows the current analyzed token), namely initial capital letter, all-lowercase letter, all-uppercase letter, all-letters, all-digits, if it contains digits, if it contains punctuation, if it is in the dictionary, token-length. Our feature selection thus contains 40 linguistic features.

For this research, we subdivided our corpus into three different datasets: 50% of the entire dataset was used to train the model, 25% to test the model and 25% to select the best hyperparameters. We made use of CRFSuite from scikit learn [28] in order to develop our CRF model. We then ran our CRF on the basis of the stochastic gradient descent optimization algorithm with a minimum frequency of 0.1, 100 possible iterations, a 10-fold cross-validation and a fixed learning rate of 0.1 to optimize our parameters, as similar methods have heretofore been used for the optimization of the model [29].

Table 1
Performance results of several models

Study	Precision	Recall	F1
Our CRF model	0.96	0.85	0.90
Liu et al. [15]	0.84	0.72	0.78
Zeng et al. [27]	0.93	0.91	0.92

4. Results

Our best model, which included both the use of the dictionary and the word embeddings, yielded a precision rate of 0.96, a recall of 0.85 and a F1 score of 0.90. We should notice that our model outperforms the results of our semi-automatic annotation (0.90 vs. 0.88), which constituted our baseline. Hence, the quality of the corpus annotation was also verified thanks to the use of specific metrics (i.e., recall, precision, and the F1 score). The performance results of the automatic annotation were the following: a recall of 0.93, a precision of 0.88 and an F1 measure of 0.90. Our results also outperform those found in [14]. It is however important to clarify that a LSTM-CRF model [27] also implemented for DNR showed a better performance than our model, which highlights the limit of the latter but also that adding a LSTM layer to our CRF could be interesting (see Table 1 for a summary of the diverse results). Other improvements could be to include both active learning and iterative corrector to our model, as it can help optimize the annotation using fewer training data and by prioritizing which data should be labelled for the training dataset, so as to yield better annotated data.

We also conducted a qualitative analysis of our drug results. We observed that hallucinogens represent the most frequent category, followed by amphetamines, cannabis, coca and cocaine, opium and opiates, central nervous system depressants, opioids and synthetic cannabinoids. Comparing the use of traditional denomination of drugs with their street names, we observed that some drug categories are more often referred to by their traditional names (i.e., opium and opiates and Central Nervous System depressants). On the contrary, other drug categories (i.e., cannabis, synthetic cannabinoid, opioids, coca and cocaine, amphetamines and hallucinogens) show a higher number of occurrences as regards their street names. These results are particularly significant considering the drug categories of cannabis (with 89.3% of occurrences for street names), opium and opiates (with 94.8% of occurrences for traditional drug terms), opioids (with 83.84% of occurrences for street names), amphetamines (with 91.6% of occurrences for street names). Generally speaking, it can be observed that street names make up for the vast majority of drug term occurrences (69.1% vs. 30.9%).

Our model enabled us to discover the presence of 232 new drug names, i.e., (1) names of new drugs, that is to say drugs that do not appear in the UNODC conventions, (2) variant names of traditional drugs but also (3) acronyms of traditional and non traditional drugs). In total, 76 new drug names (32.8% of the total of new drugs), 129 variant names of traditional drugs (55.6% of the total of new drugs) and 27 new acronyms of drugs (11.6% of the total of new drugs) were found, against the presence (more or less frequent) of 106 traditional drug names as well as their street names. As seen above, 2279 occurrences of traditional names and their street names were uncovered, while 788 occurrences of new drug names were also detected, which amount to a total of 3067 occurrences, i.e., 74.3% for already known drug names and 25.7% for new drug names. It is thus important to notice that although they are considered as "new drug names", they make up for a certain proportion of the total number of drug names. Moreover, there are also more types in the category of new drug names than in the category of traditional drugs (258 vs. 101, that is to say 69.9% and 31.1%, respectively).

5. Conclusion

In order to assist states in both their identification as well as their reporting of NPS, the UNODC decided to established the so-called Early Warning Advisory (EWA). The latter serves as a repository full of information on known NPS in order to improve the international understanding of NPS distribution and effects and thus to better understand particular health threats posed by the NPS. The latter specifically extracted both data and information that were found on the Internet. This is the reason why we decided to extract data from forum posts from the cryptomarket of Silk Road 2, as they contain user generated content that is different from simple product lists that can be normally found on cryptomarkets. We thus aimed at analyzing whether forum posts could provide useful information as regards the early appearance of drug names. The purpose of this research was also to developed a CRF-DNR model in order to analyze whether both the use of NLP techniques, such as the CRF model, and of specific linguistic features could help extract (new) drug terms.

For the purpose of this study, we decided to semi-automatically annotate our corpus, which enabled us to have access to an annotated corpus and thus to train our CRF model. It is important to emphasize that this task would be particularly time-consuming should it be done completely manually, as new posts on (cryptomarket) forums continuously appear; the latter resulting in the never-ending task of manually annotating data and thus new drug terms. Another advantage linked to our method is the fact that the model makes use of data from an already established list rather than by just looking at many random new drug terms.

Our analysis enabled us to grasp the number of occurrences of specific drug categories as well as of drugs that are enclosed in the UNODC conventions. It was observed that some drug categories have a higher number of occurrences as regards their traditional drug names (i.e., opium and opiates and Central Nervous System depressants). On the contrary, other drug categories (i.e., cannabis, synthetic cannabinoid, opioids, coca and cocaine, amphetamines and hallucinogens) show a higher number of occurrences as regards their street names. Generally speaking, it could be observed that street names make up for the vast majority of drug term occurrences.

Our model also enabled us to discover the presence of 232 new drug names (i.e., names of new drugs, that is to say drugs that do not appear in the UNODC conventions, variant names of traditional drugs but also acronyms of traditional and non traditional drugs). Hence, 76 new drug names (32.8% of the total of new drugs), 129 variant names of traditional drugs (55.6% of the total of new drugs) and 27 new acronyms of drugs (11.6% of the total of new drugs) were found, against the presence (more or less frequent) of 106 traditional drug names as well as their street names. Moreover, 2279 occurrences of traditional names and their street names were uncovered, while 788 occurrences of new drug names were also detected. It is hence important to notice that although they are considered as "new drug names", they make up for a certain proportion of the total number of drug names. Moreover, there are also more types in the category of new drug names than in the category of traditional drugs (258 vs. 101, that is to say 69.9% and 31.1%, respectively).

With respect to the other two DNR studies (i.e. [12] and [19]) that focused on forum posts, it can be observed that the vast majority of the terms found in this research were not uncovered in these previous studies. It is thus important to emphasize the fact that emerging drug terms can be both extracted and monitored thanks to online resources, such as forum posts. It should be noted that it is possible to rely on the various information that is available on these forums when wishing to grasp new drug terms. Online forums are thus promising sources for the early detection of drugs, suggesting thus that the use of an automated system could help national agencies to identify new drugs.

Our approach however has limitations that can be worked on. It is important to notice that we only made use of data from one cryptomarket forum, namely Silk Road 2. Even if it is considered as a major cryptomarket, it is not representative of all cryptomarket forums. This analysis could thus be improved by using data gathered from other cryptomarket online forums. It could also be interesting to analyze other online sources, such as websites, cryptomarket shops as well as data found in other languages but also to analyze other online sources, such as websites, cryptomarket shops. Another limitation is linked to the fact that this study made use of posts that were launched on a specific date (i.e. 2014-04-19) and that usually went on for several weeks, thereby giving us a relatively static snapshot of the language used on this specific forum at that particular time. We could thus equally focus on data extracted from other periods of time. An area of future research would be to perform a study by conducting DNR over time, that is to say over various months and years. This kind of study could help gain insight on the rise and fall of specific drug terms.

Moreover, an obvious shortcoming that is linked to our model is the fact that it performs poorly at identifying terms that are common in the language but which also have a very specific use in drug-related settings (e.g. shit). Hence, 11.34% of the semi-automatic annotation were considered as false positives, which means that 11.34% of the terms that were annotated as drug terms were not drug terms but referred to other meanings. This represents an important shortfall, as drug terms are often represented as already known and common words. One possible step to tackle this issue would be to add a further grammatical and semantic layer into the model in order to disambiguate homographs (e.g., Word to Gaussian Mixture (w2gm)). It is thus important to emphasize that our model could be improved by using both active learning and iterative corrector, as it can help optimize the annotation using fewer training data and by prioritizing which data should be labelled for the training dataset, so as to yield better annotated data. Another improvement could be to add a Bi-LSTM layer to our CRF model so as to take both context and longer relationships into account.

References

[1] F. Caudevilla, The internet and drug markets, volume 21, EMCDA, Lisbon, 2016, pp. 69–76. doi:10.2810/324608.

[2] K. Kruithof, J. Aldridge, D. Décary-Hétu, M. Sim, E. Dujso, S. Hooren, Internet-facilitated drugs trade: An analysis of the size, scope and the role of the Netherlands, RAND Corporation, Santa Monica, 2016. doi:10.7249/RR1607.

[3] Europol, How illegal drugs sustain organised crime in the eu, Business Fundamentals, Europol, 2017.

[4] J. Broséus, D. Rhumorbarbe, C. Mireault, V. Ouellette, F. Crispino, D. Décary-Hétu, Studying illicit drug trafficking on darknet markets: Structure and organisation from a canadian perspective, Forensic Science International 264 (2016) 7–14. doi:10.1016/j.forsciint.2016.02.045.

[5] D. Rhumorbarbe, L. Staehli, J. Broséus, Q. Rossy, P. Esseiva, Buying drugs on a darknet market: A better deal? studying the online illicit drug market through the analysis of digital, physical and chemical data, Forensic Science International 267 (2016) 173–182. doi:10.1016/j.forsciint.2016.08.032.

[6] J. Aldridge, D. Décary-Hétu, Cryptomarkets: The Darknet As An Online Drug Market Innovation, Technical Report, NESTA, 2015.

[7] M. J. Barratt, Discussing illicit drugs in public internet forums: Visibility, stigma, and pseudonymity, in: M. Foth (Ed.), CT '11: Proceedings of the 5th International Conference on Communities and Technologies, Paparazzi Press, Brisbane, Australia, 2011, p. 159–168. doi:10.1145/2103354.2103376.

[8] J. Buxton, T. Bingham, The Rise and Challenge of Dark Net Drug Markets, Technical Report, Global Drug Policy Observatory, 2015.

[9] UNODC, NPS Leaflet: New Psychoactive Substances, Leaflet, UNODC, 2020.

[10] F. Del Vigna, M. Avvenuti, C. Bacciu, P. Deluca, M. Petrocchi, A. Marchetti, M. Tesconi, Spotting the diffusion of new psychoactive substances over the internet, in: International Symposium on Intelligent Data Analysis, 2016. doi:10.48550/arXiv.1605.03817.

[11] P. Deluca, Z. Davey, O. Corazza, L. Di Furia, M. Farre, L. Holmefjord Flesland, M. Mannonen, A. Majava, T. Peltoniemi, M. Pasinetti, C. Pezzolesi, N. Scherbaum, H. Siemann, A. Skutle, M. Torrens, P. van der Kreeft, E. Iversen, F. Schifano, Identifying emerging trends in recreational drug use; outcomes from the psychonaut web mapping project, Progress in Neuro-Psychopharmacology and Biological Psychiatry 39 (2012) 221–226. doi:10.1016/j.pnpbp.2012.07.011.

[12] L. Kaati, F. Johansson, E. Forsman, Semantic technologies for detecting names of new drugs on darknets, in: IEEE International Conference on Cybercrime and Computer Forensic (ICCCF), 2016, pp. 1–7. doi:10.1109/ICCCF.2016.7740426.

[13] S. Simpson, N. Adams, C. Brugman, T. Conners, Detecting novel and emerging drug terms using natural language processing: A social media corpus study, JMIR Public Health Surveillance 4 (2018). doi:10.2196/publichealth.7726.

[14] S. Liu, B. Tang, Q. Chen, X. Wang, X. Fan, Feature engineering for drug name recognition in biomedical texts: feature conjunction and feature selection, Computational and Mathematical Methods in Medicine (2015). doi:10.1155/2015/913489.

[15] S. Liu, B. Tang, Q. Chen, X. Wang, Drug name recognition: Approaches and resources, Information 6 (2015) 790–810. doi:10.3390/info6040790.

[16] O. Corazza, S. Assi, G. Trincas, P. Simonato, J. Corkery, P. Deluca, Z. Davey, P. van der Kreeft Torrens, D. Zummo, F. Schifano, Novel drugs, novel solu-

tions: exploring the potentials of web-assistance and multimedia approaches for the prevention of drug abuse, Italian Journal on Addiction 1 (2011) 221–226.
[17] R. Chalapathy, E. Zare Borzeshi, M. Piccardi, An investigation of recurrent neural architectures for drug name recognition, in: R. N. Smythe, A. Noble (Eds.), Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis, volume 3 of *LAC '10*, Paparazzi Press, Milan Italy, 2016, pp. 422–431. doi:10.48550/arXiv.1609.07585.
[18] I. Segura-Bedmar, P. Martínez, M. Segura-Bedmar, Drug name recognition and classification in biomedical texts. a case study outlining approaches underpinning automated systems, Drug Discovery Today 13 (2008) 816–823. doi:10.1016/j.drudis.2008.06.001.
[19] W. Al Nabki, E. Fidalgo Fernández, J. V. Mata, Darkner: a platform for named entity recognition in tor darknet, 2019. URL: https://api.semanticscholar.org/CorpusID:203558953.
[20] R. Ferner, C. Easton, A. Cox, Deaths from medicines: A systematic analysis of coroners' reports to prevent future deaths, Drug Safety 41 (2018) 103–110. doi:10.1007/s40264-017-0588-0.
[21] A. Danielewicz-Betz, The Role of Forensic Linguistics in Crime Investigation, 1st. ed., Cambridge Scholars, Chicago, 2012, pp. 93–108.
[22] R. Sousa-Silva, Computational forensic linguistics: An overview of computational applications in forensic contexts, Language and Law 5 (2018) 221–226. URL: https://api.semanticscholar.org/CorpusID:196176347.
[23] G. Branwen, N. Christin, D. Décary-Hétu, R. Munksgaard Andersen, D. Lau, D. Kratunov, V. Cakic, Darknet market archives 2013-2015, 2015. URL: https://www.gwern.net/DNM-archives.
[24] M. Honnibal, I. Montani, spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. github, 2017.
[25] T. Ek, C. Kirkegaard, H. Jonsson, P. Nugues, Named entity recognition for short text messages, Procedia - Social and Behavioral Sciences 27 (2011) 178–187. doi:10.1016/j.sbspro.2011.10.596.
[26] J. D. Lafferty, A. McCallum, F. C. N. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in: Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001, pp. 282–289. URL: http://dl.acm.org/citation.cfm?id=645530.655813.
[27] D. Zeng, C. Sun, L. Lin, B. Liu, Lstm-crf for drug-named entity recognition, Entropy 19 (2017) 221–226. doi:10.3390/e19060283.
[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Machine learning in python, Journal of Machine Learning Research 12 (2011) 2825–2830.
[29] M. D. Zeiler, ADADELTA: an adaptive learning rate method, CoRR abs/1212.5701 (2012). URL: http://arxiv.org/abs/1212.5701. doi:10.48550/arXiv.1212.5701. arXiv:1212.5701.

Italian Crossword Generator: Enhancing Education through Interactive Word Puzzles

Kamyar Zeinalipour[1], Tommaso Iaquinta[1], Asya Zanollo[1], Giovanni Angelini[2], Leonardo Rigutini[2], Marco Maggini[1] and Marco Gori[1]

[1]Università degli Studi di Siena (UNISI), Via Roma 56, 53100 Siena, Italy
[2]expert.ai, Via Virgilio, 48/H – Scala 5 41123, Modena, Italy

Abstract
Educational crosswords offer numerous benefits for students, including increased engagement, improved understanding, critical thinking, and memory retention. Creating high-quality educational crosswords can be challenging, but recent advances in natural language processing and machine learning have made it possible to use language models to generate nice wordplays. The exploitation of cutting-edge language models like GPT3-DaVinci, GPT3-Curie, GPT3-Babbage, GPT3-Ada, and BERT-uncased has led to the development of a comprehensive system for generating and verifying crossword clues. A large dataset of clue-answer pairs was compiled to fine-tune the models in a supervised manner to generate original and challenging clues from a given keyword. On the other hand, for generating crossword clues from a given text, Zero/Few-shot learning techniques were used to extract clues from the input text, adding variety and creativity to the puzzles. We employed the fine-tuned model to generate data and labeled the acceptability of clue-answer parts with human supervision. To ensure quality, we developed a classifier by fine-tuning existing language models on the labeled dataset. Conversely, to assess the quality of clues generated from the given text using zero/few-shot learning, we employed a zero-shot learning approach to check the quality of generated clues. The results of the evaluation have been very promising, demonstrating the effectiveness of the approach in creating high-standard educational crosswords that offer students engaging and rewarding learning experiences.

Keywords
Crosswords, Natural Language Processing (NLP), Machine learning, Large Language Models (LLMs), GPT, BERT-uncased, Crossword clues, Clue-answer pairs, Supervised learning, Zero/Few-shot learning, Creativity, Fine-tuned model, Educational puzzles, Learning experiences,

1. Introduction

Crossword puzzles serve as a highly effective educational tool for numerous reasons. Firstly, they play a crucial role in enhancing children's vocabulary and spelling abilities, as solving the puzzles requires accurate word spelling [1, 2, 3]. Moreover, crossword puzzles are particularly beneficial for acquiring new lexicons in language classes and subjects that involve specialized technical terms [4, 5, 6]. Secondly, these puzzles foster problem-solving skills since students must engage in critical thinking to match clues with appropriate phrases [7, 8]. Additionally, crossword puzzles contribute to memory retention, as students need to recollect previously learned material to complete the puzzles [9, 2]. Lastly, they create an enjoyable and engaging learning experience, motivating students to continuously practice and improve their skills [10, 3]. In summary, crossword puzzles offer an enjoyable and effective approach to practice and enhance essential educational abilities [11, 6].

Creating educational crosswords requires skill, but this process can be time-consuming and limited by human resources. Recent advancements in natural language processing and machine learning offer an alternative solution: training Large Language Models (LLMs) on vast amounts of data to generate diverse and engaging crossword clues and reduce creation time.

This paper makes several contributions to the field. Our initial contribution involves the utilization of this paper to introduce an extensive dataset comprising Italian crossword clue-answer pairs, on the other hand, contributions to the field by proposing a system that uses LLMs to generate high-quality educational crossword. Our approach includes fine-tuning, zero/few-shot learning, and prompt engineering to generate clues from text and keywords. To ensure quality, we developed a set of models to filter out undesirable clues. We additionally employ an algorithm to create educational crossword schema. The resulting system can generate and filter crossword clues, creating educational crosswords with the generated clue-answer pairs.

The paper's organization is as follows: Section Two provides a comprehensive review of relevant work, and Section Three outlines the dataset used in this study. In

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
kamyar.zeinalipour2@student.unisi.it (K. Zeinalipour); tommaso.iaquinta@student.unisi.it (T. Iaquinta); a.zanollo@student.unisi.it (A. Zanollo); gangelini@expert.ai (G. Angelini); lrigutini@expert.ai (L. Rigutini); marco.maggini@unisi.it (M. Maggini); marco.gori@unisi.it (M. Gori)
0009-0009-1262-6768 (T. Iaquinta)

Section Four, we detail our investigation's approach, followed by the presentation of our test findings in Section Five. Finally, Section Six concludes this study, highlighting its implications and potential future directions.

2. Related works

The art of crafting crossword puzzle clues has been a puzzle in itself, prompting diverse strategies to tackle the challenge. Traditional methods often lean on well-established dictionaries, thesauri, or language analysis of web-retrieved texts to define clues [12, 13]. However, in a groundbreaking leap forward, Rigutini and colleagues unveiled the world's first fully automated crossword generator in 2008. Embracing the realm of natural language processing and machine learning, their innovative system autonomously generated crossword puzzle clues. The approach involved web crawling for documents, extracting word meanings, and utilizing techniques like part-of-speech tagging, dependency parsing, WordNet-based similarity measures, and classification models to rank clues by relevance, uniqueness, and readability.

Taking another path, [14] proposed an NLP-driven method for constructing crossword puzzles. They commenced by assembling a collection of texts related to the puzzle's theme. Subsequently, four critical components were built: pre-processing, candidate generation, clue production, and answer selection, altogether orchestrating a comprehensive and captivating crossword puzzle.

Venturing into the realm of Spanish language puzzles, [15] explored extracting definitions from news articles to craft crossword puzzles. They employed a two-stage process: first, identifying crucial words and phrases and extracting their meanings from a trustworthy online dictionary, followed by utilizing those definitions as clues to construct engaging crosswords.

In another linguistic context, [16] presented SEEKH, a software application employing natural language processing to extract keywords and craft crossword puzzles in a multitude of Indian languages. Combining statistical and linguistic tools, SEEKH adeptly pinpointed essential keywords, bringing to life a medley of crosswords across linguistic landscapes.

Despite extensive research efforts, effectively producing comprehensive and distinctive sets of clues and answers from linguistic corpora remains a formidable challenge, especially when dealing with the nuanced intricacies of the Italian language. To tackle these challenges head-on, we present an innovative methodology utilizing Language Models (LLMs) to craft sophisticated educational clues. Representing a pioneering endeavor, our approach successfully generates Italian educational crossword puzzles, addressing a void that previous methods have left unattended. By creating intellectually stimulating and original crossword puzzles, this novel technique enriches learners' profound comprehension of the subjects through detailed and encompassing answers. Therefore, our proposed work not only introduces novelty to the realm of Italian crossword generation but also provides a groundbreaking solution within the domain of educational tools.

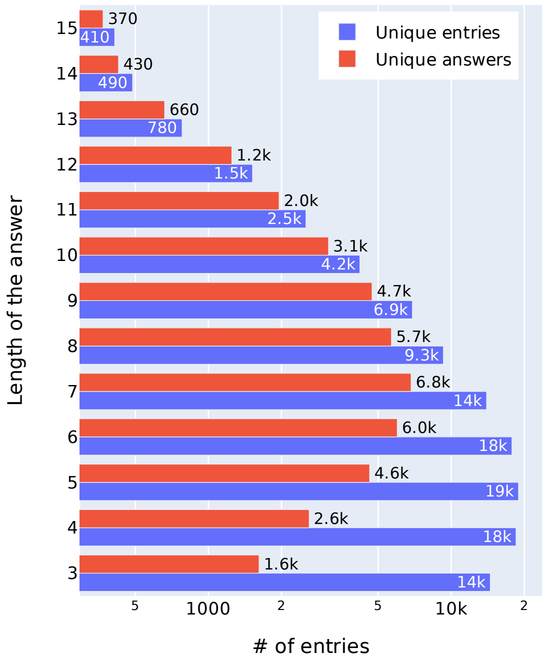


Figure 1: Distribution of the database entries by answer length, in blue the unique answer-clue pairs and in red the unique answers.

3. Dataset

To fine-tune the LLMs, we leveraged a comprehensive collection of Italian crossword clues and answers. The sources of the clues-answer pairs are both internet sites that release solutions for crossword clues as https://www.dizy.com/ and https://www.cruciverba.it/ that we scraped through apposite scripts. And also *pdf* versions of famous Italian crossword papers like *Settimana Enigmistica* and *Repubblica*, that we suitably converted to clue-answer pairs. The various sources where than cleaned, merged and the duplicates were removed. We intend to release this dataset with the support of this paper. This dataset consists of 125,600 entries that correspond to unique clue-answer pairs. It included clues related to different domains, such as history, geography, literature, and pop culture. The dataset under investigation contains a diverse array of linguistic features, including grammatical structures, syntactic patterns, and lexical elements.

A recurring structural pattern in the dataset is the usage of the phrase “known for" or “used for" to define a particular place or object. For example, the definition of a certain location might be “a place known for its historical significance" or “an object used for a specific purpose." In both cases, the answer is a specific instance of the category described in the definition. Moreover, the dataset includes instances where the definition employs clever wordplay or exploits general category definitions to arrive at a specific answer. For example, “In the middle of the Lake" might elicit the response “AK", while “An exotic legume" could be answered with “SOY" by virtue of its membership in the broader category of legumes. In figure 1 you can further go into detail regarding the distribution of the data divided by the length of the answers. Shorter answers tend to have more clues associated while as the answer gets longer the number of clues diminishes in proportion. One of the primary goals of this study was to establish the groundwork for future research by making the processed dataset publicly accessible, with the aim of encouraging other scholars to contribute to this field."[1]

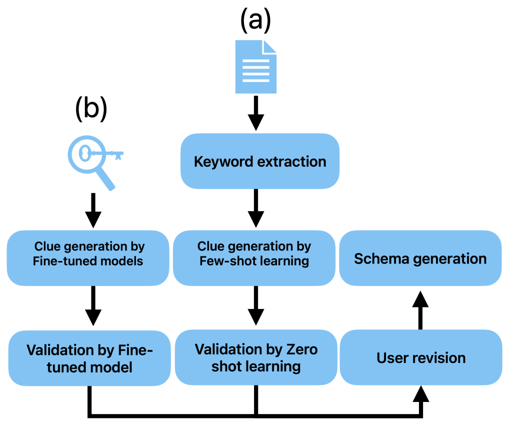


Figure 2: Overall System Architecture

4. Methodology

The system extracts clue-answer pairs from provided texts (path (a) of Figure 2), or generates clues based on given keywords (path (b) of Figure 2). As input texts we use paragraphs selected from Wikipedia pages on educational topics like science, geography, economics. Using this type of text allows us to create direct clues like definitions, appropriate for the educational usage. The system evaluates the quality of the generated clue-answer pairs using various validators. Following the generation process, users are granted the opportunity to review all the produced clue-answer pairs and select their preferred combinations. These selected pairs are then utilized by the final component of the system to generate the crossword puzzle schema.

In this segment, we will delve into the system's fundamental aspects, encompassing three essential components: the generation and validation of clue-answer pairs from provided text, the creation of clues based on given keywords, the validation of the result, and lastly, the generation of the crossword puzzle layout or schema.

4.1. Path (a)

In this section, we analyze the path (a) of Figure 2. We used a multi-step process to apply zero-shot and few-shot learning techniques to text. First, we divided the text into paragraphs and extracted precise keywords. Then, we created personalized clues inspired by the original text using those keywords. To ensure high quality, we thoroughly validated the generated clue-answer pairs. Our primary tool was the GPT-3 DaVinci base model [17]. We'll explore each step in detail in the following.

Keyword extraction: Our innovative strategy harnesses the power of zero-shot learning for an approach to our task. We meticulously craft two prompts in both Italian and English, ensuring they are well-structured with clear objectives and detailed steps to achieve them. You can access it in the appendix under the section labeled Prompts 1 and 4. This thoughtful design empowers the Language Model (LLM) to precisely extract the most relevant keywords, capitalizing on its robust zero-shot learning capabilities. By providing guidance through our prompts, we optimize the model's ability to understand and respond to the intricacies of the task at hand

Clue generation: We use a few-shot learning approach to create compelling crossword clues for each identified keyword in the paragraph. By leveraging an example educational text, crossword keywords, and valid clue examples, we empower the Language Model (LLM) to craft meaningful clues. We presented the paragraph and extracted clues as prompts to the LLM, allowing it to generate clues based on the provided text and keywords. This technique ensures precise and contextually relevant clues. We crafted prompts in both Italian and English, similar to the previous section. Two distinct types of prompts were developed, and all of them are accessible in the Appendix under Prompts 2 and 5.

Validation: We improved the quality of generated keywords and clues by implementing a multi-stage filtering process. First, we filtered out long keywords (over 3 words) as they were less suitable for crossword puzzle answers. Some generated clues inaccurately described their corresponding keywords, and some were hallucinations from the provided text. To address this, we used zero-shot learning to identify and filter out unwanted clues, resulting in a significant improvement in the final output. We created Italian and English prompts, akin to

[1]The dataset is available at https://huggingface.co/datasets/Kamyar-zeinalipour/ITA_CW

the previous section. Both prompt types can be found in the Appendix under Prompts 3 and 6.

4.2. Path (b)

Referring to pipeline (b) of Figure 2; addressing situations where users lack access to the original text and wish to generate crossword clues solely from given answers, we devised an approach to cater to this scenario. Our strategy encompassed multiple stages, each contributing to the overall effectiveness of the solution.

Initially, we focused on fine-tuning various language models specifically tailored for this unique task. Leveraging the data generated from these fine-tuned models, we then proceeded to create diverse classifiers. These classifiers were carefully designed with the primary objective of distinguishing high-quality clue-answer pairs from those that were deemed less suitable.

Fine-tuned models: In the pursuit of generating crossword clues from given answers, we undertook various fine-tuning processes of language models, using data collected from Section 3. Our selection of models comprised GPT3-DaVinci (175B parameters) and GPT3-Curie (13B parameters).

GPT3-DaVinci, with its vast parameter count, demonstrated unmatched depth, enabling it to uncover intricate patterns and craft nuanced clues. On the other hand, GPT3-Curie, while slightly smaller, proved remarkable in grasping language subtleties, further enhancing the fine-tuning process [17].

In our fine-tuning process, we employ a distinctive approach by inputting the answer and tasking the model to generate the corresponding crossword clue. This iterative method not only refines the model's ability to comprehend context but also hones its skill in crafting clues that are both challenging and contextually fitting. By continually providing the answer as input during fine-tuning, we guide the model toward a nuanced understanding of how to construct clues that align seamlessly with the given solution. This tailored training methodology further enhances the model's proficiency in delivering accurate and engaging crossword clues, solidifying its role as a versatile and effective tool in the clue-generation process.

Validation: We developed different strong classifiers using fine-tuned language models to distinguish good crossword clues from poorly crafted ones since not all generated clues fit the given answers perfectly.

In pursuit of this goal, we fine-tuned several models, each boasting unique capacities: GPT3-DaVinci (175B parameters), GPT3-Curie (13B parameters), GPT3-Babbage (1.3B parameters), GPT3-Ada (350M parameters) [17], and BERT-uncased-base (110M parameters) [18].

By harnessing the collective power of these models, each with varying parameter counts, we gained a comprehensive perspective on their effectiveness in filtering and validating the generated clues. Through this approach, our goal was to ensure that only high-quality and contextually relevant crossword clues remained, thereby elevating the overall accuracy and usability of our system.

4.3. Educational Crossword Schema Generator

Our algorithm for creating educational crosswords takes input such as answer lists, work area dimensions, and stopping criteria. It starts by randomly placing a central answer, then adds other answers nearby. The algorithm iteratively adds answers, sometimes removing recent ones or restarting. The best solution is selected based on a global score of the generated schemes. Each solution produced is evaluated using the following formula:

$$\text{Score} = (\text{FW} + 0.5 \cdot \text{LL}) \cdot \text{FR} \cdot \text{LR}$$

where FW (Filled Words) is the number of words added; LL (Linked Letters) is the number of letters that belong to two crossing words; FR (Filled Ratio) is the number of total letters divided by the minimum rectangle area used; and LR (Linked Letters Ratio) is the Linked Letters (LL) divided by the number of total letters.

The algorithm incorporates various stopping criteria, including the minimum number of answers added to the grid; reaching the threshold of minimum Filled Ratio; the limit on the number of times the grid is rebuilt from scratch, and the maximum time duration. The solution with the highest score is deemed the best. These stopping criteria play a crucial role in guiding the algorithm's decision-making process, determining when to conclude the crossword construction. Through the establishment of thresholds and limitations, we successfully ensure the efficient and effective generation of crosswords.

Within the filling process, we have the option to designate a list of "preferred answers." The algorithm places a higher priority on selecting answers from this list, increasing the probability of their incorporation into the grid.

5. Experiments

The experimental evaluation of the designed system is presented in this section, focusing on the individual components and their roles in the overall framework. The system's performance is thoroughly analyzed to assess its effectiveness and efficiency, providing insights into its strengths and weaknesses.

5.1. Experimental Evaluation: Path (a)

In our experiments, we observed variations in model output quality when altering the language of the prompts. To demonstrate this, we conducted two sets of experiments using two types of prompts: one in English and the other in Italian. Our system underwent a rigorous evaluation process using 50 paragraphs sourced from Wikipedia to

Table 1
Assessment outcomes of the clue-answer pairs generated from the provided Text.

System part	*Italian Prompt*	*English Prompt*
Acceptable keywords	79.73 %	75.60%
Acceptable clues	68.34 %	76.70 %
Validator performance	56.76 %	69.72 %

assess the performance of each component using Italian and English. Human supervision was employed, and guidelines for evaluation can be found in Appendix 6. The results of these evaluations are summarized in Table.

Initially, our focus was on keyword extraction, and we achieved promising results in our experiments. Specifically, employing the zero-shot learning approach, we obtained 79.73% and 75.60% accuracy in generating suitable keywords for crossword clues using Italian and English prompts, respectively. Subsequently, we subjected the clue-generation process to human evaluation and found that, with Italian and English prompts, 68.34% and 76.70% of the generated clues were considered acceptable, respectively. To ensure the validity of our results, we employed various approaches outlined in Section 4.1. Through this validation, we were able to identify 56.76% and 69.72% of the unacceptable clue-answer pairs generated using the Italian and English prompts, respectively. These results clearly demonstrate the effectiveness of our system in producing satisfactory crossword clues based on the evaluated text.

Figure 3 demonstrates the step-by-step process of generating crossword clue-answer pairs from input text. The image shows the various stages, such as keyword extraction, clue creation, and pair validation, and illustrates how our system converts input text into pertinent crossword clues. The results with the Italian data revealed that, when the prompt is in English, the performance of the model is better than when the prompt is in Italian.

5.2. Experimental Evaluation: Path (b)

This section delves into our experimental endeavors on generating and validating clues from keywords. Building upon the insights presented in Section 4.2, we devised and fine-tuned two distinct models GPT3-DaVinci and GPT3-Curie with a specific focus on creating clues based on given keywords. For the training phase, we selected a subset of the dataset introduced in Section 3, encompassing 50000 unique clue-answer pairs.

Once the fine-tuning phase concluded, we generated 4,000 clues from each of the fine-tuned models and subjected them to human evaluation using the guidelines provided in Appendix 6. The outcomes of this evaluation are summarized in Table 2. Remarkably, GPT-3 DaVinci outperformed GPT-3 Curie, yielding an impressive 60.1% of acceptable clues compared to Curie's 34.9%

Table 2
Assessment outcomes of the clues generated from the provided keyword.

Model	*% of acceptable clues*
GPT3-DaVinci	60.1
GPT3-Curie	34.9

To gain deeper insights into the quality of the generated clues, we meticulously assembled a collection of acceptable and unacceptable clues. These were randomly sampled from the human-supervised label dataset, offering a diverse clue for each answer. Please consult Table 3 (refer to table 5 in the Appendix for translation). This detailed analysis helps us evaluate the quality and suitability of the clues for creating engaging crossword puzzles.

We developed multiple classifiers that integrate different language models to differentiate between acceptable and unacceptable clue-answer pairs. The result of the analysis on the test set is shown in Table 4. We utilized a dataset of 6,000 human evaluations from the previous step to construct various classifiers. This is the data which we tried to evaluate GPT-3-Davinci and GPT-3-Curie by human supervision. For training and evaluation, we employed 80% of this data for fine-tuning the classifiers and reserving the remaining 20% for testing the classifiers. Within the dataset, 51% comprised acceptable clues, while the remaining 49% consisted of unacceptable clues.

The evaluation results reveal significant distinctions among the classifiers in their ability to differentiate between acceptable and unacceptable clue-answer pairs. Earning the top position, the GPT3-DaVinci model achieved an accuracy of 79.88%, solidifying its role as the most effective classifier in this task. Following closely, the GPT3-Curie base model attained a commendable 77.82% accuracy. The GPT3-Babbage model demonstrated respectable performance with 74.12% accuracy, while GPT3-Ada and BERT-uncased achieved accuracies of 69.17% and 65.62%, respectively.

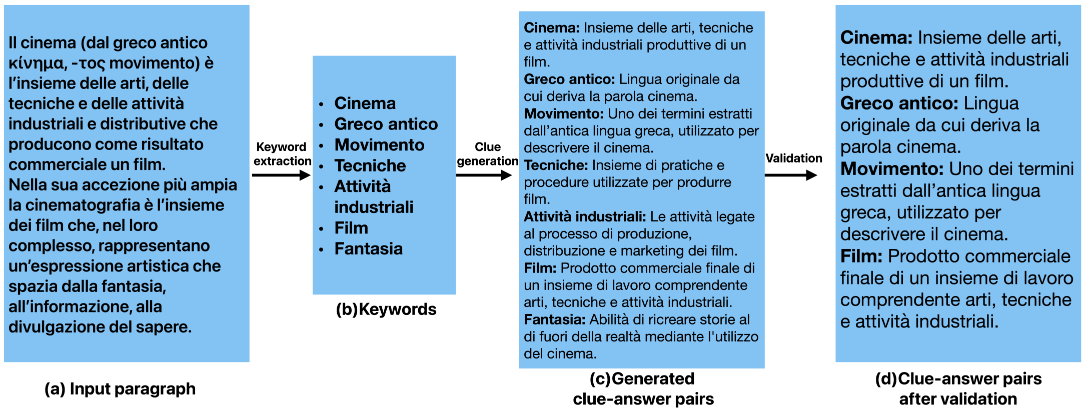


Figure 3: An concrete example of the path (a)

Table 3
Acceptable and unacceptable clues from given keywords using various models.

Clue-Answer pair	*Model*	*Accepted*
Mitologia: La conosce chi conosce i miti	DaVinci	Yes
Elettricità: Uno dei segni zodiacali	DaVinci	No
Curiosità: Il desiderio di sapere	Curie	Yes
Collaborazione: Lo si raggiunge con chiunque	Curie	No

5.3. Schema Generation

Our schema generation algorithm creates educational crosswords with diverse layouts using a single batch of words. Below is an illustration, check the Figure 4 of a comprehensive Italian educational crossword about movies produced with our system. The clue-answer pairs are both extracted from a text (path (a), see Figure 3) and generated directly from a keyword (path (b), contr-assigned with a ⋆ below).

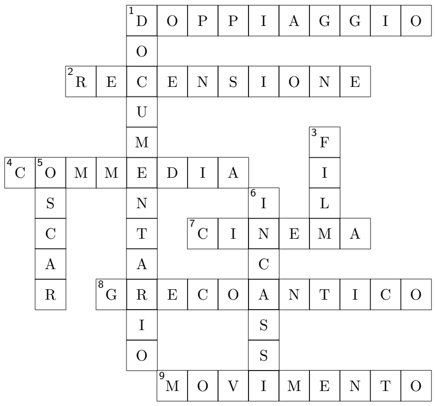
6. Conclusions

In this paper, we present various contributions, including the introduction of a substantial dataset for Italian clue-answer pairs, we developed an innovative system using Large Language Models to generate educational crossword puzzles from given texts or answers. Our approach combines human supervision and specific guidelines to ensure high-quality and relevant clues.

Our system includes a keyword extraction component (79.73% high-quality keywords) and a crossword clue generation component (76.6% relevant and acceptable

Table 4
Classifier performance on distinguishing acceptable Clue-Answer pairs

Model	*accuracy %*	*precision %*	*recall %*	*F1 Score*
GPT3-Dvinci	79.88	80.16	76.67	0.7838
GPT3-Curie	77.82	78.80	72.99	0.7578
GPT3-Babbage	74.12	72.58	73.25	0.7291
GPT3-Ada	69.17	67.77	67.06	0.6741
BERT-uncased-base	65.62	63.71	64.47	0.6409



Orizzontali 1 ⋆ Traduzione Simultanea 2 ⋆ Una valutazione critica 4 ⋆ Lo è uno spassoso racconto 7 Insieme delle arti, tecniche e attività industriali produttive di un film 8 Lingua originale da cui deriva la parola cinema 9 Uno dei termini estratti dall'antica lingua greca, utilizzato per descrivere il cinema **Verticali** 1 ⋆ Un film... come documento 3 Prodotto commerciale finale di un insieme di lavoro compredente arti, tecniche e attività industriali 5 ⋆ Un premio assai ambito 6 ⋆ Entrano nelle casse del botteghino

Figure 4: An illustrative crossword created using the newly introduced system.

clues). A validation component filters out unacceptable pairs, achieving a 69.72% detection rate. We conducted an in-depth investigation of fine-tuned generators and classifiers to enhance the quality of clues. Among the models tested, GPT3-Davinci demonstrated exceptional performance in generating clues based on given keywords, producing a remarkable 60.1% of acceptable clues. Moreover, GPT3-Davinci proved to be the most proficient classifier, accurately distinguishing between good clue-answer pairs and unacceptable ones with an impressive 79.88% accuracy.

Our algorithm for generating educational crossword schemes is efficient and produces diverse layouts. This study aims to enhance student skills and promote interactive learning. Educators can integrate our system into their instruction for more effective teaching practices.

Future research involves developing advanced models for direct clue-answer pair generation and exploring specialized models for different clue types. Our vision is to revolutionize educational crossword generation and unlock new innovations in teaching practice.

Acknowledgments

This work was supported by the IBRIDAI project, a project financed by the Regional Operational Program "FESR 2014-2020" of Emilia Romagna (Italy), resolution of the Regional Council n. 863/2021.

References

[1] W. Orawiwatnakul, Crossword puzzles as a learning tool for vocabulary development, Electronic Journal of Research in Education Psychology 11 (2013) 413–428.

[2] D. Dzulfikri, Application-based crossword puzzles: Players' perception and vocabulary retention, Studies in English Language and Education 3 (2016) 122–133.

[3] Y. D. Bella, E. M. Rahayu, The improving of the student's vocabulary achievement through crossword game in the new normal era, Edunesia: Jurnal Ilmiah Pendidikan 4 (2023) 830–842.

[4] R. Nickerson, Crossword puzzles and lexical memory, in: Attention and performance VI, Routledge, 1977, pp. 699–718.

[5] C. Sandiuc, A. Balagiu, The use of crossword puzzles as a strategy to teach maritime english vocabulary, Scientific Bulletin" Mircea cel Batran" Naval Academy 23 (2020) 236A–242.

[6] E. Yuriev, B. Capuano, J. L. Short, Crossword puzzles for chemistry education: learning goals beyond vocabulary, Chemistry education research and practice 17 (2016) 532–554.

[7] S. Kaynak, S. Ergün, A. Karadaş, The effect of crossword puzzle activity used in distance education on nursing students' problem-solving and clinical decision-making skills: A comparative study, Nurse Education in Practice 69 (2023) 103618.

[8] S. M. Dol, Gpbl: An effective way to improve critical thinking and problem solving skills in engineering education, J Engin Educ Trans 30 (2017) 103–13.

[9] S. T. Mueller, E. S. Veinott, Testing the effectiveness of crossword games on immediate and delayed memory for scientific vocabulary and concepts., in: CogSci, 2018.

[10] V. S. Zirawaga, A. I. Olusanya, T. Maduku, Gaming in education: Using games as a support tool to teach history., Journal of Education and Practice 8 (2017) 55–64.

[11] P. Zamani, S. B. Haghighi, M. Ravanbakhsh, The use of crossword puzzles as an educational tool, Journal of Advances in Medical Education & Professionalism 9 (2021) 102.

[12] L. Rigutini, M. Diligenti, M. Maggini, M. Gori, A fully automatic crossword generator, in: 2008 Seventh International Conference on Machine Learning and Applications, IEEE, 2008, pp. 362–367.

[13] L. Rigutini, M. Diligenti, M. Maggini, M. Gori, Automatic generation of crossword puzzles, International Journal on Artificial Intelligence Tools 21 (2012) 1250014.

[14] B. Ranaivo-Malançon, T. Lim, J.-L. Minoi, A. J. R. Jupit, Automatic generation of fill-in clues and answers from raw texts for crosswords, in: 2013 8th International Conference on Information Technology in Asia (CITA), IEEE, 2013, pp. 1–5.

[15] J. Esteche, R. Romero, L. Chiruzzo, A. Rosá, Automatic definition extraction and crossword generation from spanish news text, CLEI Electronic Journal 20 (2017).
[16] B. Arora, N. Kumar, Automatic keyword extraction and crossword generation tool for indian languages: Seekh, in: 2019 IEEE Tenth International Conference on Technology for Education (T4E), IEEE, 2019, pp. 272–273.
[17] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.
[18] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, The Journal of Machine Learning Research 21 (2020) 5485–5551.

Appendix

Guidelines for Validating Clue-Answer Pairs

In the course of our study, we embraced an enthralling challenge: constructing a classifier capable of discerning between acceptable and non-acceptable crossword clue-answer pairs. Crossword puzzles have held a cherished place as a beloved pastime, demanding a harmonious fusion of linguistic prowess, creative acumen, and adherence to intricate puzzle construction rules to fashion top-tier clue-answer pairs. Our pursuit of creating an automatic evaluator for generated crossword clues and their corresponding answers holds tremendous potential. This advancement promises to aid puzzle creators, enrich puzzle-solving experiences, and unlock profound insights into the subtle nuances of language and puzzle design. Ultimately, this endeavor not only elevates the world of crossword puzzles but also kindles a deeper appreciation for their linguistic artistry and cognitive allure.

To create a powerful classifier for crossword clue-answer pairs, we must establish a strong and comprehensive guideline that clearly delineates the attributes of acceptable and non-acceptable pairs. This guideline will be the cornerstone for training our classifier, enabling it to discern the defining characteristics that set apart high-quality clues from irrelevant or inappropriate ones. With strict adherence to this guideline, we can guarantee the accuracy of our classifier in assessing the quality of clue-answer pairs, ultimately leading to the creation of more captivating and enjoyable crossword puzzles.

Let us now explore the pivotal components of the guideline, essential for evaluating crossword clue-answer pairs:

Relevance and Cohesion: A top-notch crossword clue-answer pair thrives on a profound and meaningful connection between the clue and the answer. The clue should provide ample context or clever hints that smoothly lead solvers to the intended solution. Simultaneously, the answer must be directly tied to the clue, fitting flawlessly within the puzzle's theme or topic.

Wordplay and Inventiveness: Elevate your crossword clues with ingenuity and wordplay that challenge and delight solvers. Seek clues that encourage lateral thinking, incorporate witty twists, or conceal intriguing meanings. A well-crafted clue-answer pair captures the solver's imagination, transforming the puzzle into an exhilarating journey of discovery.

Clarity and Precision: Precision is key in creating crossword clues. Ensure your clues are crystal clear and unambiguous, presenting solvers with a distinct and precise solution. Avoid any ambiguity that might lead to multiple interpretations or numerous possible answers. The goal is to deliver a single correct solution that aligns perfectly with the clue's intended meaning.

Grammar and Language: Pay meticulous attention to grammar, syntax, and linguistic conventions in both the clue and the answer. Maintain grammatical correctness, coherence, and an appropriate level of complexity for a crossword puzzle.

General Knowledge and Fairness: Strike a balance between challenge and accessibility by grounding your clues in general knowledge or commonly known facts. Avoid overly obscure or specialized references that could alienate solvers. A great clue-answer pair caters to a diverse range of puzzle enthusiasts, offering a fair and engaging experience for all.

Through the adoption of this framework, a robust dataset can be generated, facilitating the development of a dependable classifier that discerns commendable crossword clue-answer pairs from incongruous or inappropriate ones. This transformative classifier holds the promise of revolutionizing crossword puzzle creation, assessment, and solving, offering invaluable revelations into the craft of constructing captivating and mentally stimulating puzzles.

Prompts

Prompt 1: **Italian, for keyword extraction**

```
prompt = f"""
Obiettivo: Il tuo compito é estrarre delle parole
  chiave, descritte nel testo proposto. Le parole
  chiave estratte saranno utilizzate per creare
  brevi definizioni di cruciverba riguardanti il
  testo da cui sono estratte le parole chiave. Le
  definizioni saranno d'aiuto per trovare la
  soluzione corrispondente e completare il
  cruciverba.

Completa l'obiettivo attraverso i seguenti
  passaggi:
```

Table 5
Translation of Table 3

Clue-Answer pair	*Model*	*Acc.*
Mythology: It is known by anyone who knows myths	DV	Yes
Electricity: One of the zodiac signs	DV	No
Curiosity: The desire to know	Curie	Yes
Collaboration: One reaches it with anyone	Curie	No

```
    1- Estrai le parole chiave piú importanti del
     testo.

    2- Controlla le parole chiave: controlla se le
     parole chiave sono descritte e definite nel
     testo o non sono descritte e definite nel testo.

    3- Parole chiave finali : sulla base del
     passaggio precedente, rimuovi tutte le parole
     chiave che non sono definite nel testo.



    Utilizza il seguente formato di output:

    Parole chiave: <Parole chiave finali>


    Text: ```{text}```
    """
```

Prompt 2: **Italian, for clue generation**

```
     prompt = f"""
    Genera brevi definizioni di cruciverba per
     ciascuna delle parole chiave fornite: {keywords}
      sulla base del seguente testo: {text}.

    Completa l'obiettivo attraverso i seguenti
     passaggi:

    1- Per ciascuna delle parole chiave fornite,
     trova il passaggio del testo contenente l'
     informazione riguardante la parola chiave.
    2- Genera brevi definizioni: per tutte le parole
     chiave genera brevi definizioni riguardanti il
     testo. Nella definizione non deve essere
     presente la parola chiave.
    3- Non usare virgolette e apostrofi nell'output.

    Segui questo esempio per completare l'obiettivo:
    "Testo: La scienza é un sistema di conoscenze
     ottenute attraverso unattivit  di ricerca
     prevalentemente organizzata con procedimenti
     metodici e rigorosi, coniugando la
     sperimentazione con ragionamenti logici condotti
      a partire da un insieme di assiomi, tipici
     delle discipline formali. Uno dei primi esempi
     del loro utilizzo lo si puó trovare negli
     Elementi di Euclide, mentre il metodo
     sperimentale, tipico della scienza moderna,
     venne introdotto da Galileo Galilei, e prevede
     di controllare continuamente che le osservazioni
      sperimentali siano coerenti con le ipotesi e i
     ragionamenti svolti.
```

```
    Parole chiave: conoscenze, ricerca, rigorosi,
     assiomi, ipotesi, Galileo
    Definizioni:
    Conoscenze: informazioni acquisite tramite
     ricerca organizzata con procedimenti metodici e
     rigorosi.
    Ricerca: attivit  organizzata prevalentemente
     con procedimenti metodici e rigorosi finalizzata
      allottenimento di conoscenze.
    Rigorosi: esatti e precisi nello svolgimento
     delle azioni.
    Assiomi: un insieme di verit  accettate come
     base dei ragionamenti logici.
    Ipotesi: assunte per comprendere le osservazioni
     sperimentali e testare le conoscenze
    Galileo : egli introdusse il metodo sperimentale
     nel processo di scienza moderna.
    "


    """
```

Prompt 3: **Italian, to auto check**

```
prompt = f"""


    Obiettivo: il tuo obiettivo é controllare se il
     contenuto di ogni definizione é presente o no
     nel testo proposto Per ciascuna definizione
     scrivi "True" se il contenuto é presente nel
     testo e "False" se il contenuto non é contenuto
     nel testo.


    Sentences: ```{clue}```

    Text: ```{text}```
    """
```

Prompt 4: **English, for keyword extraction**

```
 prompt = f"""


    Objective: Your task is to extract described
     keywords in Italian from a given Italian text.
     These keywords will be used to create Italian
     crossword short definitions based on the
     extracted text. The clues will help Italian
     solvers to find the corresponding answers and
     complete the puzzle grid.

    Please follow these steps to achieve the
     objective:
```

```
    1- Extract the most important Italian keywords in
      the Italian text.

    2- Check keywords: check if the Italian keywords
     are well Explained in the given Italian text or
     not.

    3- Final keywords : Remove all the Italian
     keywords which are not well defined in the
     Italian text based on the last step.


    Use the following output format:

    Keywords: <Final keywords>


    Text: ```{text}```
    """
```

Prompt 5: **English, for clue generation**

```
prompt = f"""


    Generate short crossword definitions in Italian
     for each provided Italian keyword: {keywords}
     based on the following Italian text: {text}.

    Follow these steps to achieve the objective:

    1- For each provided Italian keyword detect the
     part of the Italian text which contains the
     keyword information.
    2- Generate short definitions in Italian: For all
      the Italian keywords generate short definitions
      in Italian based on the Italian text, and place
      the correspondent keyword after each generated
     definition. Make sure that the Italian keyword
     is not present in the correspondent definition.
    3- Do not use quotation marks and apostrophes in
     the output.

    Follow this example to complete the task:
    "Text: La scienza é un sistema di conoscenze
     ottenute attraverso unattivit  di ricerca
     prevalentemente organizzata con procedimenti
     metodici e rigorosi, coniugando la
     sperimentazione con ragionamenti logici condotti
      a partire da un insieme di assiomi, tipici
     delle discipline formali. Uno dei primi esempi
     del loro utilizzo lo si puó trovare negli
     Elementi di Euclide, mentre il metodo
     sperimentale, tipico della scienza moderna,
     venne introdotto da Galileo Galilei, e prevede
     di controllare continuamente che le osservazioni
      sperimentali siano coerenti con le ipotesi e i
     ragionamenti svolti.
    Keywords: conoscenze, ricerca, rigorosi, assiomi,
      ipotesi, Galileo
    Clues:
    Conoscenze: informazioni acquisite tramite
     ricerca organizzata con procedimenti metodici e
     rigorosi.
    Ricerca: attivit  organizzata prevalentemente
     con procedimenti metodici e rigorosi finalizzata
      allottenimento di conoscenze.
    Rigorosi: esatti e precisi nello svolgimento
     delle azioni.
    Assiomi: un insieme di verit  accettate come
     base dei ragionamenti logici.
    Ipotesi: assunte per comprendere le osservazioni
     sperimentali e testare le conoscenze
    Galileo : egli introdusse il metodo sperimentale
     nel processo di scienza moderna.
    "

    """
```

Prompt 6: **English, to auto check**

```
prompt = f"""


    Objective: Your objective is to check whether
     each given Italian Sentence content is present
     in the provided Italian text or not. Print "True
     " if it is present in the provided Italian text
     and "False" if it is not present in the provided
      Italian text.

    Sentences: ```{clue}```

    Text: ```{text}```
    """
```

Die Rätselrevolution: Automated German Crossword Solving

Andrea Zugarini[1,*,†], Thomas Röthenbacher[2,†], Kai Klede[2], Marco Ernandes[1], Bjoern M. Eskofier[2,3] and Dario Zanca[2]

[1]expert.ai, Siena, Italy
[2]Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
[3] Institute of AI for Health, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany

Abstract

Crossword puzzles are popular word games played in various languages around the world, with diverse styles across different countries. For this reason, automated crossword solvers designed for a language, may not work well on others. In this paper, we extend Webcrow, an automatic crossword solver, to German, making it the first program for crossword solving in the German language. To address the lack of large clue-answer crossword pairs data, Webcrow combines multiple modules, known as experts, which retrieve potential answers from various resources, including the web, knowledge graphs, and linguistic rules. The system is evaluated on a collection of crosswords from variegate sources, where it is able to solve perfectly 67% of them. Additional analysis reveals that while our solver achieved commendable results, puzzles with poorly constrained schemas and original clues still presented significant hurdles. These findings shed light on the complexity of the crossword-solving problem and emphasize the need for future research to address and overcome these particular challenges effectively.

Keywords

Automated Crosswords Solving, German Crosswords, Question Answering

1. Introduction

Automated crossword solving is a challenging problem in Artificial Intelligence and Natural Language Processing (NLP) research. Solving a puzzle requires multiple skills, ranging from encyclopedic knowledge and linguistics to reasoning and constraint satisfaction. In the past, several automated crossword systems have been proposed for English [1, 2, 3]. Despite the successful results achieved by these approaches, they do not investigate crossword resolution in other languages. All those methods heavily rely on large databases of previously answered clues to retrieve and rank answer candidates, that sometimes are even re-ranked [4, 5, 6, 7]. Berkley Crossword Solver [3] make also use of multiple Language Models that have been fine-tuned to segment answers in words and to correct wrong letters. The need for such resources hinders the application of these solutions to other languages.

WebCrow [8, 9] instead, is a crossword solver that was applied also to Italian puzzles. The architecture which is composed of multiple modules, called experts, facilitates the portability to new languages. In this work, we extend WebCrow to the German language. To the best of our knowledge, we are the first to propose an automatic solver for German crosswords.

The paper is organized as follows. In Section 2, the whole WebCrow architecture is described. Then, in Section 3 we present the data gathered for German and its usage by the WebCrow experts. Experiments are outlined in Section 4. Finally, we summarize our conclusions and directions for future works in Section 5.

2. WebCrow

As showcased in Figure 2, WebCrow, similarly to prior crossword solvers like Proverb [1], works in two stages: candidate answers retrieval and constraints satisfaction. In the first stage, a list of weighted candidate answers is retrieved for each clue. The retrieval is carried out by multiple modules, called *experts*. Candidates' lists are then combined together by the merger module. In the second step, the solver fills the grid given the potential answers with the objective of maximizing the most probable solution given the constraints imposed by the grid.

2.1. The Expert Modules

WebCrow uses multiple modules to retrieve answer candidates. In general, the number of experts can vary, and

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

*Corresponding author.
†These authors contributed equally.
azugarini@expert.ai (A. Zugarini); thomas.roethenbacher@fau.de (T. Röthenbacher); kai.klede@fau.de (K. Klede); mernandes@expert.ai (M. Ernandes); bjoern@fau.de (B. M. Eskofier); dario.zanca@fau.de (D. Zanca)

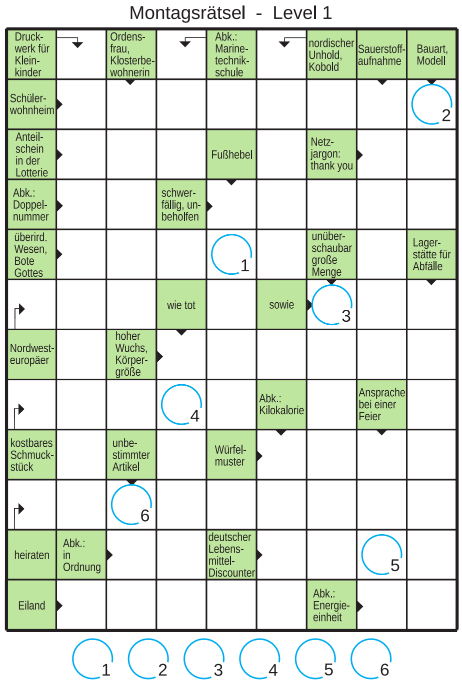


Figure 1: A German crossword puzzle in the "swedish" style.

new experts can always be added. Here, we limit the discussion only to the modules that were then used for the German language.

CWDB. A large collection of previously answered questions is paramount for any crossword solver. The CrossWord DataBase (CWDB) expert retrieves answers from a database of clue-answer pairs. Retrieval and ranking of the answers is based on the semantic similarity between a clue and the clues in the CWDB. We follow the approach in [10]. We used pre-trained encoders [11, 12, 13] to embed both the clues in the database and the ones of the crossword that is being solved.

Knowledge Graph. Ontologies are a rich source of linguistic and encyclopedic knowledge, that is frequently required in crosswords. Knowledge Graphs (KGs) contain structured information about concepts and language. The expert in Webcrow exploits KGs to collect a database analogous to the CWDB in a straightforward way: linguistic concepts in the knowledge graph are paired with their definitions, similarly to the approach followed with the clue-answer pairs from the CWDB. Following the same approach, an answer retrieval step is then applied: the definitions are encoded in a semantic representation and then ranked and retrieved accordingly to a semantic similarity between the clue and the definition in the database.

Web Search. Web Search is the expert that characterizes Webcrow, as the name suggests. This module retrieves answer candidates by searching on the web. Differently from CWDB or Knowledge Graphs, it allows the retrieval of fresh information, that can be crucial to solve some clues. For each clue, the web is queried, making use of Bing APIs.[1] The answer list can be built upon either or both the snippets and the full documents returned by the search. All the frequent words extracted from the documents are ranked according to their TF-IDF, also taking into account the rank of the documents in which they occur [8]. IDF must be pre-computed on a large collection of documents.

Lexicon. The lexicon is a large vocabulary of German words. For each answer to fill in the crossword, the dictionary module returns all the words in the lexicon that fit in length. The returned list is weighted by the n-gram model used in the Implicit Module described below.

2.2. Implicit Module

The implicit module ensures that the crossword solver also works when the correct answer is not present in the candidate list. It generates new candidates on the fly using a collection of character level n-grams, with $n = 4$, together with their relative frequencies in the answers of the CWDB. Whenever the CSP solver reaches a point where none of the answers for a given clue fit into the grid, the implicit module tries to generate the most probable sequence of characters satisfying the constraints.

2.3. Belief Propagation and Grid Filling

The current version of WebCrow looks for a solution that maximizes the expected value of clues answered correctly. To compute the posterior probabilities $q_i(a)$ for an answer a to be in slot i it uses belief propagation. These probabilities then allow us to infer letter probabilities $q_{i,n}(\lambda)$, that quantify the likelihood of a character λ at position n of the answer to clue c_i. It can be computed as in Equation 1, where $a(n) = \lambda$ means the n-th letter of answer a is λ.

$$q_{i,n}(\lambda) = \sum_{a \in D_i, \text{with } a(n)=\lambda} q_i(a). \tag{1}$$

If a cell s in the grid is the n-th cell of an horizontal clue c_i, and the m-th cell in the vertical clue c_j, then the probability $q^s(\lambda)$ becomes:

$$k \cdot q_{i,n}(\lambda) \cdot q_{j,m}(\lambda),$$

[1] https://learn.microsoft.com/en-us/rest/api/cognitiveservices-bingsearch/bing-web-api-v7-reference.

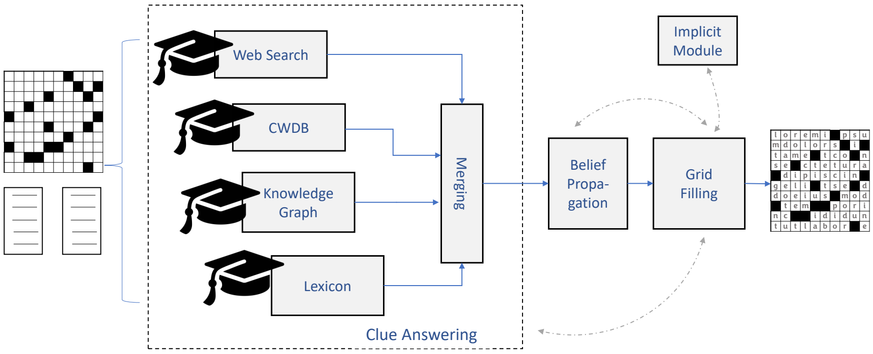


Figure 2: Sketch of the Webcrow architecture for German Crossword Puzzles.

where k is a factor that normalizes q^s to be a probability mass function. If the probability of a letter in a given cell exceeds a certain threshold, the solver "freezes" it in the solution and only allows answers that are consistent with such a letter.

A solution that maximizes the sum of the answer probabilities $q_i(a)$ is then found using Greedy Search, which fills in the missing or incomplete answers by iteratively selecting the most probable from the remaining answers that fit into the missing cells.

Table 1
The full dataset of German crosswords. Listed for each type of crossword are their dimensions, the frequency of publication in their respective newspapers, the publication date of the oldest included crossword, and the number of crosswords included in total.

Source	dims	freq	oldest	# CWs
Bild	7x7	daily	01.11.19	1188
Spiegel	9x9	daily	01.07.20	944
Focus	9x9	daily	02.11.19	1187
HA	16x16	daily	03.01.22	394
FAZ	15x15	weekly	11.03.22	46

3. Adapting WebCrow to German

3.1. Data Gathering

We collected two kinds of data: full crosswords, and a large corpus of clue-answer pairs.

Full Crosswords. Full crosswords were obtained from various newspapers that offer crossword puzzles on their websites. These include "Der Spiegel", "Bild-Zeitung", "Focus", "Frankfurter Algemeine Zeitung" (FAZ) and "Hamburger Abendblatt" (HA). The number of crosswords from each source as well as the crossword dimensions, the frequency of their publication and the earliest retrieved crossword is listed in Table 1. With the exception of FAZ, all the other crosswords schemas are in the so called Swedish style, characterized by brief definitions that goes within the black cells. An example is shown in Figure 1.

The retrieved documents were split into a training, validation, and test set. For each source except for FAZ, the training set included all crosswords published up to and including the Friday 30th September, 2022. For FAZ, an earlier cut-off date (Wednesday 31st August, 2022) was chosen to leave more crosswords for use in validation. The validation set then includes the crosswords that are not in the training set that were published up to the Thursday 24th November, 2022. The crosswords for the test set were published after that date and up to the Wednesday 31st August, 2022, however, the daily publications (every source except FAZ) were sub-sampled to include a similar number of each source. Because some crosswords, like the one from the New York Times, have differing levels of difficulty for each day of the week [1], we sampled every six crosswords, instead of every seven, to make sure all days of the week were present in the test set. This resulted in ten crosswords for FAZ and twelve for each other source.

Clue-answer pairs. We collected a large set of clue-answer pairs crawling two German online crossword web sites: kreuzwortraetsel.de (KWRDE) and kreuzwortraetsel-hilfe.com (KWRH). The download of KWRDE was covered in a period between Friday 9th

Table 2
Crossword Database composition. Clue-answer pairs were collected from different sources.

CWDB	# clue-answer	# unique clue-answer	# unique answers	# unique clues
KWRH	1,312,770	1,312,153	308,680	556,278
KWRDE	1,495,402	1,495,322	357,174	739,146
Spiegel	17247	11251	8783	10711
Bild	13721	8951	6435	8598
Focus	22357	14182	10230	13401
HA	17344	8114	6503	7645
FAZ	1029	1029	941	1029
Total	2,879,870	2,455,604	437,826	1,094,183

September, 2022 and Saturday 1st October, 2022. Similarly, KWRH was downloaded between Thursday 8th September, 2022 and Tuesday 20th September, 2022. As shown in Table 2, a total of 2.4 million unique clue-answer pairs were obtained this way, containing 438 thousand unique answers for German. To supplement the database, the clues from the crosswords in the training set were extracted and added to the database. The individual and total contributions are shown in Table 2.

3.2. Experts Adaptation

CWDB. The 2.4 million clue-answer-pairs crawled were used to build the CWDB expert. After minor preprocessing, clues were embedded with multi-lingual Universal Sentence Encoder (USE) [14] as in [10].

Knowledge Graph. To extend the coverage of our answer retrieval step, we made use of an additional proprietary German ontology[2]. Overall, it contained about 943 thousand lemmas from various web sources like encyclopedias. As for CWDB, we retrieve answer candidates based on semantic similarity [10], thus we treat each lemma as the candidate answer, and its definition is embedded with USE.

Web-Search. To find relevant words, web search experts require a database of document frequency values for the most common words. It was obtained from an online database of words and their frequencies in German movie subtitles [3]. Candidate answers were retrieved only from the web snippets.

Lexicon. The lexicon was constructed from a freely available online German dictionary [4], after romanizing umlauts (e.g. ä to ae) and excluding all non-ASCII and non-alphabetical characters.

[2]from Expert.ai.
[3]github.com/hermitdave/FrequencyWords/blob/master/content/2018/de/de_50k.txt
[4]github.com/enz/german-wordlist/blob/master/words

Implicit Module. The tetra-grams in the Implicit Module were generated from the answers in the CWDB. They include start and end tokens "$" and "^" to allow for specific n-grams at the beginning and endings of words. They are weighted based on their frequency in the corpus.

4. Experiments

In the evaluation we both measured the end-to-end performance of Webcrow and the answer retrieval capabilities of each single module.

4.1. Answer Retrieval

Although good answer ranking is clearly very important for crossword resolution, it is even more essential to have the target answer present in the candidates' list. Indeed, even poorly ranked target answers can be boosted during belief propagation and grid filling, whereas a missing answer in the list would hardly be recovered, inevitably leading to errors or incomplete solutions. Hence, besides the MRR, we also consider coverage as a performance indicator of our experts.

Results of single experts are summarized in Table 3, where we also report the quality after merging and belief-propagation modules. As we can see, CWDB is the most informative expert, with the highest MRR and coverage. Also, Web Search achieves interesting MRR scores. Lexicon and Knowledge Graph have poorer ranking quality, but they both contribute to increasing the coverage, which is the main purpose of those modules. This can be observed from the coverage after Merging, where all the lists are combined together. Clearly, there is a high overlap in the experts' lists, however, the union of all of them reduces the number of missing target answers by 0.7% absolute, almost a third of all the missing target answers.

From Table 3 we can also observe how belief propagation significantly boosts the ranking quality, enormously facilitating the grid-filling stage.

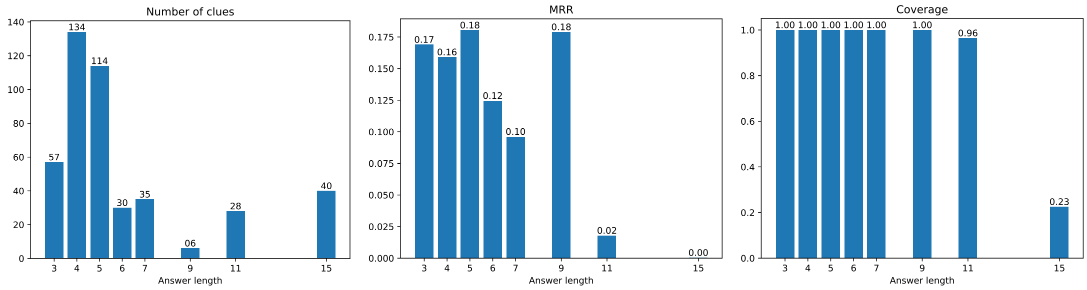


Figure 3: Number of clues, MRR, and Coverage of the merged lists divided by answer length in FAZ crosswords.

Table 3
Results of the individual experts.

Module	MRR	Coverage
Lexicon	0.005	86.8%
Web Search	0.200	75.6%
Knowledge Graph	0.086	91.3%
CWDB	0.634	97.6%
Merging	0.543	98.3%
Belief-Propagation	0.809	98.3%

4.2. Crossword Solving

The whole system was assessed on German full crosswords from the test set described in Section 3. We measured the solving quality with three indicators: the percentage of correctly inserted letters (OK letters), the percentage of correctly inserted words (OK words), and the fraction of perfectly solved crosswords (OK CWs). We report in Table 4 those metrics aggregated per crossword source. Overall 39 out of 58 (about 67% of them) crosswords were perfectly solved by German Webcrow. However, there is a strong variance between different sources. Smaller crossword grids like the ones from “Bild", were always solved without errors. Similar, satisfying performances were obtained in larger grids, also in 16x16 schema from HA. In contrast, Webcrow failed in solving FAZ puzzles, with surprisingly low results. Only 13% of words and 21% of letters were correctly answered in FAZ, far behind near-perfect crosswords from the other sources.

Performance on FAZ Crosswords. Different from other sources, FAZ puzzles are not in Swedish format. Instead, they are characterized by grids populated with many black cells that reduce the number of constraints to impose on the solution. Moreover, the grid layout is disposed in such a way that there is a significant percentage of answers longer than ten characters (see Figure 3), that in our CWDB coming are not present apart from those coming from the few FAZ crosswords in training. Thus, CWDB has little to no coverage of those clues, which are typically very important to constrain a large portion of the grid. Also, the style of the clues is remarkably different. There are many wordplays and linguistic games unusual in the rest of the data, making them challenging even for humans.

Table 4
Results of the end-to-end solving process. For each source, the number of crosswords solved completely correctly over the number of crosswords in the test set are given. Also listed are the average ratio of correct answer words and letters per tested crossword.

Source	OK CWs	OK words	OK letters
Bild	12 / 12	100%	100%
Spiegel	10 / 12	98%	99%
Focus	9 / 12	97%	99%
HA	8 / 12	98%	99%
FAZ	0 / 10	13%	21%

To delve in further, we also analyzed MRR and coverage after merging candidate answers divided by answer length in Figure 3. We can easily notice that both of them are significantly below the scores reported in Table 3 for all the crosswords. In particular, there is a non neglectable portion of long answers with poor coverage and close to zero MRR. Such a modest retrieval quality combined with a less contrained schema inevitably lead the system to fail in solving the crossword.

5. Conclusion

In this work we presented German Webcrow, the first crossword puzzle solver for the German Language. We collected both a dataset of clue-answer pairs and a set of German crosswords from different sources having various formats and styles. Webcrow achieved near-perfect word accuracy in Swedish-type crosswords, that proved to be generally easy to solve, solving overall 39/ 58 perfectly. However, our solver performed poorly on FAZ crosswords. Those puzzles were extremely challenging for multiple reasons, such as the poorly constrained schemas, due to the presence of many black cells, and the rich presence of sophisticated, original clues, involving articulated wordplays that formed words not present in the candidate answers lists.

Challenging puzzles like the ones in FAZ are a clear example of how complex the problem is, and why studying crosswords in multiple languages and formats is important for automated crossword solving. In the future we plan to address the current limitations. In particular, we plan to investigate the use of generative models, to cope with the novel unseen clues.

Acknowledgments

We thank Andreas Weber (raetsel4u.de) for allowing us to reprint one of his crossword puzzles in Figure 1 of this publication.

References

[1] M. L. Littman, G. A. Keim, N. Shazeer, A probabilistic approach to solving crossword puzzles, Artificial Intelligence 134 (2002) 23–55. URL: https://www.sciencedirect.com/science/article/pii/S000437020100114X. doi:https://doi.org/10.1016/S0004-3702(01)00114-X.

[2] M. L. Ginsberg, Dr. fill: Crosswords and an implemented solver for singly weighted csps, Journal of Artificial Intelligence Research 42 (2011) 851–886.

[3] E. Wallace, N. Tomlin, A. Xu, K. Yang, E. Pathak, M. Ginsberg, D. Klein, Automated crossword solving, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 3073–3085. URL: https://aclanthology.org/2022.acl-long.219. doi:10.18653/v1/2022.acl-long.219.

[4] G. Barlacchi, M. Nicosia, A. Moschitti, Learning to rank answer candidates for automatic resolution of crossword puzzles, in: Proceedings of the Eighteenth Conference on Computational Natural Language Learning, 2014, pp. 39–48.

[5] G. Barlacchi, M. Nicosia, A. Moschitti, A retrieval model for automatic resolution of crossword puzzles in italian language, in: The First Italian Conference on Computational Linguistics CLiC-it 2014, 2014, p. 33.

[6] M. Nicosia, G. Barlacchi, A. Moschitti, Learning to rank aggregated answers for crossword puzzles, in: European Conference on Information Retrieval, Springer, 2015, pp. 556–561.

[7] A. Severyn, M. Nicosia, G. Barlacchi, A. Moschitti, Distributional neural networks for automatic resolution of crossword puzzles, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 2015, pp. 199–204.

[8] M. Ernandes, G. Angelini, M. Gori, Webcrow: A web-based system for crossword solving, in: AAAI, 2005, pp. 1412–1417.

[9] M. Ernandes, G. Angelini, M. Gori, A web-based agent challenges human experts on crosswords, AI Magazine 29 (2008) 77–77.

[10] A. Zugarini, M. Ernandes, A multi-strategy approach to crossword clue answer retrieval and ranking, in: E. Fersini, M. Passarotti, V. Patti (Eds.), Proceedings of the Eighth Italian Conference on Computational Linguistics, CLiC-it 2021, Milan, Italy, January 26-28, 2022, volume 3033 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021. URL: http://ceur-ws.org/Vol-3033/paper72.pdf.

[11] M. Iyyer, V. Manjunatha, J. Boyd-Graber, H. Daumé III, Deep unordered composition rivals syntactic methods for text classification, in: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers), 2015, pp. 1681–1691.

[12] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, et al., Universal sentence encoder, arXiv preprint arXiv:1803.11175 (2018).

[13] T. Mikolov, W.-t. Yih, G. Zweig, Linguistic regularities in continuous space word representations, in: Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies, 2013, pp. 746–751.

[14] Y. Yang, D. Cer, A. Ahmad, M. Guo, J. Law, N. Constant, G. H. Abrego, S. Yuan, C. Tar, Y.-H. Sung, et al., Multilingual universal sentence encoder for semantic retrieval, arXiv preprint arXiv:1907.04307 (2019).

Towards a Multilingual System for Vaccine Hesitancy using a Data Mixture Approach

Oscar Araque[1], María Felipa Ledesma-Corniel[1] and Kyriaki Kalimeri[2]

[1]Universidad Politécnica de Madrid, ETSI Telecomunicación, Intelligent Systems Group, Madrid, Spain
[2]ISI Foundation, Turin, Italy

Abstract
Understanding public narratives on contentious topics like vaccination adherence is vital for promoting cooperative behaviors. During the COVID-19 pandemic, significant polarization arose from concerns about vaccines, with misinformation and conspiracy beliefs proliferating on social media. While many studies have analyzed these narratives, the focus has largely been on English-language content. This linguistic bias limits comprehensive global insights. Our study introduces a novel multilingual approach that addresses this gap. By integrating Italian examples into a primarily English dataset, we detect vaccine-hesitant language and demonstrate the model's adaptability to diverse linguistic data. Our findings highlight the importance of incorporating varied linguistic datasets for a more holistic understanding of global narratives on vaccine hesitancy.

Keywords
vaccine hesitancy, natural language processing, machine learning, transformer models

1. Introduction

Automatically understanding peoples' narratives on controversial social issues is fundamental to efficiently address the real concerns as they occur fostering collaborative, prosocial behaviours. Vaccination adherence is an exemplar case where society witnessed a notable polarisation concerning possible adverse reactions [1, 2]. Especially during the COVID-19 pandemic and despite vaccines being the most efficient and cost-effective intervention, the spread of misinformation [3], the scepticism around the scientific development of COVID-19 vaccines and the dissemination of conspiracy beliefs [4, 5], proliferated on social media platforms.

Numerous studies analysed user generated text [6, 7, 8, 9, 10], almost exclusively focusing on the English language due to the availablity of models and tools. Even if often English is universally spoken limits the analysis in specific sociodemographic groups. Lenti et al. [11] in a purely network based approach showed the existence of a global misinformation network, calling for a multilingual analysis to further understand the drivers of vaccine hesitancy in the various languages.

Here, in light of these issues, we propose a novel approach for multilingual language understanding able to deal with language unbalance. More specifically, here we progressively include Italian instances in a predominantly English dataset for the task of vaccine hesitant language detection and demonstrate the ability of the model to generalise on previously unseen data. Often, researchers and practitioners have access to large English datasets but data in other languages, such as Italian, are lacking. We show that including small datasets in different languages can improve overall performance when analyzing texts in several languages.

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
o.araque@upm.es (O. Araque); mf.ledesma@alumnos.upm.es (M. F. Ledesma-Corniel); kyriaki.kalimeri@isi.it (K. Kalimeri)
https://gsi.upm.es/oaraquem (O. Araque)
0000-0003-3224-0001 (O. Araque); 0000-0001-8068-5916 (K. Kalimeri)

2. Data and Methods

2.1. Data Collection

Although several Twitter datasets were constructed to monitor COVID-19 pandemic and are openly available to researchers, they differ in the number, timing, and language of tweets collected, as well as the search keywords used for collection [6, 12]. Here, we opted of a large multilingual dataset (MultilingTw [11]), an Italian dataset [13], while we also performed a new data collection based on a time invariant hashtag list, manually annotated as per their vaccination stance, which we share with the community.

A. Twitter-AntiVax This dataset was collected for this specific study and has been generated by capturing English Twitter messages ranging from December 2020 to March 2023. It aims to capture opinions and narratives expressed by anti-vaccination users, balancing between pro and anti stances. We collected the data using a variety of phrases and hashtags related to vaccination (e.g., "kill jab", "covid jab", "#vaccineskill", "VaccinesAreNotTheAnswer", "vaccineswork", "vaccinessavelives"), manually

Dataset	Pro-vaccine	Anti-vaccine	Total	Total no. tokens	Avg. no. tokens
A. Twitter-AntiVax (ours)	14,374	13,560	27,934	170,299	30.5
B. TwitterVax	7,210	7,150	14,360	61,312	21.3
C. MultilingTw (EN)	416	52	468	16,543	35.3
C. MultilingTw (IT)	141	45	186	6,309	33.9

Table 1
Distribution of instances for each class, number of tokens in total, and average number of tokens per document.

inspecting the relevance of a sample of the obtained messages. From these hashtags, we have identified a set of users (480 users in total) expressing pro and anti-vaccine stances. Finally, we extracted the tweets generated by these users in the considered period. The dataset with the respective annotations is freely accesible at https://github.com/gsi-upm/multilingual-vaccine-hesitancy.

B. TwitterVax dataset [13]. Originally, this dataset contains 9,068,389 Italian tweets on vaccines tweeted from 1st January 2019 to 1st June 2022. The authors annotate each captured user as anti-vaccine and pro-vaccine through network analysis. For this work, to reduce the computational load and to work with similar dataset sizes, we have selected a sub-sample of approximately 14,000 tweets. We have categorized each tweet as anti-vaccine and pro-vaccine by means of the user's annotation.

C. Multilingual Twitter dataset (MultilingTw) [11]. This dataset is composed of Twitter messages in 18 languages from October 2019 to March 2021. While the original size is around 316 million messages, we select a subsample of 1,246 tweets in English and 449 in Italian, manually labelled for their vaccine stance. To comply with the other datasets, we selected the messages labelled as pro and anti-vaccine. These sets are used as test sets.

Both the Twitter-AntiVax and TwitterVax dataset have been split into train and test sets, randomly sampling 20% of instances as test set. Some statistics for the used datasets are detailed in Table 1.

2.2. Methods.

This work is based on a multilingual approach to vaccine hesitancy analysis. To this regard, we use a DistilBERT [14] model (`distil-base-multilingual-cased`[1]). This transformer model was trained in the most common languages in Wikipedia and thus is capable of generating internal representations for a variety of languages, including English and Italian. Nevertheless, it has been shown that this kind of models do not compute language-agnostic representations but rather generates partitioned representations for each language [15]. In practice, this implies that the instances used for training in English are not directly useful for predicting vaccine hesitancy in other languages since the internal representations vary with the language.

Here, our goal is to model the effect of including small sets of data in a multilingual approach. To do so, we use the Twitter-AntiVax train set as English training data and the TwitterVax train set and Italian training data. As test sets, we use the test sets of Twitter-AntiVax and TwitterVax, as well as the Multilingual Twitter dataset in both English and Italian. To modulate the number of Italian instances included in the training set, we define the α parameter that can take values in the range $[0, 1]$. Thus, the instances in the training set are composed with the following expression:

$$\text{Train instances} = \alpha * \text{IT} + (1 - \alpha) * \text{EN}$$

where IT and EN represent the Italian and English datasets, respectively. In this way, a training set composed with $\alpha = 0$ is composed entirely of English instances, while the opposite is correct when $\alpha = 1$. Of course, with $\alpha = 0.5$, the training set would have the same number of instances for English and Italian.

Since the English and Italian datasets contain a different number of instances, this could have the undesired effect of a varying number of training instances that may affect the results. We control this to produce the same number of train instances for all possible values of α.

Evaluation. Finally, all the models are evaluated with the macro-averaged F-score of each model. This allows us to consider the effect of unbalanced data. We opted for an evaluation without label propagation via retweet networks as proposed in other studies [16, 11] since these are likely to introduce uncertainty in the groundtruth. Our evaluation is strictly based on manually annotated data regarding the vaccination stance.

3. Results

As described, the proposed experiment aims to study the effect of including Italian instances in an English dataset and train a multilingual learning model with the generated dataset. Figure 3 shows the macro-averaged f-scores obtained for an increasing number of α. The horizontal axis shows the variation of the α parameter (see Sect. 2), and the vertical axis the performance of

[1]https://huggingface.co/distilbert-base-multilingual-cased

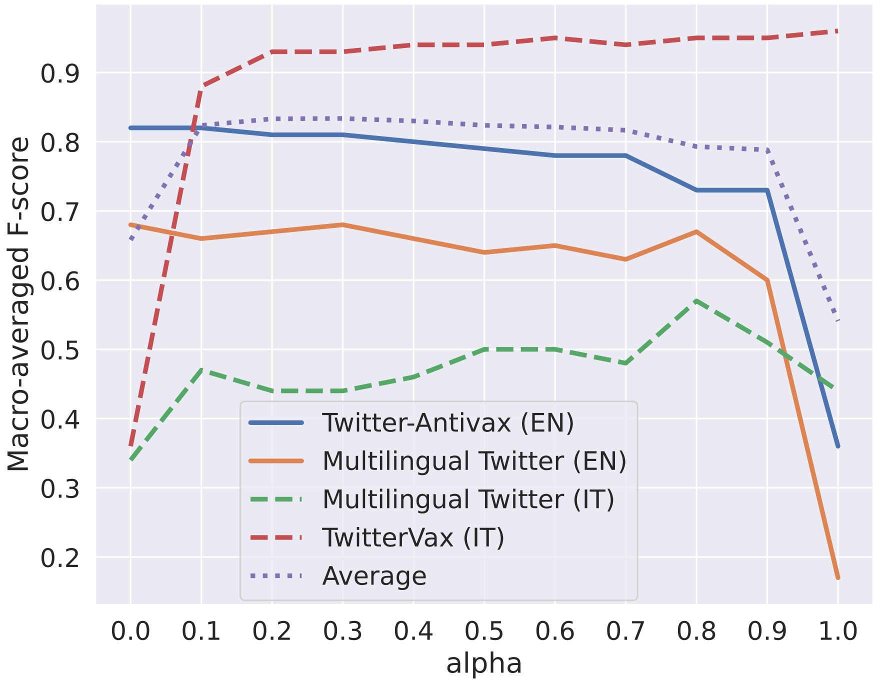


Figure 1: Macro-averaged f-scores in all test sets, both English (EN) and Italian (IT). The averaged curve is weighted with the number of instances of each test dataset.

the model in all test sets, both English and Italian. The average curve is weighted with the number of instances of each test set so that the number of correctly classified instances is better reflected.

It is worth noticing a prevalent behaviour in the obtained learning curves: the elbow evolutions with α in three of the four curves. Attending to the evolution of the performance in the Twitter-Antivax test set, we see that the best performance is obtained when $\alpha = 0$, that is, when all training instances are in English. As α increases, the performance decreases slowly. Nevertheless, when α changes from 0.9 to 1, a faster reduction is observed. The lower performance corresponds to the case where there are no English instances in the training set, negatively affecting the performance in the English language. A similar behaviour is shown by the performance on the Multilingual Twitter data in English.

In contrast, the performance in the Italian data progresses differently. We can see a large improvement in performance at the change between $\alpha = 0$ (no Italian training data) and $\alpha = 0.1$ (10% of training instances in Italian) for the TwitterVax dataset. As more Italian data is included in the composition of the training set, the performance on this dataset increases slowly. As for the Multilingual Twitter dataset in Italian, the performance tends to increase with α.

Attending to the obtained results, we can derive a general trend: the higher the percentage of training instances in a language, the higher the performance in that language. This is to be expected, as follows common experimental observations when training learning models. Besides, it is interesting to see that there is a large portion of cases where performances in both English and Italian are kept high. Practically, this situation is observed when $\alpha \in [0.1, 0.9]$ and can be better understood by attending to the averaged curve.

This interesting behavior may indicate the robustness of the proposed method to the proportion of language mixture. That is, it seems that the model successfully generalizes to a different language even when its training set is composed in a small proportion (e.g., 10%) by instances of that language. The previous observation indicates that the multilingual model may be learning to classify Italian documents while being trained with English instances, and that adding a small proportion of Italian instances facilitates such performance.

While this work is an initial attempt at describing a multilingual system trained with a mixture of data, further work should explore whether the observed behavior is maintained with more languages. How the internal representations of the evaluated model can be used for multilingual applications has yet to be thoroughly studied.

the Multilingual Twitter dataset.

4. Conclusions

Here we design and evaluate a method that achieves multilingual vaccine hesitancy detection. The experimental design considers training a multilingual classification model on a mixture of English and Italian text excerpts. Progressively varying the combination of languages in the training data, we obtain a better understanding of the of the classification problem in the two languages. Additionally, we undertook a novel data collection effort on Twitter, manually annotating content based on vaccination stance. This curated dataset is now freely accessible to the scientific community, providing a valuable resource for further research.

By adjusting the language composition in our training data, we gained deeper insights into the classification intricacies across both languages. Notably, our findings suggest that the model can effectively generalize to a different language even when its training set contains a minimal proportion (e.g., 10%) of instances from that language. This indicates the model's robustness and adaptability in handling linguistic variations with limited data.

Importantly, this approach is an important tool for researchers and practitioners who often have access to large datasets in English, but limited resources in other widely spoken languages such as Italian or Spanish. The evaluation shows that composing a mixture dataset can be effective in generating a model that classifies instances in two languages. In fact, the experimentation shows that this mixture is flexible, maintaining consistent performances across different ratios of language presence. This consistency suggests that the mixture approach is promising.

Given its language-neutral nature, our technique holds promise for broader applications across multiple languages and diverse domains. As a next step, we aim to explore various multilingual models and languages to further ascertain the scalability and adaptability of our approach.

Acknowledgments

This work has been funded by the Spanish Ministry of Science and Innovation through the COGNOS project (PID2019-105484RB-I00) and by the European Union with NextGeneration EU funds. KK gratefully acknowledges the support from the Lagrange Project of the Institute for Scientific Interchange Foundation (ISI Foundation) funded by Fondazione Cassa di Risparmio di Torino (Fondazione CRT).

The authors would like to acknowledge the support of Yelena Mejova, from ISI Foundation in Italy, for sharing the Multilingual Twitter dataset.

References

[1] A. A. Dror, N. Eisenbach, S. Taiber, N. G. Morozov, M. Mizrachi, A. Zigron, S. Srouji, E. Sela, Vaccine hesitancy: the next challenge in the fight against covid-19, European journal of epidemiology 35 (2020) 775–779.

[2] L. Betti, G. De Francisci Morales, L. Gauvin, K. Kalimeri, Y. Mejova, D. Paolotti, M. Starnini, Detecting adherence to the recommended childhood vaccination schedule from user-generated content in a us parenting forum, PLoS computational biology 17 (2021) e1008919.

[3] Y. Mejova, K. Kalimeri, Covid-19 on facebook ads: competing agendas around a public health crisis, in: Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies, 2020, pp. 22–31.

[4] K. Kalimeri, M. G. Beiró, A. Urbinati, A. Bonanomi, A. Rosina, C. Cattuto, Human values and attitudes towards vaccination in social media, in: Companion Proceedings of The 2019 World Wide Web Conference, 2019, pp. 248–254.

[5] M. G. Beiró, J. D'Ignazi, V. Perez Bustos, M. F. Prado, K. Kalimeri, Moral narratives around the vaccination debate on facebook, in: Proceedings of the ACM Web Conference 2023, 2023, pp. 4134–4141.

[6] K. Hayawi, S. Shahriar, M. A. Serhani, I. Taleb, S. S. Mathew, Anti-vax: a novel twitter dataset for covid-19 vaccine misinformation detection, Public health 203 (2022) 23–30.

[7] S. Nyawa, D. Tchuente, S. Fosso-Wamba, Covid-19 vaccine hesitancy: a social media analysis using deep learning, Annals of Operations Research (2022) 1–39.

[8] A. Fasce, P. Schmid, D. L. Holford, L. Bates, I. Gurevych, S. Lewandowsky, A taxonomy of anti-vaccination arguments from a systematic literature review and text modelling, Nature Human Behaviour (2023) 1–19.

[9] Y. Mejova, G. Crupi, J. Lenti, M. Tizzani, K. Kalimeri, D. Paolotti, A. Panisson, Echo chambers of vaccination hesitancy discussion on social media during covid-19 pandemic, in: XX ISA World Congress of Sociology (June 25-July 1, 2023), ISA, 2023.

[10] N. E. MacDonald, Vaccine hesitancy: Definition, scope and determinants, Vaccine 33 (2015) 4161–4164. URL: https://www.sciencedirect.com/science/article/pii/S0264410X15005009. doi:`https://doi.org/10.1016/j.vaccine.2015.04.036`, wHO Recommendations Regarding Vaccine Hesitancy.

[11] J. Lenti, Y. Mejova, K. Kalimeri, A. Panisson, D. Paolotti, M. Tizzani, M. Starnini, Global misinformation spillovers in the vaccination debate before and during the covid-19 pandemic: Multilingual twitter study, JMIR Infodemiology 3 (2023) e44714. doi:10.2196/44714.
[12] C. E. Lopez, C. Gallemore, An augmented multilingual twitter dataset for studying the covid-19 infodemic, Social Network Analysis and Mining 11 (2021) 102.
[13] V. Lachi, G. M. Dimitri, A. Di Stefano, P. Liò, M. Bianchini, C. Mocenni, Impact of the covid 19 outbreaks on the italian twitter vaccination debat: a network based analysis, arXiv preprint arXiv:2306.02838 (2023).
[14] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, arXiv preprint arXiv:1910.01108 (2019).
[15] J. Singh, B. McCann, R. Socher, C. Xiong, BERT is not an interlingua and the bias of tokenization, in: Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 47–55. URL: https://aclanthology.org/D19-6106. doi:10.18653/v1/D19-6106.
[16] F. Gargiulo, F. Cafiero, P. Guille-Escuret, V. Seror, J. K. Ward, Asymmetric participation of defenders and critics of vaccines to debates on french-speaking twitter, Scientific reports 10 (2020) 6599.

TA*ll*: a new Shiny app of Text Analysis for *All*

Massimo Aria[1,4], Corrado Cuccurullo[2,4], Luca D'Aniello[1,4], Michelangelo Misuraca[3,4] and Maria Spano[1,4]

[1] *University of Naples Federico II, 80126 Napoli, Italy*
[2] *University of Campania "Luigi Vanvitelli", 81043 Capua (CE), Italy*
[3] *University of Calabria, 87036 Arcavacata di Rende (CS), Italy*
[4] *K-Synth spin-off, University of Naples Federico II, 80126 Napoli, Italy*

Abstract

The rapid technological advancements in recent years allowed to process different kinds of data to study several real-world phenomena. Within this context, textual data has emerged as a crucial resource in numerous research domains, opening avenues for new research questions and insights. However, many researchers lack the necessary programming skills to effectively analyze textual data, creating a demand for user-friendly text analysis tools. While languages such as R and python provide powerful capabilities, researchers often face constraints in terms of time and resources required to become proficient in these languages.

This paper introduces TA*ll* - Text Analysis for All, an R Shiny app that includes a wide set of methodologies specifically tailored for various text analysis tasks. It aims to address the needs of researchers without extensive programming skills, providing a versatile and general-purpose tool for analyzing textual data. With TA*ll*, researchers can leverage a wide range of text analysis techniques without the burden of extensive programming knowledge, enabling them to extract valuable insights from textual data in a more efficient and accessible manner.

Keywords

text analysis, shiny app, web app

1. Introduction

In the era of big data, researchers across various disciplines are increasingly faced with the challenge of analyzing vast amounts of textual data.

Textual data, such as research articles, social media posts, customer reviews, and survey responses, hold valuable insights that can contribute to the advancement of knowledge in fields ranging from social sciences to healthcare and beyond.

Researchers seek to analyze textual data to uncover patterns, identify trends, extract meaningful information, and gain deeper insights into various phenomena. By employing advanced natural language processing (NLP) techniques and machine learning algorithms, researchers can explore the semantic and syntactic structures of texts, perform topic detection, polarity detection, and text summarization among other analyses. Moreover, the advent of digital platforms and the proliferation of online content have generated vast amounts of textual data that were previously inaccessible or challenging to obtain. Researchers can tap into these resources to explore new research questions, validate existing theories, and generate novel insights.

By harnessing the power of computational tools and techniques, researchers can efficiently process and analyze large volumes of text, significantly reducing the time and effort required compared to manual analysis. Moreover, there is a growing recognition of the need for text analysis tools that cater to individuals who may not possess extensive programming skills. While programming languages like R and python provide powerful capabilities for data analysis, not all researchers have the time or resources to acquire proficiency in these languages.

This paper presents the first version of TA*ll* - Text analysis for All - a new R Shiny app that brings together all the major advancements in text analysis developed in recent years. For researchers who lack programming skills, TA*ll* offers a viable solution, providing an intuitive interface that allow researchers

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

maria.spano@unina.it (M. Spano); massimo.aria@unina.it (M. Aria); corrado.cuccurullo@unicampania.it (C. Cuccurullo); luca.daniello@unina.it (L. D'Aniello); michelangelo.misuraca@unical.it (M. Misuraca)

0000-0002-3103-2342 (M. Spano); 0000-0002-8517-9411 (M. Aria); 0000-0002-7401-8575 (C. Cuccurullo); 0000-0003-1019-9212 (L. D'Aniello); 0000-0002-8794-966X (M. Misuraca)

to interact with data and perform analyses without the need for extensive programming knowledge.

TA*ll* offers a comprehensive workflow for data cleaning, pre-processing, statistical analysis, and visualization of textual data, by combining state-of-the-art text analysis techniques into an R Shiny app.

2. Discovering TA*ll* workflow

First TA*ll* combines the functionality of a set of R packages developed for NLP tasks (see: https://cran.r-project.org/web/views/NaturalLanguageProcessing.html) with the ease of use of web apps using the Shiny package environment. TA*ll* workflow aims to facilitate the discovery and analysis of text data by systematically processing and exploring the content.

Figure 1 shows the three main steps of a TA*ll* workflow.

annotation process (i.e., tokenization, PoS tagging, and lemmatization).

TA*ll* utilizes pre-trained models provided by Universal Dependencies Treebanks. Universal Dependencies (https://universaldependencies.org) is a framework for consistent annotation of grammar (Part-of-Speech, morphological features, and syntactic dependencies) across different human languages. UD is an open community effort with over 500 contributors producing over 200 treebanks in over 100 languages. By using these models TA*ll* supports the analysis of texts written in 60 different languages.

Each text is parsed into individual tokens (words) and tagged with its respective Part-of-Speech (PoS) label to better understand word usage patterns. All the subsequent statistical analyses could be performed alternatively on tokens or lemmas. Moreover, users can define and load custom lists of words for various research purposes (e.g., to substitute synonyms, remove domain stopwords, and semantically tag specialized lexicons'terms).

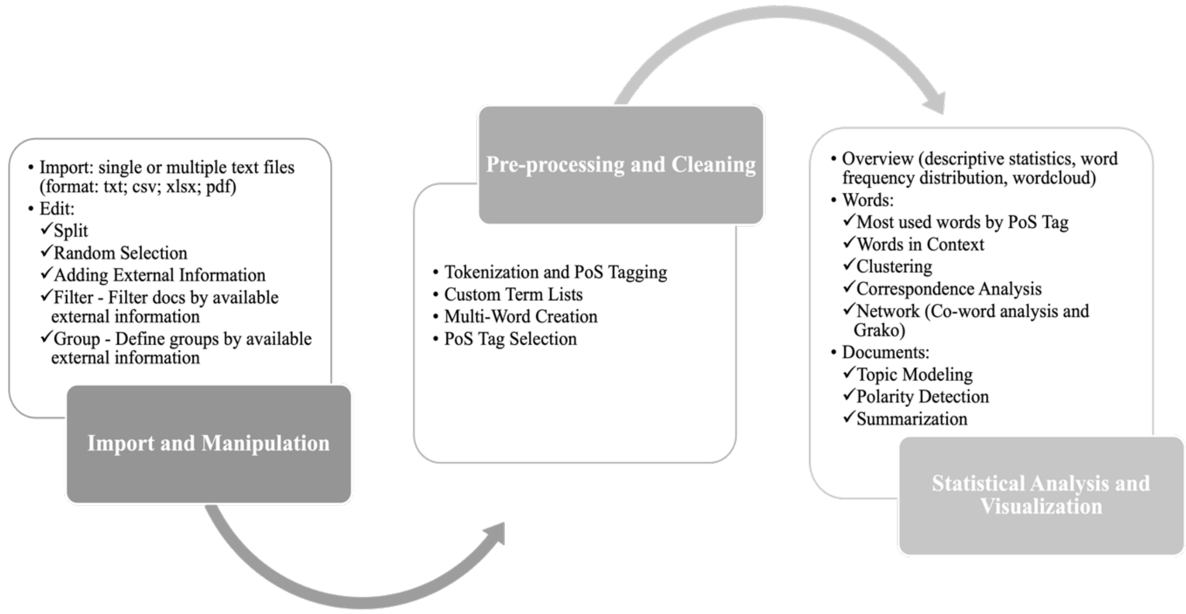


Figure 1: Discovering TA*ll* tabs, methods and workflow

The first step **Import and Manipulation** involves importing one or multiple text files in various formats, such as txt, csv, xlsx, and pdf, allowing easy loading of a diverse range of textual data. Subsequently, texts could be subjected to several editing actions, including the division into smaller segments, such as chapters or paragraphs, or the selection of texts' subsets for sampling or random analysis purposes. Users can supplement the imported texts with additional external information (e.g., author, publication date, rating) attached to the texts or added during the analysis. Concerning both the aim of analysis and the availability of external variables, texts could be filtered, enabling to focus on specific subsets or grouped for comparison purposes.

Before beginning the **Pre-processing and Cleaning** step, a language model was necessary for the

A crucial aspect when we deal with text analysis is to identify and handle multi-word expressions and collocations. To face this issue, TA*ll* performs *Rapid Automatic Keyword Extraction* (RAKE) algorithm [1] that uses a delimiter-based approach to identify candidate keywords and scores them using word co-occurrences that appear in the candidate keywords.

At the end of pre-processing and cleaning phase, users can select specific PoS tags to focus their analyses considering only content-bearing words (e.g., nouns, adjectives, verbs, collocations).

Statistical analysis and Visualization step opens the opportunity of exploring cleaned texts by performing one or more approaches as listed in Figure 1. Descriptive statistics (e.g., number of tokens, types, sentences, lexical measures), word frequency distribution, and wordclouds provide an initial overview of the text corpus. TA*ll* tabs are then organized by considering two levels of analysis: words and documents.

Detailed analysis of words includes a set of statistical methods mainly devoted to topic detection. The most intuitive approach is to identify and visualize through dynamic plots the most frequently used words for each PoS tag, looking obviously at their absolute frequency but also considering more complex weighting schemes as *term frequency/inverse document frequency* (TF-IDF) [2] to uncover words with the highest discriminative power. Despite the simplicity, often analyzing the frequency distribution of words gives a general idea of the contents in the text collection, but it is not enough to identify topics. A topic can be represented as a set of meaningful words with syntagmatic relatedness [3].
Following this definition, the three methods most widely shared in the literature [4] are implemented in TA*ll*:

- *Clustering* [5, 6] to group similar words based on their usage patterns or context;
- *Correspondence Analysis* [7, 8] to explore semantic relationships among words, identifying the latent structure of the text collection;
- *Network* (Co-word analysis and Grako) [9] to analyze co-occurrence patterns of words within texts, highlighting subsets of words strictly related through community detection algorithms [10].

The documents tab includes a set of statistical methods to cope with specific tasks where the focus is properly on the entire documents:

- *Topic Modeling* to identify both prominent topics and their distribution within documents using the well-known Latent Dirichlet Allocation (LDA) algorithm [11]. Moreover, TA*ll* estimates the number of topics automatically through the measures proposed in [12, 13, 14, 15], but users can also explore different solutions by setting the number of the desired topics;
- *Polarity Detection* to determine the polarity (positive, negative, neutral) of documents by choosing among different lexicons (i.e., Hu & Liu [16], Loughran & McDonald [17], nrc [18]);
- Summarization to concisely summarize each text to capture key insights rapidly. TA*ll* performs TextRank algorithm [19], based on applying Google's PageRank [20] to the network of sentences for extracting the most relevant ones.

This comprehensive workflow provides users with the statistical methods to process texts efficiently and share their results and workflows with collaborators by downloading plots and reports from TA*ll*, facilitating and speeding up all analysis steps. paragraph in every section does not have first-line indent. Use only styles embedded in the document.

3. Conclusion and remarks

This paper presented a brief overview of the first version of TA*ll*, a new shiny app for importing, pre-processing, and analyzing textual data.
Our idea stems from the now growing need to analyze textual content to today's ever-increasing number, offering the opportunity to explore it quickly and efficiently, even for those without programming skills. Using a user-friendly text analysis tool, researchers can focus more on their domain expertise and research questions rather than spend significant time learning programming languages or writing complex code. Tools like TA*ll* democratize text analysis, making it accessible to a broader audience and promoting interdisciplinary collaboration.
Moreover, general-purpose software can be used in every research field and encourages reproducibility and transparency in research. paragraph in every section does not have first-line indent. Use only styles embedded in the document.

References

[1] S. Rose, D. Engel, N. Cramer, W. Cowley, 2010. Automatic keyword extraction from individual documents, pages 1–20. Wiley Online Library.

[2] G. Salton, C. Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.

[3] H. Schütze, 1993. A vector model for syntagmatic and paradigmatic relatedness. In Making sense of words: 9th annual conference of the UW Centre for the New OED and Text Research.

[4] M. Misuraca, M. Spano, 2020. *Unsupervised Analytic Strategies to Explore Large Document Collections*, pp. 17–28. Heidelberg: SPRINGER, 06.

[5] A. K. Jain, M. N. Murty, P. J. Flynn. 1999. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323.

[6] D. Xu and Y. Tian, 2015. A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2:165–193.

[7] J. P. Benzécri, 1982. *Histoire et préhistoire de l'analyse des données*. Dunod, Paris.

[8] L. Lebart, A. Salem, L. Berry, 1997. *Exploring textual data*, volume 4. Springer Science & Business Media.

[9] M. Callon, J.-P. Courtial, W. A. Turner, S. Bauin, 1983. From translations to problematic networks: An introduction to co-word analysis. *Social science information*, 22(2):191–235.

[10] S. Fortunato, D. Hric, 2016. Community detection in networks: A user guide. *Physics Reports*, 659:1–44, nov.

[11] D. M. Blei, A. Y. Ng, M. I Jordan, 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993– 1022.

[12] T. L. Griffiths, M. Steyvers, 2004. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235.

[13] R. Deveaud, E. Sanjuan, P. Bellot. 2014. Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique*, 17:61–84, 06.

[14] J. Cao, T. Xia, J. Li, Y. Zhang, S. Tang, 2009. A density-based method for adaptive lda model selection. *Neurocomputing*, 72(7):1775–1781. Advances in Machine Learning and Computational Intelligence.

[15] R. Arun, V. Suresh, C. E. Veni Madhavan, M. N. Narasimha Murthy, 2010. On finding the natural number of topics with latent dirichlet allocation: Some observations. In Mohammed J. Zaki, Jeffrey Xu Yu, B. Ravindran, and Vikram Pudi, editors, *Advances in Knowledge Discovery and Data Mining*, pages 391–402, Berlin, Heidelberg. Springer Berlin Heidelberg.

[16] M. Hu, B. Liu, 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pp. 168–177, New York, NY, USA. Association for Computing Machinery.

[17] T. Loughran, B. McDonald, 2016. Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4):1187–1230.

[18] S. Mohammad, P. Turney, 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, pp. 26–34, Los Angeles, CA, June. Association for Computational Linguistics.*

[19] R. Mihalcea, P. Tarau, 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 404–411, Barcelona, Spain, July. Association for Computational Linguistics.

[20] L. Page, S. Brin, R. Motwani, T. Winograd, 1998. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford Digital Library Technologies Project.

On the impact of Language Adaptation for Large Language Models: A case study for the Italian language using only open resources

Pierpaolo Basile[1], Pierluigi Cassotti[1], Marco Polignano[1], Lucia Siciliani[1] and Giovanni Semeraro[1]

[1]University of Bari Aldo Moro, Department of Computer Science, via E. Orabona, 70125, Bari, Italy

Abstract

The BLOOM Large Language Model is a cutting-edge open linguistic model developed to provide computers with natural language understanding skills. Despite its remarkable capabilities in understanding natural language by capturing intricate contextual relationships, the BLOOM model exhibits a notable limitation concerning the number of included languages. In fact, Italian is not included among the languages supported by the model making the usage of the model challenging in this context. Within this study, using an open science philosophy, we explore different *Language Adaptation* strategies for the BLOOM model and assess its zero-shot prompting performance across two different downstream classification tasks over EVALITA datasets. It has been observed that language adaptation followed by instruction-based fine-tuning is shown to be effective in correctly addressing a task never seen by the model in a new language learned on a few examples of data.

Keywords

Natural Language Processing, Language Adaptation, Large Language Model

1. Introduction

As language diversity becomes increasingly important in the digital age, the capability of a Natural Language Understanding model to handle a wide array of languages gains significance. Large Language Models (LLMs) have emerged as excellent approaches for comprehending, generating, and manipulating human language with unprecedented accuracy and fluency [1].

They can grasp nuances, idioms, and even ambiguous phrases, enabling more accurate sentiment analysis, question answering, and information retrieval tasks. This enhanced understanding contributes to more effective communication between humans and machines, fostering seamless interactions across various applications. LLMs possess remarkable generalization capabilities, allowing them to perform well on tasks they were not explicitly trained for, also in a multilingual fashion. Among the largest and most effective Large Language Models can be found BLOOM [2], a 176B-parameter open-access language model designed and built thanks to the collaboration of hundreds of researchers. BLOOM is a decoder-only Transformer language model that was trained on a large corpus comprising hundreds of sources in 46 natural and 13 programming languages, culminating in a comprehensive dataset that spans 59 languages in total. Nevertheless, it excludes some of the world's most widely spoken languages, including Russian, Korean, and Italian, raising the need for a more inclusive linguistic approach. Training an effective LLM focused solely on a particular language is a prohibitive challenge, given the substantial volumes of data and resources required for such a task. At the same time, tackling downstream tasks in a specific language effectively necessitates a model with a comprehensive understanding of that language.

Our hypothesis focuses on the Language Adaptation methodology, which is particularly fascinating for addressing the challenge of transferring knowledge from a pre-trained Language Model (LM) to a specific application language. In this context, we aim to adapt BLOOM models to work with a new language, such as Italian, using only a limited sample size, i.e., 100,000 samples.

Indeed, we evaluated the adapted models after a phase of instruction-based fine-tuning on two different classification tasks using Italian data. Our experiments demonstrate that the Language Adaptation process improves the zero-shot ability of the model if executed for the same language of the evaluating data. One of the most important aims of our work is the development of all the methodologies using an open-science approach without using private data created or elaborated by no open-source tools. In addition to this, all data and models used in this work are under an open-source license, reflecting our

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

pierpaolo.basile@uniba.it (P. Basile); pierluigi.cassotti@uniba.it (P. Cassotti); marco.polignano@uniba.it (M. Polignano); lucia.siciliani@uniba.it (L. Siciliani); giovanni.semeraro@uniba.it (G. Semeraro)

0000-0002-0545-1105 (P. Basile); 0000-0001-7824-7167 (P. Cassotti); 0000-0002-3939-0136 (M. Polignano); 0000-0002-1438-280X (L. Siciliani); 0000-0001-6883-1853 (G. Semeraro)

commitment to transparency and collaborative attitude towards the scientific community. We want to prove that it is possible to foster innovation and build effective LLMs using only open resources.

2. Language Adaptation Approaches

LLMs, such as GPT [3], Vicuna [4], LLaMA [5], or BLOOM [2], are trained on vast amounts of text data from diverse sources, which gives them a broad understanding of language and context. Nonetheless, it is important to note that the general knowledge inherent in these models might not be optimised for a particular language [6]. For this reason, Language Adaptation can strongly support the model's capacity to effectively navigate and address downstream tasks in a specific language. Language Adaptation of LLMs refers to the process of tuning a pre-trained LM to work effectively with a specific target language. In the scientific literature, different approaches for Language Adaptation have been recently proposed [7]. Among them, we can distinguish i) continuing the pre-training on new data [8], ii) creating a model adapter [9], iii) training a random subset of the model parameters [10].

In this work, we focus mainly on the MAD-X strategy [11] that has already been applied to BLOOM and proved to perform well in several languages as reported in [7]. Adapters were originally applied in NLP for parameter efficiency and quick fine-tuning of a base pre-trained Transformer model to new domains and tasks. [12] for the first time exploited adapters for transferring a pre-trained monolingual model to an unseen language by relying on learning new token-level embeddings. However, this solution does not scale to a large number of languages. Essentially, it is possible to adapt a pre-training multilingual LM to another unseen language L_u by 1) fine-tuning the model directly to a specific task in L_u; 2) using the obtained model for performing inference in L_u.

Due to the multilingual nature of the original model, it is not possible to obtain good performance in L_u since the model tends to balance many languages. On the contrary, the MAD-x strategy tries to tailor the original model to the target language by using an adapter. The idea is to fine-tune the original model using the masked language learning strategy instead of fine-tuning it on a specific task. This allows the use of unlabeled data written using the target language for fine-tuning the model to that specific language. The invertible adapters strategy provided by the MAD-X configuration facilitates the adaptation of BLOOM to the Italian language. The language adapter, located within each Transformer block, consists of a bottleneck adapter with down- and up-projection feedforward layers. Meanwhile, the invertible adapter works in the embedding layers to address the discrepancies between the vocabularies of the original and newly introduced languages.

3. Adaptation Pipeline

Starting from BLOOM, we build three models, i.e.:

- the BLOOM model with language adaptation for Italian and fine-tuned using the EVALITA training data and the instruction-based dataset Dolly (**B-it-D-E**);
- the BLOOM model with language adaptation fine-tuned using only the EVALITA training data (**B-it-E**);
- the BLOOM model without language adaptation fine-tuned with the EVALITA training data (**B-E**).

The adaptation and fine-tuning process is sketched in Figure 1.

To reduce computational costs, we decided to use the 1B version of the BLOOM model[1] (i.e., "BLOOM-1b7"). It has not been trained on any dialogue instruction as the counterpart BLOOMZ, and it does not contain the training data documents written in the Italian language. We follow the hypothesis that instruction-based fine-tuning should be performed after a phase of Language Adaptation, with instructions provided in the specific language of interest.

As a language adaptation strategy, we use MAD-X [7]. To produce a valuable model, we follow the suggestions of the authors of the paper, using default script parameters and selecting a sample of 100,000 sentences in Italian. We decided to sample data from the Filtered Oscar Dataset [13] for the Italian Language[2] released by [14].

Over the language-adapted models, we perform a general-purpose instruction-based fine-tuning step. Specifically, we use a version of the Dolly Instruction Dataset [15], which was adequately translated into Italian. For the translation, we opt to use an open-source tool[3] instead of a closed software. Dolly[4] is made of 15k high-quality human-generated prompt/response pairs specifically designed for instruction tuning LLMs. The dataset was authored by more than 5,000 Databricks[5] employees during March and April of 2023, and instructions are not copied from the web or other LLMs.

The instructions are mostly about Open/Closed Q&A, ExtractSummarize information, Brainstorming, Classification, and Creative writing. This fine-tuning step has

[1] https://huggingface.co/bigscience/bloom-1b7
[2] https://huggingface.co/datasets/gsarti/clean_mc4_it
[3] https://pypi.org/project/argostranslate/
[4] https://huggingface.co/datasets/databricks/databricks-dolly-15k
[5] https://www.databricks.com/

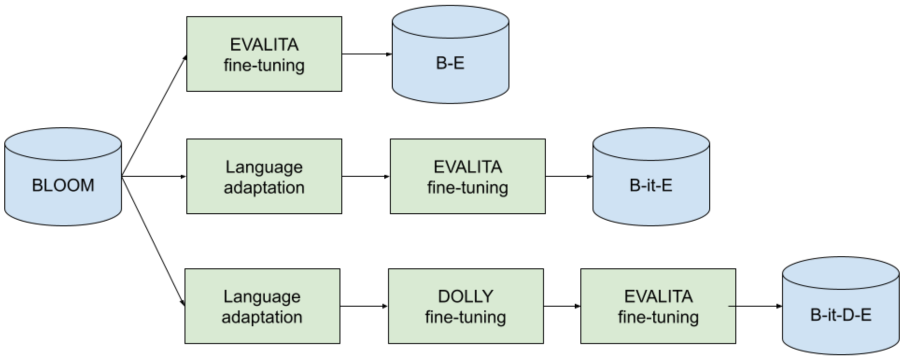


Figure 1: The adaptation pipeline

been performed by adapting the Python script released through GitHub [6].

Finally, we opt to fine-tune the models over two classification task prompts. To deal with this step, we decided to use data from two well-known EVALITA tasks, i.e., AMI2020 [16] and HaSpeeDe-v2-2020 [17]. AMI (Automatic Misogyny Identification) is aimed at automatically identifying misogyny in Tweets written in Italian. More specifically, Subtask A is focused on predicting Misogyny and Aggressiveness independently, while Subtask B is focused only on Misogyny but the dataset has been enriched with synthetic template-generated text. The HaSpeeDe task is focused on Hate Speech detection. The whole task is built on Tweets in Italian and Subtask A is aimed at determining whether the message contains Hate Speech or not, while Subtask B consists in determining whether the message contains Stereotype. All these tasks are structured as binary classification problems, where the label can be either `true` or `false`.

To use these resources to fine-tune our models, we first transformed the training data of the two tasks into an LLM prompt following a template. In particular, for the AMI task, we used the following template: *"instruction": "Nel testo seguente si esprime odio contro le donne? Rispondi sì o no.", "input": <training_text>, "output": <sì/no>*. Similarly, for HASPEEDE we used: *"instruction": "Il testo seguente incita all'odio? Rispondi sì o no.", "input": <training_text>, "output": <sì/no>*. To fill these templates, we mapped the label "1" with the word "sì" and the label "0" with the word "no", <training_text> is just the sentence from the dataset to classify. The fine-tuning step has been performed by using the same script as the previously described Dolly adaptation process.

[6]https://github.com/hyintell/BLOOM-fine-tuning/tree/main

3.1. Data Release

Following the open-science principles, we release all the models on HuggingFace. The available models are:

- **B-E**: the BLOOM model fine-tuned on EVALITA data without language adaptation[7];
- **B-it-E**: the BLOOM model adapted on the Italian language and fine-tuned on EVALITA data[8];
- **B-it-D-E**: the BLOOM model adapted on the Italian language and fine-tuned on both Dolly and EVALITA data[9].

It is important to underline that when you use the adapted LLM or one of its fine-tuned models is necessary to use the tokenizer of the adapted model. The BLOOM model adapted to the Italian language is available on HuggingFace[10].

The data used for fine-tuning are:

- the Italian translation of the Dolly dataset[11];
- the instructions generated from EVALITA data for training and test[12].

4. Validation and Discussion of Results

For the evaluation of the zero-shot abilities of the obtained models, we used the test data of AMI2020 and HASPEEDE-v2-2020. Also, in this case, the datasets have been translated into LLM prompts using the previously

[7]https://huggingface.co/basilepp19/bloom-1b7-evalita
[8]https://huggingface.co/basilepp19/bloom-1b7-it-evalita
[9]https://huggingface.co/basilepp19/bloom-1b7-it-dolly-evalita
[10]https://huggingface.co/basilepp19/bloom-1b7_it
[11]https://huggingface.co/datasets/basilepp19/dolly-15k-it
[12]https://huggingface.co/datasets/basilepp19/evalita2020-AH-instr

reported templates. The model output “sì/no” has been mapped back to the original labels 1/0 to allow a standard model evaluation setting.

Task	B-E	B-it-E	B-it-D-E	Baseline
AMI				
Subtask A	0.702	**0.730**	0.714	0.665
Subtask B	0.695	**0.785**	0.762	0.602
Haspeede2				
Task A (news)	0.518	0.555	**0.579**	0.621
Task A (tweets)	**0.706**	0.670	0.667	0.721
Task B (news)	0.584	0.623	**0.650**	0.669
Task B (tweets)	0.672	**0.686**	0.658	0.715

Table 1
Results

In Table 1, we report the results for AMI and Haspeede 2 tasks for all the three models: **B-E**, **B-it-E** and **B-it-D-E**. The evaluation metrics used are the Average Macro F1-score (F1) for AMI Subtask A and the Macro F1-score for Haspeede 2 tasks. The AMI Subtask A required predicting *Misogyny* and *Aggressiveness* classes independently using the Macro F1-score. The final score is obtained by averaging the two macro F1 scores. The results show that the MAD-X language adaptation with EVALITA fine-tuning (**B-it-E)** achieved the highest Average Macro F1-score of 0.730. The model that exploits Dolly fine-tuning obtains slightly worse results with an Average Macro F1-score of 0.714. In all the configurations, the model overcomes the task baseline. AMI Subtask B ranks model runs based on a weighted combination of AUC scores from the test raw dataset and three per-term AUC-based bias scores from the synthetic dataset, considering the performance for specific identity terms (e.g. “girlfriend” and “wife”). The results for Subtask B of the AMI task indicate that the BLOOM model with MAD-X language adaptation and EVALITA fine-tuning (**B-it-E**) achieved the highest Macro F1-score of 0.785, outperforming the other models and surpassing the task baseline significantly.

In Haspeede2 Task A, the models were required to classify *hateful* content, while Task B aimed to identify the presence of *stereotypes* related to the same targets, such as immigrants, Muslims and Roma. For the Haspeede2 Task A on news, BLOOM alone (**B-E**) achieved a moderate score of 0.522. The combination of MAD-X and EVALITA (**B-it-E**) improved the performance to 0.540, and adding Dolly fine-tuning (**B-it-D-E**) further increased it to 0.589. BLOOM alone yielded a reasonable score of 0.706 for tweets on Task A. While the other models which use language adaptation slightly decreased the performance.

In the Haspeede2 Task B evaluation, the MAD-X adaptation has proven to be remarkably effective for both news and tweets. Specifically, when applied to news data, it yielded outstanding results of 0.650 using both EVALITA and Dolly fine-tuning. Meanwhile, for tweets, the MAD-X adaptation achieved an even higher performance, reaching a score of 0.686 when using only the EVALITA fine-tuning. These findings highlight the adaptability and superiority of language adaptation (MAD-X) in handling different data types. All models cannot overcome the baselines in Haspeede2, but in three cases, the language adaptation provides the best result.

For fine-tuning and testing our models, we use a single NVIDIA A6000 GPU with 48 GB of RAM. The language adaptation steps require about 15 hours, while the fine-tuning of EVALITA 7 hours, and the Dolly fine-tuning only 5 hours.

5. Conclusions

In this paper, we explored a language adaptation strategy for the BLOOM model to address the challenge of handling languages not covered during the training.

Our approach is distinguished by its reliance on open-source data, software, and models, aligning with a commitment to transparency and accessibility in research. Despite the remarkable capabilities of the BLOOM model in understanding natural language for widely spoken languages, it showed limitations when applied to languages which are not included in the original training set, such as Italian. To overcome this limitation, we conducted experiments using the MAD-X language adaptation approach followed by instruction-based fine-tuning on Italian data.

The outcomes of our research demonstrate the effectiveness of language adaptation in significantly improving the zero-shot ability of the BLOOM model for Italian. The combination of MAD-X language adaptation with EVALITA fine-tuning achieved the highest performance on both the AMI2020 and HASPEEDE 2 tasks, showcasing the importance of the adaptation process for downstream classification tasks in Italian. In future work, we plan to evaluate our approach to more large BLOOM models and more recent tasks for the Italian.

The proposed methodology can be adapted for other languages and requires few examples for obtaining satisfying results. The adapted models can be easily fine-tuned on several tasks providing proper instructions. Our future research will extend to testing this approach with other languages and diverse adaptation strategies, contributing to the broader landscape of language model adaptability.

Acknowledgments

We acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), Spoke 6 - Symbiotic AI (CUP H97G22000210007) under the NRRP MUR program funded by the NextGenerationEU.

References

[1] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Barnes, A. Mian, A comprehensive overview of large language models, arXiv preprint arXiv:2307.06435 (2023).

[2] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, et al., Bloom: A 176b-parameter open-access multilingual language model, arXiv preprint arXiv:2211.05100 (2022).

[3] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[4] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, et al., Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, See https://vicuna. lmsys. org (accessed 14 April 2023) (2023).

[5] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).

[6] K. Nowakowski, M. Ptaszynski, K. Murasaki, J. Nieuważny, Adapting multilingual speech representation model for a new, underresourced language through multilingual fine-tuning and continued pretraining, Information Processing & Management 60 (2023) 103148.

[7] Z.-X. Yong, H. Schoelkopf, N. Muennighoff, A. F. Aji, D. I. Adelani, K. Almubarak, M. S. Bari, L. Sutawika, J. Kasai, A. Baruwa, et al., Bloom+ 1: Adding language support to bloom for zero-shot prompting, arXiv preprint arXiv:2212.09535 (2022).

[8] E. C. Chau, L. H. Lin, N. A. Smith, Parsing with multilingual bert, a small corpus, and a small treebank, arXiv preprint arXiv:2009.14124 (2020).

[9] R. Wang, D. Tang, N. Duan, Z. Wei, X. Huang, G. Cao, D. Jiang, M. Zhou, et al., K-adapter: Infusing knowledge into pre-trained models with adapters, arXiv preprint arXiv:2002.01808 (2020).

[10] A. Ansell, E. M. Ponti, A. Korhonen, I. Vulić, Composable sparse fine-tuning for cross-lingual transfer, arXiv preprint arXiv:2110.07560 (2021).

[11] J. Pfeiffer, I. Vulić, I. Gurevych, S. Ruder, MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 7654–7673. URL: https://aclanthology.org/2020.emnlp-main.617. doi:10.18653/v1/2020.emnlp-main.617.

[12] M. Artetxe, S. Ruder, D. Yogatama, On the cross-lingual transferability of monolingual representations, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 4623–4637. URL: https://aclanthology.org/2020.acl-main.421. doi:10.18653/v1/2020.acl-main.421.

[13] J. Abadji, P. O. Suarez, L. Romary, B. Sagot, Towards a cleaner document-oriented multilingual crawled corpus, arXiv preprint arXiv:2201.06642 (2022).

[14] G. Sarti, M. Nissim, It5: Large-scale text-to-text pretraining for italian language understanding and generation, arXiv preprint arXiv:2203.03759 (2022).

[15] M. Conover, M. Hayes, A. Mathur, J. Xie, J. Wan, S. Shah, A. Ghodsi, P. Wendell, M. Zaharia, R. Xin, Free dolly: Introducing the world's first truly open instruction-tuned llm, 2023. URL: https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm.

[16] E. Fersini, D. Nozza, P. Rosso, et al., Ami@ evalita2020: Automatic misogyny identification, in: Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020), (seleziona...), 2020.

[17] M. Sanguinetti, G. Comandini, E. Di Nuovo, S. Frenda, M. Stranisci, C. Bosco, T. Caselli, V. Patti, I. Russo, Haspeede 2@ evalita2020: Overview of the evalita 2020 hate speech detection task, Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (2020).

Annotating Homeric Emotions by a Domain-Specific Language

Federico Boschetti[1], Laura Chilla[2], Maria Konstantinidou[3] and John Pavlopoulos[4]

[1]CNR-ILC
[2]Ca' Foscari University of Venice, Italy
[3]Laboratory of Papyrology and Palaeography, Department of Greek Philology, Democritus University of Thrace, Greece
[4]Athens University of Economics and Business, Greece

Abstract
In this study, we introduce a novel approach to understanding the emotive content present in ancient literature, specifically focusing on the first Book of Homer's *Iliad*. Recognizing the challenges inherent in interpreting emotions from ancient texts, we developed a Domain-Specific Language (DSL) tailored for this purpose. This DSL not only allows for the annotation of basic sentiments such as positive, negative, neutral, or mixed but also facilitates the identification and categorization of specific emotions. To ensure the relevance and applicability of our annotations, we mapped the extracted emotions to some authoritative domain ontologies. This mapping process aids in bridging the gap between ancient emotional expressions and contemporary understanding. Our preliminary results, which we discuss in detail, highlight the potential of our approach in offering deeper insights into the emotional landscape of ancient texts. We believe that our methodology can serve as a foundation for future studies aiming to decode emotions in historical literature.

Keywords
sentiment analysis, digital philology, collaborative annotation, domain-specific languages

1. Introduction

This study follows the investigations of Pavlopoulos [1] for the annotation of sentiment and emotions in the first Book of Homer's *Iliad*, translated in modern Greek. In this second step, the ancient Greek text is analysed and the main focus is on the expressiveness of the annotation system to capture multiple aspects [2] of the textual units under observation.

Like in the previous work, annotators are asked to indicate both the sentiment (i.e. positive, negative, neutral or mixed) and the specific emotions (from an open set of possibilities). But they can annotate at any level of granularity (from a single word to several verses), both on the paradigmatic (i.e. words outside context) and the syntagmatic (i.e. textual units in context) axes, from the perspective of different experiencers (e.g. the character and the ancient audience), towards different participants to the scene.

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
federico.boschetti@ilc.cnr.it (F. Boschetti); lauradchilla@gmail.com (L. Chilla); mkonst@helit.duth.gr (M. Konstantinidou); annis.pavlo@gmail.com (J. Pavlopoulos)
https://www.ilc.cnr.it/people/federico-boschetti-2/ (F. Boschetti); https://helit.duth.gr/author/mkonst/ (M. Konstantinidou); https://ipavlopoulos.github.io/ (J. Pavlopoulos)
0000-0002-7810-7735 (F. Boschetti); 0009-0009-0160-4467 (L. Chilla); 0000-0002-8744-1444 (M. Konstantinidou); 0000-0001-9188-7425 (J. Pavlopoulos)

2. Related work

Sprugnoli et al. [3] apply sentiment analysis to the *Odes* of Horace at sentence level. One of the main interests is to test *LatinAffectus* [4] the lexical resource for Latin in which words are associated to their polarity: positive, neutral, negative, or mixed. Concerning the ancient Greek Literature, Yeruva et al. [5] studied the inter-annotation agreement of human annotators and machines on an English translation of Aeschylus's tragedies. Luraghi and Sausa [6] study the construal of emotions in Homeric verbs.

Annotating emotions mentioned in ancient literary texts is a complex task because there are no native speakers (as pointed out by Sprugnoli about LatinAffectus), there is not a perfect match between emotions defined in different languages and cultures, and there is abundant secondary literature to take into account (such as exegetical commentaries, lexica, translations).

Sini et al. [7] demonstrate that different languages (in their case: Neo-Latin) structure the lexicon of the emotions in similar but not identical ways.

Kleinginna and Kleinginna [8] and Williams et al. [9] discuss multiple definitions for the category of emotions and suggest how to merge or harmonise different tables of them.

Studies on the emotions in the ancient world are necessary to keep the correct historical distance between the contemporary audience or the annotators and the text or the ancient audience: for instance, Braund and Most [10] on ancient anger (text) or Becker [11] on Stoic emotions (ancient, but late, audience).

3. Method

A group of 15 volunteers (students and scholars) are asked to annotate the first book of the Iliad on the Euporia [12] web platform developed by CoPhiLab.[1] and they are provided with detailed guidelines, based on the syntax and semantics of the Domain-Specific Language (DSL) created for this task. A DSL is a formal language, usually defined by a Context-Free Grammar, which is compact and familiar to the user because it is optimized for limited purposes. Each participant annotates the same portion of the text, resulting in the first book being annotated in its entirety 15 times. As shown in Fig. 3, an annotation is constituted by a reference to the Homeric text and a sequence of one or more structured sentences, ending with a punctuation mark. If we consider the first line of the bottom box, 1.11*ētímasen* (which means: [he, i.e. Agamemnon] dishonored, disdained [Chryses]) is the reference to the text, followed by two sentences: a lexical observation and a situational observation. The structure of a sentence extends the syntax of the scripting language *turtle*: it is constituted by subject(s), predicate(s) and object(s), each of which can be preceded by hashtags to categorise them; the object(s) can be followed by a recipient (or cause) of the emotion. In our example, the first sentence is the lexical observation: `#lex this expresses #quite_neg shame, humiliation.` In this case the annotator asserts that *ētímasen* expresses two quite negative emotions: shame and humiliation. The second sentence is the situational observation: `#character Chryses feels #neg humiliation.` As pointed out by Luraghi and Sausa [6], an emotion is a complex experience that involves an experiencer (the subject), an expertum (the emotion) and a stimulus (the cause of the emotion, that can be mentioned or not). The classifiers for experiencers are: `#character` (an Iliadic hero), `#narrator` (that can be `the_poet`, or an internal narrator), `#recipient` (that can be `the_ancient_audience` or `the_contemporary_audience`), and `#annotator` (that is always expressed by "`I`"). In this way, multiple perspectives can be captured: e.g. different characters may experience different emotions in the same scene. The polarity of the emotions are graded on a scale of seven degrees from `#very_neg` to `#very_pos`.

The annotations are parsed with a compiler compiler[2] and converted in XML or json to facilitate data analysis.

[1] https://cophilab.ilc.cnr.it/ A prototype of the platform, which is an app for eXist-db, is available at https://github.com/CoPhi/euporia

[2] We use the compiler compiler ANTLR: https://www.antlr.org/

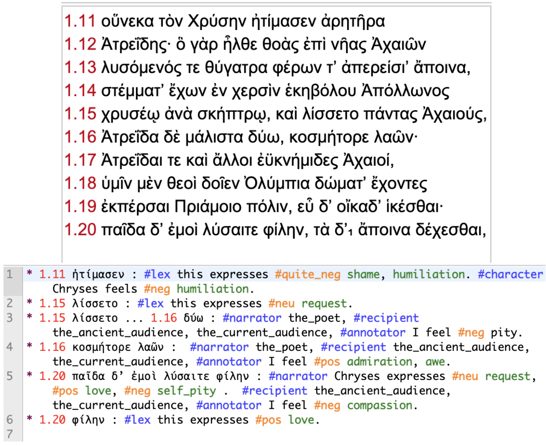


Figure 1: The user interface of Euporia

3.1. Mapping annotations to existing ontologies

x' Because the annotators are free to add new emotion terms to their list, we structure these terms *a posteriori* by mapping them to the following ontologies: the Emotion Ontology (*MFOEM*), the Semanticscience Integrated Ontology (*SIO*), Visualized Emotion Ontology (*VEO*), and the National Cancer Institute Thesaurus ontology (*NCIT*).
MFOEM [13] applies a cognitive perspective and builds upon MF (Mental Functioning) and BFO (Basic Format Ontology) and considers affective phenomena, e. g. emotions, moods, appraisals as well as subjective feelings.
SIO [14] uses a simpler approach but assigns positive and negative polarities to the emotions.
VEO [15] builds on MFOEM, but aims at the visual representation of emotions.
NCIT [16] applies a clinical perspective and distinguishes between emotions and feelings.

We take into account also the Time Event Ontology, *TEO*, [17] to shape temporal aspects of the annotations, such as the phases of a complex event that involves multiple emotions.
To better clarify the differences between the different ontologies, we shall now describe how *Anger* is classified in each:

- MFOEM, anger < emotion < affective process < mental process < bodily process < process < occurrent < entity;
- SIO, anger < disgust < hostility < negative emotion < emotion < behaviour < process < entity;
- VEO, anger < emotion < affective process < mental process < behaviour < action < bodily process < process < occurrent < continuant <

entity < consequence < aspect;

- NCIT, anger < emotion < mental process < neurologic process < organismal process < biological process.

We decided to use MFOEM as the main reference ontology, because it is more suitable for our purposes.In fact, both MFOEM and our ontologies have a cognitive perspective. We then proceeded to map the list of emotions extracted from the annotations to MFOEM, SIO, VEO and NCIT.

Only 7 of these emotions - *anger, fear, hope, joy, sadness, satisfaction* - were included in all the ontologies of emotions, but other 32 terms can be mapped to one or another of them. Besides 7 unfound terms, all the remaining terms were classifiable through MultiWordNet[3], but as hyponyms of the following synsets: *feeling* (6), *speech act* (3), *cognitive state, state of mind* (3), *trait* (2), *emotion* (2), *human action* (2), *feeling, cognitive state, state of mind* (1), *sentiment* (1), *emotion, feeling* (1), *communication* (1), *state* (1), *human action, feeling* (1). Furthermore, for each emotion the following attributes have been instantiated:

- time, indicating when the emotion is perceived, with values *present, future*;
- agent, meaning who perceives the emotion, with the values *oneself, external.* There is also one instance - *submission* - of oneself toward external;
- valence, with the values, *positive, negative, ambiguous*;
- consequence, i.e. when the result of the emotion will take place and what type of result will be, with the values *unpleasant, pleasant, expected, unexpected, actual, future.* In one instance - *suspense* - it was not possible to determine the effect of the emotion.

For example, *empathy* is classified as somebody's reaction for an actual consequence to an event happened toward another agent Therefore, it is perceived in the present by oneself with a negative and it has an actual, unpleasant consequence.

By comparing the list of emotions in Pavlopoulos et al. [1] and the list extracted from the current annotations, 2 terms are missing: *guilt, loneliness* and 10 are new entries: *scorn, threat, acknowledgement, warlike, sadness, betrayal, contempt, disrespect, emotion, rage.*

4. Current results

The most frequently annotated emotion was anger (Figure 4), with 97 occurrences. The emotions of respect, aggression, and fear followed with less than 50 occurrences each. On the other hand, the most infrequent annotations regarded encouragement, gratitude, and shock. In 62% of the annotations of anger, the most frequently annotated emotion, the polarity was negative (very negative in 5, quite negative in one) while in the rest it was neutral. As can be seen in Figure 4, the number of polarity-carrying emotions per verse varies. One of the highest peak was observed in verse 474, where Apollo is satisfied by the song sang to him by the Achaeans. All annotations were positive.

Currently only 4 annotators out of 15 have completed their tasks.

Figure 2: Frequency of annotated emotions

[3]https://multiwordnet.fbk.eu

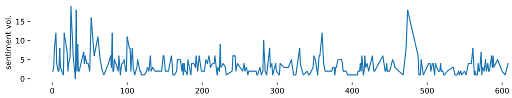


Figure 3: Number of annotations per verse, not distinguishing regarding polarity.

The DSL context-free grammar, the reference text of the first book of the *Iliad* encoded in XML-TEI, the updated annotations and the script to convert the DSL in XML are available at https://github.com/CoPhi/emohomer.

5. Conclusions

This study developed a domain-specific language for the annotation of emotive content of the first Book of Homer's *Iliad*. We mapped the list of emotions we extracted from the annotations and we discussed the results. Next steps comprise the study of polarity for more emotions, an exploration of verses with contradicting polarity, and sentiment analysis based on the subject's role (annotator, character, or narrator). Also, we plan to employ more annotators, in order to measure inter-annotator agreement, and study verses which provoke consistent and diverse emotions to the different annotators.

Acknowledgments

We are grateful to Giota Nikolopoulou and to the students and scholars that are annotating the emotional aspects related to the first book of the *Iliad*.

References

[1] J. Pavlopoulos, A. Xenos, D. Picca, Sentiment Analysis of Homeric Text: The 1st Book of Iliad, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 7071–7077. URL: https://aclanthology.org/2022.lrec-1.765.

[2] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, SemEval-2014 task 4: Aspect based sentiment analysis, in: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Association for Computational Linguistics, Dublin, Ireland, 2014, pp. 27–35. URL: https://aclanthology.org/S14-2004. doi:10.3115/v1/S14-2004.

[3] R. Sprugnoli, F. Mambrini, M. C. Passarotti, G. Moretti, Sentiment Analysis of Latin Poetry: First Experiments on the Odes of Horace, in: Proceedings of the Eighth Italian Conference on Computational Linguistics (CLiC-it 2021). Milan, Italy, January 26-28, 2022, Accademia University Press, Torino, 2021, pp. 314–320. URL: https://api.semanticscholar.org/CorpusID:245125493.

[4] R. Sprugnoli, M. Passarotti, D. Corbetta, A. Peverelli, Odi et Amo. creating, evaluating and extending sentiment lexicons for Latin., in: Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 3078–3086. URL: https://aclanthology.org/2020.lrec-1.376.

[5] V. K. Yeruva, M. ChandraShekar, Y. Lee, J. Rydberg-Cox, V. Blanton, N. A. Oyler, Interpretation of Sentiment Analysis in Aeschylus's Greek Tragedy, in: Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, International Committee on Computational Linguistics, Online, 2020, pp. 138–146. URL: https://aclanthology.org/2020.latechclfl-1.17.

[6] S. Luraghi, E. Sausa, Hate and anger, love and desire: The construal of emotions in Homeric Greek, in: D. Haug (Ed.), Historical Linguistics 2013, John Benjamins, Amsterdam/Philadelphia, 2015, pp. 233–255. URL: https://api.semanticscholar.org/CorpusID:148416336.

[7] B. Sini, C. Tinti, D. Galati, Semantic Structure of Emotional Lexicon, in: A. C. Michalos (Ed.), Encyclopedia of Quality of Life and Well-Being Research, Springer Netherlands, Dordrecht, 2014, pp. 5821–5828. URL: https://doi.org/10.1007/978-94-007-0753-5_4197. doi:10.1007/978-94-007-0753-5_4197.

[8] P. R. Kleinginna, A. M. Kleinginna, A categorized list of emotion definitions, with suggestions for a consensual definition, Motivation and Emotion 5 (1981) 345–379. URL: http://link.springer.com/10.1007/BF00992553. doi:10.1007/BF00992553.

[9] L. Williams, M. Arribas-Ayllon, A. Artemiou, I. Spasić, Comparing the Utility of Different Classification Schemes for Emotive Language Analysis, Journal of Classification 36 (2019) 619–648. URL: http://link.springer.com/10.1007/s00357-019-9307-0. doi:10.1007/s00357-019-9307-0.

[10] S. Braund, G. W. Most, Ancient anger: Perspectives from Homer to Galen, volume 32, Cambridge University Press, 2004.

[11] L. C. Becker, Stoic emotion, in: S. K. Strange, J. Zupko (Eds.), Stoicism: Traditions and Transformations, Cambridge University Press, 2004,

p. 250–276. doi:10.1017/CBO9780511498374.014.
[12] G. Mugelli, F. Boschetti, A. Bellandi, R. Del Gratta, A. F. Khan, A. Taddei, Annotating ritual in ancient greek tragedy: a bottom-up approach in action, Digital Humanities Quarterly 15 (2021) 1–11.
[13] A. D. Spear, W. Ceusters, B. Smith, Functions in Basic Formal Ontology, Applied Ontology 11 (2016) 103–128. URL: https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/AO-160164. doi:10.3233/AO-160164.
[14] M. Dumontier, C. J. Baker, J. Baran, A. Callahan, L. Chepelev, J. Cruz-Toledo, N. R. Del Rio, G. Duck, L. I. Furlong, N. Keath, D. Klassen, J. P. McCusker, N. Queralt-Rosinach, M. Samwald, N. Villanueva-Rosales, M. D. Wilkinson, R. Hoehndorf, The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery, Journal of Biomedical Semantics 5 (2014) 14. URL: https://doi.org/10.1186/2041-1480-5-14. doi:10.1186/2041-1480-5-14.
[15] R. Lin, M. T. Amith, C. Liang, R. Duan, Y. Chen, C. Tao, Visualized Emotion Ontology: a model for representing visual cues of emotions, BMC Medical Informatics and DecisionMaking 18 Suppl. 2 (2017) 102–113.
[16] OBO Foundry, NCI Thesaurus OBO Edition, 2022. URL: https://obofoundry.org/ontology/ncit.html, https://obofoundry.org/ontology/ncit.html Accessed = 2023-07-12.
[17] F. Li, J. Du, Y. He, H.-Y. Song, M. Madkour, G. Rao, Y. Xiang, Y. Luo, H. W. Chen, S. Liu, L. Wang, H. Liu, H. Xu, C. Tao, Time event ontology (TEO): to support semantic representation and reasoning of complex temporal relations of clinical events, Journal of the American Medical Informatics Association 27 (2020) 1046–1056. URL: https://doi.org/10.1093/jamia/ocaa058. doi:10.1093/jamia/ocaa058.

How good is NLLB-200 for low-resource languages? A study on Genoese

Davide Buscaldi[1], Paolo Rosso[2]

[1]LIPN, CNRS UMR 7030, Université Sorbonne Paris Nord, 99 av. Jean-Baptiste Clément, 93430, Villetaneuse, France
[2]PRHLT, Universidad Politécnica de Valencia, Camino de Vera s/n, 46022, Valencia, Spain

Abstract
English. In this paper we analyze the performance of the NLLB-200 models from Meta AI on a manually built parallel corpus of Ligurian (specifically, the Genoese variant), consisting in 283 sentences and their respective Italian translation. Our experiments highlight some issues with NLLB-200, especially regarding local knowledge, deriving from some choices done for the training process.
Italiano. In quest'articolo analizziamo la performance del modello NLLB-200 di Meta AI su un corpus parallelo, costruito manualmente, di 283 frasi in genovese e la loro rispettiva traduzione in italiano. Mostriamo i punti deboli di NLLB-200, in particolare il trattamento dei toponimi ed altri termini in relazione con un contesto locale ligure, evidenziando alcuni problemi derivati dalle scelte fatte nel training di questo modello.

Keywords
NLLB, Machine Translation, Genoese, Endangered Languages

1. Introduction

NLLB (No Languages Left Behind) [1] is a collection of language models created by Meta AI to fill the void left in Machine Translation (MT) for some low- and very low-resource languages. NLLB-200 is the latest model and is able to provide MT for 200 languages, including some that had never been considered before. One of these languages is Ligurian, an endangered language that is spoken mainly in the Liguria region in Italy, Monaco (where it is called Monegasque), and some small islands in the Mediterranean Sea (Carloforte and Calasetta in Sardinia).

The content in Ligurian on the web is very scarce. The main source is Wikipedia, which has only $11,172$[1] articles in Ligurian, many of them being a bit more than drafts. In comparison, the number of articles in Welsh, which has an estimated equivalent number of native speakers ($500,000$) is twenty times as much as Ligurian.[2] This difference is easily explained by the fact that Welsh is an official language, supported by the local government while Ligurian is mostly orally spoken.

The rarity of content in Ligurian is not the only problem that may affect MT tools and methods. In particular, the syntax of Ligurian has not been completely standardised: many variants of the same word may exist, even when they are pronounced in the same way, due to various reasons. First of all, the local variants of Ligurian, but also because the language has been passed down from a generation to another one mostly in an oral way. For instance, in Monegasque, the word "white" is written as *giancu* while in Genoese (the predominant variant) it is written as *gianco*[3]. This problem has been well exposed in the work of [2], which also cite the lack of regulatory bodies as one of the sources of variations. In their study, they propose a corpus of normalized and unnormalized texts in Ligurian to train a neural model for the normalization of Ligurian texts. The example in Figure 1 allows to appreciate the high variability of Ligurian spelling.

Unna	rondaniña	affammâ	a s' é pösâ in	sciô	teito de coppi
Ûnn-a	rōndaninn-a	affammâ	a s' é pösâ in	sciö	teito de cōppi
Ûnn-a	rondaninn-a	affammâ	a s' è pösâ in	sciö	teito de coppi
Ûnn-a	rondaninn-a	affamâ	a s' è pösâ in	sce-o	teito de coppi
Ûña	rundaniña	affammä'	a s' é pösä in	sce o	téyto de cuppi

Figure 1: Examples of 4 variants of Ligurian from [2], with the reference standardised spelling on top. In our work, we did not standardise the texts but used them "as they are".

Given this premise, it is important to evaluate whether the NLLB-200 model is able to deal with these problems. For this reason, we conducted an evaluation by composing a test dataset in Genoese that was not used for NLLB training. The copyright-free subset of this

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
buscaldi@lipn.fr (D. Buscaldi); prosso@dsic.upv.es (P. Rosso)
https://lipn.fr/~buscaldi (D. Buscaldi); https://www.prhlt.upv.es/paolo-rosso/ (P. Rosso)
0000-0003-1112-3789 (D. Buscaldi); 0000-0002-8922-1242 (P. Rosso)

[1]as of 20/07/2023
[2]https://meta.wikimedia.org/wiki/List_of_Wikipedias_by_speakers_per_article
[3]https://fr.wikipedia.org/wiki/Monegasque

dataset is available at the following address: https://github.com/dbuscaldi/zeneize.

In the following section we describe the NLLB-200 model and how the training has been carried out. In Section 3 we show the experiments carried out on our data set and discuss the results. Finally, in Section 4 we draw some conclusions of this analysis and propose some ideas to improve the model on the basis of this experimentation.

2. NLLB-200

To be able to interpret the results we must first take a better look at NLLB-200: the dataset on which it was trained, and the characteristics of NLLB-200 and the models used in our experiments.

2.1. Training Data

The training data for Ligurian were created by composing a set of 6, 193 professionally-translated sentences in the Wikipedia domain, named *NLLB-Seed*[4] [1]. Data for NLLB-Seed was sampled from Wikimedia's "List of articles every Wikipedia should have", a collection of 10,000 Wikidata IDs corresponding to notable topics in different fields of knowledge and human activity. These are split into 11 categories such as People, History, Philosophy and Religion, Geography. NLLB developers note that half the data for Ligurian were first translated from English to Italian, then translated from Italian to Ligurian while the other half was translated directly from English. It can be noted that this process is covering English language domain knowledge rather than knowledge related to the local language. we will come back to this aspect later during our evaluation.

2.2. Models

The NLLB-200 model is an encoder-decoder model that makes the most out of the LASER-3 embeddings [3]. LASER-3 are multilingual embeddings that focus on training multiple language *family-specific* representations. This means that embeddings trained for Italian will still have a degree of similarity to other embeddings in the same family of languages (i.e., Romance languages). The final translation models come in various configurations, distilled and non-distilled. The non-distilled models have 3.3B and 1.3B parameters, while the distilled ones have 1.3B and 600M parameters. Distillation for these models is based on online word-level distillation [4], which means that the student model is trained on the training data but with an additional objective: to minimize the cross-entropy with respect to the word-level distribution of the teacher model.

3. Experiments and Results

We collected a dataset of texts in Genoese from three different sources. First of all, 95 lines from two of the most famous songs of Fabrizio De André, "Crêuza de mä" (small path to the sea) and "'A çimma" (the cima is a typical Genoese dish) with their respective translation in Italian found on the official Fondazione De André page[5] (retrieved on 2022-08-13), and the popular song "Trilli trilli". Then, 188 sentences from "Zêna e contórni", a translation in Genoese from Charles Dickens section on Genoa, Italy, from his "Pictures of Italy" work. We used the wikisource text[6] and an Italian translation obtained from the English one with Deepl[7].

We applied the NLLB-200 models on this dataset, obtaining the results in Tables 1 and 2. We calculated the results using SacreBLEU [5], in particular the measures spBLEU [6] with flores-200 tokenization (as in the NLLB paper), the character n-gram based measure chrF [7], and TER (Translation Error Rate) [8].

The time required to run the translation varied considerably from the dist-600M (about 20 minutes) to the 3.3B model (about 3 hours and 45 minutes). The intermediate size models, 1.3B parameters took about 45 minutes on average to process the whole dataset. All these values were obtained on a CPU 2,6 GHz Intel Core i7 (no GPU acceleration used) and 16GB RAM.

As it can be seen, the results are quite appalling even with the largest model, casting some doubts on the usability of the NLLB-200 model for Genoese. It can be observed that all models are having more problems with the translation from Italian to Genoese than in the opposite direction. As expected, in most cases, the larger the model, the better the results, although the improvements in the Italian-Genoese translation are lower than in Genoese-Italian. TER values higher than 100 indicate that the models are overgenerating, producing sequences that are longer than the reference ones. This is particularly evident in the songs subset, in the Italian to Genoese direction.

An inspection of the results in Genoese shows some interesting outputs of the models. Toponyms are often translated incorrectly. For instance, let's consider the Ligurian capital, Genova, which is mentioned 9 times in our dataset. In the Italian-Genoese direction, only NLLB 3.3B translates it correctly in 3 out of 9 cases. NLLB 1.3B once translates "Genova" into "Genoa" instead of "Zena" and it always keeps "Genova" elsewhere. NLLB

[4] https://github.com/facebookresearch/fairseq/tree/nllb

[5] http://www.fabriziodeandre.it/

[6] https://lij.wikisource.org/wiki/Zêna_e_contórni

[7] https://www.deepl.com/translator

Table 1
Genoese to Italian results

	De André + Trilli			Dickens			Full Dataset		
Model	**spBLEU**	**chrF**	**TER**	**spBLEU**	**chrF**	**TER**	**spBLEU**	**chrF**	**TER**
NLLB dist-600M	11.2	35.1	74.4	14.2	40.5	73.6	15.2	39.9	73.7
NLLB dist-1.3B	16.3	**38.3**	67.7	18.6	**44.9**	68.0	**20.2**	**44.2**	**67.9**
NLLB 1.3B	14.0	37.9	**66.2**	18.9	44.0	68.9	18.4	43.3	68.5
NLLB 3.3B	**14.2**	35.9	71.4	**20.9**	44.8	**67.4**	**20.2**	43.8	**67.9**

Table 2
Italian to Genoese results

	De André + Trilli			Dickens			Full Dataset		
Model	**spBLEU**	**chrF**	**TER**	**spBLEU**	**chrF**	**TER**	**spBLEU**	**chrF**	**TER**
NLLB dist-600M	4.2	24.5	144.4	2.4	25.3	96.3	4.4	25.2	102.4
NLLB dist-1.3B	4.1	26.4	108.6	3.8	**26.9**	**93.4**	5.5	**26.8**	95.3
NLLB 1.3B	7.8	24.9	**98.8**	4.8	25.9	93.5	5.3	25.8	**94.2**
NLLB 3.3B	**9.0**	**27.7**	102.2	**5.3**	26.6	95.8	**5.9**	26.7	96.6

dist-600M is never able to translate correctly "Genova" into "Zena", it always translates it the same as in Italian (Genova). Finally, NLLB dist-1.3B correctly translates it 6 times. The problem seems also to affect other proper nouns, such as Saint Peter (San Pietro in Italian) which is correctly translagted in Genoese as "San Pê". In fact, is translated as "San Peixe", which is Portuguese for fish, in both distilled models. The 3.3B model translates it as "San Pêo" and the 1.3B one translates it as "San Peçio". Both these translations make no sense in Genoese.

In the Genoese-Italian direction, only NLLB dist-1.3B translates "Zêna" correctly in 3 out of 9 cases. Both 1.3B models translate in one case it as "Ginevra" (Geneva, in Switzerland). Other spelling errors show "Giena" by the dist-600M model and "Gênes" (in French instead of Italian) by the 3.3B model. Looking into the tokenizer we observed that "Genova" is not in the dictionary and is tokenized as *Gen-ova*, and "Zêna" is tokenized as *Z-êna*. On the other hand, "San Pê" is correctly translated by all models. Due to the output occurring sometimes in different languages than the target ones, we suspect that the previous errors may result from the LASER-3 embeddings which are language-family based.

Both distilled models fall into repetitions. For instance, in the Dickens text, NLLB dist-600M translates "Ma, per il momento, gironzolo qui intorno, in tutti i buchi e gli angoli del quartiere, in un perpetuo stato di forzata sorpresa" ("But, as yet, I stroll about here, in all the holes and corners of the neighbourhood, in a perpetual state of forlorn surprise") into "*Ma, pe-o momento, o l'é in sciâ çitæ, in tutti i buchi e in tutti i cantoni do quartiere, in un stato de sorpresâ forçâ pe pe pe pe pe pe pe pe pe...*" ("But, for now, he is on top of the city, in all the holes and corners of the neighbourhood, in a state of forced surprise for for for for..."). The larger model (dist-1.3B) is not immune to this behaviour although it happens only 2 times instead of 9. The non-distilled models don't present this problem. The fact that the models fall into this kind of repetition could be due to the lack of sufficient training data for the word-based online distillation process. Therefore the probability distribution for the tokens is skewed towards some frequent words ("pe" - *for*, "ti" - *you*, "ben" - *well*). We observed that the minimum frequency in the NLLB-seed dataset of words that are repeated is 39 (for the word "sciâ": probably as part of "in sciâ" - on top of).

4. Conclusions

From our preliminary analysis, carried out on a dataset specific to the Genoese culture, we can affirm that currently NLLB-200 is not good enough to deal with Genoese texts or to translate text into Genoese. In particular, we found out that local toponyms are difficult to translate: how good is an MT tool that is not able to correctly translate the name of the largest city where the language is spoken or the name of the language itself? Given the information provided regarding NLLB-200 models, we can identify two main elements explaining this behaviour. The first one is the training data: they do not cover local information, but general English Wikipedia articles, so they lack to provide the context in which Genoese is usually spoken. The second one is the tokenization process and the LASER-3 embeddings: given the high spelling variability of the Ligurian language, we suspect that the tokenization process may not be precise and that it may map some tokens into a position in the embedding space that does not correspond to their actual "meaning", maybe also because of a sort of interference from other Western Romance languages that are very close to Ligurian.

However, NLLB-200 is a big step forward making endangered languages such as Ligurian and its variants available to everyone. From our point of view, we think that NLLB-200 could be improved in various ways, for instance fine-tuning the model on more “local" datasets; and possibly including knowledge regarding Out-Of-Vocabulary words that are often named entities, for instance with the methods proposed by [9], or integrating dictionaries to deal with named entities.

Acknowledgments

This work is supported/ partially supported by a public grant overseen by the IdEx Université Paris Cité (ANR-18-IDEX-0001) as part of the Labex Empirical Foundations of Linguistics - EFL.

References

[1] M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, et al., No language left behind: Scaling human-centered machine translation, arXiv preprint arXiv:2207.04672 (2022).

[2] S. Lusito, E. Ferrante, J. Maillard, Text normalization for low-resource languages: the case of ligurian, in: Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages, 2023, pp. 98–103.

[3] K. Heffernan, O. Çelebi, H. Schwenk, Bitext mining using distilled sentence representations for low-resource languages, arXiv preprint arXiv:2205.12654 (2022).

[4] Y. Kim, A. M. Rush, Sequence-level knowledge distillation, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas, 2016, pp. 1317–1327. URL: https://aclanthology.org/D16-1139. doi:10.18653/v1/D16-1139.

[5] M. Post, A call for clarity in reporting bleu scores, arXiv preprint arXiv:1804.08771 (2018).

[6] T. Kudo, J. Richardson, Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2018, pp. 66–71.

[7] M. Popović, chrf: character n-gram f-score for automatic mt evaluation, in: Proceedings of the tenth workshop on statistical machine translation, 2015, pp. 392–395.

[8] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, J. Makhoul, A study of translation edit rate with targeted human annotation, in: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers, 2006, pp. 223–231.

[9] J. Waldendorf, A. Birch, B. Hadow, A. V. Micele Barone, Improving translation of out of vocabulary words using bilingual lexicon induction in low-resource machine translation, in: Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track), Association for Machine Translation in the Americas, Orlando, USA, 2022, pp. 144–156. URL: https://aclanthology.org/2022.amta-research.11.

Debunker Assistant: a support for detecting online misinformation

Arthur Thomas Edward Capozzi Lupi[1,2], Alessandra Teresa Cignarella[1,2], Simona Frenda[1,2], Mirko Lai[1,2], Marco Antonio Stranisci[1,2] and Alessandra Urbinati[1,3]

[1] *aequa-tech, Turin, Italy*
[2] *Computer Science Department, University of Turin, Turin, Italy*
[3] *MOBS, Northeastern University, Boston, USA*

Abstract

This paper describes the framework developed for the *Debunker-Assistant*, an application that allows users and newspapers to assess the trustworthiness of a news item starting from its headline, body of text and URL. The Debunker-Assistant adapts ideas from Natural Language Processing and Network Science to counter the spread of online misinformation. Its centerpiece is a set of four *News Misinformation Indicators* based on linguistically engineered features, models, network analysis metrics (Echo Effect, Alarm Bell, Sensationalism, and Reliability). In this short contribution, we describe the back-end structure on which the indicators are implemented.

Keywords

Misinformation, Debunker-Assistant, Linguistic Features, Web Domain Network

1. Introduction and Background

Fake news threatens democracies, public health, and news outlets' credibility. For this reason, tackling misinformation is an open challenge faced by governments, private companies, and the scientific communities [1].

There are many proposed approaches, some based on AI methods, others on fact-checking by human experts, still others on a combination of the two [2, 3]. However, the fact that fake news detection algorithms are often owned by private social media companies, and additionally the adoption of "black-boxed" algorithms contribute to the lack of transparency in the fake news identification and filtering process.

Debunker-Assistant (D-A) is a back-end AI tool which supports the analysis and detection of online misinformation. The D-A tool takes as an input the link of a news article and returns its misinformation profile based on Natural Language Processing (NLP) and Network Analysis (NA) features. Such an approach is inspired by the survey provided by [4].

D-A works with Italian and is not bounded on a specific topic: it is designed as a general purpose tool that can extract relevant features for assessing the quality of information[1].

Its main purposes are:

1. displaying the indicators to deal with misinformation;
2. de-biasing the mechanisms to make trustworthy the internet;
3. showing insights about a certain context to aid the search and discovery of information.

In this paper, we present the various features of NLP and NA embedded in the D-A tool and how these features have been designed and developed. As represented in Figure 1, D-A allows users to search for news and compare them against a given set of 4 macro-indicators of misinformation: Echo Effect, Alarm Bell, Sensationalism, and Reliability. These indicators are designed on the basis of specific linguistic and network features, such as the absence of sources, non-authority of references, the presence of specific figures of speech or flames, and other stylistic characteristics.

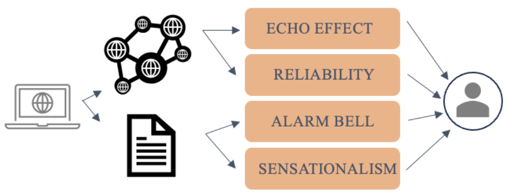


Figure 1: Overall pipeline of the D-A tool.

[1]The access to the D-A tool is offered through API, a public available version can be found at https://github.com/AequaTech/DebunkerAssistant

All the authors contributed equally.

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

arthurtomasedward.capozzi@unito.it (A. T. E. C. Lupi); alessandrateresa.cignarella@unito.it (A. T. Cignarella); simona.frenda@unito.it (S. Frenda); mirko.lai@unito.it (M. Lai); marcoantonio.stranisci@unito.it (M. A. Stranisci); a.urbinati@northeastern.edu (A. Urbinati)

To operationalize our objectives, we focus on three main actions: **Inform** D-A can be freely used by citizens who want to check if a content they have read includes potential misinformation signals; **Support** D-A can also be adopted by journalists who work in the field of debunking, empowering their activity; **Teach** teachers are able to use D-A for designing information literacy activities in their classrooms.

2. News Misinformation Indicators

2.1. Echo Effect

Echo Effect encodes the coupled aspects of origin and dissemination. There are news sources that occupy a key role in the spread of misinformation [5]. This class aims to identify the most important domains in the news ecosystem, and measure the impact and range of these sources. We employ the network science framework, modeling each domain as a node of a network model, $G = (V, E, \varpi)$, where V is a set of nodes of N domains, E it the set of link $e_{ij} = (i, j)$ that encodes the pairwise interactions between two domains, and $\varpi : V \times V \to \mathbb{Y}$ is a function, defined for each pair of nodes $i, j \in V$, that maps each link e_{ij} into a weight w_{ij}, that stands for the total number of interactions from the domain i to j.

A well established method to assess the importance of a node in a network is to measure the value of "centrality" that a node has with respect to all the other nodes. There are many possible definitions of importance and so many centrality measures. In the context of news, we adopted the *hyperlink-induced topic search* [6] algorithm to infer the "origin" and "destination" of the echo effect, i.e., the set of nodes responsible for the amplification of signals traversing the network and the set of nodes in which the signal resonates the most. While the *betweenness centrality* [7] measures the extent to which a node can reach other locations in the network[2].

2.2. Alarm Bell

Alarm Bell takes into account the presence of possible pragmatic implications like flaming and ironic language in the news. Specifically, it analyzes whether the headline and news content contain these elements, and outputs an average score of the various probabilities.

To obtain these values, we designed specific classifiers to account for the presence in the text of irony, hate speech and stereotypes. For training our models, we used two well-known benchmark datasets: IronITA [8] and the extended version of the HaSpeeDe2 [9, 10]. These datasets are annotated for all the dimensions necessary to our purposes: IronITA contains Italian tweets about political and immigration issues, annotated for irony; the extended version of HaSpeeDe2 contains Italian tweets and news headlines about the integration of minority communities, annotated for hate speech, stereotypes, and irony.

To create the models, we fine-tuned the base version of BERT for Italian[3], optimizing basic hyperparameters like learning rate and batch size for each phenomenon. To make the models smalls and easy to call at each request of the API, we reduced the number of trainable parameters, adopting the LoRA technique implemented by HuggingFace [11]. If the aim of the training is to minimize the loss function[4] on the validation set, the evaluation of the models is focused on the analysis of f1-macro on the test sets of news headlines of HaSpeeDe2 (0.722 for hate speech and 0.749 for stereotype) and of tweets of IronITA (0.769 for irony)[5].

The tool returns values in a range from 0 to 1, representing the probabilities that these phenomena (irony, hate speech and stereotypes) are present in the news. The presence of them could be an alarm bell on the seriousness and trustworthiness of the news [12], as well as of their malicious intentions [4].

2.3. Sensationalism

Sensationalism includes three groups of features that are indicative of the diffusion of clickbait, especially while observing online headlines: informal style, syntactic complexity, emotion profile. Inspired by [13], who studied the linguistic and typological features commonly associated with clickbait in online news headlines, we designed the following features.

1. `informal_style`:
   use of upper case - The ratio between number of uppercase words and total number of words in a news headline;
   repeated letters - The ratio between words with a repeated letter for $n > 3$ (e.g., *SVEGLIAAAAA!* or *diiiiici*) and the total number of words in a news headline.
   distinctive punctuation patterns - In particular the count of 'weird' punctuation such as ! ... * + = $, and the count of 'normal' punctuation such as . , ; : ?
   presence of emojis - The normalized count of emojis in a news headline.

[2] See Appendix for details.

[3] https://huggingface.co/dbmdz/bert-base-italian-cased

[4] The function used in our experiments is the Binary Cross Entropy with different weight for each class, considering the constant imbalance of the positive class in all the datasets.

[5] The models are available in https://github.com/AequaTech/DebunkerAssistant

2. `syntactic_complexity`:
   avg length - A metric that computes the textual length of the news headline and compares it with the average textual length of a news headline stored in a proprietary database. The value is normalized between 0 and 1.
   shortest/longest - A metric that shows the comparison of the length of the headline and collocates it in one of the four quartiles of the database with all the other news headlines previously analyzed.

3. `affective_profile`:
   overall emotion - Inspired by the study Vosoughi et al. [14] about the presence of specific emotions in false stories, this feature returns the averaged score about the presence of emotions in the news. This value is based on the identification of the eight primary emotions—anger, fear, sadness, disgust, surprise, anticipation, trust, and joy— following the psycho-evolutionary taxonomy proposed by [15] through the multilingual NRC lexicon of [16]. Emotional signals have also been explored to detect credibility by Giachanou et al. [17].
   overall sentiment - As the previous feature, it is a normalized value of the presence of sentiment, exploiting the intensity score from Sentix lexicon [18]. The sentiment score to detect misinformation has already been used in previous works such as Baly et al. [19], and Ghanem et al. [20].

2.4. Reliability

Inspired by the different source cues that influence people when making credibility evaluation decisions [21], we also leverage the context of certain sites, because they could impact the type of content spread. In this line, this class aims to quantify its overall "reliability". Firstly, supported by debunking websites, we compiled two separate lists summarizing the established positions of specific URL domains regarding the spread of misinformation: the *white list* containing sites considered mostly safe, and the *black list* containing sites known for disseminating intentionally misleading information. Secondly, to frame the context, we assess two aspects linked to a URL domain.

1. `neighborhood`: We employ the network analysis framework to evaluate the neighborhood of a web page. We retain the same model described in Section 2.1 to detect special agglomerations of nodes, or communities (locally dense connected sub-graphs), such that the nodes that belong to the same community have a higher probability to be linked than the nodes in different communities. We employed the stochastic block model [22], a generative model that regards in a formal context the actual presence of a specific not random map into which the network can be partitioned. Once having the different communities, we characterized them by assigning to each a label, *white* or *black*, counting the majority of their nodes (web pages) belonging to either the white or the black list.

2. `solidity`: To quantify solidity, we take advantage of the "whois" metadata attached to each URL domain, which is the country of registration, creation date, and expiration date. We calculated the amount of time (days) that a specific domain maintains the same characteristics.

3. Conclusions

D-A is a tool developed in support of detecting online misinformation. Its social impact is twofold. First, it contributes to counter misinformation and improve the quality of information. In fact, the tool is aimed at supporting who deal every day with this issue. Second, it provides a better understanding of how these artificial intelligence technologies work in the context of the news media.

Given the complex and ever-changing nature of content creation and information dissemination, there are several directions for improvement. For example, users could be involved in providing anonymous feedback on the news itself and on the characterization of the evaluated articles, improving the overall performing skill of the tool, more so for those features that are less explored in the literature. In addition, this type of interaction makes the user think about important aspects of the online information, thus increasing awareness. Over time, as users search for new URLs, the core data that feed the models will expand to cover larger and more diverse sets of domains, incorporating a richer perspective on news consumption. To help the above research directions, we plan to develop a user-friendly interface and evaluate the general user experience. Finally, a future challenge would be to scale the model for other languages starting from English.

Ethics Considerations D-A provides the user with a set of characteristics about the article and a set of information about the domain hosting the article. Thus, the output generated does not consist of a binary classification of the truthfulness of an online newspaper article. Nonetheless, we are aware of the ethical issues surrounding the characterization and evaluation of online news.

First of all, the fallibility of NLP models must be taken into account, secondly some aspects concerning the world of information can have shades of subjectivity and be sensitive especially for some users.

Acknowledgments

This project has received funding from the European Union's Horizon Europe research and innovation program via the NGI Search grant agreement (ID: 101069364), linked with the Next Generation Internet Initiative.

References

[1] S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online, Science 359 (2018) 1146–1151. URL: https://www.science.org/doi/abs/10.1126/science.aap9559. doi:10.1126/science.aap9559.

[2] P. Nakov, D. P. A. Corney, M. Hasanain, F. Alam, T. Elsayed, A. Barr'on-Cedeno, P. Papotti, S. Shaar, G. D. S. Martino, Automated fact-checking for assisting human fact-checkers, in: International Joint Conference on Artificial Intelligence, 2021.

[3] S. I. Manzoor, J. Singla, Nikita, Fake news detection using machine learning approaches: A systematic review, in: 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), 2019, pp. 230–234. doi:10.1109/ICOEI.2019.8862770.

[4] G. Ruffo, A. Semeraro, A. Giachanou, P. Rosso, Studying fake news spreading, polarisation dynamics, and manipulation by bots: A tale of networks and language, Computer science review 47 (2023) 100531.

[5] S. Vilella, A. Semeraro, D. Paolotti, G. Ruffo, Measuring user engagement with low credibility media sources in a controversial online debate, EPJ data science 11 (2022) 29.

[6] J. M. Kleinberg, Hubs, authorities, and communities, ACM computing surveys (CSUR) 31 (1999) 5.

[7] L. C. Freeman, D. Roeder, R. R. Mulholland, Centrality in social networks: Ii. experimental results, Social networks 2 (1979) 119–141.

[8] A. T. Cignarella, S. Frenda, V. Basile, C. Bosco, V. Patti, P. Rosso, Overview of the EVALITA 2018 task on irony detection in Italian tweets (IronITA), in: Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2018) co-located with the Fifth CLiC-it, volume 2263, 2018, pp. 1–6.

[9] M. Sanguinetti, G. Comandini, E. Di Nuovo, S. Frenda, M. Stranisci, C. Bosco, T. Caselli, V. Patti, I. Russo, Haspeede 2@ evalita2020: Overview of the evalita 2020 hate speech detection task, in: Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online. CEUR. org, 2020.

[10] S. Frenda, V. Patti, P. Rosso, Killing me softly: Creative and cognitive aspects of implicitness in abusive language online, Natural Language Engineering (2022) 1–22. doi:10.1017/S1351324922000316.

[11] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-rank adaptation of large language models, in: International Conference on Learning Representations, 2022. URL: https://openreview.net/forum?id=nZeVKeeFYf9.

[12] V. L. Rubin, N. Conroy, Y. Chen, S. Cornwell, Fake news or truth? using satirical cues to detect potentially misleading news, in: Proceedings of the second workshop on computational approaches to deception detection, 2016, pp. 7–17.

[13] R. Kemm, The linguistic and typological features of clickbait in youtube video titles, Social Communication 8 (2022) 66–80.

[14] S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online, Science 359 (2018) 1146–1151. doi:10.1126/science.aap9559.

[15] R. Plutchik, A general psychoevolutionary theory of emotion, in: Theories of emotion, Elsevier, 1980, pp. 3–33.

[16] S. M. Mohammad, P. D. Turney, Crowdsourcing a word-emotion association lexicon, Computational Intelligence 29 (2013) 436–465. URL: https://doi.org/10.1111/j.1467-8640.2012.00460.x.

[17] A. Giachanou, P. Rosso, F. Crestani, Leveraging emotional signals for credibility detection, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'19, Association for Computing Machinery, New York, NY, USA, 2019, p. 877–880. URL: https://doi.org/10.1145/3331184.3331285. doi:10.1145/3331184.3331285.

[18] V. Basile, M. Nissim, Sentiment analysis on Italian tweets, in: Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 2013, pp. 100–107.

[19] R. Baly, G. Karadzhov, D. Alexandrov, J. Glass, P. Nakov, Predicting factuality of reporting and bias of news media sources, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 3528–3539. URL: https://aclanthology.org/D18-1389. doi:10.18653/v1/D18-1389.

[20] B. Ghanem, S. P. Ponzetto, P. Rosso, F. Rangel,

FakeFlow: Fake news detection by modeling the flow of affective information, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, 2021, pp. 679–689. URL: https://aclanthology.org/2021.eacl-main.56. doi:10.18653/v1/2021.eacl-main.56.

[21] M. J. Metzger, A. J. Flanagin, Psychological Approaches to Credibility Assessment Online, John Wiley & Sons, Ltd, 2015, pp. 445–466. doi:https://doi.org/10.1002/9781118426456.ch20.

[22] T. P. Peixoto, Bayesian stochastic blockmodeling, Advances in network clustering and blockmodeling (2019) 289–332.

Appendix - Centrality measures

Here are the details of Section 2.1. The hyperlink-induced topic search algorithm (also known as HITS or *hubs and authorities*) is defined for directed networks, and computes the authority centrality and the hub centrality, which quantify vertices' prominence in the two roles, as a "receiver" or as a "provider" of information. The hub vector $\vec{h}_t = (h_{1,t}, \ldots, h_{|V|,t})^\intercal$ and the authority vector $\vec{a}_t = (a_{1,t}, \ldots, a_{|V|,t})^\intercal$ in $t \in T$ of $G = (V, T, \varpi)$ are defined by the limit of the following set of iterations:

$$\vec{h}_t(x+1) = c_t(x) W_t \vec{a}_t(x+1) \tag{1}$$

and

$$\vec{a}_t(x+1) = d_t(x) W_t^\intercal \vec{h}_t(x), \tag{2}$$

where $c_t(x)$ and $d_t(x)$ are normalization factors to make the sums of all elements become unity, i.e., $\sum_{i=1}^{|V|} h_{i,t}(x+1) = 1$ and $\sum_{i=1}^{|V|} a_{i,t}(x+1) = 1$. The initial values of the scores are $h_{i,t}(0) = \frac{1}{|V|}$ and $a_{i,t}(0) = \frac{1}{|V|}$ for all $i \in V$.

The betweenness centrality measures the extent to which a node lies on paths between other vertices. We define the betweenness centrality of a node $i \in V$ at time $t \in T$ as

$$c_b(i,t) = \sum_{\substack{s,e \in V \\ i \neq s \neq e}} \frac{\sigma_{se,t}(i)}{\sigma_{se,t}}, \tag{3}$$

where $\sigma_{se,t}$ is the total number of shortest paths from node s to node e at time t, and $\sigma_{se,t}(i)$ is the number of such paths passing through node i.

Towards a Multi-Level Annotation Format for the Interoperability of Automatic Term Extraction Corpora

Nicola Cirillo[1], Daniela Vellutino[1]

[1]University of Salerno, 132 Via Giovanni Paolo II, Fisciano (SA), 84084, Italy

Abstract

English. The main corpora used as benchmarks in Automatic Term Extraction are represented in different formats. Unfortunately, none of these formats covers the wide range of linguistic phenomena related to terminology. To address this issue, we propose to encode Automatic Term Extraction corpora in RDF using the OntoLex-Lemon and the NLP Interchange Format ontologies. Furthermore, we developed a small Italian corpus on waste management legislation to provide an example of the proposed formalization.

Italiano. I corpora principali impiegati nella valutazione degli algoritmi di Estrazione Automatica di Termini sono codificati in formati diversi. Purtroppo, nessuno di questi formati permette di rappresentare l'ampia gamma di fenomeni linguistici legati alla terminologia. Per affrontare la questione, proponiamo di codificare i corpora di Estrazione Automatica di Termini in RDF usando le ontologie OntoLex-Lemon e NLP Interchange Format. Inoltre, abbiamo sviluppato un piccolo corpus italiano riguardante la legislazione della gestione dei rifiuti per fornire un esempio della formalizzazione proposta.

Keywords

Terminology, Automatic Term Extraction, Linguistic Linked Data, OntoLex-Lemon

1. Introduction

Automatic Term Extraction - ATE is an NLP task that involves recognizing terms in specialized corpora. As with most NLP tasks, ATE research benefits from annotated corpora that are employed as training data and evaluation benchmarks. Nevertheless, existing term annotation schemata are far from capturing the complex organization that characterizes the terminology of specialized languages [1, 2, 3]. Most ATE studies, with a few exceptions [4, 5, 6], overlook the complex organization of terms in specialized languages assuming that the terms contained in a corpus belong to a single domain. At best, they draw a difference between *domain terms* (that belong to the investigated domain) and *out-of-domain terms* (that belong to different domains). Unfortunately, this assumption is too simplistic since every specialized corpus contains terms from different subject fields. Moreover, in the interest of reusability, researchers who use terminology corpora in their work must be able to define the subject fields of interest according to their needs.

Furthermore, ATE corpora do not adhere to standard formats used to encode terminological data like TermBase eXchange, an ISO standard [7], and OntoLex-Lemon, a W3C standard. This lack of standardization poses interoperability issues and hinders the evaluation of ATE tools. The adoption of a standard format will provide at least three main benefits:

- It will grant the interoperability of termbases. Therefore if a term is already present in an existing termbase, it could be imported.
- It will grant the interoperability of corpora, meaning that multiple corpora could be combined to cover different languages and subject fields.
- It will ease the effort made to evaluate ATE tools from both sides developers and users.

In this paper, we propose a custom form design of multi-level annotation to formalize ATE corpora in RDF format by using the OntoLex-Lemon[1] and the NLP Interchange Format - NIF[2] ontologies to represent termbases and corpora, respectively. Moreover, we develop a small annotated corpus to provide a proof-of-concept. The corpus and the code employed in its formalization are publicly available on GitHub[3]

The remainder of this paper is organized as follows. Section 2 lays out the main feature of terms. Section 3 gives an overview of the main ATE corpora. Section 4 illustrates the proposed formalization schema. In Section 5 we describe the corpus annotation experiment. Finally, Section 6 provides conclusions.

2. Features of terms

According to ISO, a term is a "designation that represents a general concept by linguistic means" [8]. Therefore,

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

nicirillo@unisa.it (N. Cirillo); dvellutino@unisa.it (D. Vellutino)
0000-0002-2107-1313 (N. Cirillo); 0000-0002-2525-7940 (D. Vellutino)

[1]https://www.w3.org/community/ontolex/wiki/Final_Model_Specification

[2]https://nif.readthedocs.io/en/latest/

[3]https://github.com/nicolaCirillo/lod4term

corpus	discontinuous terms	nested terms	corpus format	termbase format
GENIA	yes	yes	XML	none
ACL-RD TEC	no	no	XML; vert	TSV
ACTER	yes	only in the termbase	TSV	TSV

Table 1
Features of main ATE corpora

terms have linguistic and conceptual features and a sound annotation schema must account for both. The most relevant are illustrated above.

Nested terms A term is nested when it is contained into another (longer) term. For example, the term *competent authority of dispatch* contains both the terms *competent authority* and *dispatch*, joined by the preposition *of*.

Discontinuous terms A term is discontinuous when there is unrelated linguistic material between its words. Sometimes, discontinuous terms are also nested. For example, the term *prevention of pollution* is discontinuous when it appears inside the term *integrated prevention and control of pollution*.

Term variants A term variant is a term that expresses the same concept as other terms. For example, the terms *air pollution* and *atmospheric pollution* are term variants because they both refer to the "contamination of the indoor or outdoor environment by any chemical, physical or biological agent that modifies the natural characteristics of the atmosphere".[4] **Acronym** and **abbreviations** are specific kinds of term variants. Resolving abbreviations is one of the goals of the Simple Text Track at CLEF 2023.[5]

Terminology layer A terminology layer is a set of terms belonging to a given subject field. For example, the terminology layer of waste management comprises terms such as *incineration plant*, *separate collection*, and *landfill*. Some ATE techniques focus on isolating terminology layers [5, 6, 9]

Translation equivalent A translation equivalent is a term of a natural language that denotes the same concept as another term of another natural language. For example, the term *autorità competente di spedizione* is the Italian equivalent of the English term *competent authority of dispatch*. Finding translation equivalents from comparable corpora is an ATE subtask [10].

3. Related Work

With regards to the format of ATE benchmark corpora, there is no agreed standard. The most popular corpora are encoded in different formats as summarized in Table 1.

The GENIA corpus [11] is composed of 2000 English abstracts taken from the MEDLINE database. It is focused on the biology domain, specifically on transcription factors in human blood cells. The corpus is encoded in XML with each occurrence of a term being enclosed in the `<term>` tag. Discontinuous and nested terms are allowed. From the conceptual perspective, each term is an instance of a class defined in the GENIA ontology (e.g. the term *fibroblastic tumour* is an instance of the *Tissue* class).

Being constituted of 300 abstracts from the ACL Anthology Reference Corpus, the ACL RD-TEC corpus [12] has been developed with the intent of providing a term extraction corpus on which computational linguists are experts themselves. It is available in XML and in a vertical format (i.e. one token per line). Discontinuous and nested terms are not allowed. From the conceptual perspective terms are categorized following the guidelines (e.g. *technology and method*, *tool and library*, *language resource*, etc.).

The ACTER corpus [10] is composed of multiple subcorpora covering four different subject fields and three languages. ACTER is specifically made to test ATE tools on different topics and languages while retaining a consistent annotation and, thus, comparable results. It is available in TSV (one token per line) with IOB (Inside, Outside, Beginning) or IO (Inside, Outside) tags. The list of terms found in the corpus is also made available. Discontinuous and nested terms are allowed but the latter are not represented in the IOB and IO formats. From the conceptual perspective terms are classified according to a domain-independent annotation schema composed of four labels: *specific term*, *common term*, *out-of-domain term*, and *not term*. Moreover, it distinguishes terms from Named Entities.

[4] https://iate.europa.eu/entry/result/3567909/en
[5] http://simpletext-project.com/2023/clef/tasks

4. Proposed Format

An ATE corpus has two main components: a termbase and the actual corpus. The termbase contains the list of unique terms (*types*) that appear in the corpus and provides information for each of them. Conversely, the corpus contains contextualized instances (*tokens*) of the terms in the termbase. We propose to formalize both resources using formats based on RDF/OWL on account of their interoperability. Besides, linked data formats have already improved the benchmarking of Named Entity Recognition [13].

4.1. Termbase Representation

We propose to use the OntoLex-Lemon ontology to encode the termbase, for various reasons. First of all, OntoLex-Lemon is already a standard among terminologists [14, 15, 16]. In addition, it can represent many linguistic and conceptual information that are of interest to ATE and its subtasks.

In OntoLex-Lemon, there are three main entities: *entries*, *concepts*, and *senses*. Entries are instances of the `LexicalEntry` class. They are linguistic units with one or more forms. For example, the term *heap* is an entry with a singular form (*heap*) and a plural form (*heaps*). Concepts are instances of the `LexicalConcept` class. They represent units of thought. For example, the concept *heap* is defined as "engineered facility for the deposit of solid waste on the surface".[6] Finally, senses are instances of the `LexicalSense` class. They are entry-concept pairs. For example, the entry *heap* has multiple senses one of which couples this entry with the concept defined above. Entries are related to senses through the `sense` property and concepts through the `lexicalicalizedSense` property. Moreover, entries can also be directly linked to concepts through the `evokes` property.

Furthermore, Ontolex-Lemon allows the representation of many linguistic and conceptual features that are of interest to ATE and its subtasks (see Section 2). Term variants are easy to identify because they are entries referring to the same concept. However, OntoLex-Lemon also allows to directly link entries and senses and specifies their relation via the LexInfo[7] ontology. Namely, the `synonym` property of LexInfo links two senses with the same meaning, `abbreviationFor` links an abbreviation to its full form, and `translation` links two terms that are translations of each other. OntoLex-Lemon can also represent nested terms by means of the `subterm` property of its decomposition module. Lastly, terminology layers can be handled by assigning senses and concepts to the respective subject field through the `subject` property of the Dublin Core ontology. Moreover, to grant interoperability, we propose to employ DBPedia categories [17] to represent subject fields. In this way, a term belonging to a given subject field (e.g. waste management), automatically belongs also to the subject fields it descends from (e.g. waste; sustainability and environmental management; economy and the environment), in a multi-level fashion.

[6]https://iate.europa.eu/entry/result/3504812/en

[7]https://lexinfo.net/

4.2. Corpus Representation

To formalize the corpus, we propose to use the NIF and the POWLA[8] ontologies. NIF is based on RDF/OWL and has been developed to achieve interoperability between NLP tools, language resources, and annotations. It provides multiple benefits. First of all, it does not rely strictly on tokenization like TSV formats (see Section 3). It provides support for terminology annotation and has already been used for this purpose within the FREME project [18]. The only drawback of NIF is that it cannot represent discontinuous terms. To this end, we use POWLA nodes to join NIF strings, as suggested in [19]. Then, we link each POWLA node to the corresponding `LexicalSense` in the termbase to produce unambiguous annotations (see Appendix A).

5. Example Corpus

In order to test our approach and provide a proof-of-concept, we run an annotation experiment on a European directive, namely the Italian version of the *Directive 2006/21/EC of the European Parliament and of the Council of 15 March 2006 on the management of waste from extractive industries and amending Directive 2004/35/EC* (26,882 tokens).

5.1. Annotation

Two non-expert annotators carried out the annotation. They were instructed to identify terms in the corpus and classify them according to the subject field (i.e. *law*, *EU law*, *waste management*, *waste management law*, *environment*, *other*). Particular attention has been paid to the identification of nested and discontinuous terms (see Section 2). After the annotation phase, we asked annotators to revise the list of unique terms they found (i.e. the termbase) to delete incorrect ones and revise nested terms. Finally, we kept in the corpus only the annotations of terms that were in the revised termbases and standardized the annotation of nested terms. Namely, we removed their manual annotations and automatically tagged them according to the subterms provided in the revised termbase, thus ensuring consistency.

[8]https://github.com/acoli-repo/powla

	termbase (F-score)	corpus (F-score)
before revision	0.440	0.470
after revision	0.474	0.619
subject fields (Fleiss' k)	0.707	

Table 2
Inter-annotator agreement

To estimate the inter-annotator agreement on term identification, we computed the F-score measure, similar to [12], for both the corpus and the termbase, before and after the revision process (see Table 2). Moreover, to estimate the agreement on subject fields, we computed the Fleiss' k only on terms identified by both annotators.

Inter-annotator agreement scores confirm the benefits of the revision process. Even though the agreement on the termbase shows only a little improvement after the revision (+0.034), the effect on the corpus is much more relevant (+0.149) as a result of the standardization of nested terms.

In the final dataset, we joined the annotations of both annotators and linked the resulting termbase to IATE[9] by associating each concept with the respective IATE entry when it exists.

6. Conclusions and Future Work

The lack of standardization in the representation of ATE corpora constitutes a bottleneck for the evaluation of ATE tools. To address this issue, we proposed an RDF-based formalization that employs OntoLex-Lemon to represent termbases and NIF to represent corpora. We showed that these formats are able to represent the wide range of linguistic and conceptual phenomena that characterize terminology. In addition, we developed a small corpus about waste management legislation in order to provide an example of the proposed formalization.

In future, we plan to convert the major ATE corpora into the proposed format, to further improve ATE standardization. Moreover, we intend to increase the size and quality of the small ATE corpus we developed.

Acknowledgments

The authors contributed to this paper as follows. sections 1, 2, and 6 are attributed to Daniela Vellutino while sections 3, 4, and 5 are attributed to Nicola Cirillo.

[9] https://iate.europa.eu/home

References

[1] D. Vellutino, R. Maslias, F. Rossi, Verso l'interoperabilità semantica di iate. studio preliminare per il dominio "gestione dei rifiuti urbani", Terminologie specialistiche e diffusione dei saperi (2016) 1–240.

[2] D. Vellutino, R. Maslias, F. Rossi, C. Mangiacapre, M. P. Montoro, Verso l'interoperabilità semantica di iate. studio preliminare sul lessico dei fondi strutturali e d'investimento europei, Diversité et Identité Culturelle en Europe/Diversitate si Identitate Culturala in Europa (2016) 1–254.

[3] D. Vellutino, L'italiano istituzionale per la comunicazione pubblica, il Mulino, 2018.

[4] A. Lenci, S. Montemagni, V. Pirrelli, G. Venturi, Ontology learning from italian legal texts, in: Law, Ontologies and the Semantic Web, IOS Press, 2009, pp. 75–94.

[5] F. Bonin, F. Dell'Orletta, S. Montemagni, G. Venturi, A contrastive approach to multi-word extraction from domain-specific corpora, in: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta, 2010. URL: http://www.lrec-conf.org/proceedings/lrec2010/pdf/553_Paper.pdf.

[6] P. Drouin, M.-C. L'Homme, B. Robichaud, Lexical profiling of environmental corpora, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 2018.

[7] ISO, ISO 30042:2019 - Management of terminology resources - TermBase eXchange (TBX), International Organization for Standardization, 2019.

[8] ISO, ISO 1087:2019 - Terminology work and terminology science, International Organization for Standardization, 2019.

[9] N. Cirillo, Isolating terminology layers in complex linguistic environments: a study about waste management (short paper), in: Proceedings of the 2nd International Conference on Multilingual Digital Terminology Today (MDTT 2023), volume 3427, CEUR Workshop Proceedings, 2023. URL: https://ceur-ws.org/Vol-3427/short3.pdf.

[10] A. Rigouts Terryn, V. Hoste, E. Lefever, In no uncertain terms: a dataset for monolingual and multilingual automatic term extraction from comparable corpora, Language Resources and Evaluation 54 (2020) 385–418. doi:10.1007/s10579-019-09453-9.

[11] J. D. Kim, T. Ohta, Y. Tateisi, J. Tsujii, Genia corpus - a semantically annotated corpus for biotextmining, Bioinformatics 19 (2003). doi:10.1093/bioinformatics/btg1023.

[12] B. Qasemizadeh, A.-K. Schumann, The acl rd-

tec 2.0: A language resource for evaluating term extraction and entity recognition methods, European Language Resources Association (ELRA), 2016, pp. 1862–1868. URL: https://aclanthology.org/L16-1294.

[13] M. Röder, R. Usbeck, A.-C. N. Ngomo, Gerbil – benchmarking named entity recognition and linking consistently, Semantic Web 9 (2018) 605–625. doi:10.3233/SW-170286.

[14] P. Martín-Chozas, T. Declerck, Representing multilingual terminologies with ontolex-lemon, in: Proceedings of the 1st International Conference on Multilingual Digital Terminology Today (MDTT 2022), CEUR Workshop Proceedings, 2022.

[15] S. Piccini, F. Vezzani, A. Bellandi, Tbx and 'lemon': What perspectives in terminology?, Digital Scholarship in the Humanities 38 (2023) i61–i72. URL: https://doi.org/10.1093/llc/fqad025. doi:10.1093/llc/fqad025.

[16] M. Fiorelli, A. Stellato, T. Lorenzetti, A. Turbati, P. Schmitz, E. Francesconi, N. Hajlaoui, B. Batouche, Towards ontolex-lemon editing in vocbench 3, AIDAinformazioni, Rivista di scienze dell'informazione (2018).

[17] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. V. Kleef, S. Auer, C. Bizer, Dbpedia - a large-scale, multilingual knowledge base extracted from wikipedia, Semantic Web 6 (2015) 167–195. doi:10.3233/SW-140134.

[18] M. Dojchinovski, F. Sasaki, T. Gornostaja, S. Hellmann, E. Mannens, F. Salliau, M. Osella, P. Ritchie, G. Stoitsis, K. Koidl, M. Ackermann, N. Chakraborty, Freme: Multilingual semantic enrichment with linked data and language technologies, volume 8, European Language Resources Association (ELRA), 2016, pp. 4180–4183. URL: https://aclanthology.org/L16-1660.

[19] P. Cimiano, C. Chiarcos, J. P. McCrae, J. Gracia, Linguistic Linked Data Representation, Generation and Applications, 1 ed., Springer Cham, 2020. doi:10.1007/978-3-030-30225-2.

A. Example of RDF files

```
@prefix dbc: <https://dbpedia.org/page/Category:> .
@prefix dct: <http://purl.org/dc/terms/> .
@prefix decomp: <http://www.w3.org/ns/lemon/decomp#> .
@prefix lexinfo: <http://www.lexinfo.net/ontology/2.0/lexinfo#> .
@prefix ontolex: <http://www.w3.org/ns/lemon/ontolex#> .
@prefix termbase: <http://example.com/termbase/> .

termbase:entry_rifiuto_delle_industrie_estrattive a ontolex:MultiwordExpression ;
  decomp:subterm termbase:entry_industria_estrattiva ,
    termbase:entry_rifiuto ;
  ontolex:canonicalForm termbase:form_rifiuto_delle_industrie_estrattive ;
  ontolex:otherForm termbase:form_rifiuti_delle_industrie_estrattive ;
  ontolex:sense termbase:rifiuto_delle_industrie_estrattive_sense1
  ontolex:evokes termbase:concept_rifiuto_delle_industrie_estrattive .

termbase:form_rifiuto_delle_industrie_estrattive a ontolex:Form ;
  lexinfo:gender lexinfo:masculine ;
  lexinfo:number lexinfo:singular ;
  ontolex:writtenRep ”rifiuto delle industrie estrattive”@it .

termbase:form_rifiuti_delle_industrie_estrattive a ontolex:Form ;
  lexinfo:gender lexinfo:masculine ;
  lexinfo:number lexinfo:plural ;
  ontolex:writtenRep ”rifiuti delle industrie estrattive”@it .

termbase:rifiuto_delle_industrie_estrattive_sense1 a ontolex:LexicalSense ;
  ontolex:isLexicalizedSenseOf termbase:concept_rifiuto_delle_industrie_estrattive ;
  ontolex:isSenseOf termbase:entry_rifiuto_delle_industrie_estrattive ;
  dct:subject dbc:Waste_management ;
  lexinfo:synonym termbase:rifiuto_derivante_dalle_industrie_estrattive_sense1 ,
    termbase:rifiuto_generato_dalle_industrie_estrattive_sense1 ,
    termbase:rifiuto_prodotto_dalle_industrie_estrattive_sense1 ,
    termbase:rifiuto_proveniente_dalle_industrie_estrattive_sense1 .

termbase:concept_rifiuto_delle_industrie_estrattive a ontolex:LexicalConcept ;
  dct:subject dbc:Waste_management ;
  ontolex:lexicalizedSense termbase:rifiuto_delle_industrie_estrattive_sense1 ,
    termbase:rifiuto_derivante_dalle_industrie_estrattive_sense1 ,
    termbase:rifiuto_generato_dalle_industrie_estrattive_sense1 ,
    termbase:rifiuto_prodotto_dalle_industrie_estrattive_sense1 ,
    termbase:rifiuto_proveniente_dalle_industrie_estrattive_sense1 ;
  ontolex:isEvokedBy termbase:entry_rifiuto_delle_industrie_estrattive_sense1 ,
    termbase:entry_rifiuto_derivante_dalle_industrie_estrattive ,
    termbase:entry_rifiuto_generato_dalle_industrie_estrattive ,
    termbase:entry_rifiuto_prodotto_dalle_industrie_estrattive ,
    termbase:entry_rifiuto_proveniente_dalle_industrie_estrattive .
```

Figure 1: Example of the termbase.

```
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix itsrdf: <http://www.w3.org/2005/11/its/rdf#> .
@prefix nif: <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix powla: <http://purl.org/powla/powla.owl#> .
@prefix termbase: <http://example.com/termbase/> .
@prefix corpus: <http://example.com/corpus/> .

_:terms a powla:Root .

_:term1 a powla:Node;
  powla:string "rifiuti" ;
  powla:hasParent _:terms ;
  itsrdf:term "yes" ;
  itsrdf:termInfoRef termbase:rifiuto_sense1 .

_:term2 a powla:Node;
  powla:string "industrie estrattive" ;
  powla:hasParent _:terms ;
  itsrdf:term "yes" ;
  itsrdf:termInfoRef termbase:industria_estrattiva_sense1 .

_:term3 a powla:Node;
  powla:string "rifiuti delle industrie estrattive" ;
  powla:hasParent _:terms ;
  itsrdf:term "yes" ;
  itsrdf:termInfoRef termbase:rifiuto_delle_industrie_estrattive_sense1 .

corpus:doc1 a nif:Context,
    nif:OffsetBasedString ;
    nif:isString "Direttiva 2006/21/CE del Parlamento europeo e del Consiglio ... " .

corpus:doc1#offset_105_112 a nif:OffsetBasedString,
    nif:Word,
    powla:Node ;
  nif:anchorOf "rifiuti" ;
  nif:referenceContext corpus:doc1 ;
  powla:hasParent _:term1,
    _:term3 .

corpus:doc1#offset_113_118 a nif:OffsetBasedString,
    nif:Word,
    powla:Node ;
  nif:anchorOf "delle" ;
  nif:referenceContext corpus:doc1 ;
  powla:hasParent _:term3 ;
  powla:next corpus:doc1#offset_113_118 .

corpus:doc1#offset_113_118 a nif:OffsetBasedString,
    nif:Word,
    powla:Node ;
  nif:anchorOf "industrie" ;
  nif:referenceContext corpus:doc1 ;
  powla:hasParent _:term2,
    _:term3 ;
  powla:next corpus:doc1#offset_129_139 .

corpus:doc1#offset_129_139 a nif:OffsetBasedString,
    nif:Word,
    powla:Node ;
  nif:anchorOf "estrattive" ;
  nif:referenceContext corpus:doc1 ;
  powla:hasParent _:term2,
    _:term3 .
```

Figure 2: Example of the corpus

The PBSDS: A Dataset for the Detection of Pseudoprofound Bullshit

Evan D. DeFrancesco[1], Carlo Strapparava[1,2]

[1]*Università degli Studi di Trento, Italy*
[2]*Fondazione Bruno Kessler*

Abstract
This paper introduces the PBSDS, a dataset of tweets containing pseudoprofound bullshit—statements designed to appear profound but lacking substantive meaning. The PBSDS serves as a resource for studying pseudoprofound bullshit, exploring potential linguistic factors in perceiving bullshit. The dataset's creation and experiments with classifiers show promising results, despite limitations such as selection bias and subjective annotation.

Keywords
pseudoprofound bullshit, stylistic analysis, pragmatics

1. Introduction

"Bullshit" refers to communication that is designed to impress but is constructed without concern for truth [1]. Bullshit differs from lying in that the liar deliberately manipulates and subverts truth (usually with the intent to deceive), while the bullshitter is simply unconcerned with what is true and what is false. A liar needs to know the truth value of a proposition; the bullshitter simply does not care.

Although bullshit comes in different forms, in this project, we focused specifically on what is referred to as "pseudoprofound bullshit," which is designed to convey some sort of potentially profound meaning but is actually semantically vacuous [2], e.g., "Hidden meaning transforms unparalleled abstract beauty." Table 1 reports further examples of pseudoprofound bullshit and non-pseudoprofound bullshit sentences from our dataset.

The goal of this project is to construct a dataset of tweets that contain pseudoprofound bullshit in English (the PBSDS).[1] Operating under the assumption that bullshit is similar to spam email, we hypothesize that it should be possible to detect pseudoprofound bullshit using relatively simple classification algorithms.

Pseudopf BS?	*Sentence*
yes	The unpredictable is a reflection of humble excellence.
no	You must be good to yourself if you are ever going to be any good for others.
yes	The law of attraction is always responding to your thoughts. You are attracting in every moment of your life.
yes	Evolution is an ingredient of subjective excellence.
yes	Our consciousness is a reflection of the door of balance.
no	A garden is a zoo for plants.
no	Scientists are simply adults who retained and nurtured their native curiosity from childhood.

Table 1
Examples of pseudoprofound bullshit and non-pseudoprofound bullshit from the PBSDS.

2. Related work and motivation

Pennycook et al. [2] first explored the psychological nature of pseudoprofound bullshit, establishing an index of bullshit receptivity. They found that a tendency to judge pseudoprofound bullshit statements as profound was correlated with relevant variables such as an intuitive cognitive style and belief in the supernatural. They also found that detecting bullshit was not simply a matter of skepticism but rather of discerning deceptive vagueness in impressive-sounding claims. Walker et al. [3] established a link between illusory pattern perception and the propensity to rate pseudo-profound bullshit statements as profound. Later research by Pennycook and Rand [4] has found that low pseudoprofound bullshit receptivity correlates positively with perceptions of fake news accuracy and negatively with the ability to distinguish fake and real news. Littrell and Fugelsang [5] extended this understanding by exploring individuals' susceptibility to misleading information and its association with re-

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
evan.defrancesco@studenti.unitn.it (E. D. DeFrancesco); strappa@fbk.eu (C. Strapparava)

[1]The dataset is freely available upon request from the first author.

duced engagement in reflective thinking. They found that both highly receptive and highly resistant individuals exhibited limited awareness of their detection abilities for pseudo-profound bullshit. Turpin et al. [6] investigated the influence of different types of titles on the perceived profoundness of abstract art, revealing that pseudo-profound bullshit titles specifically enhanced the perceived profundity of the artwork. Nilsson et al. [7] found an association between pseudoprofound bullshit receptivity and social conservatism and economic progressivism. Relatedly, Evans et al. [8] examined scientific bullshit receptivity, which demonstrated positive correlations with pseudo-profound bullshit receptivity, belief in science, conservative political beliefs, and faith in intuition. They found that scientific literacy moderated the relationship between the two types of bullshit receptivity. These studies collectively shed light on the nature of pseudo-profound bullshit, its reception, and the underlying cognitive mechanisms. However, the development of a dedicated dataset of pseudoprofound bullshit can further facilitate comprehensive investigation and understanding of this phenomenon, contributing to future research endeavors.

Such a dataset could provide researchers with a standardized and reliable resource to study and analyze the phenomenon of pseudoprofound bullshit systematically. It would allow for the exploration of various linguistic, cognitive, and contextual factors that contribute to the perception of profoundness in nonsensical statements. Additionally, an annotated dataset could serve as a benchmark for developing and evaluating computational models and algorithms aimed at detecting and combating pseudoprofound bullshit. It would enable the training and testing of automated systems to recognize and classify instances of pseudoprofound bullshit accurately. This could be instrumental in building tools and technologies to enhance critical thinking, identify deceptive information, and improve media literacy.

3. Data

3.1. Scraping Twitter

We used snscrape[2], an easy-to-use Python package, to crawl the Twitter[3] profiles of six accounts and return the 2,000 most recent tweets from each account. The accounts were scraped on 8 August 2023. We selected accounts that, we hoped, would provide a mix of pseudoprofound bullshit, non-pseudoprofound bullshit, profound philosophy and generic statements. For the initial dataset, we chose accounts that were associated with alternative medicine, pseudoscience, new age spirituality, philosophy and scientific communication. In particular, we scraped the following accounts, from which we collected a total of 12,000 tweets:

- **@DeepakChopra**: Deepak Chopra is a new-age author and alternative medicine promoter. His writing has been described as "incoherent babbling strewn with scientific terms."[4]
- **@WisdomofChopra**: WisdomOfChopra is operated by a bot that produces tweets that are meant to replicate the tone and structure (but not necessarily the content) of Deepak Chopra. The tweets are generated by a simple algorithm: words and phrases are contained within four PHP arrays. The first array contains sentence subjects; the second array contains verb phrases; the third contains determiner phrases and adjectives; the fourth contains nouns. Words and phrases from each array are then combined to generate tweets.
- **@TheSecret**: The Secret's Twitter account is largely composed of messages that promote the pseudoscientific "law of attraction," which claims that positive thoughts attract positive experiences and negative thoughts attract negative experiences.
- **@realNDWalsche**: Neale Donald Walsch is an American new-age writer and speak whose work has appeared in a film version of *The Secret*. His own writing consists primarily of new-age spirituality texts.
- **@kate_manne**: Kate Manne is an associate professor of philosophy at Cornell University. Her research focuses on moral philosophy, metaethics, moral psychology, feminist philosophy and social philosophy. In 2019, Manne was named one of the world's top fifty thinkers.[5]
- **@neiltyson**: Neil deGrasse Tyson is an astrophysicist and science communicator.

We recognize that the decision to include artificially generated content from @WisdomofChopra may be seen as a controversial one. However, the distinction between human and artificial origins of the content was secondary for our purposes. What remained paramount was the essence of the content itself: its pseudoprofound nature.

3.2. Data cleaning

From the initial 12,000 tweets collected, we excluded: duplicate tweets; single-word tweets; tweets that were composed only of hashtags; tweets that were direct replies

[2]https://github.com/JustAnotherArchivist/snscrape
[3]As of 2023, called X.

[4]https://www.washingtonpost.com/news/answer-sheet/wp/2015/05/15/scientist-why-deepak-chopra-is-driving-me-crazy/
[5]https://www.prospectmagazine.co.uk/magazine/prospect-worlds-top-50-thinkers-2019

to other Twitter users; tweets that contained URLs; and tweets that contained emojis. We also removed the hashtag (#) and at-sign (@) from tweets. Finally, we decided to remove tweets that explicitly referenced a personal and individual deity (represented in the tweets as "God"), as we did not wish to cause any inadvertent offense by labelling religious beliefs as pseudoprofound bullshit. After data cleaning, we were left with 5,196 tweets, comprising the initial PBSDS.

3.3. Annotation

Two volunteer annotators provided judgments of whether a tweet constituted pseudoprofound bullshit. The annotators were both students in their mid-20s and were previously not familiar with the concept of pseudoprofound bullshit. The annotators were provided with a working definition of pseudoprofound bullshit (*i.e., statements that sound profound and meaningful but that are actually semantically vacuous; pseudoprofound bullshit may use grandiose terms to deceive people*) as well as several examples of sentences that constituted pseudoprofound bullshit and that did not constitute pseudoprofound bullshit. The working definition was left purposefully vague, given the general difficulty of defining pseudoprofound bullshit. After all, what one person may consider to be pseudoprofound, another person might consider to be actually profound. Annotators were instructed to label the tweet '1' if they believed that it constituted pseudoprofound bullshit and '0' if they did not. Perhaps reflecting the difficulty of arriving at a single sense of pseudoprofound bullshit, Cohen's kappa was calculated at 0.52, indicating moderate inter-rater reliability [9]. The first author of this paper adjudicated disagreements between the two annotators' judgments.

4. Dataset description

After annotation, the PBSDS contains 2756 tweets judged as pseudoprofound bullshit (53.04% of the total dataset) and 2440 tweets judged as non-pseudoprofound bullshit (46.96% of the total dataset). Although the two classes are reasonably well-balanced, pseudoprofound bullshit may be disproportionately represented in the dataset compared to its overall occurrence in natural language. However, this is not unexpected, given that the dataset was sourced primarily from Twitter accounts that were likely to include a large amount of pseudoprofound bullshit.

5. Experiments and Results

We trained six machine learning classifiers and compared the performance to test the validity of the dataset. The six classifiers selected for the task were the Support Vector Classifier (SVC), K-nearest Neighbors (KNN), Multinomial Naive Bayes (MNB), Decision Tree Classifier (DTC), Logistic Regression Classifier (LRC) and Random Forest Classifier (RFC). All models were implemented via the scikit-learn library [10].

The tweets were vectorized using tf-idf vectorization, and the data was split into a training set (85%) and a testing set (15%).

In order to evaluate and compare the results of the six classifiers, we used the standard metrics in text classification: Precision (*P*), Recall (*R*), F-score (*F1*) and Accuracy (*Acc*). The results achieved with the six classifiers are reported in Table 2.

Classifier	*P*	*R*	*F1*	*Acc*
SVC	0.9307	0.7943	0.8571	0.8564
KNN	0.8406	0.8227	0.8315	0.8192
MNB	0.9008	0.8156	0.8561	0.8513
DTC	0.8719	0.8203	0.8453	0.8372
LRC	**0.9435**	0.7896	0.8597	0.8603
RFC	0.9309	**0.8274**	**0.8761**	**0.8731**

Table 2
Results obtained from the six classifiers, reported in terms of precision, recall, F-score and accuracy.

6. Limitations

The PBSDS has several limitations that could be addressed in future versions of the dataset. The dataset was collected from specific Twitter accounts presumed to contain pseudoprofound bullshit. This may have resulted in an overrepresentation of pseudoprofound content compared to its overall occurrence in natural language. The dataset thus may not fully capture the range and diversity of pseudoprofound bullshit found in other contexts. Relatedly, the PBSDS's reliance on tweets from specific Twitter accounts limits its generalizability to other platforms or sources of pseudoprofound bullshit. The characteristics and patterns observed in the dataset may not be representative of pseudoprofound content found elsewhere. Future versions of the PBSDS could address this concern by diversifying the sources of data collection. This would involve not only expanding the range of Twitter accounts under examination but also branching out to other social media platforms, blogs, articles, printed publications and even, perhaps, spoken word content. By incorporating a broader spectrum of sources, the dataset would provide a more comprehensive and varied representation of pseudoprofound bullshit.

Additionally, defining and identifying pseudoprofound bullshit can be challenging and subjective. The annotation process relied on the judgments of two annotators, which may have introduced inherent biases and variations in interpretations. Although efforts were made to establish guidelines, the subjective nature of the task may have affected the consistency of annotations. While the inter-rater reliability between the annotators was measured to be moderate, there was still inherent subjectivity and disagreement in determining whether a tweet constituted pseudoprofound bullshit. The resolution of disagreements by a single adjudicator introduced another layer of subjectivity. Introducing a multi-rater system, in which multiple individuals assess the content's (pseudo)-profundity, could add layers of reliability and objectivity to the dataset.

Finally, the PBSDS comprises 5,196 tweets, which is relatively small in comparison to other text corpora. This limited size may restrict the scope and statistical power of analyses, potentially impacting the generalizability of findings derived from the dataset.

7. Conclusion

Despite its limitations, the PBSDS offers valuable insights into the phenomenon of pseudoprofound bullshit and its detection. The dataset provides a foundation for further research, enabling comprehensive investigations into linguistic patterns, cognitive biases, and societal implications associated with pseudoprofound bullshit. By better understanding and identifying pseudoprofound bullshit, researchers can develop tools and strategies to enhance critical thinking, combat deceptive communication, and promote media literacy in an increasingly complex information landscape.

Acknowledgments

We acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU.

References

[1] H. G. Frankfurt, On bullshit, in: On Bullshit, Princeton University Press, 2005.

[2] G. Pennycook, J. Allan Cheyne, N. Barr, D. J. Koehler, J. A. Fugelsang, On the reception and detection of pseudo-profound bullshit, Judgment and Decision Making 10 (2015) 549–563. doi:10.1017/S1930297500006999.

[3] A. C. Walker, M. H. Turpin, J. A. Stolz, J. A. Fugelsang, D. J. Koehler, Finding meaning in the clouds: Illusory pattern perception predicts receptivity to pseudo-profound bullshit, Judgment and Decision Making 14 (2019) 109–119.

[4] G. Pennycook, D. G. Rand, Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking, Journal of Personality 88 (2020) 185–200.

[5] S. Littrell, J. A. Fugelsang, Bullshit blind spots: The roles of miscalibration and information processing in bullshit detection, Thinking & Reasoning (2023) 1–30.

[6] M. H. Turpin, A. C. Walker, M. Kara-Yakoubian, N. N. Gabert, J. A. Fugelsang, J. A. Stolz, Bullshit makes the art grow profounder, Judgment and Decision making 14 (2019) 658–670.

[7] A. Nilsson, A. Erlandsson, D. Västfjäll, The complex relation between receptivity to pseudo-profound bullshit and political ideology, Personality and Social Psychology Bulletin 45 (2019) 1440–1454.

[8] A. Evans, W. Sleegers, Ž. Mlakar, Individual differences in receptivity to scientific bullshit, Judgment and Decision Making 15 (2020) 401–412.

[9] J. R. Landis, G. G. Koch, The measurement of observer agreement for categorical data, Biometrics (1977) 159–174.

[10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Machine learning in python, The Journal of Machine Learning Research 12 (2011) 2825–2830.

A Post-Modern Approach to Automatic Metaphor Identification

Dario Del Fante[1,*,†], Federico Manzella[1,*,†], Guido Sciavicco[1,*,†] and Eduard I. Stan[2,*,†]

[1]University of Ferrara (Italy)
[2]Free University of Bozen-Bolzano (Italy)

Abstract
This paper provides the theoretical bases for a symbolic approach to text classification, particularly metaphor identification, that generalizes the existing ones and is inspired by similar generalizations of symbolic approaches to learning models for non-text-related tasks.

Keywords
Automatic metaphor detection and interpretation, Symbolic learning, NLP, Modal logic

1. Introduction

Metaphors involve talking and, potentially, thinking of one thing in terms of another; the two things are different, but we can perceive sets of correspondences between them. In other words, a metaphor corresponds to using a word or phrase from the context in which it is expected to occur to another context, where it is not expected to occur [1]. Metaphors are ubiquitous in language [2]: they cannot only be considered a pure artistic ornament that exclusively pertains to literary discourse, but they are essential for the development of language and culture [3, 4, 5].

Much work has been devoted to discussing the metaphor identification and interpretation process, as in [6]. In this sense, a qualitative approach represents the safest methodology since metaphors regard an aspect of language that occasionally can be ambiguous. For example, two speakers from the same linguistic and cultural context can interpret the same metaphor differently. However, this approach is time-consuming and requires at least more than two human coders to be effectively reliable. Despite this phenomenon, it remains a computationally hard task given the many structural problems that make automatic identification not quickly effective [7]. Scholars between digital humanities and computational linguistics have developed different approaches to support automatic identification. Indeed, the recent improvements regarding artificial intelligence and machine learning might consistently impact metaphor research regarding the time and quantity of analyzed text [8].

From a computational point of view, metaphor identification is a particular case of text classification. The recent literature on general text classification, particularly metaphor identification, is quite broad and includes both *top-down* approaches [9] and *bottom-up* ones. Top-down approaches start from a human-designed theory of the phenomenon, which is later digitalized to provide automatic identification. Bottom-up, or *data-driven* ones, on the other hand, aim to perform identification starting from a dataset of examples. Bottom-up strategies can be, in turn, separated into *symbolic* and *sub-symbolic* approaches. Sub-symbolic approaches, commonly realized via several types of neural networks, produce black-box models which in some cases can be very accurate [10]. Along with the application of pre-trained and large language models they currently are a de-facto standard for text-related learning tasks, and quite a lot of results exist even in the narrow field of metaphor identification (see, among many others, [11, 10, 12, 13, 14]). Conversely, the purpose of a symbolic approach is to provide an identification model *and* a statistically validated theory of the phenomenon, written in a suitable logical language. While symbolic systems are sometimes used for text-related tasks in general, their application to the case of metaphor identification needs to be addressed.

In this paper, we provide the theoretical bases for a symbolic approach to text classification, particularly metaphor identification, that generalizes the existing ones and is inspired by similar generalizations of symbolic approaches to learning models for non-text-related tasks.

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
*Corresponding author.
†These authors contributed equally.
dario.delfante@unife.it (D. Del Fante); federico.manzella@unife.it (F. Manzella); guido.sciavicco@unife.it (G. Sciavicco); ioneleduard.stan@unibz.it (E. I. Stan)
0000-0002-1650-273X (D. Del Fante); 0000-0002-4944-2163 (F. Manzella); 0000-0002-9221-879X (G. Sciavicco); 0000-0001-9260-102X (E. I. Stan)

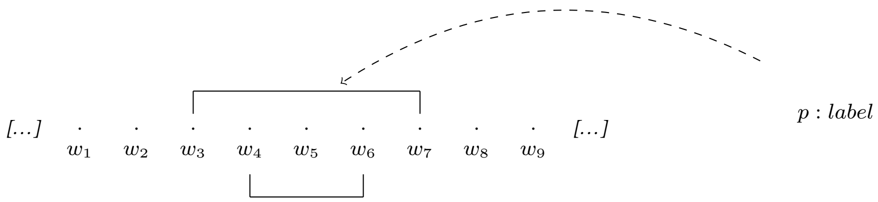


Figure 1: Example of generalized 2-gram.

2. A Logic-Based Post-Modern Approach

Symbolic and sub-symbolic approaches to text-related tasks are different in spirit. In both cases, the key idea is to provide a representation of the text later used for learning. However, in the case of sub-symbolic strategies, such a representation, usually referred to as *embedding*, is numerical. The most famous examples of sub-symbolic representations are (all variants of) *vectorizations* of tokens (i.e., words, sentences, or paragraphs). Each token is mapped to a point of a high-dimensional space so that mathematical tools can be used to reason about texts, and a learned model, for example, for metaphor identification, takes the form of a mathematical function.

In symbolic approaches, on the other hand, we encode a token (typically, an entire sentence or paragraph) as a logical model. In the most uncomplicated cases, following the so-called *bag-of-words* methodology, a text is encoded starting from a fixed (arbitrarily long) dictionary; it is translated into a binary vector of length N, being N the size of the dictionary, where the i-th component takes the value 1 if and only if the i-th word of the dictionary occurs in the text. Text-based encodings are easily generalized along two directions: bag-of-words become *bag-of-n-grams*, and vector components become counters so that the i-th component takes value m if and only if the i-th n-gram of the fixed n-grams vocabulary occurs exactly m times in the text (in this context, n-grams are not used in their canonical, probabilistic version, that is, to predict the n-th element from the previous $n-1$ ones, but, instead, in their crisp one, that is, a straightforward generalization of single words). In most cases, the experiments show that using 2-grams attain the best compromise between the computational complexity of the tasks and the performances of the learned models. The logical interpretation of symbolic encoding emerges by introducing propositional letters to represent the text by the presence of relevant n-grams. Simplifying, a symbolic encoding classification model can be described by (sets of) rule(s) of the type:

If 'flood of immigrants' occurs then metaphor.

In the above example, *'flood of immigrants'* is a 2-gram (before tokenization, stemming, and stop words elimination), and the rule that has been learned checks whether or not that particular 2-gram occurs. Towards an abstract representation, 2-grams can be encoded into propositional letters, which can represent not only their occurrence but also other interesting properties, such as the number of times that they occur. In the end, a text is represented as a model of propositional logic, and a (set of) propositional rule(s) can be statistically learned from a dataset of texts.

A further generalization of symbolic text-based encodings requires two steps: generalizing the concepts of n-gram and increasing the expressive power of the logic that we use to describe texts. Both ideas are simple. Focusing on 2-grams, specifically, the most natural generalization consists of eliminating the constraint of two words being one next to the other to form a 2-gram. So a *generalized* 2*-gram* can be defined as any pair of successive, non-consecutive words. Such a generalization has two main consequences: first, the label of a generalized 2-gram may be much richer than the label of a standard one, and second, the encoding of a text using generalized 2-grams can be much more expressive than the encoding of the same text using standard ones.

Let us focus on labeling. As explained above, a standard 2-gram is logically labeled using (the number of times) that it *occurs*. A generalized 2-gram, on the other hand, can be labeled using the occurrences of the words in between. In Fig. 1, we see the abstract idea of a generalized 2-gram: the pair of words w_3, w_7 form a generalized 2-gram (they are two, possibly non-consecutive, words) and, in the encoding, they are represented by a propositional letter (in the example, p). The meaning of such a propositional letter is no longer limited to depend on the occurrence of w_3, w_7, either separately or together. On the contrary, one can use the entire sentence between w_3 and w_7 to build p; examples may range from the topic of the sentence, to its length, the semantic category of any word between the extremes of the generalizes 2-gram, and so on.

Concerning the expressive power of the encoding, *These people. They arrive forming a continuous wave, an endless flow that changes societies at all levels, swirling together different and irreconcilable cultures. These are the migrants, often considered a problem.*

$$\langle L\rangle(\text{topic 'migrants'} \wedge \langle D\rangle \text{topic 'fluid'})$$
$$\Downarrow$$
metaphor

These people, the migrants, arrive on dilapidated boats at the mercy of the waves and flows. They risk their lives and when they arrive they are often rejected, because, it is believed, they risk changing societies at all levels, including cultural ones.

$$\langle L\rangle(\text{topic 'migrants'} \wedge \neg[D] \text{topic 'fluid'})$$
$$\Downarrow$$
not a metaphor

Figure 2: Example of generalized 2-grams in text, and of metaphor identification via their qualitative relationship.

now observe that in the standard text-based approaches, the relative ordering of the original sentences is lost, while only the order of constituents of each n-words is preserved. Generalized 2-words, instead, are naturally linked to a qualitative, more-than-propositional logic that allows one to preserve the ordering in a very expressive way. The key idea is that a sentence can be seen as a *linearly ordered* sequence of words, which entails, in turn, a *temporal order*, as also proposed in other models, such as BiLSTMs [15, 16]. Thus, a generalized 2-words is an *interval* in such a order, and any two intervals on a linear order can be qualitatively related to each other in exactly one of thirteen ways. The family of logics that allow one to describe propositional properties of intervals on a linear order is called *interval temporal logics*, and they belong to the more general category of *modal logics*. Originally studied by Allen in the early 80s, interval temporal logic have been formalizes a few years later, and the most representative language for expressing propositional properties of intervals is the *modal logic of time intervals*, or HS [17]. In HS, each of the possible binary relations that may exist between two intervals becomes an accessibility relation; it can be immediately verified that they are, in fact, thirteen: *after* (capturing an interval that starts at the end of the current one, usually denoted by $\langle A\rangle$), *later* (capturing an interval that starts past the end of the current one, $\langle L\rangle$), *overlaps* (capturing an interval that starts during the current one and ending after it, $\langle D\rangle$), *during* (capturing an interval that starts and ends within the current one, $\langle D\rangle$), *begins* (capturing an interval that starts at the start of the current one and ends before it, $\langle B\rangle$), and *ends* (capturing an interval that starts within the current one and ends with it, $\langle E\rangle$). Working with the relations/operators as they were originally introduced may not be always suitable; interval temporal logics such as HS have been simplified for specific tasks in several ways. Among them, the most relevant proposals include the so-called *topological* versions of interval temporal logic, in which the relations are, in fact, disjunctions of Allen's relations. So, for example, in the case of HS_3 [18], two intervals can just *have at least one point in common* or can be *completely separated*; languages such as this one may in fact be designed, and its expressive power be modulated, depending on the task. Symbolic learning algorithms for interval temporal logic have been recently studied [19] and used for learning interval temporal properties in very different contexts, mostly, but not exclusively, in the medical sciences (see [20, 21], among others); in those cases, the object being encoded are multi-variate temporal series, via a process that eventually produces interval temporal models from which rules are ultimately learned. It is of notice how such diverse contexts, including text-related tasks, can in fact be approached with the same methodology. Continuing with the example in Fig. 1, the generalized 2-gram w_4, w_6 is *during* the generalized 2-gram w_3, w_7

In Fig. 2, we show how, in a text, relevant generalized 2-grams are identified; in both texts, two generalized 2-grams are identified. Focusing on the top paragraph, the first generalized 2-gram, in red color, is captured by the words *people* and *migrants*; the entire text in between (even ignoring the full stop, thus ignoring that they belong to two different sentences) is categorized as *topic 'migrants'*, thus imitating a human reader who, reading the complete text, can identify when the writer starts referring to some category of persons, when he/she stops doing that, and which one this category is. The second generalized 2-gram, in blue color, is captured by the words *wave* and *swirling*, and the entire text in between is categorized as *topic 'fluid'* (observe the frequencies of words that refer to fluids, and water in particular, that occur in the blue-highlighted text). The bottom paragraph shows similar words in a similar but identical order. Both topics are still present and identified in the same way. However, the two topics are in a different topological order. On the right-hand side, we propose a possible rule linking the topics' topological order to distinguish between metaphoric and non-metaphoric text written in propositional HS. Most interestingly, ChatGPT (version 3.5, consulted on the prompt in September 2023) classifies both texts as metaphoric, probably because metaphors linking fluids and migrants are statistically common.

3. Conclusions

This work represents an initial attempt to approach symbolic learning for text-related tasks like metaphor detection. A symbolic approach can extract a theory from a specific linguistic phenomenon, which raises at least three problems: first, determining whether a theory of a phenomenon should exist and in what terms; second, finding the appropriate logic for the extraction process; and third, ensuring the existence of an automatic method for extracting the theory in that logic. In this work, we have attempted to address the first and second points, and we did so using a logical formalism for which a solution to the third one already exists. Should this approach be successful, it can be used to address other text-related challenges, such as all variants of text classification. Additionally, our generalized 2-gram encoding can be further generalized to partially benefit from well-known word-to-vec approaches without compromising its symbolic essence.

We will further verify our hypotheses by conducting some tests on an annotated newspaper corpus, which was human-labeled as *metaphor* or *non-metaphor*, consisting of 13,000 tokens and 2,000 different words. The label pertains to the entire text, and the task will regard recognizing its metaphoric expressions.

References

[1] J. Charteris-Black, Corpus Approaches to Critical Metaphor Analysis, Palgrave Macmillan UK, 2004. URL: http://link.springer.com/10.1057/9780230000612. doi:10.1057/9780230000612.

[2] G. Lakoff, M. Johnson, Metaphors we live by, University of Chicago Press, Chicago, 1980.

[3] R. Gibbs, The Poetics of Mind: Figurative Thought, Language, and Understanding, transferred to digital printing ed., Cambridge Univ. Press, Cambridge, 1994.

[4] R. Trim, Metaphor and the Historical Evolution of Conceptual Mapping, 1. publ ed., Palgrave Macmillan, 2011. OCLC: 760134631.

[5] E. Semino, Metaphor in Discourse, Cambridge University Press, Cambridge, 2008.

[6] S. Nacey, G. Reijnierse, T. Krennmayr, A. Dorst, Metaphor Identification in Multiple Languages: MIPVU Around the World, number 22 in Converging Evidence in Language and Communication Research, John Benjamins Publishing Company, Philadelphia, 2019.

[7] T. Veale, E. Shutova, B. Klebanov, Metaphor: A Computational Perspective, Springer International Publishing, Cham, 2016. URL: https://link.springer.com/10.1007/978-3-031-02160-2. doi:10.1007/978-3-031-02160-2.

[8] X. Tong, E. Shutova, M. Lewis, Recent advances in neural metaphor processing: A linguistic, cognitive and social perspective, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 4673–4686. URL: https://aclanthology.org/2021.naacl-main.372. doi:10.18653/v1/2021.naacl-main.372.

[9] Y. Neuman, D. Assaf, Y. Cohen, M. Last, S. Argamon, N. Howard, O. Frieder, Metaphor identification in large texts corpora, PLOS ONE 8 (2013) 1–9. URL: https://doi.org/10.1371/journal.pone.0062343.

[10] R. Mao, Computational Metaphor Processing, Ph.D. thesis, University of Aberdeen, UK, 2020. URL: https://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.834151.

[11] Y. Shlomo, M. Last, MIL: automatic metaphor identification by statistical learning, in: Proceedings of the Workshop on Interactions between Data Mining and Natural Language Processing (DMNLP), volume 1410 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2015, pp. 19–30. URL: https://ceur-ws.org/Vol-1410/paper4.pdf.

[12] M. Babieno, M. Takeshita, D. Radisavljevic, R. Rzepka, K. Araki, MIss RoBERTa WiLDe: Metaphor identification using masked language model with wiktionary lexical definitions, Applied Sciences 12 (2022). URL: https://www.mdpi.com/2076-3417/12/4/2081. doi:10.3390/app12042081.

[13] S. Rai, S. Chakraverty, D. Tayal, Supervised metaphor detection using conditional random fields, in: Proceedings of the Fourth Workshop on Metaphor in NLP, Association for Computational Linguistics, San Diego, California, 2016, pp. 18–27. URL: https://aclanthology.org/W16-1103. doi:10.18653/v1/W16-1103.

[14] Y. Bizzoni, M. Ghanimifard, Bigrams and BiLSTMs two neural networks for sequential metaphor detection, in: Proceedings of the Workshop on Figurative Language Processing, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 91–101. URL: https://aclanthology.org/W18-0911. doi:10.18653/v1/W18-0911.

[15] C. Wu, F. Wu, Y. Chen, S. Wu, Z. Yuan, Y. Huang, Neural metaphor detecting with CNN-LSTM model, in: Proceedings of the Workshop on Figurative Language Processing, (Fig-Lang@NAACL-HLT), Association for Computational Linguistics, 2018, pp. 110–114.

[16] R. Mao, C. Lin, F. Guerin, End-to-end sequential metaphor identification inspired by linguistic theo-

ries, in: Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL), Association for Computational Linguistics, 2019, pp. 3888–3898.
[17] J. Halpern, Y. Shoham, A propositional modal logic of time intervals, Journal of the ACM 38 (1991) 935–962.
[18] E. Muñoz-Velasco, M. Pelegrín-García, P. Sala, G. Sciavicco, I. Stan, On coarser interval temporal logics, Artif. Intell. 266 (2019) 1–26. URL: https://doi.org/10.1016/j.artint.2018.09.001. doi:10.1016/j.artint.2018.09.001.
[19] D. Della Monica, G. Pagliarini, G. Sciavicco, I. Stan, Decision trees with a modal flavor, in: Proceedings of 21st International Conference of the Italian Association for Artificial Intelligence (AIxIA), volume 13796 of *Lecture Notes in Computer Science*, Springer, 2022, pp. 47–59.
[20] F. Manzella, G. Pagliarini, G. Sciavicco, I. Stan, The voice of COVID-19: breath and cough recording classification with temporal decision trees and random forests, Artif. Intell. Medicine 137 (2023) 102486. URL: https://doi.org/10.1016/j.artmed.2022.102486. doi:10.1016/j.artmed.2022.102486.
[21] S. Mazzacane, M. Coccagna, F. Manzella, G. Pagliarini, V. A. Sironi, A. Gatti, E. Caselli, G. Sciavicco, Towards an objective theory of subjective liking: a first step in understanding the sense of beauty, PLOS ONE 18 (2023) e0287513.

Building a corpus on Eating Disorders from TikTok: challenges and opportunities

Melissa Donati, Ludovica Polidori, Paola Vernillo and Gloria Gagliardi

Alma Mater Studiorum - University of Bologna, Italy

Abstract

We present two synchronic corpora of Eating Disorders (ED) related discourse on Social Media. PAC (i.e., ProAna/Anorexia Corpus) and RAC (i.e., Recovery from Ana/Anorexia Corpus) resources focus on the contents posted on TikTok, respectively, by communities promoting anorectic behavior and users sharing experiences concerning the process of recovery from their ED. We report on the corpus statistics and creation process, focusing specifically on the methodological issues raised by this novel Social Media platform.

Keywords

Eating Disorders, Corpus Linguistics, TikTok

1. Introduction

It was only 20 years ago that one of the darkest sides of Eating Disorders (ED) was revealed through the proliferation of websites, blogs, and social networks, in which a growing number of adolescents and young adults started sharing information about their eating experiences with like-minded users. Among these pro-ED communities, researchers and clinicians showed particular concern for pro-Ana (i.e., "pro-anorexia") groups, i.e., web-based communities of anorexic (or aspiring anorexic) individuals engaged in the promotion of their Eating Disorder [1]. Interestingly, one of the most horrific and dangerous aspects of pro-Ana groups is that Anorexia Nervosa (AN) is not presented as a psychiatric disorder associated with pathological body image dissatisfaction [2], but more as a way of living with its own rules and rituals to be respected. While over the last years, much has been done to prevent the circulation of pro-ED content on social media (e.g., TikTok's adoption of measures to obscure harmful contents: [3]), a new but specular phenomenon recently took the toll, that is, the spread of pro-recovery accounts of individuals who are in the process of healing from an ED and are willing to share their eating experience to help other online users [4]. From a linguistic perspective, research on ED has been very limited and became an object of study only in recent years [5, 6, 7, 8, 9, 10] as opposed to other psychopathologies, such as schizophrenia [11, 12], personality disorder [13], and depression [14, 15, 16, 17]. This already problematic picture has been further compromised by the inhomogeneous representation of linguistic data in the literature, where the majority of studies have been dedicated to the linguistic profiling of ED-affected individuals in a Germanic language (English, German, Norvegian) [18]. This paper represents a small step towards the reversal of this tendency but a crucial part of two larger projects (Metaphan[1] and RaAM project 2022[2]) aiming at identifying, by the adoption of different NLP techniques and tools, potential lexical and semantic patterns in anorectic individuals. To this end, in the current research, we show the data collection process (i.e. oral and written productions) from ED communities on TikTok, currently representing the most widely used social media among young people and adolescents, namely the population groups at greater risk for EDs. In the following paragraphs, we give a brief overview of the literature on the topic (Section 2), then we describe the process of creating the corpus and discuss the methodological issues that were met (Section 3) and to conclude we provide few insights for future works (Section 4).

2. Related Works

In recent years, we have witnessed exponential growth in the use of Social Media (SM), especially by adolescents and young people. The community-building nature and the interactive dynamics of these platforms, as well as the less direct way of communicating, encourage users to openly discuss a wide variety of topics [19]. In turn, this makes available huge amount of data that can be used for different purposes (e.g. extract actionable patterns, form conclusions about users, conduct research, etc.). For this reason, Social Media Mining (SMM), i.e., the process of extracting big data from SM, now constitutes

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

melissa.donati@studio.unibo.it (M. Donati); ludovica.polidori@studio.unibo.it (L. Polidori); paola.vernillo@unibo.it (P. Vernillo); gloria.gagliardi@unibo.it (G. Gagliardi)

[1] https://site.unibo.it/metaphan/en

[2] https://site.unibo.it/metaphan/en/connected-research-activities

a well-established methodology to collect large samples of data in different research areas [20]. This approach has proved particularly fruitful for collecting data on EDs as people suffering from these disorders seem to overcome the self-protective nature of their ED to engage in ED-related discourse with online users sharing similar experiences [21]. Indeed, in the last decade, many studies have used different SM platforms as a source of data to analyze EDs [22, 23, 24, 25, 21, 26, 27, 28, 29]. However, the state-of-the-art on ED-discourse on SM currently presents two main limitations: i) the majority of the analysis was carried out on small datasets built ad-hoc for the purpose of the work (with the only exception of [30]), and ii) they mostly focused on the English language. As a matter of fact, in the Italian framework there have been very little research on the representation of EDs on SM, and that little was mostly focused on Anorexia-Nervosa and did not target EDs in general [31, 32, 33].

3. Corpus Creation: Methodological Issues

Against this background and intending to fill this gap, we created a collection of English and Italian ED-related data that could be used for different types of research (from purely linguistic and content analyses that could help pinpointing the features and characteristics of ED-related discourse, to various computational techniques that could be used to implement systems of automatic detection of ED-related content on SM). We selected TikTok as a source of data as it currently represents the most widely used SM, especially among young people and adolescents, namely the at-risk population for EDs [34].

To achieve this goal, we first needed to define the nature and characteristics of the corpus itself. As far as the linguistic features are concerned, our corpus is specialized (i.e., is focused on the topic of EDs discourse on TikTok), synchronic (i.e., refers to a specific point in time that is the moment the data were downloaded), and targets both written and spoken language (TikTok videos contain spoken and/or written text). We did not set *a priori* a target dimension to be reached, because this feature is totally dependent upon the possibility of extracting the data automatically (Section 3.1). Conversely, following the common practice in the domain of SMM, we assumed that 'there is no data like more data' and intended to download as many videos as possible. To maximize the corpus representativity, we tried to balance the sample with respect to the types of videos being collected but we could not do so concerning the users' gender, because for both corpora the vast majority of profiles were of female individuals (see Section 3.3 for more details). The target population consisted of those profiles that identify themselves in one of the two following categories: i) supporters of anorectic behaviors (for the English PAC corpus); ii) witnesses and motivators for the recovery process (for the Italian RAC corpus). Such profiles were identified based on the linguistic and non-linguistic (i.e., emojis) information present in their profile bio. The selection criteria will be presented in Section 3.1, prior to the description of the data collection process and the discussion of the related issues that were encountered.

Before getting further into the methodology, it is necessary to make an ethical consideration concerning the collection of data from SM. Broadly speaking, SM posts that are publicly accessible are treated as belonging to the public domain, therefore, according to common practice, consent from the creators is not deemed necessary to download such data. This is strengthened by the fact that, upon registration, TikTok asks its users to consent to a set of terms of service that make the data available for access to third parties [35]. In addition, when creating and managing their accounts and contents users can decide to make them publicly accessible or private (i.e. only viewable by accepted followers); at any time, they can also restrict access to some of their contents through privacy settings and choose whether to make them downloadable. For the above reasons, given that for the purpose of this work only public and downloadable data was analysed, we did not seek users' consent to collect the posts. In compliance with similar SM analysis [26], no reference to any identifying information, such as usernames, will be made.

3.1. Data Collection

As explained above, the selection criteria adopted to identify the target profiles was based on the information present in the profiles' bio. However, to track the target profiles, we needed to start from a list of ED-related hashtags that could lead us to such profiles via a keyword-based search. The hashtags that were used herein were generated both by brainstorming and by exploring the platform for a couple of weeks, noting down the most popular trends and the most widely used hashtags (see Table 1 for an overview). Following this hashtag-driven search, we noticed that there was very little -if any- pro-Ana content produced in Italian, that is why for this type of ED-related content we decided to collect a small sample of English data. On the other hand, we found quite some profiles representing the ED-recovery community.

Among these profiles, we selected those having at least 10k followers (some of them exceed 2M followers) and at least 10 ED-related posts, so that we could maximize the chance of gathering interesting and relevant linguistic information. We then used the ED-related hashtags to conduct a within-profile research to select only the ED-related videos in each profile in order to extract them.

At this point, the next step consisted of extracting the

Table 1
List of pro-Ana and pro-Recovery hashtags that were used to search for TikTok profile that share ED-related content.

Pro-Ana hashtags	Pro-Recovery hashtags
#weightloss (w3ightl0ss) #unhealthyweightloss (+ lexical variations) #kpop[3]	#dcarecovery (dcar3covery) #dca [4] #dcaitalia #fiocchettolilla [5] #dcafighting

identified ED-related videos from the selected profiles. For the sake of time and efficiency, we wanted to download the data automatically. However, differently from other popular SM, TikTok has not yet released any official API that can be used by researchers and developers to automate the process of accessing and extracting the data. In addition, even if unofficial APIs exist, they get outdated almost immediately after their release because TikTok is constantly updating the anti-bot system preventing automatic access from the same IP. To get around this, we looked for a reliable and cost-effective proxy provider for TikTok scraping, but we could not find any viable solution.

Therefore we decided to proceed with the manual downloading of the data. The main drawback of this way of proceeding is that due to time and resource constraints we could not collect a very large number of videos (see Table 2). On the bright side, however, the manual downloading allowed us to i) enhance the content filtering process and ii) notice that TikTok videos have different formatting styles that might be worth distinguishing not only to ease the ensuing transcription process but also to conduct separate content analysis and compare the different results. Based on our observations about the different formatting styles, we grouped the TikTok videos into 4 subcorpora: 1)*Speech-onl*y videos: in which the user was talking in the absence of background music and/or written text; 2) *Playback*: in which the user lip-sync over a song or an extract from a movie or tv shows; 3) *Text-only*: in which there is neither background music nor the users themselves speaking, but only written text superposed on the video; and 4) *Mixed*: in which the above-mentioned features are present in various combinations.

[3]K-pop (for Korean-pop) is a popular genre of music originating from South Korea that has been hugely influential in the 'diet scene' because young people want to look like their favourite K-pop stars that are known for their extreme diets, indeed many young artists have left behind the K-pop world in order to focus on eating disorder treatment.

[4]Disturbo del Comportamento Alimentare (Eating Disorder).

[5]The Lilac Ribbon is the official international symbol against Eating Disorders.

3.2. Transcription

Organizing the videos into 4 categories was particularly useful for the transcription phase as it allowed to adopt different strategies and techniques based on the input characteristics. As for the downloading phase, although we intended to automatize the transcription process as much as possible, the high complexity of the data has, in some cases, made human intervention necessary.

For speech-only and playback videos automatic transcription was performed using the Google Web Speech API, which is easily accessible through the SpeechRecognition Library [36]. To assess the quality of the automatic transcription, a random sample of videos (n=10) for each category was extracted, transcribed manually and then compared with the machine-based transcription. For speech-only videos, a high agreement score was obtained between human and machine transcription (>90%) which confirmed the viability of the method adopted. Conversely, playback videos emerged as more problematic, thus manual correction was needed because both singing and the music accompaniment adversely impacted on intelligibility.

Automatic transcription was also attempted for text-only videos by means of Optical Character Recognition (OCR) using the Tesseract OCR engine [37], but we obtained poor results due to the high visual complexity of the input data, more specifically to the extreme variability of font type, size, and color, the lack of adequate contrast with the background, the non-hierarchical spatial organization of texts, and the presence of non-textual graphical elements (e.g., lexical variations of words, where letters are substituted by numbers or emojis to prevent the platform's censorship and filtering system from blocking the content as potentially harmful, e.g., 'starving' written replacing star with the corresponding emojis, or 'disorder' written as 'd1s0rder'). The same issue, boosted to the maximum, was observed with mixed videos, where speech, music, and written text were mingled. Therefore, for these two categories of videos, we could only perform the transcription manually.

We reported below, as an example of the type of ED-related content that was selected, the transcription of two videos, one for each of the two datasets.

> [from RAC]
>
> *"questo video è davvero davvero difficile da registrare per me ma lo faccio perché voglio condividere tutta la mia vita con voi e voglio aiutare delle persone che si trovano nella mia stessa situazione parlando del mio problema dovete sapere che io sono stata prima anoressica sono arrivata a pesare 36 kg e vi parlerò poi te la causa scatenante poi riscoperto il cibo ho iniziato ad abbuffarmi in una maniera assurda a sentirmi in colpa e quindi poi a vomitare questa*

si chiama bulimia ovviamente alternavo momenti digiuno quindi magari non mangio proprio per giorni a momenti in cui il tuo corpo ha bisogno di cibo e quindi ti abbuffi e mangi qualsiasi cosa volevo solo dirvi che ieri è successa un'altra volta il fatto è che io me lo vedo subito in faccia cioè mi vedo 10 volte più grossa e mi sento davvero super gonfia che senti ma sono riuscita a non vomitare perché io sono più forte sono con tutte voi[6]"

[from PAC]

*"i'm *** i'm a new member stats starter weight 140.1 ibs goal weight 100 ibs ultimate goal weight 90 ibs for now i binge eat when i'm bored so i gained a lot of weight in the past months i'm trying to limit myself on eating i am currently 4'10 and i'm overweight for my height age i listen to subliminal and trying to workout also i hate exercising but i realized it is healthy for me and my body 33"*

3.3. Corpus Statistics

In Table 2, we reported an overview of the statistics for the two corpora in terms of number of videos, number of words, and number of users from whose profiles the data were extracted.

The two corpora are registered in CLARIN [7], but not publicly accessible for the moment.

Table 2
Statistics for the two corpora.

	PAC	RAC
n videos	250	1000
n words	13169	116261
n users	14 (all F)	27 (26 F, 1 M)

[6] [our translation] *"making this video is really really hard for me but I am doing it because I want to share everything about my life with you and I want to help those who are experiencing the same situation by talking about my problem you must know that I have suffered first from anorexia I ended up weighting 36 kg and I will tell you about the trigger then I rediscovered food and started insanely binging and feeling guilty and then as a consequence throwing up this is called bulimia obviously I alternated periods of fasting so peraphs I would not eat for days with periods in which my body needed food and I would eat anything and I just wanted to tell you that yesterday it happened again and the thing is that I see it immediately on my face that is I see myself 10 times bigger and I fell really extremely bloated that you know but I managed not to throw up because I am stronger I am with you all"*

[7] http://hdl.handle.net/20.500.11752/OPEN-997

4. Conclusion and Future Works

The aim of this work was twofold: on the one hand, we wanted to present two corpora on EDs, the English pro-Ana corpus (PAC) and the Italian pro-Recovery corpus (RAC), that were both built by extracting data from the popular SM TikTok; on the other, we wanted to discuss some methodological issues related to building a corpus using this platform as a source of data. More specifically, we pointed out that the absence of an official API does not allow the automatic extraction of the videos and requires manual work, which is highly time-consuming and does not allow to collect a very large sample of data. This, in turn, might impede the application of more complex computational analysis and limit the generalizability of the results. In addition, we raised the issue related to the transcription of the videos to text. In this case, implementing automatic approaches is not always feasible because of the extreme visual complexity and variability of TikTok videos.

Given the highly interactive nature of this SM and its unprecedented success, we believe that TikTok constitutes an extremely interesting source of linguistic and non-linguistic data that could be used to analyze other complex social and psychological phenomena and we hope that this work paves the way for further research in this direction.

CRediT authorship contribution statement

MD Conceptualization, Methodology, Software, Data Curation (i.e., download, automatic transcription, annotation), writing (§2,3,4)
LP Data Curation (i.e., manual transcription)
PV Conceptualization, Data Curation (i.e., download), Writing (§1)
GG Supervision, Funding acquisition.

Funding

This work was partially funded by the RaAM Association (project "How about metaphors for dinner? A digest of metaphorical conceptualizations in pro-Ana communities") and the University of Bologna (AlmaIdea 2022 - "MetaphAN" project).

References

[1] N. Boero, C. J. Pascoe, Pro-anorexia communities and online interaction: Bringing the pro-ana body online, Body & Society 18 (2012) 27–57.

[2] J. B. Williams, M. First, Diagnostic and statistical manual of mental disorders, in: Encyclopedia of social work, 2013.

[3] S. Marsh, Tiktok investigating videos promoting starvation and anorexia, The Guardian 7 (2020).

[4] A. K. Greene, H. N. Norling, L. M. Brownstone, E. K. Maloul, C. Roe, S. Moody, Visions of recovery: a cross-diagnostic examination of eating disorder pro-recovery communities on tiktok, Journal of Eating Disorders 11 (2023) 109.

[5] C. F. Bates, "i am a waste of breath, of space, of time" metaphors of self in a pro-anorexia group, Qualitative Health Research 25 (2015) 189–204.

[6] O. Knapton, Pro-anorexia: Extensions of ingrained concepts, Discourse & Society 24 (2013) 461–477.

[7] E. J. Lyons, M. R. Mehl, J. W. Pennebaker, Pro-anorexics and recovering anorexics differ in their linguistic internet self-presentation, Journal of psychosomatic research 60 (2006) 253–256.

[8] F. Skårderud, Eating one's words, part ii: The embodied mind and reflective function in anorexia nervosa—theory, European Eating Disorders Review: The Professional Journal of the Eating Disorders Association 15 (2007) 243–252.

[9] F. Skårderud, Eating one's words, part i:'concretised metaphors' and reflective function in anorexia nervosa—an interview study, European Eating Disorders Review: The Professional Journal of the Eating Disorders Association 15 (2007) 163–174.

[10] M. Wolf, F. Theis, H. Kordy, Language use in eating disorder blogs: Psychological implications of social online activity, Journal of Language and Social Psychology 32 (2013) 212–226.

[11] V. Bambini, G. Arcara, M. Bechi, M. Buonocore, R. Cavallaro, M. Bosia, The communicative impairment as a core feature of schizophrenia: Frequency of pragmatic deficit, cognitive substrates, and relation with quality of life, Comprehensive psychiatry 71 (2016) 106–120.

[12] J. De Boer, M. Van Hoogdalem, R. Mandl, J. Brummelman, A. Voppel, M. Begemann, E. Van Dellen, F. Wijnen, I. Sommer, Language in schizophrenia: relation with diagnosis, symptomatology and white matter tracts, npj Schizophrenia 6 (2020) 10.

[13] A. Arntz, L. D. Hawke, L. Bamelis, P. Spinhoven, M. L. Molendijk, Changes in natural language use as an indicator of psychotherapeutic change in personality disorders, Behaviour research and therapy 50 (2012) 191–202.

[14] J. D. Bernard, J. L. Baddeley, B. F. Rodriguez, P. A. Burke, Depression, language, and affect: an examination of the influence of baseline depression and affect induction on language, Journal of Language and Social Psychology 35 (2016) 317–326.

[15] T. Brockmeyer, J. Zimmermann, D. Kulessa, M. Hautzinger, H. Bents, H.-C. Friederich, W. Herzog, M. Backenstrass, Me, myself, and i: self-referent word use as an indicator of self-focused attention in relation to depression and anxiety, Frontiers in psychology 6 (2015) 1564.

[16] N. Ramirez-Esparza, C. Chung, E. Kacewic, J. Pennebaker, The psychology of word use in depression forums in english and in spanish: Testing two text analytic approaches, in: Proceedings of the international AAAI conference on web and social media, volume 2, 2008, pp. 102–108.

[17] J. Zimmermann, T. Brockmeyer, M. Hunn, H. Schauenburg, M. Wolf, First-person pronoun use in spoken language as a predictor of future depressive symptoms: Preliminary evidence from a clinical sample of depressed patients, Clinical psychology & psychotherapy 24 (2017) 384–391.

[18] V. Cuteri, G. Minori, G. Gagliardi, F. Tamburini, E. Malaspina, P. Gualandi, F. Rossi, M. Moscano, V. Francia, A. Parmeggiani, Linguistic feature of anorexia nervosa: a prospective case–control pilot study, Eating and Weight Disorders-Studies on Anorexia, Bulimia and Obesity (2021) 1–9.

[19] A. Lenhart, K. Purcell, A. Smith, K. Zickuhr, Social media & mobile internet use among teens and young adults. millennials., Pew internet & American life project (2010).

[20] P. Gundecha, H. Liu, Mining social media: a brief introduction, New directions in informatics, optimization, logistics, and production (2012) 1–17.

[21] T. E. Kenny, S. L. Boyle, S. P. Lewis, # recovery: Understanding recovery from the lens of recovery-focused blogs posted by individuals with lived experience, International Journal of Eating Disorders 53 (2020) 1234–1243.

[22] M. Lukač, et al., Down to the bone: A corpus-based critical discourse analysis of pro-eating disorder blogs, Jezikoslovlje 12 (2011) 187–209.

[23] L. Mullany, C. Smith, K. Harvey, S. Adolphs, 'am i anorexic?'weight, eating and discourses of the body in online adolescent health communication, Communication & medicine 12 (2016).

[24] M. Moessner, J. Feldhege, M. Wolf, S. Bauer, Analyzing big data in social media: Text and network analyses of an eating disorder forum, International Journal of Eating Disorders 51 (2018) 656–667.

[25] B. K. Bohrer, U. Foye, T. Jewell, Recovery as a process: Exploring definitions of recovery in the context of eating-disorder-related social media forums, International Journal of Eating Disorders 53

(2020) 1219–1223.
[26] S. S. Herrick, L. Hallward, L. R. Duncan, "this is just how i cope": An inductive thematic analysis of eating disorder recovery content created and shared on tiktok using# edrecovery, International journal of eating disorders 54 (2021) 516–526.
[27] G. L. Jordan, M. D. Garcìa, B. L. Dìez, P. M. Sànchez, J. G. Del Barrio, R. Ayesa-Arriola, Facebook as a pro-ana and pro-mia resource, European Psychiatry 64 (2021) S703–S703.
[28] C. González-Nuevo, M. Cuesta, J. Muñiz, Concern about appearance on instagram and facebook: Measurement and links with eating disorders, Cyberpsychology: Journal of Psychosocial Research on Cyberspace 15 (2021).
[29] M. Minadeo, L. Pope, Weight-normative messaging predominates on tiktok—a qualitative content analysis, Plos one 17 (2022) e0267997.
[30] M. Donati, C. Strapparava, CorEDs: A corpus on eating disorders, in: Proceedings of the RaPID Workshop - Resources and ProcessIng of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric/developmental impairments - within the 13th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 80–85. URL: https://aclanthology.org/2022.rapid-1.10.
[31] V. Richichi, A. Chinello, F. Parma, L. E. Zappa, E. Mazzoni, F. Monti, Anoressia nervosa e internet. uno studio sui blog pro-ana in italia, Psicologia clinica dello sviluppo 22 (2018) 499–514.
[32] N. L. Bragazzi, G. Prasso, T. S. Re, R. Zerbetto, G. Del Puente, A reliability and content analysis of italian language anorexia nervosa-related websites, Risk management and healthcare policy (2019) 145–151.
[33] G. Gagliardi, "odio tutto ciò, voglio le ossa": Una prima indagine sulle caratteristiche linguistiche delle pagine social pro-ana in lingua italiana, Italiano LinguaDue 13 (2021) 520–536.
[34] A. Sherman, Tiktok reveals detailed user numbers for the first time, Retrieved October 2 (2020) 2020.
[35] 2023. URL: https://www.tiktok.com/legal/page/eea/privacy-policy/en.
[36] Y. Zhang, M. Pezeshki, P. Brakel, S. Zhang, C. L. Y. Bengio, A. Courville, Towards end-to-end speech recognition with deep convolutional neural networks, arXiv preprint arXiv:1701.02720 (2017).
[37] J. Ooms, tesseract: Open Source OCR Engine, 2023. Https://docs.ropensci.org/tesseract/ (website) https://github.com/ropensci/tesseract (devel).

A preliminary release of the Italian Parliamentary Corpus

Valentino Frasnelli[1,†], Alessio Palmero Aprosio[2,*,†]

[1] Università di Trento, Via Giuseppe Verdi 26, I-38122 Trento, Italy
[2] Fondazione Bruno Kessler, Via Sommarive 18, I-38121 Trento, Italy

Abstract

English. Political debates have been used for years in political and social studies on languages and their cultures. In this paper, we release a preliminary version of the Italian Parliamentary Corpus, a dataset containing 1.2 billion words that includes the political debates in the Italian Parliament from 1848 to 2018. The data has been collected applying an Optical Character Recognition (OCR) software to the original documents, available in PDF format on the websites of *Camera dei Deputati* and *Senato della Repubblica*.

Italian. I dibattiti politici vengono usati da anni in studi sociali e politici sulle lingue e le loro culture. In questo articolo, rilasciamo una versione preliminare dell'Italian Parliamentary Corpus, un dataset contenente 1.2 miliardi di parole che include i dibattiti politici del Parlamento Italiano dal 1848 al 2018. I dati sono stati collezionati applicando un software di Optical Character Recognition (OCR) ai documenti originali, disponibili in formato PDF sui siti web della *Camera dei Deputati* e del *Senato della Repubblica*.

Keywords

Parliamentary Corpus, Political debates, OCR post-correction, Italian Parliament

1. Introduction

The analysis of parliamentary debates is very important from many research perspectives. Apart from political science, this kind of data can be used to understand how a language and its culture evolves in history. In particular, in the last two centuries the Italian society has changed under a lot of points of view. Since the transition from the absolute monarchy to the parliamentary monarchy, that took place in 1848, Italy went through historical events such as two world wars, the fascist dictatorship, the exile of the royal family, the universal suffrage, the accession to the European Union, and much more. Such important milestones, along with all the rest of the Italian political and social life, are traced in the parliamentary reports.

Most research groups around the world have already collected and released corpora of political debates in various languages, used in diversified fields, such as religion [1] and gender [2] studies, multilinguality [3], and so on. GerParCor [4] is a dataset containing the German-language parliamentary protocols from three centuries and four countries. Similarly, siParl [5], DutchParl [6], and the Polish Parliamentary Corpus [7] are collection of political debates, in Slovenian, Dutch, and Polish languages respectively. In addition, since the creation of the European Union, political debates of the European Parliament have been made available in multiple languages, becoming a precious resource for machine translation [8].

In this paper, we present the preliminary version of the Italian Parliamentary Corpus, a collection of documents covering 200 years and containing all the documents redacted by the two houses of the bicameral Italian Parliament (*Camera dei Deputati*, the lower house, and *Senato della Repubblica*, previously *Senato del Regno*, the upper house).

The rest of this article is structured as follows. In Section 2 we describe how the raw data has been collected. Section 3 we show the steps performed to get the clean texts. Section 4 contains some statistics of the dataset. Finally, both the source code and the dataset are available for download, as described in Section 5.

2. Data collection

We downloaded all the available documents available online on the websites of the two houses of the Italian Parliament.

While each website is managed by a different administration, both of them released the data in structured format (RDF for the *Camera dei Deputati*, and CSV/JSON/XML for the *Senato della Repubblica*). The *Camera dei Deputati* website contains complete catalogue of digital data and documents from the Legislature of the Kingdom of Sardinia to all the data of the Republic. Differently, for the same data belonging to the Senato della Repubblica we could directly download only documents produced after 1948. Since the debates in the 1848-1940

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
*Corresponding author.
† These authors contributed equally.
valentino.frasnelli@studenti.unitn.it (V. Frasnelli); aprosio@fbk.eu (A. Palmero Aprosio)
0000-0002-1484-0882 (A. Palmero Aprosio)

time interval have already been digitalized, but not yet published at the time of writing, we could obtain them thanks to the precious help from the *Servizio dei Resoconti e della Comunicazione istituzionale del Senato della Repubblica*.

In both cases, documents dated before 1996 were not produced natively in a digital format, therefore are available only in PDF scanned format. Starting from 1996 (Republic Legislature number XIII), debates have been published also in text format on the web.

3. Processing

To convert PDF scanned documents to text, we used Optical Character Recognition (OCR), in particular Tesseract [9], a software originally developed by Hewlett-Packard, and subsequently released as open source. Tesseract is free to use and can support more than 100 languages out-of-the-box (among them, Italian).

After the conversion, the data is cleaned using some rule-based heuristics: headers, footers and indexes are removed, hyphenated words are joined, and pages are merged.

Finally, we needed to test the OCR output quality. To do this, we compiled a gold standard consisting of 30 pages manually transcribed, taken from different legislatures spanning from 1848 to 1996.

To evaluate the accuracy of the extraction, we use two metrics: word error rate (WER), and character error rate (CER). The error rates are derived from Levhenstein distance [10] and quantify the number of operation – insertions, deletions and substitutions – needed to transform one string in the other. They are common metrics for evaluating the performance of speech recognition and machine translation systems, but are often used also for OCR [11].

They are computed as follows:

$$\mathrm{WER/CER} = \frac{I + S + D}{N}$$

where I, S, and D represent the number of insertions, substitutions, and deletions respectively. N is the total number of instances (words or character, depending on which metric is considered). The lower the value, the higher the accuracy.

As a baseline, we first evaluated the accuracy of the extraction on the output of Tesseract. Then, we applied the spell-checker software SymSpell.[1] Since SymSpell only works on words (or word-like strings), we removed all the punctuation marks from the text. We also ignore case and consider every word as lowercase.

SymSpell makes use of dictionaries for the correction of documents in the format `<word> <frequency>` for all words one wants to insert in the dictionary. Since SymSpell Italian default dictionary is build on top of recent and general purpose texts, we attempted to create dictionaries using the lexicon present in the documents themselves, trying to filter out those words containing errors. The idea is to create custom dictionaries for each legislature, containing only words coming from the time period of that legislature, in order to better capture the historical nuances for each legislature. To avoid as much as possible inserting words with spelling errors into the dictionaries, only words with a Tesseract confidence score over a user-set threshold (meaning that their recognition is likely accurate) were inserted in the dictionary. Furthermore, in order to make its creation more robust, the dictionary for a specific legislature is merged with those chronologically adjacent, meaning that dictionaries contained words from both its legislature of origin and a user-selected window of adjacent legislatures (for instance, a span of 7 legislatures mean the dictionary having on average a span of around 35 years). Figure 1 shows how the windowed dictionary system works. In theory, this allowed SymSpell to have access to both more domain specific and historically realistic lexicon in the dictionaries, instead of the Italian dictionary that comes out-of-the-box with the software.

By looking at the error made by SymSpell, it seems that most of the problems belong to proper names (such

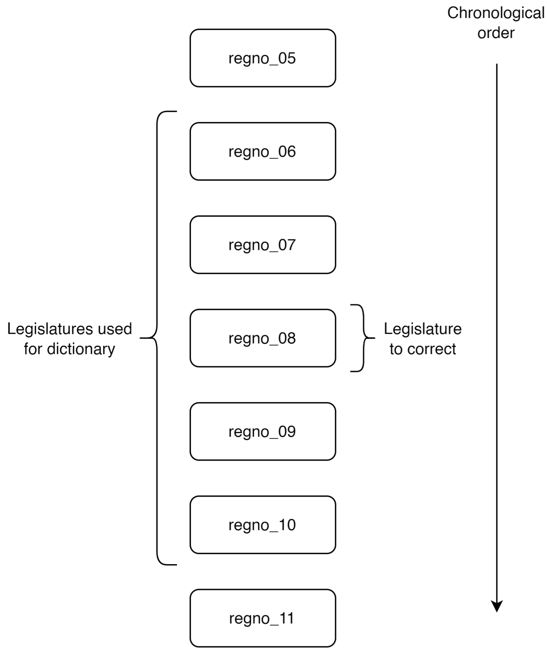


Figure 1: Example of how the dictionary used to correct the VIII legislature of the Kingdom of Italy would be constructed, with the parameter window set to 5.

[1]https://github.com/wolfgarbe/SymSpell

Correction method	CER	WER
Original	0.030	0.071
SymSpell	0.036	0.121
Windowed	0.033	0.102
Windowed lower cased	0.031	0.087

Table 1
Mean CER and WER against the test set (the lower, the better).

as persons and geographical entities), that often are not included into the dictionary and are replaced by existing words very close to the apparently-misspelled term.

We then compare four different approaches: OCR plain output from Tesseract, SymSpell with the original dictionary, Symspell with the windowed dictionary, Symspell with the windowed dictionary applied only to lower-cased words.

Table 1 shows the results of the four configurations. The CER and WER value calculated without applying SymSpell are lower than the other ones, resulting in a more accurate extraction. However, the use of the custom frequency list and the removal of proper nouns seems promising when compared to SymSpell applied with the original model.

By looking at the data, we can infer some useful insights. First of all, the raw text returned by Tesseract is already very precise: the Italian documents are written in a very clear font, and the digitalization has been done at a good level. The errors show that SymSpell replaced right words with wrong ones in case of proper names and very technical words, as expected.

In this first release, then, we will not use any spelling correction software, and provide the raw text extracted by Tesseract.

4. Dataset statistics

Table 2 shows some statistics of the dataset. In particular, for each legislature, one can see the number of words, pages and documents. In recent legislatures (since 1996) data is published in HTML format on the web, therefore the number of pages is not available.

5. Release

Both the data and the scripts (written in Python) are free to use and released on Github.[2]

The data contained in the *Camera dei Deputati* and *Senato della Repubblica* websites is released under the Creative Commons Attribution 3.0.[3] We use the same policy and distribute the text data under the same license.

[2] https://github.com/valefras/Italian_Parliament_Symspell
[3] https://creativecommons.org/licenses/by/3.0/

6. Conclusion and Future Work

In this paper we describe a preliminary version of the Italian Parliamentary Corpus, containing the Italian Parliament debates since 1848. In total, around 1.2 billion words have been collected.

In the future, we will further investigate OCR post-correction solutions to get cleaner data. We will also complete the data collection, by downloading and processing attachments to the parliamentary sessions, bulletins, law proposals, and reports of the Standing Committees, already available on the Italian Parliament houses websites.

We are also planning to assign each speech to the corresponding politician, and release the dataset so that anyone can use the tagging to make comparative and social studies.

References

[1] J. E. Cheng, Islamophobia, muslimophobia or racism? parliamentary discourses on islam and muslims in debates on the minaret ban in switzerland, Discourse & Society 26 (2015) 562–586.

[2] A. Paoletti, La presenza femminile nelle assemblee parlamentari: Per un'analisi comparata, Il Politico 56 (1991) 77–96.

[3] P. Bayley, Cross-cultural perspectives on parliamentary discourse, Cross-Cultural Perspectives on Parliamentary Discourse (2004) 1–390.

[4] G. Abrami, M. Bagci, L. Hammerla, A. Mehler, German parliamentary corpus (gerparcor), in: Proceedings of the Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 1900–1906.

[5] A. Pancur, T. Erjavec, The siParl corpus of Slovene parliamentary proceedings, in: Proceedings of the Second ParlaCLARIN Workshop, European Language Resources Association, Marseille, France, 2020, pp. 28–34.

[6] M. Marx, A. Schuth, DutchParl. the parliamentary documents in Dutch, in: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta, 2010.

[7] M. Ogrodniczuk, Polish Parliamentary Corpus, in: D. Fišer, M. Eskevich, F. de Jong (Eds.), Proceedings of the LREC 2018 Workshop *ParlaCLARIN: Creating and Using Parliamentary Corpora*, European Language Resources Association (ELRA), Paris, France, 2018, pp. 15–19.

[8] P. Koehn, Europarl: A parallel corpus for statistical machine translation, in: Proceedings of Machine Translation Summit X: Papers, Phuket, Thailand, 2005, pp. 79–86.

Table 2
Statistics of the dataset.

Legislature		Camera dei Deputati Words	Pages	Docs	Senato della Repubblica (del Regno) Words	Pages	Docs
Regno 01	8 May 1848 - 30 Dec 1848	1,479,030	1,365	125	345,124	329	47
Regno 02	1 Feb 1849 - 30 Mar 1849	661,745	628	59	123,546	123	22
Regno 03	30 Jul 1849 - 20 Nov 1849	1,444,180	1,344	87	345,144	331	36
Regno 04	20 Dec 1849 - 20 Nov 1853	13,028,979	11,841	691	2,891,139	2,821	319
Regno 05	19 Dec 1853 - 25 Oct 1857	9,357,294	8,702	496	1,700,528	1,758	196
Regno 06	14 Dec 1857 - 21 Jan 1860	3,059,049	3,324	180	473,000	555	63
Regno 07	2 Apr 1860 - 17 Dec 1860	1,244,804	1,159	76	261,535	304	36
Regno 08	18 Feb 1861 - 7 Sep 1865	18,131,690	19,286	809	5,026,679	5,919	451
Regno 09	18 Nov 1865 - 13 Feb 1867	3,320,700	3,801	161	532,399	612	54
Regno 10	22 Mar 1867 - 2 Nov 1870	13,281,102	15,291	603	2,656,412	3,131	229
Regno 11	5 Dec 1870 - 20 Sep 1874	12,673,461	14,827	566	3,932,011	5,305	273
Regno 12	23 Nov 1874 - 3 Oct 1876	5,895,113	7,518	245	2,247,260	3,235	135
Regno 13	20 Nov 1876 - 2 May 1880	12,637,227	16,246	530	3,646,076	5,427	272
Regno 14	26 May 1880 - 2 Oct 1882	9,465,573	12,102	396	2,030,825	3,286	150
Regno 15	22 Nov 1882 - 27 Apr 1886	13,980,213	18,326	586	2,973,795	4,737	212
Regno 16	10 Jun 1886 - 22 Oct 1890	14,748,962	19,784	633	4,672,145	7,358	314
Regno 17	10 Dec 1890 - 27 Sep 1892	6,292,409	8,633	246	1,998,816	3,116	124
Regno 18	23 Nov 1892 - 8 May 1895	8,159,069	11,820	321	2,339,573	3,751	149
Regno 19	10 Jun 1895 - 2 Mar 1897	6,092,924	8,794	233	1,979,404	3,131	125
Regno 20	5 Apr 1897 - 17 May 1900	10,369,144	14,942	432	3,460,008	5,427	247
Regno 21	16 Jun 1900 - 18 Oct 1904	16,167,079	22,135	594	4,833,899	7,842	337
Regno 22	30 Nov 1904 - 8 Feb 1909	17,480,096	25,020	574	5,774,328	9,804	287
Regno 23	24 Mar 1909 - 29 Sep 1913	19,179,900	26,890	588	6,706,423	11,711	337
Regno 24	27 Nov 1913 - 29 Sep 1919	15,438,966	21,444	394	3,184,990	5,154	201
Regno 25	1 Dec 1919 - 7 Apr 1921	6,964,613	9,728	194	2,477,546	3,609	123
Regno 26	11 Jun 1921 - 25 Jan 1924	8,050,058	11,150	243	3,599,413	5,695	173
Regno 27	24 May 1924 - 21 Jan 1929	6,770,726	9,778	246	5,909,658	11,330	216
Regno 28	20 Apr 1929 - 19 Jan 1934	6,562,034	9,616	239	4,084,703	7,001	208
Regno 29	28 Apr 1934 - 2 Mar 1939	4,017,010	5,628	150	3,174,981	4,460	138
Regno 30	23 Mar 1939 - 5 Aug 1943	424,944	626	28	350,342	562	23
Consulta Nazionale	25 Sep 1945 - 1 Jun 1946	732,609	1,012	44			
Assemblea Costituente	25 Jun 1946 - 31 Jan 1948	9,377,227	12,866	621			
Repubblica 01	8 May 1948 - 24 Jun 1953	32,044,382	42,385	1,114	26,357,526	39,282	984
Repubblica 02	25 Jun 1953 - 11 Jun 1958	27,444,116	36,615	738	17,231,858	26,559	653
Repubblica 03	12 Jun 1958 - 15 May 1963	27,924,486	37,418	789	19,417,065	31,667	697
Repubblica 04	16 May 1963 - 4 Jun 1968	33,501,680	45,096	844	27,874,392	45,368	804
Repubblica 05	5 Jun 1968 - 24 May 1972	24,326,333	33,912	549	18,257,542	29,670	597
Repubblica 06	25 May 1972 - 4 Jul 1976	19,351,600	27,820	483	15,978,505	25,972	572
Repubblica 07	5 Jul 1976 - 19 Jun 1979	17,601,367	28,463	418	10,195,047	16,975	395
Repubblica 08	20 Jun 1979 - 11 Jul 1983	35,750,707	63,837	674	17,913,360	30,014	617
Repubblica 09	12 Jul 1983 - 1 Jul 1987	29,672,360	55,945	639	17,320,038	29,961	597
Repubblica 10	2 Jul 1987 - 22 Apr 1992	42,453,808	96,131	769	22,650,712	52,571	665
Repubblica 11	23 Apr 1992 - 14 Apr 1994	11,642,821	22,920	312	11,328,792	28,313	287
Repubblica 12	15 Apr 1994 - 8 May 1996	9,778,180	20,029	326	12,897,363	30,619	310
Repubblica 13	9 May 1996 - 29 May 2001	34,825,006		878	27,613,322		1,059
Repubblica 14	30 May 2001 - 27 Apr 2006	33,929,341		757	39,427,915		964
Repubblica 15	28 Apr 2006 - 28 Apr 2008	11,492,227		278	13,066,129		283
Repubblica 16	29 Apr 2008 - 14 Mar 2013	28,685,376		739	41,579,478		859
Repubblica 17	15 Mar 2013 - 22 Mar 2018	29,946,098		863	48,645,737		923
Total		726,857,818		22,560	471,486,483		16,763

[9] A. Kay, Tesseract: An open-source optical character recognition engine, Linux J. 2007 (2007) 2.

[10] V. I. Levenshtein, Binary codes capable of correcting deletions, insertions and reversals., Soviet Physics Doklady 10 (1966) 707–710. Doklady Akademii Nauk SSSR, V163 No4 845-848 1965.

[11] S. Schulz, J. Kuhn, Multi-modular domain-tailored OCR post-correction, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 2716–2726. doi:10.18653/v1/D17-1288.

Extracting an expectation-based lexicon for UD treebanks

Matteo Gay and Cristiano Chesi

IUSS, P.zza Vittoria 15, Pavia, 27100, Italy

Abstract

In this paper, we present a refinement of the deterministic lexicon extraction algorithm discussed in [1]. This new version is able to infer categorial expectations from any University Dependency (UD) treebank [2] and recast non-trivial backward dependencies in terms of movement operations [3,4]. The improvements with respect to a preliminary version of the algorithm are related to a more granular definition of the functional categories and a significant reduction in lexical ambiguity. These results might constitute a substantial baseline for (very) large language models assessment in terms of descriptive adequacy [5].

Keywords

universal dependencies, lexicon, minimalist grammars, parsing, lexical and structural ambiguity

1. Introduction

The advent of large language models (LLMs), especially those based on pre-trained transformers (GPT-like, [6]) demonstrated the impressive capability of next-token prediction task in training. This prediction is largely based on expectations of various kinds (morphosyntactic, semantic, and inferential) that, in virtue of a well-tuned attention mask mechanism can generalize categorial constraints across word sequences, apparently, without significant distance or directionality limitations among tokens, though the number of tokens accepted in input at each iteration remains fixed. This represents a clear difference with respect to recurrent networks training (e.g., LSTM, [7]) which benefits from a more cognitively plausible working memory mechanism that tries to preserve a sound incremental processing also modeling "backward dependencies" (i.e. when β depends on α, but β linearly precedes α). It has been demonstrated however that these new generation attention-based transformers are computationally very intensive [8] and their performance on various linguistic datasets demonstrates the implausibly of their linguistic generalizations [9]. In a recent study, Wilcox, Futrell & Levy [10] demonstrate the impressive ability of certain recurrent networks (JRNN and GRNN) to model islands constraints in a way that is comparable (or even better, sometimes) to that of larger LMs such as GTP-3. These results, apparently, not only challenge the poverty of stimulus hypothesis [5], but also suggest that, as far as linguistic competence is concerned, models with a smaller number of parameters (and a different architecture) might be more efficient in learning a human-like linguistic competence. What is then the minimum model size we might afford to maintain this high level of meta-linguistic competence?

In this paper, to address the "minimum model size problem", we stress the relevance of a cognitively plausible expectation mechanism (model architecture) that benefits from explicit modeling of (at least) certain categorial constraints (minimum number of parameters needed).

Expectation-based Minimalist Grammars (e-MGs, [11]), for instance, simply rely on the notion of satisfied local selection (through categorial features) to identify a successful local dependency or the requirement to be satisfied through a non-local one. The advantages are (i) the complete transparency of this model compared to a cognitively plausible grammatical theory [12] and (ii) the computational complexity cost in recognition which grows polinomially to the length of the sentence (unless specific parameterization to extend empirical coverage is adopted, [1]). A possible disadvantage of this approach is related to the fact that a fully explicit lexicon is required to run the recognition or the generation algorithm. Based on [13], this work extends the proposed algorithm to retrieve a large-scale lexicon from annotated datasets such as UD treebanks. In Section 2, we will introduce the critical aspects of this procedure, essentially focusing on the directionality of each dependency and on the inherent lexical ambiguity obtained by applying a naïve lexicon extraction procedure. In Section 3, we propose a refinement of the extraction approach, and we will calculate the efficiency of this new method in terms of reduction of lexical and syntactic ambiguity and, overall, lexicon size. Section 4 will conclude this paper by suggesting further improvements to be evaluated in the future.

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

matteo.gay@iusspavia.it (M. Gay); cristiano.chesi@iusspavia.it (C. Chesi)

0009-0003-6066-4036 (M. Gay); 0000-0003-1935-1348 (C. Chesi)

2. Categorial selection in UD

Categorial morphosyntactic selection is the standard constraint used in (e-)MGs to deal both with local (“[the dog]”) and non-local (“[what] did John [eat $_{what}$]?”) dependencies through feature-matching [11,14]; in Minimalist Grammars, both kinds of dependencies are expressed by destructive feature checking operations targeting specific features (*select* and *base* for local dependencies; *licensee* and *licensor* for non-local ones), while only local dependencies are feature-destructive in e-MGs. In both formalisms, we dub “Merge” the syntactic operation establishing a local dependency as informally expressed in **Errore. L'origine riferimento non è stata trovata.** (“X” is a categorial morphosyntactic feature expressed by β and selected/expected by α, that is $\alpha_{=expect}\,{}_{expected}\beta$):

(1) Merge ($\alpha_{=X}$, ${}_{X}\beta$) = [$\alpha_{\cancel{=X}}$ [${}_{\cancel{X}}\beta$]]

When a non-local dependency must be established (i.e., a link between words/tokens spanning over other tokens/words, as in most wh- dependencies in languages like English or Italian), e-MGs postulate that an unsatisfied local Merge operation forces the unsatisfied categorial expectation storage into a memory buffer. This keeps the unexpected item’s features in the computation and forces its re-Merge as soon as an appropriate categorial expectation is processed, as informally expressed in (2):

(2) i. MERGE ($\alpha_{=X}$, ${}_{X\,Y}\beta$) = [$\alpha_{\cancel{=X}}$ [${}_{\cancel{X}\,Y}\beta$]]
ii. MOVE ([${}_{\cancel{X}\,Y}\beta$]) ⇒ memory([${}_{Y}(\beta)$])
iii. MERGE ($\gamma_{=Y}$, memory([${}_{Y}(\beta)$]) =
[$\alpha_{=X}$ [${}_{X\,Y}\beta$] … $\gamma_{\cancel{=Y}}$ [${}_{\cancel{Y}}(\beta)$]]

The advantage of this approach is that all non-local dependencies (of the “movement” type) are (i) forward, (ii) increase monotonically the complexity of the computation and (iii) recast syntactic ambiguity (attachment) at the lexical level. One relatively simple algorithm to extract a general-purpose lexicon from annotated UD treebanks [15] is described in [1] and can be summarized as follows.

Assuming that α is a distinct token in UD:

(i) for each distinct UPOS associated with at least one occurrence of α in UD, add a lexical entry α and associate the UPOS string to the category to be expected to license a Merge operation (*expected* feature);
(ii) for each node dependent on α, add its UPOS feature as an expectation of α (*expect* feature);
(iii) repeat (ii), and duplicate the α lexical entry, for any occurrence of α that introduces different dependents (i.e., if one occurrence of α has just one dependent X and another occurrence of α has also one extra dependent Y, then add both $\alpha_{=X}$ and $\alpha_{=X\,=Y}$ to the lexicon).

The procedures guarantee that lexical ambiguity (POS ambiguity, i.e., ${}_{X}\alpha$ and ${}_{Y}\alpha$ are both in the lexicon, α is the same token, and X and Y are two distinct UPOSs) is preserved (step i.) and that syntactic ambiguity (node attachment) is represented directly in the lexicon (step iii.). The advantage of this approach is that it makes transparent the level of morphosyntactic ambiguities a parser should deal with, distinguishing between local and non-local dependencies as well as backward and forward dependencies: each time a dependent is resolved within an adjacent item, the dependency is considered “trivial” and can be readily solved by the recognition (or the generation) algorithm discussed in [11]. If the dependency is local and “forward” (i.e. α → β, where α immediately precedes β in the text), then MERGE ($\alpha_{=X}$, ${}_{X}\beta$) = [$\alpha_{\cancel{=X}}$ [${}_{\cancel{X}}\beta$]]; if the local dependency is “backward” (i.e. β → α, where α immediately follows β in the text), then MOVE trivially applies (as in (2)), deriving in two steps [${}_{X}\beta$ [$\alpha_{\cancel{=X}}$ [${}_{\cancel{X}}(\beta)$]].

A preliminary extraction experiment following this procedure was implemented on four treebanks adopting the UPOS tagset, two for “head initial” languages [16], UD English GUM for English [17], and UD Italian ISDT [18], two for “head-final” languages, namely the UD Turkish PENN treebank [19], and the UD Japanese GSD treebank [20,21]. Although, as noticed by an anonymous reviewer, there is not yet consensus on the criteria to assign UPOS tags and this might surely affect our extraction results, UD treebanks still represent the most reliable repositories to run an automatic explorative study as the one we conducted here. The results of this extraction experiment are reported in Table 1:

Table 1
Ambiguity ratio and dependency locality in the extracted lexica English (EN), Italian (IT), Turkish (TR), and Japanese (JP)

	EN	*IT*	*TR*	*JP*
Tokens	126530	294403	166514	168333
Lexicon size (Types)	13757	27021	33036	20140
Ambiguity ratio	0.38	0.28	0.34	0.33
Lexical	*0.33*	*0.22*	*0.20*	*0.25*
Morpho.	*0.17*	*0.03*	*0.08*	*< 0.01*
Depend.	*0.50*	*0.75*	*0.72*	*0.75*
Backward depend. ratio	0.69	0.61	0.83	0.48
Locality ratio	*0.54*	*0.60*	*0.69*	*0.32*

The estimated levels of ambiguity (Lexical, i.e., POS-related; Morphological, i.e., related to a morphosyntactic featural specification such as agreement features; dependency-based, i.e., variance in terms of number and kind of dependents), as well as the amount of "backward dependencies" (i.e., those triggering movement in e-MGs), are indicated with a specification of the locality of this last kind of dependencies. We concluded that a critical issue was related to the number of “non-trivial” (i.e., non-local) backward dependencies that, especially in English, constitute a robust 15% of the lexical entries. The exploration is however incomplete at least for the following three reasons: (i) UPOS tags are relatively poor and underspecified, especially for those functional items that behave in a very different syntactic way (e.g. articles and quantifiers); (ii) the directionality of the dependencies assumed in UD is the reverse of the one presumed in other generative approaches: e.g. the determiner selects the noun in (e-

)MGs, that is, NOUN depends on DET and not the way around (this is critical, since, as predicted by e-MGs, an argument is not licensed in its thematic position unless properly determined); (iii) empty elements and multiple dependencies are absent from UD. These issues have been preliminarily addressed in [13] and expanded in the following Section.

3. A better extraction algorithm

Following the critics addressed in [22], we first inverted the original directionality of the UD dependencies between lexical and functional items, then we decorated the items extracted with specific categories and added them to the lexicon according to the following procedure in five steps:

(i) When, in UD, a functional category (e.g., a preposition, annotated as ADP) is locally dependent on a lexical one (e.g., a VERB, as in Figure 1), the first (ADP) becomes the expected feature of the functional category and the second the expectation (i.e., ADV *per* =VERB in e-MG format).

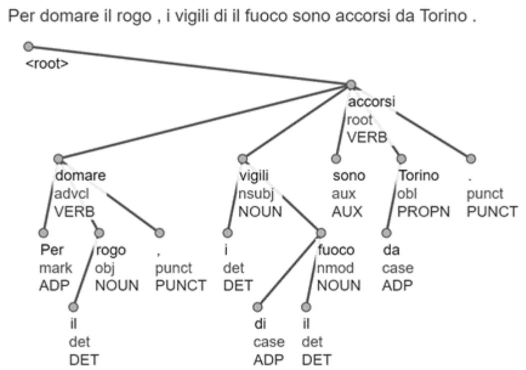


Figure 1: Example of annotation extracted from ISDT [18]

When in UD a functional category is not locally dependent on a lexical one (e.g. “del fuoco”, that is “di il fuoco”, “of the fire”), we preserve linear order and transform the dependencies as a left-right cascade of expectations (i.e. ADP *di* =DET, DET *il* =NOUN, NOUN *fuoco*); this way, many instances of spurious non-local dependencies are all reduced to local dependencies.

(iii) Because of the finiteness of functional categories [23–25], their universal ordering [26,27] and relative (monotonic) optionality [1] (if C depends on B which depends on A, C can directly follows A), we refine each UPOS category and modify the directionality and nature of the dependency accordingly (including phonetically empty lexical items such as null complementizers);

(iv) Phonetically empty pronominal elements (pro, PROs, [28]) are associated, and merged as empty items, directly with the verbal inflection through parameterization (in Italian, finite inflection can license an empty subject, but not in English);

(v) Unselected adjuncts (e.g., locative modifications, relative clauses) are not supposed to be “expected”, so, no select feature is added to the lexical categories they modify (i.e. an optional modification is not included in the lexicon, but predicted by the e-MG implementation of Merge: in case of a relative clause modification, for instance, a raising analysis is implemented, triggered by the relative complementizer/pronoun that behaves as a DP/PP within the relative clause).

The lexicon extracted from the sentence in Figure 1 is then the following one (the format used is expected α =expect; agreement features and morphemic decomposition are ignored for the sake of compactness):

(3) e-MG extracted Lexicon = { ADP.MODE *per* =VERB, VERB *domare* =DET, DET.DEF *il* =NOUN, NOUN *rogo*, DET.DEF *i* =NOUN, NOUN *vigili*, ADP.ARG *di* =DET, NOUN *fuoco*, AUX.BE *sono* =VERB.UNACC, VERB.UNACC *accorsi* =DET, ADP.LOCATIVE.FROM *da* =DET, DET *Torino*}

3.1. Results

The application of the algorithm to ISDT treebank produces a significant reduction of the number of non-local dependencies (-58%) as well as the overall number of non-local critical dependencies (inducing structural ambiguities, -36%) compared to the first version discussed in [1].

The results of our experiment are reported in Table 2 (sentences including token with UPOS 'INTJ', 'X', 'SYM', 'NUM' are removed):

Table 2
Ambiguity ratio and dependency locality using the revised extraction procedure in Italian (ISDT treebank)

	ISDT (original)	ISDT (revised)
Sentences processed	10616	=
Tokens/Types	193062/ 22709	=
Type/Token Ratio	0,118	=
Backward Dependencies (BD)	107382	45157
Local BD	58317	13669
Locality ratio in BD	0.543	0.303
Final Lexicon	24809	=

4. Discussion

These preliminary results are encouraging and demonstrate, on the one hand, that structural ambiguity can be reasonably reduced by assuming properly crafted linguistic generalizations, on the other hand, valency-based theories [29] are less accurate and induce an unwanted level of ambiguity that can be readily avoided.

More work needs to be done at least in the following direction: lexical items are considered as atomic entities. As demonstrated by the efficacy of word embeddings based on sub-words units (notably the byte-pair encoding, [30,31]) this step is more than necessary (and doable in the current e-MG framework, see [32]).

Back to the original consideration on how these results can serve as a baseline for LLM assessment, there is at least one possible road we consider: if explicit categorial expectations are implicitly represented in LLM, then the number of features needed in the lexicon should be considered as the minimal dimensionality needed for word embedding: as far as structural constraints are considered, this should represent the lowest number of parameters (i.e. levels of abstraction/representation) of the lexical item. Although no lexical semantic consideration is addressed here (even though the categorial approach is, in principle, tenable also from the semantic compositional perspective), specific generalization should be obtained both with e-MGs and with LLM of comparable size in terms of parameters (i.e., roughly corresponding to the number of categories in the expect(ed) required by the e-MGs lexicon). The phenomena to be tested should include islands violations (e.g. "*what did John read the book that was talking about?"), thematic selection constraints (e.g. "*John put", "*John eats a sandwich to Mary"), etc. (on the line, for instance, of [33] or [9]).

If more parameters (where "more" corresponds to at least one order of magnitude) are needed to perform similarly to e-MGs on these constraints, we might conclude that LLMs still need some "optimization" since they do not qualify (yet) as efficiently, and descriptively adequate.

Acknowledgements

This project is partially founded by BEXT (New Behavioral EXperimental approaches to complexity perception and stress assessment in linguistic and cognitive research) MIUR PRO3 2021-2023 and T-GRA2L (Testing GRAdeness and GRAmmaticality in Linguistics) PRIN 2022 (code: 202223PL4N) Projects.

References

[1] C. Chesi, Parameters of cross-linguistic variation in expectation-based Minimalist Grammars (e-MGs), IJCoL. 9 (2023) 21.

[2] J. Nivre, Ž. Agić, L. Ahrenberg, L. Antonsen, M.J. Aranzabe, M. Asahara, L. Ateyah, M. Attia, A. Atutxa, L. Augustinus, others, Universal Dependencies 2.1, (2017).

[3] E. Stabler, Two Models of Minimalist, Incremental Syntactic Analysis, Top Cogn Sci. 5 (2013) 611–633. https://doi.org/10.1111/tops.12031.

[4] N. Chomsky, The minimalist program, MIT press, Cambridge, MA, 1995.

[5] N. Chomsky, Aspects of the Theory of Syntax, MIT press, 1965.

[6] T.B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D.M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language Models are Few-Shot Learners, arXiv:2005.14165 [Cs]. (2020). http://arxiv.org/abs/2005.14165 (accessed April 21, 2021).

[7] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Computation. 9 (1997) 1735–1780.

[8] J. Vanian, K. Leswing, ChatGPT and generative AI are booming, but the costs can be extraordinary, CNBC News. (2023).

[9] C. Chesi, F. Vespignani, R. Zamparelli, Modelli generativi e sintassi generativa, Sistemi Intelligenti. 2 (2023).

[10] E.G. Wilcox, R. Futrell, R. Levy, Using Computational Models to Test Syntactic Learnability, Linguistic Inquiry. (2023) 1–44. https://doi.org/10.1162/ling_a_00491.

[11] C. Chesi, Expectation-based Minimalist Grammars, arXiv:2109.13871 [Cs]. (2021). http://arxiv.org/abs/2109.13871 (accessed November 2, 2021).

[12] N. Chomsky, Three factors in language design, Linguistic Inquiry. 36 (2005) 1–22.

[13] M. Gay, From Universal Dependencies to Expectation-Based Minimalist Grammars: automatic lexicon extraction and ambiguity issues., Bachelor, IUSS, 2023.

[14] E. Stabler, Computational Perspectives on Minimalism, in: C. Boeckx (Ed.), The Oxford Handbook of Linguistic Minimalism, Oxford University Press, 2011. https://doi.org/10.1093/oxfordhb/9780199549368.013.0027.

[15] M.-C. de Marneffe, C.D. Manning, J. Nivre, D. Zeman, Universal Dependencies, Computational Linguistics. 47 (2021) 255–308. https://doi.org/10.1162/coli_a_00402.

[16] M.C. Baker, The atoms of language, 1st ed, Basic Books, New York, 2001.

[17] A. Zeldes, The GUM Corpus: Creating Multilayer Resources in the Classroom, Language Resources and Evaluation. 51 (2017) 581–612. http://dx.doi.org/10.1007/s10579-016-9343-x.

[18] C. Bosco, F. Dell'Orletta, S. Montemagni, The Evalita 2014 Dependency Parsing Task, in: Proceedings of the First Italian Conference on Computational Linguistics CLiC-It 2014 and of the Fourth International Workshop EVALITA 2014 9-11 December 2014, Pisa, pisa university press, 2014. https://doi.org/10.12871/clicit201421.

[19] K. Oflazer, B. Say, D.Z. Hakkani-Tür, G. Tür, Building a Turkish Treebank, in: A. Abeillé (Ed.), Treebanks, Springer Netherlands, Dordrecht, 2003: pp. 261–277. https://doi.org/10.1007/978-94-010-0201-1_15.

[20] T. Tanaka, Y. Miyao, M. Asahara, S. Uematsu, H. Kanayama, S. Mori, Y. Matsumoto, Universal Dependencies for Japanese, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), European Language Resources Association (ELRA),

Portorož, Slovenia, 2016: pp. 1651–1658. https://aclanthology.org/L16-1261.

[21] M. Asahara, H. Kanayama, T. Tanaka, Y. Miyao, S. Uematsu, S. Mori, Y. Matsumoto, M. Omura, Y. Murawaki, Universal Dependencies Version 2 for Japanese, in: N.C. (Conference chair), K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, T. Tokunaga (Eds.), Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan, 2018.

[22] T. Osborne, K. Gerdes, The status of function words in dependency grammar: A critique of Universal Dependencies (UD), Glossa: A Journal of General Linguistics. 4 (2019). https://doi.org/10.5334/gjgl.537.

[23] G. Cinque, ed., Functional structure in DP and IP, Oxford University Press, Oxford ; New York, 2002.

[24] L. Rizzi, ed., The structure of CP and IP, Oxford University Press, Oxford ; New York, 2004.

[25] A. Belletti, Structures and Beyond: The Cartography of Syntactic Structures, Volume 3, Oxford University Press, 2004.

[26] G. Cinque, Deriving Greenberg's Universal 20 and Its Exceptions, Linguistic Inquiry. 36 (2005) 315–332. https://doi.org/10.1162/0024389054396917.

[27] G. Cinque, Adverbs and functional heads: A cross-linguistic perspective, Oxford University Press, Oxford (UK), 1999.

[28] L. Rizzi, Null objects in Italian and the theory of pro, Linguistic Inquiry. 17 (1986) 501–557.

[29] L. Tesnière, Elements of structural syntax, John Benjamins Publishing Company, Amsterdam ; Philadelphia, 2015.

[30] P. Gage, A new algorithm for data compression, C Users Journal. 12 (1994) 23–38.

[31] R. Sennrich, B. Haddow, A. Birch, Neural Machine Translation of Rare Words with Subword Units, (2016). http://arxiv.org/abs/1508.07909 (accessed July 24, 2023).

[32] G.M. Kobele, Minimalist Grammars and Decomposition, in: G. Kleanthes K., E. Leivada (Eds.), The Cambridge Handbook of Minimalism, Cambridge University Press, Cambridge (UK), 2023.

[33] A. Warstadt, A. Singh, S.R. Bowman, Neural Network Acceptability Judgments, arXiv Preprint arXiv:1805.12471. (2018).

An Analysis of Visually Grounded Instructions in Embodied AI Tasks

Marco Grazioso[1,2,*,†], Alessandro Suglia[3,4,†]

[1]Interdepartmental Research Center Urban/Eco, University of Naples Federico II, Italy
[2]Logogramma s.r.l, Italy
[3]Heriot-Watt University, Scotland, UK
[4]AlanaAI, Edinburgh, United Kingdom

Abstract

Thanks to Deep Learning models able to learn from Internet-scale corpora, we observed tremendous advances in both text-only and multi-modal tasks such as question answering and image captioning. However, real-world tasks require agents that are embodied in the environment and can collaborate with humans by following language instructions. In this work, we focus on ALFRED, a large-scale instruction-following dataset proposed to develop artificial agents that can execute both navigation and manipulation actions in 3D simulated environments. We present a new Natural Language Understanding component for Embodied Agents as well as an in-depth error analysis of the model failures for this challenge, going beyond the success-rate performance that has been driving progress on this benchmark. Furthermore, we provide the research community with important directions for future work in this field which are essential to develop collaborative embodied agents.

Keywords

embodied AI, situated interaction, visual grounding, deep learning

1. Introduction

In recent years, we experienced tremendous improvements in Natural Language Understanding (NLU) tasks thanks to powerful Large Language Models (e.g., [1, 2, 3]). These models are trained by leveraging internet-scale textual data. However, by having access to text only, they leverage only a part of the rich multi-modal training data that can be derived from interaction with the world and with other agents [4]. Embodied Artificial Intelligence (EAI) is the field of AI that aims at developing agents that can perceive the environment via multi-modal inputs, and that can execute actions in the world.

Many benchmarks have been proposed so far in EAI. For instance, Vision+Language Navigation [5] aims at studying the capabilities of EAI agents to follow natural language instruction in 3D simulated environments. However, the agent can only output navigation actions limiting the richness of concepts that the agent can learn. To simulate a scenario that is closer to the real-world usage of these systems, Shridhar et al. [6] proposes ALFRED, a new instruction-following benchmark that facilitates the study of both situated language understanding as well as visual memory, commonsense reasoning, as well as long-term action planning.

So far, progress on ALFRED has been driven by accuracy-based metrics on the official leaderboard (e.g., [7, 8, 9, 10]). However, considering that the success rate on this benchmark is still below production-level performance (∼40%), this calls for a more in-depth analysis of model failures. In this paper, we provide two main contributions: 1) we train a novel Natural Language Understanding component for an EAI agent trained using multi-task learning that has a 0.117 error rate on the validation unseen of ALFRED, an improvement over the one proposed by Min et al.[7]; 2) we provide an in-depth analysis of our model's failures highlighting lack of important situated language understanding capabilities that are key for an EAI agent such as referential expression resolution, and conversational grounding [11].

2. The ALFRED dataset

In this study, we use ALFRED [6], a benchmark aimed at assessing the ability of embodied agents to learn from natural language instructions and egocentric vision to generate sequences of actions for household tasks. The ALFRED dataset comprises 25,743 human-annotated language directives corresponding to 8,055 expert demonstration episodes. Each directive includes a high-level goal and a set of step-by-step instructions. Directives fall under one of the following seven tasks parameterised

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

*Corresponding author.
†These authors contributed equally.
marco.grazioso@unina.it (M. Grazioso); a.suglia@hw.ac.uk (A. Suglia)
0000-0002-4056-544X (M. Grazioso); 0000-0002-3177-5197 (A. Suglia)

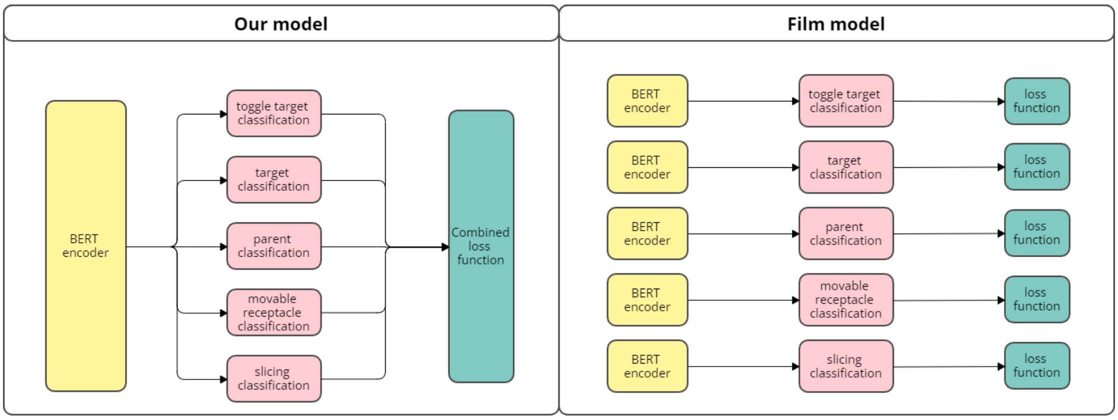


Figure 1: Comparison between our model and the one proposed in the FILM paper.

by 84 object classes in 120 scenes: pick and place, stack and place, pick two and place, clean and place, heat and place, cool and place, and examine in light. In addition to the task category, for each instruction, the dataset also provides a few relevant annotations: 1) **target object**: the principal object involved in the interaction; 2) **parent object**: the final destination of the target object (sink, counter, and similar); 3) **movable receptacle**: a movable object containing the target one, e.g. a spoon in a mug; 4) **slicing action**: true or false respectively if the target object must be cut or not; 5) **toggle target**, indicating an object to toggle on/off (oven, microwave, etc.).

To better estimate the models' capability to generalise to new environments, the validation set is composed of two subsets called seen and unseen, respectively. In the former, the agent has to complete tasks in rooms/scenarios that have already been seen during training, while the latter provides examples in unseen scenarios to assess the ability of the agent to generalise.

3. Baseline model and error analysis

We implemented the solution proposed in the FILM paper [7] which is the base approach for many of the state-of-the-art models for ALFRED (e.g., [8]). FILM is a modular architecture that is composed of a trained natural language component that derives a semantic representation represented in terms of intents and slot values akin to conventional NLU systems (e.g., [12]). This representation is in turn converted into an action plan by a rule-based component.

In this paper, we focus on improving the language understanding component of FILM, which is essential for instruction interpretation. Concretely, the FILM implementation casts instruction understanding as a classification task and fine-tunes one BERT-based model [13] for each of the following tasks: 1) **task classification**: the instruction sequence is classified into one of the seven task categories; 2) **target classification**: the instruction sequence is classified into one of the allowed target objects; 3) **movable receptacle classification**: the instruction sequence is classified into one of the allowed movable receptacle objects; 4) **parent classification**: the instruction sequence is classified into one of the allowed parent objects; 5) **slicing classification**: the instruction sequence is binary classified to be a slicing/non-slicing action.

However, training five different models each one having its own BERT encoder can be suboptimal because 1) it has a high computational cost, and 2) it does not consider the semantic relationships between a task and the objects required to solve that task. To take advantage of these relationships, we implemented a multi-task model by fine-tuning a single BERT encoder on all the tasks [14] (see Figure 1). As shown in Table 2, thanks to this multi-task setup, our model obtains an improvement in the overall language understanding performance measured using the error rate which we define as the proportion of examples for which no mistakes are made (i.e., neither on the high-level task nor on the single slots). Additionally, we report in Table 1, our model performance on specific high-level tasks measured using F1-score.

Despite the superior performance of our multitask model, the capabilities of this model were still limited. Therefore, we performed a manual error analysis based on 821 instructions from the validation unseen split of the ALFRED dataset. Particularly, as shown in Table 3, the most common errors are about a wrong object classification and, even if the object was correctly classified, a

Class	Ours F1-score	FILM F1-score
look_at_obj_in_light	0.994	1
pick_and_place_simple	0.985	0.91
pick_and_place_with_movable_recep	0.991	0.94
pick_clean_then_place_in_recep	0.991	0.98
pick_cool_then_place_in_recep	1	0.93
pick_heat_then_place_in_recep	0.996	0.97
pick_two_obj_and_place	0.988	0.9

Table 1
Comparison between our model scores and FILM's model scores in task classification on unseen validation set.

Model	Language processing error rate
FILM	0.196
Ours	**0.117**

Table 2
Comparison between our model error rate and FILM's model error rate on all language processing tasks on unseen validation set.

Error type	Subtype	Rate
Referential ambiguity	Mismatching	40/821
	Underspecification	24/821
	Others	32/821
Target object search	Object not visible	171/821
	Spatial understanding	106/821
Others interaction errors		218/821

Table 3
Error rate for each error type derived from our error analysis.

failure to find it in the environment. Specifically, we can categorise errors in two main classes namely REFERENTIAL AMBIGUITY, and TARGET OBJECT SEARCH, which we further divide into the following classes:

MISMATCHING OBJECT REFERENCE: the user refers to an object with a non-conventional name or a particular linguistic form due to visual ambiguity (*brown ball* or *potato* instead of *egg*).

UNDERSPECIFIED OBJECT REFERENCE: the user refers to an object using a name which could be ambiguous because not precise enough (typically "soap" is used to refer to a *soap bar* or a *soap bottle*).

OBJECT NOT FOUND BECAUSE NOT VISIBLE: this can happen when the target object is contained in other objects (e.g., spoons are contained in drawers).

SPATIAL UNDERSTANDING: the user gives nuanced spatial references for the object but the system does not understand them (e.g., *pick up the salt which is inside the cabinet under the coffee machine*).

Finally, we use a third class (OTHERS) which includes other interaction errors that do not depend on the language understanding component.

4. Challenges for embodied instruction following

Thanks to our error analysis, we derive that an embodied agent faces several challenges when fusing multiple modalities. Moreover, it must take care of the basic concepts of human-to-human communication [11].

In this context, the agent's reasoning can be seen as a sequential process in which it implements a set of strategies to follow the current language instruction. An embodied agent must rely on visual context, commonsense knowledge, and interactive skills. For instance, when the user asks for an object, e.g., soap, the agent must be able to understand that "soap", "soap bar" and "soap bottle" share enough features to define them as similar objects. Additionally, it should take advantage of the visual context to resolve ambiguities (if the only soap in the agent's field of view is a soap bar, this should be the target). Therefore, multi-modal information becomes crucial to understanding visually grounded instructions, going from spatial language instructions to multi-modal input ones [15]. Finally, if no other strategy resulted in a solution, it should ask for human intervention, e.g. using clarification strategies [16]. Furthermore, integrating commonsense knowledge can result in better interpretation (e.g., by leveraging knowledge graphs [17]) as well as better action plans by reasoning over pre-conditions and post-conditions of the actions.

In collaborative tasks [18, 19, 20], agents have to build common ground to successfully complete their tasks and adapt to new situations [11]. Therefore, negotiating meanings becomes a fundamental skill that allows the agent to learn how the user refers to the environment, and understand user preferences which will lead to a more effective interaction.

5. Conclusion

In this work, we used the ALFRED dataset as a benchmark to investigate the language understanding abilities of state-of-the-art EAI models. We started by improving the model originally proposed by Min et al. [7] by training using multi-task learning and we showed that even by using the new model several issues remain unsolved. We categorised these problems into different classes to facilitate our analysis. This classification led us to the conclusion that an EAI agent must leverage multi-modal signals, commonsense knowledge, and interaction with the user to solve embodied problems in an effective way.

According to Schlangen [21], *situated interaction is a direct, purposeful encounter of free and independent but similar agents.* Following this definition, in the ALFRED tasks there are two different agents: a follower and a leader. The leader is intended as an oracle that provides instructions in one go without conversing with the follower. Moreover, the leader assumes that the follower has perfect capabilities to follow the provided instructions without considering the notion of uncertainty or potential mistakes. Finally, there is no concept of conversational grounding intended as a joint activity in which the two agents have to negotiate meanings that are required to solve the task effectively and efficiently. In this sense, even if the ALFRED dataset still represents a challenging task, it is far from providing a benchmark that can be used to develop artificial agents able to collaboratively solve tasks using natural language.

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[2] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288 (2023).

[3] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, et al., Bloom: A 176b-parameter open-access multilingual language model, arXiv preprint arXiv:2211.05100 (2022).

[4] Y. Bisk, A. Holtzman, J. Thomason, J. Andreas, Y. Bengio, J. Chai, M. Lapata, A. Lazaridou, J. May, A. Nisnevich, et al., Experience grounds language, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 8718–8735.

[5] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, A. Van Den Hengel, Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 3674–3683.

[6] M. Shridhar, J. Thomason, D. Gordon, Y. Bisk, W. Han, R. Mottaghi, L. Zettlemoyer, D. Fox, ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020. URL: https://arxiv.org/abs/1912.01734.

[7] S. Y. Min, D. S. Chaplot, P. Ravikumar, Y. Bisk, R. Salakhutdinov, Film: Following instructions in language with modular methods, 2021. arXiv:2110.07342.

[8] Y. Inoue, H. Ohashi, Prompter: Utilizing large language model prompting for a data efficient embodied instruction following, arXiv preprint arXiv:2211.03267 (2022).

[9] A. Suglia, Q. Gao, J. Thomason, G. Thattai, G. Sukhatme, Embodied bert: A transformer model for embodied, language-guided visual task completion, arXiv preprint arXiv:2108.04927 (2021).

[10] A. Pashevich, C. Schmid, C. Sun, Episodic transformer for vision-and-language navigation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 15942–15952.

[11] H. H. Clark, S. E. Brennan, Grounding in communication., in: Perspectives on socially shared cognition., American Psychological Association, 1991, pp. 127–149.

[12] Q. Zhu, Z. Zhang, Y. Fang, X. Li, R. Takanobu, J. Li, B. Peng, J. Gao, X. Zhu, M. Huang, Convlab-2: An open-source toolkit for building, evaluating, and diagnosing dialogue systems, arXiv preprint arXiv:2002.04793 (2020).

[13] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[14] X. Liu, P. He, W. Chen, J. Gao, Multi-task deep neural networks for natural language understanding, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 4487–4496. URL: https://aclanthology.org/P19-1441. doi:10.18653/v1/P19-1441.

[15] M. Grazioso, A. S. Podda, S. Barra, F. Cutugno, Natural interaction with traffic control cameras through multimodal interfaces, in: International Conference on Human-Computer Interaction, Springer, 2021, pp. 501–515.

[16] V. Russo, A. Mancini, M. Grazioso, M. Di Bratto, Graph-based representations of clarification strategies supporting automatic dialogue management, IJCoL. Italian Journal of Computational Linguistics 8 (2022).

[17] A. Origlia, M. Di Bratto, M. Di Maro, S. Mennella, A multi-source graph representation of the movie domain for recommendation dialogues analysis, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 1297–1306.

[18] H. De Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, A. Courville, Guesswhat?! visual

object discovery through multi-modal dialogue, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5503–5512.
[19] N. Ilinykh, S. Zarrieß, D. Schlangen, Meet up! a corpus of joint activity dialogues in a visual environment, in: Proceedings of the 23rd Workshop on the Semantics and Pragmatics of Dialogue-Full Papers, 2019.
[20] A. Suhr, C. Yan, J. Schluger, S. Yu, H. Khader, M. Mouallem, I. Zhang, Y. Artzi, Executing instructions in situated collaborative interactions, 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), 2019, pp. 2119–2130.
[21] D. Schlangen, What a situated language-using agent must be able to do: A top-down analysis, arXiv preprint arXiv:2302.08590 (2023).

Exploring sentiments in summarization: SentiTextRank, an Emotional Variant of TextRank

Md. Murad Hossain[1,*,†], Luca Anselma[1,†] and Alessandro Mazzei[1,†]

[1] *Department of Computer Science, University of Turin, Corso Svizzera 185, 10149 Torino, Italy*

Abstract

English. A summary that aims at preserving the emotions of the original text can be interesting in certain application scenarios, such as in the generation of metareviews, both in academic and commercial domains. TextRank is a well-studied algorithm for automatic extractive summarization. This work introduces SentiTextRank, an emotional variant of TextRank, to enhance the extractive technique for both single-document and multi-document summarization. SentiTextRank incorporates emotions into the summarization process by classifying sentences into the eight emotional categories used in SenticNet. The preliminary evaluation of SentiTextRank yields encouraging results. In particular, our method generates informative summaries composed of sentences that preserve the emotional content of the original document.

Italian. Un riassunto che miri a preservare le emozioni del testo originale può essere interessante in alcuni scenari applicativi, come ad esempio nella generazione di meta-recensioni sian nel dominio accademico che in quello commerciale. TextRank è un algoritmo per il riassunto automatico estrattivo molto studiato. Questo lavoro introduce SentiTextRank, una variante emozionale di TextRank, per potenziare la tecnica estrattiva sia per il riassunto di singoli documenti che per il riassunto di documenti multipli: SentiTextRank integra le emozioni nel processo di sintesi, classificando le frasi nelle otto categorie emotive utilizzate in SenticNet. La valutazione preliminare di SentiTextRank produce dei risultati incoraggianti. In particolare, il nostro metodo produce dei riassunti informativi formati da frasi che rispettano il contenuto emozionale del documento originale.

Keywords

Extractive summarization, SentiTextRank, emotional variant, single and multi-document Summary, emotional content.

1. Introduction

Summarization is the process of reducing a larger body of information into a concise and coherent summary that captures the essential points and main ideas. Extractive summarization involves selecting and combining sentences or phrases directly from the source text to form the summary [1] and plays an important role in condensing news articles into concise summaries, allowing readers to quickly grasp the key information. Traditional extractive methods primarily rely on lexical word distance to select important sentences for summarization. In many cases the emotional aspects found in the documents are not considered in summarization, and this can affect how readers engage with and understand the information.

Sentiment analysis is the extraction of subjective information from text, encompassing emotions and opinions, and the classification based on the expressed emotions, such as happiness, sadness, anger, fear, or surprise, to capture the overall emotional sentiment [2] and [3]. Sentiment analysis had a huge impact on many applications of NLP in the last years, but there is still space for understanding the details of its implementations [4].

The current study employs SenticNet [3], a multidisciplinary approach to opinion mining that lies at the intersection of affective and common-sense computing. This approach integrates elements from semiotics, psychology, linguistics, and machine learning. Unlike statistical sentiment analysis, Sentic computing focuses on preserving the semantic representation of natural language concepts and sentence structure. The foundation of SenticNet is the Hourglass of Emotions, an emotion categorization model designed to accurately express the affective information present in natural language text.

To the best of our knowledge, sentiment generation is an understudied argument in the field of automatic summarization. Despite the advancements in text summarization techniques, there is a gap in research when it comes to considering emotions in the process. However, we believe that there are a number of applications in which the emotions in a summary should correspond to the emotions of the original document(s). For instance, this is the case of meta-reviews in conference management or the summarization of product reviews in the case of e-commerce. For these application domains, it is

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

* Corresponding author.
† These authors contributed equally.
mdmurad.hossain@unito.it (Md. M. Hossain); luca.anselma@unito.it (L. Anselma); alessandro.mazzei@unito.it (A. Mazzei)
https://shorturl.at/bEGZ0 (Md. M. Hossain); http://www.di.unito.it/~anselma/ (L. Anselma); http://www.di.unito.it/~mazzei/ (A. Mazzei)
0000-0002-8224-3246 (Md. M. Hossain); 0000-0003-2292-6480 (L. Anselma); 0000-0003-3072-0108 (A. Mazzei)

important to consider emotions to create summaries that truly reflect the essence of the source texts.

In this paper, by incorporating sentiment scoring between sentences, we generate summaries that capture the emotional tone and impact of the original text. We believe that exploring this aspect further would lead to more comprehensive and effective text summarization methods.

This paper has two main goals. First, we define a new algorithm called *SentiTextRank*, which is an *emotional* variant of TextRank [5]. Second, we provide an initial evaluation of SentiTextRank by considering two automatic metrics based on content distance.

Note that modern LLMs showed some abilities in summarizing texts by using a specific style, with some limitations in producing a summary that is truly extractive. Moreover, LLMs showed also a big impact from the point of view of the required computational resources. We believe that the work presented in this paper, that requires just few hours to conduct all the experiments, can be seen as a cheap (in many senses) alternative to the use of modern expensive (in many senses) LLMs[1].

The paper is structured as follows. In Section 2, we define the new SentiTextRank algorithm, in Section 3 we report the result of a first experimental evaluation of the SentiTextRank algorithm and the Section 4 ends the paper pointing out to work in progress.

2. SentiTextRank: a variant of TextRank accounting for emotions

TextRank is a popular algorithm for extractive summarization which constructs a graph of sentences or words from a text and assigns scores to each node based on their importance in the graph structure. Finally, it ranks the nodes and selects the top-ranked sentences or words as the summary [5]. The TextRank algorithm is based on the PageRank algorithm, where the sentences of the documents play the role of web pages, and a similarity score plays the role of hyperlink connectivity. Our approach enhances traditional TextRank by incorporating emotions. In particular, we categorize sentences of the original source(s) on the basis of emotions using SenticNet [6]. On the basis of this classification, we obtain a number of distinct emotion sets of sentences. The main idea is to build one single final summary by merging in a selective way the results of TextRank on each one of these emotions sets.

So, the proposed *SentiTextRank* algorithm generates extractive summaries with emphasis on emotion categories through the following steps:

[1]We thank an anonymous reviewer for pointing out this point.

SentiTextRank: Input=Source, **Output**=Sum_F

1. Set compression ratio parameter C between the source(s) and the final summary Sum_F.
2. Classify sentences of the source(s) into different SenticNet emotion categories CAT_{em} with $em \in$ {joy, admiration, surprise, fear, disgust, anger, sadness, interest}.
3. Generate a summary Sum_{em} for each emotion category CAT_{em} by using TextRank.
4. Build Sum_F by picking a number of sentences proportional to C from each Sum_{em} maintaining the original sentence order of the source document.

3. Experimental Result and Discussion

In this section, we present the experimental results of single-document summarization using two datasets, the CNN/Daily Mail dataset (CNN) and the DUC2001 single document dataset (D01), as well as the results of multi-document summarization using two datasets, the DUC2001 multi-document dataset (MD01) and the DUC2004 multi-document dataset (MD04).

The DUC 2001 single document and DUC 2001 Multi-document datasets were collected from the website[2] and consist of news datasets. For our experiments with single documents, we utilized a sample data set of 54 documents. The DUC 2004 multi-document summarization dataset[3] includes 50 items with multiple files and four reference files per item, from which we utilized the first reference for each item. Additionally, we used the CNN/Daily mail dataset[4], where we considered the "highlights" column as the reference summary. Our experiments were conducted on the first 100 rows of text in the CNN/Daily mail dataset. Since the datasets provide just abstractive gold summaries, in order to provide a fair comparison we have converted the abstractive summaries into extractive summaries. This procedure has been proposed in [7, 8]. An extractive reference summary should yield the highest Rouge score when compared to the gold abstractive summary. As finding the globally optimal subset of sentences that maximize the Rouge score is computationally intractable, we adopt a greedy approach: we iteratively add one sentence at a time to the summary, ensuring that the Rouge score of the current set of selected sentences is maximized in relation to the entire gold summary. We repeat this process until there are no more candidate sentences that could enhance the Rouge score when added

[2]https://duc.nist.gov/data.html
[3]https://rb.gy/gp1gbt
[4]https://rb.gy/v4u2g

Original Text	Ever noticed how plane seats appear to be getting smaller and smaller? With increasing numbers of people taking to the skies, some experts are questioning if having such packed out planes is putting passengers at risk.
Gold Abstractive Summary	Experts question if packed out planes are putting passengers at risk. U.S consumer advisory group says minimum space must be stipulated.
Reference Extractive Summary	Ever noticed how plane seats appear to be getting smaller and smaller?. This week, a U.S consumer advisory group set up by the Department of Transportation said at a public hearing that while the government is happy to set standards for animals flying on planes, it doesn't stipulate a minimum amount of space for humans.
Lead	Ever noticed how plane seats appear to be getting smaller and smaller? With increasing numbers of people taking to the skies, some experts are questioning if having such packed out planes is putting passengers at risk.
TR	They say that the shrinking space on aeroplanes is not only uncomfortable - it's putting our health and safety in danger. This week, a U.S consumer advisory group set up by the Department of Transportation said at a public hearing that while the government is happy to set standards for animals flying on planes, it doesn't stipulate a minimum amount of space for humans.
STR	They say that the shrinking space on aeroplanes is not only uncomfortable - it's putting our health and safety in danger. 'It is time that the DOT and FAA take a stand for humane treatment of passengers.

Table 1
An excerpt from the Original Text from the CNN dataset, the existing reference summary (Gold Abstractive Summary), the generated reference summary (Reference Extractive Summary), the lead baseline (Lead), the summary generated by TextRank (TR), and the summary generated by SentiTextRank (STR).

to the current summary set. The subset of sentences that we have at this point is then considered the extractive reference summary for the evaluation. In Table 1 we report an example of summaries generated with the different methods.

Table 1 reports an excerpt from the CNN dataset (Original Text) and the corresponding reference summary (Gold Abstractive Summary). Moreover, Table 1 contains the corresponding generated reference extractive summary (Reference Extractive Summary), the prefix baseline (Lead), the text generated with the TextRank baseline (TR) and, finally, the SentiTextRank generated summary (STR).

Dataset	Algorithm	RL-F1	BERT-F1
D01	Lead	**0.600**	**0.729**
	TR	*0.382*	*0.649*
	STR	0.366	0.605
CNN	Lead	**0.711**	**0.794**
	TR	0.345	*0.642*
	STR	*0.372*	0.608
MD01	Lead	**0.802**	**0.851**
	TR	0.061	*0.560*
	STR	*0.163*	0.505
MD04	Lead	**0.511**	**0.683**
	TR	0.123	*0.575*
	STR	*0.227*	0.542

Table 2
The results of summarization experiments. Lead = Lead Baseline, TR = TextRank, STR = SentiTextRank.

Table 2 presents the experimental results of different summarization methods, namely the Baseline (Lead), TextRank (TR), and our proposed method, SentiTextRank (STR) evaluated on single-document datasets DUC-2001 (D01) and CNN, and the multi-document datasets DUC-2001 (MD01) and DUC-2004 (MD04). As a baseline, we selected the leading sentences from the original documents based on the compression ratio. We evaluated the summaries using two measures: Rouge-L F1 (RL-F1) and BERT F1. ROUGE (Recall Oriented Understudy for Gisting Evaluation) is frequently used to assess how well summarization techniques perform. Rouge-L computes ROUGE for the longest sequence of n-grams [9]. BERT F1 score is a metric commonly used in text classification tasks. It measures the token-level similarity between the generated summary and the reference summary, considering both precision and recall [10].

The results consistently indicate that the Lead method outperforms the other methods across the datasets, showcasing its superiority in generating high-quality summaries. Specifically, Lead achieves the highest scores in Rouge-L F1 and BERT-F1 for D01, CNN, MD01, and MD04. The TR and STR methods exhibit moderate performance in specific evaluation metrics.

Note that the better performance of the Lead method can be attributed to the fact that all the experiments were conducted using news datasets; indeed this result is consistent with the results reported in the literature, where a baseline composed of the leading sentences frequently outperforms extractive and abstractive models on news datasets [11]. However, we think that the comparison between original TR and STR shows encouraging results.

Indeed, the fact that using emotions does not degrade the performance with regard to TextRank shows that we can produce a summary that represents the content as well as the emotions of the source documents. In order to experimentally prove this intuition, we need to formalize an *emotional distance* between summaries and source documents. We plan to develop this point in the future by both (1) using LLM and (2) considering human evaluation.

Further research is necessary to evaluate the performance of our proposed STR method on another domain dataset to provide a comprehensive understanding of its effectiveness.

4. Conclusion and Future Work

This paper introduces the SentiTextRank algorithm, which integrates emotions into the extractive summarization process to create more informative and emotionally rich summaries. The experimental results are encouraging with respect the effectiveness of SentiTextRank in capturing factual information.

The ongoing work on SentiTextRank is following different directions.

First, we want to design a new version of the algorithm that will not be based on the classification of a sentence in one single prevalent emotion. The idea that we want to develop is to define one single measure that combines both *content* and *emotion* similarities. By using this combined measure, we can apply the original TextRank algorithm on the entire set of sentences from the source(s) and obtain one single ranking structure accounting for both content and emotion.

Second, we want to conduct more extensive experiments also on datasets from different domains. In particular, we are considering medical applications since the affective component of medical information can represent a relevant biopsychosocial feature [12].

Third, we are aware that automatic metrics not always measure a real *quality* of the summarized text with respect to human judgment [13, 14]. So, we plan in future to conduct human-based evaluation too.

References

[1] A. Nenkova, K. McKeown, et al., Automatic summarization, Foundations and Trends® in Information Retrieval 5 (2011) 103–233.

[2] M. Wankhade, A. C. S. Rao, C. Kulkarni, A survey on sentiment analysis methods, applications, and challenges, Artificial Intelligence Review 55 (2022) 5731–5780.

[3] E. Cambria, D. Das, S. Bandyopadhyay, A. Feraco, Affective computing and sentiment analysis, A practical guide to sentiment analysis (2017) 1–10.

[4] K. Kenyon-Dean, E. Ahmed, S. Fujimoto, J. Georges-Filteau, C. Glasz, B. Kaur, A. Lalande, S. Bhanderi, R. Belfer, N. Kanagasabai, et al., Sentiment analysis: It's complicated!, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018, pp. 1886–1895.

[5] R. Mihalcea, P. Tarau, Textrank: Bringing order into text, in: Proceedings of the 2004 conference on empirical methods in natural language processing, 2004, pp. 404–411.

[6] E. Cambria, Y. Li, F. Z. Xing, S. Poria, K. Kwok, Senticnet 6: Ensemble application of symbolic and subsymbolic ai for sentiment analysis, in: Proceedings of the 29th ACM international conference on information & knowledge management, 2020, pp. 105–114.

[7] R. Nallapati, F. Zhai, B. Zhou, Summarunner: A recurrent neural network based sequence model for extractive summarization of documents, in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI'17, AAAI Press, 2017, p. 3075–3081.

[8] M. Isonuma, T. Fujino, J. Mori, Y. Matsuo, I. Sakata, Extractive summarization using multi-task learning with document classification, in: Proceedings of the 2017 Conference on empirical methods in natural language processing, 2017, pp. 2101–2110.

[9] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text summarization branches out, 2004, pp. 74–81.

[10] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[11] Y. Liu, M. Lapata, Text summarization with pretrained encoders, arXiv preprint arXiv:1908.08345 (2019).

[12] D. Caldo, S. Bologna, L. Conte, M. S. Amin, L. Anselma, V. Basile, M. M. Hossain, A. Mazzei, P. Heritier, R. Ferracini, et al., Machine learning algorithms distinguish discrete digital emotional fingerprints for web pages related to back pain, Scientific Reports 13 (2023) 4654.

[13] J. Novikova, O. Dušek, A. Cercas Curry, V. Rieser, Why we need new evaluation metrics for nlg, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 2241–2252. doi:10.18653/v1/D17-1238.

[14] F. Moramarco, A. Papadopoulos Korfiatis, M. Perera, D. Juric, J. Flann, E. Reiter, A. Belz, A. Savkov, Human evaluation and correlation with automatic metrics in consultation note generation, in: Proceedings of the 60th Annual Meeting of the Asso-

ciation for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 5739–5754. URL: https://aclanthology.org/2022.acl-long.394. doi:10.18653/v1/2022.acl-long.394.

An Italian Verb Lexicon for Sentiment Inference

Manfred Klenner[1], Dylan Massey[1] and Susanna Tron[2]

[1]Department of Computational Linguistics (Zurich university), Zurich, Switzerland
[2]Alumna of Department of Computational Linguistics (Zurich university), Zurich, Switzerland

Abstract
Verb-centered sentiment analysis provides a fine-grained perspective on the polar impact a situation has on its participants. It is good for a person to be honored for e.g. some achievement and we know in advance that such an achievement is expected to have positive polarity. Also, there is a positive relation between the honorer and the honoree. We introduce an Italian verb lexicon that specifies the polar relations a verb expresses and the effects and expectations that the roles of the verb bear.

Keywords
Sentiment inference, polarity lexicon, verb-based sentiment analysis,

1. Introduction

A sentence is not just positive or negative, it moreover sheds a (polar) light on its participants. Given a sentence like *I mourn my freedom* (It. Piango la mia libertà), a negative effect is cast on the subject and we expect a positive filler of the direct object. Furthermore, we might conclude that the subject is in favour of the object. We have created an Italian verb lexicon comprising 307 verbs with up to 3 frames per verb in order to enable such kind of inferences. Such a lexicon might be useful for stance detection or even for lexicon induction (e.g. from the object expectation of *mourn* we can conclude that *libertà* is a positive concept).

In this short paper, we discuss the ingredients of our lexicon both from a practical and theoretical point of view.

2. Related Work

There are quite a number of approaches to Italian sentiment analysis, be it for social media [1], stance detection [2] or in the context of aspect-based sentiment analysis [3]. Recently, deep learning approaches have conquered the field, but there are lexicon-based approaches as well. To the best of our knowledge, our lexicon is the first for Italian sentiment inference based on a verb resource. In the past, we have created resources and prototypes e.g. [4], [5] and [6] for German sentiment inference[1] and we have started to augment this to the other national Swiss languages: French and Italian. We have also applied our model to gender profiling in order to find out how male or female denoting nouns are distributed in the context of polar verbs [7]. For English, various approaches for sentiment inference are available, see [8], [9], but also [10] and [11]. Also LLMs like ChatGPT are capable of doing it, see [12].

3. Theoretical Framework

A polar verb often expresses a polar relation like *in favour* or *against*. Given the sentence *Lei lo critica* (Eng. She criticizes him), we conclude that the referent of *lei* (Eng. she) is (in a situation-specific way) against the referent of *lo* (Eng. him). We can fix this directed relation by qualifying the subject (the agent) of the verb as the source and the direct object (patient or theme) as the target. We do not claim that all of these relations are intentional, i.e. express a deliberate polar opinion of the source towards the target. Our ultimate goal is to enable (transitive) sentiment inference on the basis of our lexical specifications. Take the sentence *She is against the contract because it prevents a solution.* From this we can directly learn that a) she is intentionlly against the contract and that b) the contract is (non-intentionally directed) against a solution. Only the presence of b) allows us to draw the transitve inference that she is in favour of a solution. The underlying rule is: if X is against Y and Y is against Z, then X is in favour of Z. Thus, the source and target can be animate or non-animate. If it is animate, than the relation might indicate an intentional polar relation originating from the source towards the target. A non-animate source just counts as a cause.

Some polar verbs do not cast a polar relation, but yield a positive or negative perspectivation. For instance, the intransitive usage of the verb *soffrire* (Eng. suffer) like in *Lui soffre* (Eng. He suffers) just tells us that the referent

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
klenner@cl.uzh.ch (M. Klenner); massey@ifi.uzh.ch (D. Massey); tron.susanna@gmail.com (S. Tron)
https://www.cl.uzh.ch/de/people/team/compling/klenner.html (M. Klenner)

[1]For a German demo see: https://pub.cl.uzh.ch/demo/stancer/index.py.

of *lui* is experiencing a negative effect. A polar effect is a positive or negative connotation associated with a verb role (comparable to *good/bad for* effects of [13]). A verb expressing a polar relation might also cast effects, but not in any case. In *Lei lo punisce* (Eng. she punishes him), besides the negative relation there is also a negative effect casted onto *lo* (Eng. him). However in *Lui soffre del suo capo* (Eng. he suffers from his boss) although a negative relation from *Lui* towards *suo capo* holds, there is no negative effect on it. This becomes even clearer if we replace the animate filler by a non-animate one like in *Lui soffre per questa situazione* (Eng. He suffers from the situation). It makes no sense to claim a negative effect on the situation.

In sum: our annotation scheme for polar verbs specifies which roles of a verb do have a positive or negative (directed) relation given an affirmative usage of the verb. Also, we specify whether a verb casts a positive or negative effect on its roles.

So far in our discussion, we have used dependency labels or grammatical functions in order to refer to the roles of a verb. We claim, for instance, that the subject of *criticare* (Eng. criticise) is against the direct object. Even if we assume some normalization, namely passive voice normalization, this might lead to problems. In both cases, *The famous chef cooks* and *The rice cooks* there is a subject, but in the first case it realises an agent in the second a theme. Although in principle semantic role labelling (SRL) might be regarded a better basis for such specification, we have found in [14] that at least for German the available SRL approaches are not suited. Among others, the problem is that the agent role is assigned not exclusively to animate objects (or metonomically used inanimate objects) but also to plain physical objects. For example, the semantic role labeller InVeRo [15] using VerbAtlas[2] assigns the theme role to both (German) subjects: a) *Vergeltung droht ihm* (Eng. retribution threatens him) and b) *Er (animate) droht ihm* (Engl. He threatens him). Only in b) a negative relation should hold (i.e. He against him). For our German model, we sticked with dependency labels and specified coarse-grained selectional restrictions for each verb frame. We then used an animacy classifier, see [16], in order to select the right verb frame. For Italian, the quality of semantic labeling has to be evaluated in order to come to a final decision about the role inventory. This is future work.

Detecting polar relations and effects in sentences on the basis of a lexicon (and a parser) might be regarded antiquated in the age of deep learning. However, there is no need to learn these patterns from annotated data. If the trigger conditions of a verb (frame) are met, then the relations and effects can safely be asserted. Nevertheless, lexicon-based approaches suffer from out-of-vocabulary (oov) cases. Here deep learning (e.g. word embeddings) might actually be useful, since the similarity of oov verbs to verbs that have been seen in the training phase could be exploited.

[2]https://verbatlas.org

4. Corpus: Data Format

We have created a json version of our resource and placed it at Zenodo[3]. For each of the 307 verbs[4] a dictionary is given with the verb lemma, all (subcategorization) frames, the verb polarity, the relations expressed, the effects and expections and one or more sample sentences. For instance the verb *interrompere* (Eng. e.g. cut off) (only one frame is shown) is described in Fig. 1. The English

```
{"verb": "interrompere",
"frame": ["Subj", "Obj"],
"polarity": "NEG",
"relation":[("Subj","against","Obj")]
"effects": [("Obj", "neg")],
"expectations": [],
"example": "Hanno interrotto gli aiuti
           umanitari"}
```

Figure 1: Sample entry for *interrompere*

translation of the sample sentence is *They have cut off humanitarian aid.* According to the verb frame, a subject and an object are required. Given this, a negative effect is cast on the object (humanitarian aid) and an *against* relation is drawn to the object. In the case of non-overt pronouns the arc of the finite verb is annotated, see the *against* relation in Fig. 2.

Some roles of a verb seem to carry a polar expectation. For instance if we hear that somebody is suffering from something, we immediately know that this something is (regarded as) negative. We have integrated expectations in our lexicon, since it might be useful for e.g. lexicon induction.

5. Further Details and Restrictions

First of all, verb disambiguation cannot be totally replaced by subcategorization checks. Take an English example: the two sentences a) *he won the race* and b) *he won the trophy* have the same subcategorization frame (also VerbAtlas assigns the same semantic roles), but the (polar) meaning is different due to verb ambiguity. In a) *he* receives a positive effect, but has no relation to the race, while in b) we can conclude that *he* is in favour of

[3]https://zenodo.org/record/8195009
[4]See the appendix for a list of all modelled verbs.

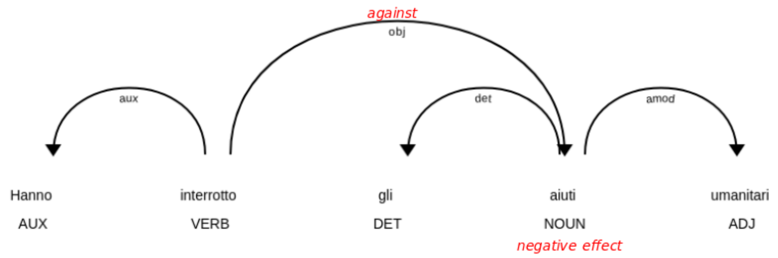


Figure 2: Annotated example (Spacy parse tree)

the trophy. This is to say that using our resource in a system demands additional functionalities.

Some verbs need further checks for them to get triggered. Take for instance the verb *impedire* (Eng. impede). If someone impedes something negative (a struggle), this is good of her (a positive effect and negative relation), but if it is positive (e.g. a solution), it is bad of her (a negative effect and - again - a negative relation). We can determine polar restrictions on the basis of a traditional polarity lexicon. But sometimes also sentiment composition might be needed as in *He stopped the unjustified promotion of the minister* . Here *unjustified promotion* is negative according to a simple compostition rule that says that a negative adjective followed by a positive noun gives a negative phrase. Our manually created Italian polarity lexicon[5] (3,247 lemmas) can be used for such sentiment composition. Nouns and adjectives have been specified on the basis of the appraisal theory [17]: a word might be positive or negative wrt. to one of three dimensions: moral, emotion, appreciation. This allows even to distinguish between cases where a negative effect depends on the appraisal type of the filler. *sick* and *cheating* are both negative, but the former is an appreciation while the latter is morally negative adjective. While the sentence *she admires her sick colleague* proves that she is compassionate, the sentence *she admires her cheating colleague* makes her a morally questionable person. Table 1 shows the distribution of word and polarities in our polarity lexicon.

PoS	positive	negative
noun	370	1242
adjectives	421	1060
adverbs	76	78

Table 1
Polarity Lexicon: distribution of word polarities

Many Italian polarity lexicons[6] specifying the word level polarity of nouns, adjectives and verbs are available. However, there is no fine-grained lexicon at the verb level comparable with ours.

[5]https://zenodo.org/record/8194906

[6]https://www.ai-lc.it/en/affective-lexica-and-other-resources-for-italian

6. Conclusion

In this paper, we have described an Italian verb lexicon and the theory behind it. The lexicon specifies for each Italian verb the underlying relation (in favour or against) and for each verb argument the effects and expectations (if any). The resource has been released together with a traditional polaritiy lexicon (nouns, adjectives). The verb lexicon can be used for sentiment inference, finding out who is in favour of or against who or what. This is useful e.g. for the profiling of political parties: are they in favour or against some topic. Also, the polar perspective which we call effects can be determined: to answer the question whether a text is positive or negative towards an animate or inimate object. This enables to determine the stance of a text author, but also can be used in order to find out how a person or a product is framed e.g. in a particular text or text collection.

Acknowledgements

We like to thank Michi Amsler for the coordination of the development of our lexical resources for German, French and Italian. He also was involved in the annotation of the German lexicon.

References

[1] F. Bianchi, D. Nozza, D. Hovy, FEEL-IT: Emotion and sentiment classification for the Italian language, in: Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Association for Computational Linguistics, Online, 2021, pp. 76–83. URL: https://aclanthology.org/2021.wassa-1.8.

[2] M. Moraca, G. Sabella, S. Morra, Uninastudents @ sardistance: Stance detection in italian tweets

- task a (short paper), EVALITA Evaluation of NLP and Speech Tools for Italian - December 17th, 2020 (2020). URL: https://api.semanticscholar.org/CorpusID:229292804.

[3] P. Basile, D. Croce, V. Basile, M. Polignano, et al., Overview of the evalita 2018 aspect-based sentiment analysis task (absita), in: EVALITA Evaluation of NLP and Speech Tools for Italian, CEUR, 2018, pp. 1–10.

[4] M. Klenner, S. Tron, M. Amsler, N. Hollenstein, The Detection and Analysis of Bi-polar Phrases and Polarity Conflicts, De Gruyter, 2015, pp. 161–172. URL: https://doi.org/10.5167/uzh-99629.

[5] M. Klenner, M. Amsler, Sentiframes: A resource for verb-centered German sentiment inference, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), European Language Resources Association (ELRA), Paris, France, 2016.

[6] M. Klenner, D. Tuggener, S. Clematide, Stance detection in Facebook posts of a German right-wing party, in: LSDSem 2017/LSD-Sem Linking Models of Lexical, Sentential and Discourse-level Semantics, 2017. URL: https://doi.org/10.5167/uzh-136567.

[7] M. Klenner, Sentiment inference for gender profiling, in: Proceedings of the 4th Conference on Language, Data and Knowledge, in press, Viena, Austria, 2023.

[8] A. Neviarouskaya, H. Prendinger, M. Ishizuka, Semantically distinct verb classes involved in sentiment analysis, in: IADIS AC (1), 2009, pp. 27–35.

[9] L. Deng, J. Wiebe, Sentiment propagation via implicature constraints, Meeting of the European Chapter of the Association for Computational Linguistics (EACL-2014) (2014).

[10] H. Rashkin, S. Singh, Y. Choi, Connotation frames: A data-driven investigation, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 311–321. URL: https://aclanthology.org/P16-1030. doi:10.18653/v1/P16-1030.

[11] K. Park, Z. Pan, J. Joo, Who blames or endorses whom? entity-to-entity directed sentiment extraction in news text, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 4091–4102. URL: https://aclanthology.org/2021.findings-acl.358. doi:10.18653/v1/2021.findings-acl.358.

[12] B. Zhang, D. Ding, L. Jing, How would stance detection techniques evolve after the launch of ChatGPT?, CoRR abs/1207.0016 (2023).

[13] L. Deng, Y. Choi, J. Wiebe, Benefactive/malefactive event and writer attitude annotation, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Sofia, Bulgaria, 2013, pp. 120–125. URL: https://aclanthology.org/P13-2022.

[14] M. Klenner, A. Göhring, Semantic role labeling for sentiment inference: A case study, in: Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022), KONVENS 2022 Organizers, Potsdam, Germany, 2022, pp. 144–149. URL: https://doi.org/10.5167/uzh-220537.

[15] S. Conia, R. Orlando, F. Brignone, F. Cecconi, R. Navigli, InVeRo-XL: Making cross-lingual Semantic Role Labeling accessible with intelligible verbs and roles, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, ACL, Online and Punta Cana, Dominican Republic, 2021, pp. 319–328. URL: https://aclanthology.org/2021.emnlp-demo.36. doi:10.18653/v1/2021.emnlp-demo.36.

[16] M. Klenner, A. Göhring, Animacy denoting german nouns: Annotation and classification, in: Proceedings of the Language Resources and Evaluation Conference, European Language Resources Association (ELRA), Marseille, France, 2022, pp. 1360–1364. URL: https://aclanthology.org/2022.lrec-1.145.

[17] J. R. Martin, P. R. R. White, Appraisal in English, Palgrave, London, 2005.

Appendix: Verbs and Frame Size (#)

verb	#frame	verb	#frame
abbandonare	2	abbassarsi	1
abbattere	2	abolire	1
abusare	1	accettare	1
accusare	2	adattare	2
adeguare	2	adempiere	1
adorare	2	affascinare	1
affaticare	2	affidare	2
affligere	2	affondare	1
aggravare	2	aggredire	1
agitare	2	aiutare	2
alimentare	3	alleare	2
alleviare	1	allontanare	3
amare	2	ammazzare	2
ammettere	1	ammirare	1
animare	1	annullare	2
apostrofare	1	applaudire	1
apprezzare	1	approfittare	2
approvare	1	arraffare	3
arrendersi	1	arrestare	1
arricchire	2	assalire	1
assaltare	1	assassinare	2
assicurare	2	attaccare	1
augurare	2	aumentare	1
autorizzare	1	avvantaggiare	2
battere	2	beneficiare	1
bloccare	2	bruciare	3
cacciare	2	calmare	2
cancellare	2	castigare	1
cedere	2	celebrare	1
cessare	1	chiudere	2
colpire	2	coltivare	1
combattere	4	condannare	1
confortare	2	congratularsi	1
conquistare	1	consacrare	2
conseguire	1	consentire	4
conservare	1	consigliare	2
consolare	2	contentare	2
contestare	2	contrastare	1
contribuire	1	convalidare	1
convincere	1	correggere	2
costare	2	costringere	2
crescere	1	criticare	1
curare	2	danneggiare	2
declinare	2	dedicare	2
deludere	1	denunciare	2
deplorare	1	desiderare	1
desistere	1	detestare	2
deviare	1	dibattersi	1
difendere	2	diffidare	2
dimenticare	2	dimettersi	1
diminuire	1	dimostrare	1
disgustare	2	dissimulare	2

verb	#frame	verb	#frame
distruggere	2	disturbare	1
divertire	3	domare	1
dubitare	1	eliminare	1
entusiasmare	2	esaltare	2
esaudire	1	escludere	3
espandere	1	espellere	1
esplodere	2	evolversi	1
facilitare	1	fallire	1
favorire	1	felicitarsi	1
ferire	2	festeggiare	1
forzare	2	fuggire	2
garantire	3	godere	2
ignorare	3	illuminarsi	1
imbrogliare	2	impedire	2
impegnarsi	1	imporre	2
incantare	2	incitare	1
incolpare	2	incoraggiare	3
indebolire	2	indignare	2
infrangere	2	ingannare	2
inquietare	2	integrarsi	1
intensificarsi	1	interessare	1
interrompere	2	invadere	1
irritare	1	isolare	3
ispirare	2	lacerare	2
lamentare	2	lasciare	1
liberare	2	licenziare	1
limitare	2	lottare	2
mancare	3	mantenere	1
mentire	2	meritare	2
migliorare	3	minacciare	1
modernizzare	2	molestare	1
moltiplicare	2	morire	1
nascondere	2	naufragare	1
negare	4	nuocere	1
nutrire	3	obbligare	2
occultare	2	odiare	1
offendere	2	offrire	5
onorare	2	opporre	2
ostacolare	1	pagare	2
patire	1	penalizzare	1
perdere	3	perdonare	2
perire	1	permettere	3
perseguitare	1	piacere	2
piangere	1	picchiare	2
precipitare	2	preoccupare	2
preservare	1	privare	2
privilegiare	1	progredire	1
promettere	1	promuovere	1
proteggere	2	protestare	1
provocare	1	punire	1
raccommandare	2	raddoppiare	1
rafforzare	2	rallegrare	2
rassicurare	2	ratificare	1
recuperare	1	regolare	1
resistere	2	respingere	1

verb	#frame	verb	#frame
ricompensare	1	riconoscere	2
ridere	2	ridurre	4
rifiutare	2	rimpiangere	1
rimproverare	3	rinchiudersi	1
ringraziare	1	rinnovare	2
rinunciare	1	riposare	2
rischiare	1	risentire	3
risolvere	2	risparmiare	1
rispettare	1	ristabilire	2
ritirare	3	riuscire	2
rivelarsi	2	rompere	4
rovesciare	2	rubare	3
sacrificare	2	salutare	1
salvare	3	sanzionare	1
sbagliare	3	sbarazzare	2
scappare	1	scatenare	2
schiacciare	2	scioccare	1
sciupare	2	scombussolare	1
scomparire	1	scompigliare	2
sconvolgere	1	scoppiare	2
scoraggiare	2	scordare	2
scuotere	2	sedurre	1
separarsi	1	sfidare	1
sfuggire	1	sgomberare	2
sgridare	1	smarrire	3
smentire	4	smettere	1
snobbare	2	soddisfare	1
soffrire	3	sopportare	1
sopprimere	1	sorridere	2
sospendere	1	sospettare	3
sottomettere	2	sottrarsi	1
sperare	1	spezzare	4
stancare	2	stigmatizzare	1
stimare	1	strappare	2
stravolgere	1	stuprare	1
subire	1	suggerire	1
suicidarsi	1	suscitare	1
temere	3	tormentare	1
tradire	2	trascinare	2
trascurare	2	trasgredire	1
truffare	2	turbare	1
ubbidire	1	uccidere	2
urtare	2	vantare	2
vegliare	1	vietare	1
vincere	1	violare	1

The Inherence of Telicity: Unveiling Temporal Reasoning in Video Question Answering

Olga Loginova[1], Raffaella Bernardi[2]

[1]DISI, University of Trento, Via Sommarive, 9, 38123 Povo TN, Italy
[2]CIMeC, University of Trento, Corso Bettini 31, 38068 Rovereto TN, Italy

Abstract

Video question answering (VQA) requires models to understand video-related questions and generate natural language answers. In multiple-choice VQA, models must associate visual content with one of several predetermined answers. As videos often encompass intricate events and actions unfolding over time, these models must possess the ability to reason across multiple frames and discern the relationships between them with respect to the answers. This paper focuses on the Answerer component of a multiple-choice VQA model, which predicts answers using language-infused key frames. We hypothesise that the Answerer's capacity for temporal reasoning is closely intertwined with its understanding of aspectuality. To investigate this, we augment NeXT-QA, a VQA dataset for causal and temporal reasoning, with annotations for telicity. We then delve into the performance evaluation of SeViLA, a state-of-the-art multiple-choice VQA model, on it. Our findings demonstrate that the model generally exhibits correct handling of aspects, albeit with a bias that is inherent in human nature.

Keywords

video question answering, temporal reasoning, aspect, telicity

1. Introduction

Temporal ordering of actions and events is not solely determined by time; it is also influenced by causality. The organisation of activities in episodic memory is established based on contingency, where one activity triggers another [1]. Recognising cause-effect relationships is essential for temporal understanding, as causes typically precede effects. A cause that has reached its culmination induces the effect.

In language, linguistic aspects play a role in how activities unfold and whether they have culminated. The concept of telicity marks the endpoint of an activity: a verb phrase with a clear endpoint is considered telic (e. g., "to pick up something"), while an atelic one is ongoing, without a specific endpoint (e. g., "to clap"). In descriptions of a sequence of activities with the resultative structure there is an evident human bias towards telic interpretation [2].

Previous research explored telicity for textual transformer-based [3] models, showing that they can classify activities based on duration and telicity with an accuracy surpassing 80% [4]. Such performance at a level comparable to humans, even with limited training data, indicates their ability to capture temporal reasoning through aspect classification.

Our work extends this line of research to video-language models, where video content comes with text labels assigned to key frames or the whole video. Ordering of events corresponds to changing frames, making the correct key frame extraction critical for temporal reasoning. Action timestamps to the frames provide additional cues for temporal reasoning. We propose a study that focuses on contemporary video question-answering (VQA) models in order to explore the relevance of telicity for answering temporal questions related to simultaneous and consecutive activities. We consider the aspects of question's both main and dependent clauses.

To achieve this, we annotate[1] the test set of NExT-QA [5], widely used for causal and temporal reasoning benchmarks, with telicity and evaluate the SeViLA model [6] on this annotated dataset. To the best of our knowledge, this is the first such endeavor in the VQA field.

Q: what is the girl doing with her hands (A)
as the man moved his head sideways at the end of the video (T)

a0: smiles (A)	a1: jamming (A)	a2: pick up something (T)	$\boxed{\text{a3: clap (A)}}$	**a4: pick up toy (T)**

Figure 1: Example of a temporal question and answer options in NexT-QA augmented with our annotation for telic (T) and atelic (A) actions.

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
olga.loginova@unitn.it (O. Loginova); raffaella.bernardi@unitn.it (R. Bernardi)
https://github.com/ologin (O. Loginova); http://disi.unitn.it/~bernardi/ (R. Bernardi)
0009-0006-1885-3759 (O. Loginova); 0000-0002-3423-1208 (R. Bernardi)

[1]The annotations for the dataset are publicly available on GitHub: https://github.com/ologin/Telicity-on-NExT-QA.

Figure 1 provides an example of NExT-QA temporal questions with our telicity annotation: SeViLA selects the telic answer “pick up toy" (in bold) that does not match the ongoing nature of the question's main clause “is ... doing", while in the correct answer “clap" (boxed) there is a match of atelic activities.

Our findings demonstrate that the VQA model SeViLA can effectively handle telicity. Furthermore, when making a mistake in prediction the model, like humans, tends to adopt a telic-prone approach.

2. Related Literature

Numerous transformer-based models tackle the challenge of video question answering [7, 8, 9, 10, 11, 12, 6, 13, 14]. These models process both the visual and textual modalities by incorporating video, captions or subtitles, and fuse these streams to generate the final answer. They have showed impressive performance in modelling multimodal VQA. However, they were never assessed for telicity. SeViLA [6], selected for our experiment, consists of two modules: Localizer, for action recognition within videos, and Answerer. The modules are fine-tuned based on BLIP-2 [15]. The model has proved the best results in comparison to other similar models on several datasets, such as STAR [16], NExT-QA [5], How2QA [17], and TVQA [18].

We examined datasets that offer multiple-choice answer options where models must choose the correct answer from a set of candidates. CausalQA [19], Social-IQ [20], CLEVRER [21], STAR [16], and NExT-QA [5] are specifically designed to explore temporal dynamics and the role of causal relationships. NExT-QA proved to be particularly suitable for our experiment, as it is the most comprehensive and emphasises the real-world scenarios.

3. Annotation

The NExT-QA test set comprises 1000 videos with 8564 question-answer pairs supported by five answer options each. From a range of 1 to 15 questions with an average of 9-10 questions per video, we selected solely temporal (T-type) questions. We further excluded closed questions and questions that do not involve two distinct temporally-linked activities, such as “did the baby get hurt after putting out the candle" or “what are the people in this video doing". Thus, the refined total set (RTS) consists of 2060 question-answer pairs.

Notably, RTS questions pertaining to the following activities are in the absolute majority, while the ones concerning preceding actions are very few.[2]

[2]More details on the dataset are in Section A of the Appendix.

3.1. Aspect Annotation

We divided all question activities into two groups: activities of the main clause (MCA) and activities of the dependent clause (DCA). We annotated independently both question groups, as well as the target and predicted answers, with the following labels of the internal temporal structure:

- **T (telic)** for activities implying an endpoint (e. g., “what happened", “pick up camera", “after the door opens"),
- **A (atelic)** for enduring processes (e. g., “how is the person in black positioned", “smiles", “while watching"), and
- **U (undefined)** for activities lacking clear telicity and duration (e. g., “what does the dog do", “do the same", “to man's action to him").

Additionally, an **I (irrelevant)** marker was assigned to answers unrelated to aspectuality, such as “astonished" or “nothing". This marker appears among the target answers too in response to questions like “how did the boy react to..." or “what does the person do while...".

Table 1
Telicity of all activities in RTS: questions' main and dependent clauses, as well as the correct (target) answers

Activity Group	T	A	U	I
`MCA`	40	159	1861	0
`DCA`	1254	801	5	0
`Answer (target)`	758	1283	0	19

From Table 1 it is evident that the question's main clause rarely impose a definitive telic label, setting the model free to explore temporal relations without predefined constraints. The majority of DCAs are telic and, considering that the most of RTS questions center around the following activity, this affirms the cause-effect nature of the dataset, where the cause predominantly culminates in an endpoint.

4. Experiment and Results

We ran zero-shot SeViLA setting on the test dataset decreasing the batch size down to 2. The obtained results revealed the overall accuracy of 63.18% and the T-type question accuracy of 60,18%. On RTS, we obtained 58.1% of matching predicted and target answers.

We further calculated the telicity precision, recall, F1 score and accuracy on the annotated RTS.

4.1. Results

SeViLA selected 781 telic (T) and 1261 atelic (A) responses, alongside 2 instances marked as undefined (U), and 16 responses classified as irrelevant (I). [3]

As demonstrated in Table 2, the results verify that the model attains an accuracy rate exceeding 80%.

Table 2
Telicity precision, recall, F1 score and accuracy results on RTS

Metric	Value
Precision	0.76
Recall	0.74
F1 score	0.75
Accuracy	0.81

The confusion matrix shown in Figure 2 indicates a higher frequency of atelic answers. The majority of atelic responses might initially prompt an inference of an atelic predisposition of the model. Upon closer examination, however, we observed that the incidence of erroneous allocations from atelic to telic responses is more pronounced than in the inverse direction. Thus, the model exhibits a clear inclination towards selecting telic values instead of the target atelic ones: in 26,12% of the target atelic answers it chooses the telic ones, while there are only 14.45% of the opposite cases.

		TARGET	
		T	A
SeViLA	T	574	203
	A	182	1077

Figure 2: Confusion matrix for telicity classification in RTS. U and I labels are excluded as uninformative.

4.2. Qualitative analysis

The SeViLA Answerer employs a top-k frame extraction strategy to evaluate each frame's probability and determine the optimal choice for answering a question. The erroneous answers often come from the model's misjudgment in instructive key frames.

As shown in Figure 3, the telicity cues may have their origins in the question's both MCA and DCA. As much as in the TN-question (top) SeViLA disregards the DCA's telic action, it also struggles to correspond with the atelic activities of the MCA in the answer for the TC-question (bottom).

[3]Additional data regarding SeViLA's predictions in the context of RTS can be found in Section B of the Appendix.

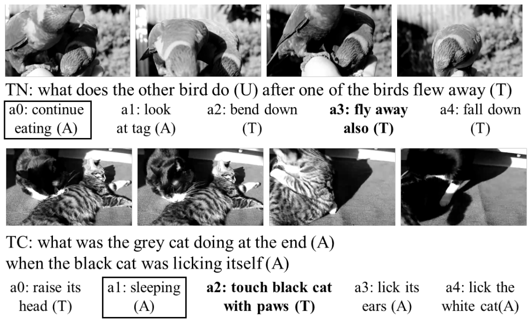


Figure 3: Instances of the key frame misjudgement for next (TN) and current (TC) activities: the SeViLA answers (in bold) and the target ones (boxed).

5. Limitations

While NExT-QA is distinguished as a versatile dataset, it has limitations in representing temporal expressions from a linguistic perspective. Primarily, its questions use a limited set of temporal conjunctions, including *after*, *before*, *during*, *as*, *while*, and *whenever*. A dataset with a broader array of temporal constructions related to both time and telicity could introduce variations, potentially altering model's outcomes.

Another source of result variations can stem from the number of annotators. The annotations were created by a professional linguist in a pilot version, but it is important to acknowledge a potential subjective bias. To mitigate the bias, at least three annotators are suggested for each question-answer pair.

6. Conclusion

The linguistic models grounded in cognitive research highlight a tendency for individuals to remember causally linked activities. Sequential actions and events are associated with the idea that the culmination of one activity sets off another. This culmination is closely tied to the internal structure of the activity which is expressed in language through aspects and, in particular, telicity.

Using NExT-QA dataset, we revealed that VQA models, such as SeViLA, generally capture the contrast in durative and endpoint activities at a human level. Whereas they mostly tend to predict correct telicity for causal and temporal reasoning, their inherent erroneous implication of culminated activity, in essence, aligns with human intuition.

This revelation prompts us to answer the follow-up question: to what extent the improvement in matching telicity in questions and answers will amplify the key frame extraction for correct answering in multiple-choice VQA models.

Acknowledgments

The second author extends her gratitude to Amazon Alexa for their research donation, which significantly supported her work.

References

[1] M. Moens, M. Steedman, Temporal ontology and temporal reference, Comput. Linguist. 14 (1988) 15–28.

[2] Y. Zhao, J. G. Ngui, L. Hall Hartley, S. Bethard, Do pretrained transformers infer telicity like humans?, in: Proceedings of the 25th Conference on Computational Natural Language Learning, Association for Computational Linguistics, Online, 2021, pp. 72–81. doi:10.18653/v1/2021.conll-1.6.

[3] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Neural Information Processing Systems, 2017.

[4] E. Metheniti, T. Van De Cruys, N. Hathout, About time: Do transformers learn temporal verbal aspect?, in: Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 88–101. doi:10.18653/v1/2022.cmcl-1.10.

[5] J. Xiao, X. Shang, A. Yao, T.-S. Chua, Next-qa: Next phase of question-answering to explaining temporal actions, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 9777–9786.

[6] S. Yu, J. Cho, P. Yadav, M. Bansal, Self-chained image-language model for video localization and question answering, ArXiv abs/2305.06988 (2023).

[7] L. Zhu, Y. Yang, Actbert: Learning global-local video-text representations, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 8743–8752.

[8] Z. Yang, N. Garcia, C. Chu, M. Otani, Y. Nakashima, H. Takemura, Bert representations for video question answering, in: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1545–1554. doi:10.1109/WACV45572.2020.9093596.

[9] A. U. Khan, A. Mazaheri, N. da Vitoria Lobo, M. Shah, Mmft-bert: Multimodal fusion transformer with bert encodings for visual question answering, in: Findings, 2020.

[10] J. Lei, L. Yu, T. L. Berg, M. Bansal, Tvqa+: Spatio-temporal grounding for video question answering, ArXiv abs/1904.11574 (2019).

[11] R. Zellers, X. Lu, J. Hessel, Y. Yu, J. S. Park, J. Cao, A. Farhadi, Y. Choi, Merlot: Multimodal neural script knowledge models, in: Neural Information Processing Systems, 2021.

[12] Y. Zhong, W. Ji, J. Xiao, Y. Li, W. Deng, T.-S. Chua, Video question answering: Datasets, algorithms and challenges, ArXiv abs/2203.01225 (2022).

[13] J. Xiao, P. Zhou, T.-S. Chua, S. Yan, Video graph transformer for video question answering, in: European Conference on Computer Vision, 2022.

[14] J. Xiao, P. Zhou, A. Yao, Y. Li, R. Hong, S. Yan, T.-S. Chua, Contrastive video question answering via video graph transformer, IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (2023) 13265–13280.

[15] J. Li, D. Li, S. Savarese, S. C. H. Hoi, Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, ArXiv abs/2301.12597 (2023).

[16] B. Wu, S. Yu, Z. Chen, J. B. Tenenbaum, C. Gan, Star: A benchmark for situated reasoning in real-world videos, in: Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS), 2021.

[17] L. Li, Y.-C. Chen, Y. Cheng, Z. Gan, L. Yu, J. Liu, Hero: Hierarchical encoder for video+language omni-representation pre-training, in: Conference on Empirical Methods in Natural Language Processing, 2020.

[18] J. Lei, L. Yu, M. Bansal, T. L. Berg, Tvqa: Localized, compositional video question answering, in: Conference on Empirical Methods in Natural Language Processing, 2018.

[19] A. Bondarenko, M. Wolska, S. Heindorf, L. Blübaum, A.-C. Ngonga Ngomo, B. Stein, P. Braslavski, M. Hagen, M. Potthast, CausalQA: A benchmark for causal question answering, in: Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 3296–3308.

[20] A. Zadeh, M. Chan, P. P. Liang, E. Tong, L.-P. Morency, Social-iq: A question answering benchmark for artificial social intelligence, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8799–8809. doi:10.1109/CVPR.2019.00901.

[21] K. Yi, C. Gan, Y. Li, P. Kohli, J. Wu, A. Torralba, J. B. Tenenbaum, Clevrer: Collision events for video representation and reasoning, in: International Conference on Learning Representations, 2020.

A. NExT-QA RTS Dataset Statistics

This section provides more details on the dataset used for the experiment.

A.1. Types of Questions

There is prominent imbalance among all T-type questions in RTS with the majority of TN questions.

Table 3
All types of temporal questions in RTS: questions that ask previous (TP), next (TN) and current (TC) activities

Type	Number	Percentage
TP	91	4,42
TN	1333	64,71
TC	636	30,87

A.2. Question Structure

The detailed overview of RTS questions shows that the dataset has predominately questions with the "what did S do after..." structure.

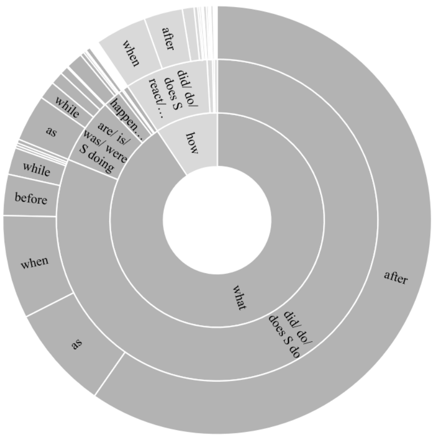


Figure 4: NeXT-QA RTS question structure. S signifies subject of the main clause.

B. SeViLA's performance on RTS

This section presents the details concerning the data predicted by the model.

B.1. Matching in Absolute Numbers and Percentage

Target answers do not have underdetermination with little data irrelevant from the aspect point of view.

Table 4
The amount of correct and incorrect predictions of SeViLA on RTS

Matching	T	A	U	I
correct	425	758	0	12
incorrect	356	504	1	4

Table 5
The percentage of correct and incorrect predictions of SeViLA on RTS

Matching	T	A	U	I
correct	54.42	60.06	0	75
incorrect	45.48	39.94	100	25

B.2. Predicted vs. Target Answers

The examination of the most frequently predicted and target answers reveals a significant number of matches, predominantly characterised by atelic labels.

Table 6
Top 10 predicted answers

Answer	Telicity Label	Amount
smile	A	21
walk away	T	18
walks away	T	17
look around	A	17
look at camera	A	16
stand up	T	15
laugh	A	14
clap	A	11
turn around	T	10
smiling	A	8

Table 7
Top 10 target answers

Answer	Telicity Label	Amount
walk away	T	27
stand up	T	19
laugh	A	17
smile	A	17
dance	A	15
look at camera	A	15
turn around	T	12
clap	A	9
look around	T	9
smiling	A	9

Identification of Multiword Expressions: comparing the performance of a Conditional Random Fields model on corpora of written and spoken Italian

Ilaria Manfredi[1], Lorenzo Gregori[1]

[1]University of Florence, P.zza San Marco 4, 50121 Florence, Italy

Abstract

This paper describes an experiment that compares the performance of a Conditional Random Fields model on identification of Multiword expressions in corpora of spoken and written Italian. The model is trained on a corpus of spoken language and a corpus of written language annotated with Multiword expressions, then tested on two other corpora (one written and one spoken). This methodology provides very good results regarding Precision.

Keywords

Multiword Expressions, Conditional Random Fields, Spoken corpora

1. Introduction

"Multiword expression" (MWE) is a term used to refer to groups of words that display formal or functional idiosyncratic properties with respect to free word combinations, and therefore behave like a unit [1]. This notion encompasses a wide set of linguistic phenomena, of both semantic and syntactic nature, like idioms, verb-particle constructions, complex nominals, and support verb constructions. The computational treatment of MWEs notoriously poses a challenge in NLP [2], but in recent years a lot of effort has been put into the development of techniques and tools for the identification of MWEs in corpora. These are almost exclusively derived from, and tested on, written corpora. This leaves the study of MWEs in spoken varieties of languages, including Italian, a rather unexplored field.

Given the major differences between spoken and written language, we deemed it important to establish how an MWEs automatic extraction tool trained on written corpus performs on a spoken one, also considering the lack of specific resources for spoken corpora. We have decided to conduct an experiment training a Conditional Random Fields (CRF) model [3] to identify MWEs. The model was trained on both a corpus of spoken and one of written Italian; the two models obtained were then tested on corpora of spoken and written Italian, and their performances were evaluated. In § 2 we give an overview of existing research on MWEs and related resources for Italian; in § 3 we describe the resources used to build the training and test corpora; in § 4 we described the methodology followed to annotate the training corpora with MWEs and the testing; results of the experiment are presented in § 5 and discussed in § 6.

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
ilariamanfredi3@gmail.com (I. Manfredi);
lorenzo.gregori@unifi.it (L. Gregori)

2. Related work

Identification of MWEs in corpora is essential for various NLP tasks such as machine translation and parsing, so a lot of research has been done on automatic acquisition of MWEs, both in general and for specific languages [4]. Many studies have explored the use of Association Measures for MWEs identification [5, 6, 7]; methodologies based on parallel corpora have also been investigated [8]. More recently, the use of different AI models has been tested for this task [9, 10]. Among these, CRF models has been used successfully in NLP for various sequence labeling tasks, including MWEs identification [11, 12, 13]. Given that, we have decided to use one of the CRF models available for our experiment (see § 4). As already mentioned all of these studies have been conducted on written corpora only, and so are the resources derived (mainly MWE annotated corpora and gold standard lists).

As for MWEs in spoken corpora, Strik et al. investigated possible ways of automatically identifying MWEs in Dutch speech corpora based on pronunciation characteristics; Trotta et al. built PoliSdict, a dictionary of Italian MWEs extracted from a corpus of political speech. To the best of our knowledge, this is the only resource of speech language MWEs existing for Italian. Other resources for Italian MWEs are PARSEME-It, a written corpus annotated with verbal MWEs [16, 17], and a validated dataset of MWEs from written corpora compiled by Masini et al. [19].

This brief overview highlights the gap in existing literature regarding MWEs from spoken language; hence, our experiment seeks to evaluate the performance of one of the tools available, up to now tested only on written corpora.

3. Resources

For the experiment, we have used two training corpora and two test corpora (described in § 4.1) derived from the following resources.

KIParla [20] is a spoken corpus containing more than 112 hours of speech recorded in various settings from speakers of different areas of Italy, and is currently composed of two modules. The KIP module [21] contains speech of students and professors recorded in the Universities of Bologna and Turin.

IMAGACT is a corpus of approximately 1.8 million tokens[1] used for the creation of the IMAGACT Visual Ontology resource [22]; it contains texts of spoken Italian derived from LABLITA Corpus of Spontaneous Italian, LIP corpus, and the spoken section of CLIPS corpus. The materials contained are heterogeneous from a diaphasic, diastratic, and diatopic point of view (see Gagliardi for a detailed description).

CorDIC-scritto is a web corpus created within the RIDIRE project [24] containing written texts pertaining to five different semantic and functional domains: creative, bureaucratic, news, arts, economy[2].

PAISÀ [25] is a web corpus of approximately 250 million tokens containing documents from web pages. Part of the documents was obtained by retrieving pages using pairs words from the Italian basic vocabulary list as queries; others were derived from the Italian versions of various Wikimedia Foundation projects.

4. Methodology

This work has been conducted making use of the `mwetoolkit` software [26] for the extracting, filtering and annotating of the MWEs; the CRF model we have used is the one implemented in the CRFsuite software [27] and provided within the toolkit.

4.1. Training and test corpora

We have used the KIP module[3] of KIParla as the spoken training corpus and CorDIC-scritto as the written training corpus. As the spoken test corpus we have used IMAGACT. Lastly, for the written test corpus we have sampled PAISÀ to have approximately the same number of words of IMAGACT, in order for them to be comparable in size. Table 1 shows numbers of words and tokens of each of the corpora.

Table 1
Training and test corpora with number of words and tokens.

	Name	Words	Tokens
Spoken training	KIP	559,816	637,867
Written training	CorDIC	502,665	589,036
Spoken test	IMAGACT	1,366,305	1,870,272
Written test	PAISÀ	1,366,313	1,686,217

All of the corpora have been POS-tagged and lemmatized with Treetagger [28] using Baroni's parameter file [4].

4.2. Annotation of the training corpora

The first step to annotate the training corpora was the extraction of candidates, obtained by searching the corpora with sets of POS-patterns (see Ramish and Lenci et al. for an assessment of the method). The chosen POS-patterns were derived from the work of Masini et al., who provided a dataset of 1682 validated Italian MWEs extracted from written corpora with the POS-pattern method. We chose to use the top 20 POS-patterns in the dataset ranked by number of MWEs. Since the patterns in the dataset are provided according to the ISST-Tanl tagset[5], we first "translated" the tags to their respective ones in Baroni's tagset. The tagsets are not symmetrical (for example ISST-Tanl tags *RD* 'determinative article' and *RI* 'indeterminative article' are both *ART* 'article' in Baroni's tagset) so we computed again frequency of MWEs for each pattern and then took the top 20. The 20 POS-patterns used are bigrams and trigrams of adjectival, nominal, verbal, adverbial and prepositional patterns.

Using `mwetoolkit` functions, the corpora were searched and for every POS-pattern a list of candidates was obtained; each corpus was searched independently and the lists of candidates were examined separately. As a second step, all the lists of the candidates were filtered by number of occurrences: only candidates with a frequency of 4 or more were kept. Lists containing a high number of candidates were further filtered, before being manually examined: for KIP, lists having more than 150 candidates were ranked by LogLikelihood and the top 100 were examined; for CorDIC, lists with more than

[1] Here tokens are intended as single graphic units that include punctuation, symbols and words, as usual in computational linguistics
[2] See `http://cordic.lablita.it/`
[3] Compared to the original resource, available on `https://kiparla.it/search/`, our corpus lacks the documents BOC1006, BOD2008, TOA3005, TOD1005bis.

[4] https://home.sslmit.unibo.it/ baroni/collocazioni/itwac.tagset.txt.
[5] http://www.italianlp.it/docs/ISST-TANL-POStagset.pdf

Table 2
Candidates extracted (f $>$ 3) and candidates examined for each POS-pattern.

POS-pat	cK	aK	cC	aC
A-N	189	100	263	100
PreArt-A-N	25	25	48	48
PreArt-N	729	100	1781	100
PreArt-N-Pre	37	37	240	100
N-A	258	100	697	100
N-PreArt-N	56	56	284	100
N-N	36	36	27	27
N-Pre-N	108	108	228	100
N-V	108	108	134	100
Pre-A-N	15	15	38	38
Pre-Art-N	115	115	255	100
Pre-DInd-N	15	15	28	28
Pre-N	664	100	1216	100
Pre-N-Pre	52	52	143	100
V-A	106	106	104	100
V-Adv	439	100	151	100
V-Art-N	148	148	69	69
V-PreArt-N	16	16	42	42
V-N	109	109	84	84
V-Pre-N	50	50	48	48
Total	3275	1496	5980	1584

100 candidates were ranked by LogLikelihood[6] and the top 100 were examined. In lists having less candidates than that, all of the candidates were examined. This way there is approximately the same number of candidates to be examined for each corpus: 1496 for KIP and 1584 for CorDIC.

Table 2 shows, for each POS-pattern, the number of candidates with frequency $>$ 3 in KIP (candK) and CorDIC (candC) and the number of candidates examined in each corpus (anK and anC). POS are abbreviated like this: A = adjective, N = noun, Pre-Art = articulated preposition, Pre = preposition, V = verb, Art= article, DInd = indefinite determiner, Adv = adverb.

As the final step, the remaining candidates from all the lists were manually examined. Candidates who showed some type of idiomaticity, fixedness, or were characterized by high familiarity of use were annotated as MWEs: in total, 214 MWEs for KIP and 204 for CORDIC. MWEs were tagged in their respective corpora using the IOB format [32]. In this process, attention has been put to only tag MWEs when they are in an idiomatic context, and not where they have a literal meaning.

[6] To calculate LogLikelihood for trigrams we have used the Ngram Statistics Package [30, 31]

Table 3
Occurrences of MWEs and Precision for each model on each corpus

	MWEs	Pr
S model IMAGACT	7508	0,974
S model PAISÀ	3337	0,908
W model IMAGACT	6291	0,978
W model PAISÀ	5047	0,946

4.3. Training and testing

The model was trained on MWE annotated KIP and CorDIC independently, using the functions of `mwetoolkit`; the training script was not modified and the features were kept as provided[7].

So we obtained two models, one trained on KIP (the 'spoken model') and one trained on CorDIC (the 'written model'). We used each of them to identify MWEs from IMAGACT and PAISÀ, with the aim to compare the results and determine if the best performance on spoken corpus comes from a spoken o written model, and vice versa.

5. Results

The spoken model tagged 7508 occurrences of MWEs in IMAGACT and 3337 in PAISÀ; the written model tagged 5047 occurrences of MWEs in PAISÀ and 6291 in IMAGACT. For a full evaluation of the models we need to compute Precision and Recall of the annotated corpora. Computation of Recall needs all the false negatives in test corpora to be identified; for that, we would need to manually annotate the entire corpora which is a very time-consuming task that requires multiple trained annotators. Another element of complexity for this task is to provide annotators with a precise definition of what to consider a MWE, as the distinction between MWEs and other types of word combinations is not always clear-cut. So, evaluation has been performed by manually computing Precision on a sample of 500 MWEs from each batch of results. Table 3 shows occurrences of MWEs and Precision at 500 for spoken and written models on each corpus.

6. Discussion

Results obtained show a great performance overall for both of the models, given the high value for Precision for all four of the corpora tagged. However, considering also the number of MWE occurrences tagged, we

[7] See https://gitlab.com/mwetoolkit/mwetoolkit3/-/blob/master/resources/default-config/listFeatures.txt

can see that the spoken model performed the worst on PAISÀ, having the lowest Precision and number of occurrences, while better results are achieved on the same corpus by the written model. On IMAGACT, both of the models performed very well, with the written model having the best Precision overall but slightly fewer occurrences of MWEs found. We have also counted the number of MWEs tagged (per lemmas) in IMAGACT, and how many of these were "new" compared to the ones annotated in the training corpora. The spoken model tagged 222 MWEs (per lemmas) of which 63 were new (28.4%) and the written model tagged 224 MWEs (per lemmas), 64 being new (28.6%), so the models performed similarly in this regard too. A slight difference in performance can be noted comparing Precision in tagging new MWEs: new MWEs found by spoken model account for a total of 119 occurrences, 46 of which results correctly tagged; new MWEs found by written model account for 123 occurrences, 60 of which are correctly tagged.

In conclusion, the results of this experiment show that on spoken corpora 'written models' perform similarly to 'spoken models'; this looks really promising, considering the lack of resources dedicated to MWEs in spoken language. Future works in this line of research include the computing of Recall for the models and qualitative evaluation of the MWEs extracted.

References

[1] F. Masini, Multi-word expressions and morphology, 2019. doi:10.1093/acrefore/9780199384655.013.611.

[2] I. A. Sag, T. Baldwin, F. Bond, A. A. Copestake, D. Flickinger, Multiword expressions: A pain in the neck for NLP, in: A. F. Gelbukh (Ed.), Computational Linguistics and Intelligent Text Processing, Third International Conference, CICLing 2002, Mexico City, Mexico, February 17-23, 2002, Proceedings, volume 2276 of *Lecture Notes in Computer Science*, Springer, 2002, pp. 1–15. doi:10.1007/3-540-45715-1_1.

[3] J. D. Lafferty, A. McCallum, F. C. N. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in: Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01, Morgan Kaufmann Publishers Inc., San Francisco, California, 2001, p. 282–289.

[4] C. Ramish, A generic and open framework for multiword expression treatment: from acquisition to applications, Ph.D. thesis, Universitade Federal do Rio Grande do Sul, 2012.

[5] P. Pecina, A machine learning approach to multiword expression extraction, in: Proceedings of the LREC 2008 Workshop Towards a Shared Task for Multiword Expressions, 2008, pp. 54–57.

[6] A. Fazly, S. Stevenson, Distinguishing subtypes of multiword expressions using linguistically-motivated statistical measures, in: Proceedings of the Workshop on A Broader Perspective on Multiword Expressions, Association for Computational Linguistics, Prague, Czech Republic, 2007, pp. 9–16. doi:10.3115/1613704.1613706.

[7] G. I. Lyse, G. Andersen, Collocations and statistical analysis of n-grams: Multiword expressions in newspaper text, in: Exploring newspaper language: Using the web to create and investigate a large corpus of modern Norwegian, 2012, pp. 79–110. doi:10.1075/scl.49.05lys.

[8] Y. Tsvetkov, S. Wintner, Extraction of multi-word expressions from small parallel corpora, Natural Language Engineering 18 (2012) 549–573. doi:10.1017/S1351324912000101.

[9] W. Gharbieh, V. Bhavsar, P. Cook, Deep learning models for multiword expression identification, in: Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 54–64. doi:10.18653/v1/S17-1006.

[10] R. Swaminathan, P. Cook, Token-level identification of multiword expressions using pre-trained multilingual language models, in: Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023), Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 1–6. doi:10.18653/v1/2023.mwe-1.1.

[11] A. Maldonado, L. Han, E. Moreau, A. Alsulaimani, K. D. Chowdhury, C. Vogel, Q. Liu, Detection of verbal multi-word expressions via conditional random fields with syntactic dependency features and semantic re-ranking, in: Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017), Association for Computational Linguistics, Valencia, Spain, 2017, pp. 114–120. doi:10.18653/v1/W17-1715.

[12] K. Nongmeikapam, D. Laishram, N. B. Singh, N. M. Chanu, S. Bandyopadhyay, Identification of reduplicated multiword expressions using crf, in: A. F. Gelbukh (Ed.), Computational Linguistics and Intelligent Text Processing, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, pp. 41–51. doi:10.1007/978-3-642-19400-9_4.

[13] M. Scholivet, C. Ramisch, Identification of ambiguous multiword expressions using sequence models and lexical resources, in: Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017), Association for Computational Linguistics, Valencia, Spain, 2017, pp. 167–175. doi:10.18653/v1/

W17-1723.
[14] H. Strik, M. Hulsbosch, C. Cucchiarini, Analyzing and identifying multiword expressions in spoken language, Language Resources and Evaluation 44 (2010) 41–58. doi:10.1007/s10579-009-9095-y.
[15] D. Trotta, T. Albanese, M. Stingo, R. Guarasci, A. Elia, Multi-word expressions in spoken language: Polisdict, in: Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-it 2018: 10-12 December 2018, Accademia University Press, Torino, 2018. doi:10.4000/books.aaccademia.3654.
[16] J. Monti, M. P. Di Buono, F. Sangati, Parseme-it corpus: An annotated corpus of verbal multiword expressions in italian, in: Proceedings of the Fourth Italian Conference on Computational Linguistics CLiC-it 2017: 11-12 December 2017, Rome [online], Accademia University Press, 2017, pp. 228–233. doi:10.4000/books.aaccademia.2433.
[17] A. Savary, C. Ramisch, B. Guillaume, A. Hawwari, A. Walsh, A. Fotopoulou, A. Bielinskienė, A. Estarrona, A. Gatt, A. Butler, A. Rademaker, A. Maldonado, A. Villavicencio, A. Farrugia, A. Muscat, A. Gatt, A. Antić, A. De Santis, A. Raffone, A. Riccio, A. Pascucci, A. Gurrutxaga, A. Bhatia, A. Vaidya, A. Miral, B. QasemiZadeh, B. Priego Sanchez, B. Griciūtė, B. Erden, C. Parra Escartín, C. Herrero, C. Carlino, C. Pasquer, C. Liebeskind, C. Wang, C. Ben Khelil, C. Bonial, C. Somers, C. Aceta, C. Krstev, E. Bejček, E. Lindqvist, E. Erenmalm, E. Palka-Binkiewicz, E. Rimkute, E. Petterson, F. Cap, F. Hu, F. Sangati, G. Wick Pedro, G. Speranza, G. Jagfeld, G. Blagus, G. Berk, G. Attard, G. Eryiğit, G. Finnveden, H. Martínez Alonso, H. de Medeiros Caseli, H. Elyovich, H. Xu, H. Xiao, I. Miranda, I. Jaknić, I. El Maarouf, I. Aduriz, I. Gonzalez, I. Matas, I. Stoyanova, I.-P. Jazbec, J. Busuttil, J. Waszczuk, J. Findlay, J. Bonnici, J. Šnajder, J.-Y. Antoine, J. Foster, J. Chen, J. Nivre, J. Monti, J. McCrae, J. Kovalevskaitė, K. Jain, K. Simkó, K. Yu, K. Azzopardi, K. Adalı, L. Uria, L. Zilio, L. Boizou, L. van der Plas, L. Galea, M. Sarlak, M. Buljan, M. Cherchi, M. Tanti, M. P. Di Buono, M. Todorova, M. Candito, M. Constant, M. Shamsfard, M. Jiang, M. Boz, M. Spagnol, M. Onofrei, M. Li, M. Elbadrashiny, M. Diab, M.-M. Rizea, N. Hadj Mohamed, N. Theoxari, N. Schneider, N. Tabone, N. Ljubešić, O. Vale, P. Cook, P. Yan, P. Gantar, R. Ehren, R. Fabri, R. Ibrahim, R. Ramisch, R. Walles, R. Wilkens, R. Urizar, R. Sun, R. Malka, S. A. Galea, S. Stymne, S. Louizou, S. Hu, S. Taslimipoor, S. Ratori, S. Srivastava, S. R. Cordeiro, S. Krek, S. Liu, S. Zeng, S. Yu, Š. Arhar Holdt, S. Markantonatou, S. Papadelli, S. Leseva, T. Kuzman, T. Kavčič, T. Lynn, T. Lichte, T. Pickard, T. Dimitrova, T. Yih, T. Güngör, T. Dinç, U. Iñurrieta, V. Tajalli, V. Stefanova, V. Caruso, V. Puri, V. Foufi, V. Barbu Mititelu, V. Vincze, V. Kovács, V. Shukla, V. Giouli, X. Ge, Y. Ha-Cohen Kerner, Y. Öztürk, Y. Yarandi, Y. Parmentier, Y. Zhang, Y. Zhao, Z. Urešová, Z. Yirmibeşoğlu, Z. Qin, Stank, M. Cristescu, B.-M. Zgreabăn, E.-A. Bărbulescu, R. Stanković, PARSEME corpora annotated for verbal multiword expressions (version 1.3), 2023. URL: http://hdl.handle.net/11372/LRT-5124, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.
[18] F. Masini, M. Micheli, A. Zaninello, S. Castagnoli, M. Nissim, Multiword expressions we live by: A validated usage-based dataset from corpora of written italian, in: Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020, Bologna, Italy, March 1-3, 2021, volume 2769, CEUR-WS.org, 2020. doi:10.4000/books.aaccademia.8710.
[19] F. Masini, M. S. Micheli, A. Zaninello, S. Castagnoli, M. Nissim, Mwe_combinet_release_1.0, 2020. URL: https://amsacta.unibo.it/id/eprint/6506/.
[20] C. Mauri, S. Ballarè, E. Goria, M. Cerruti, S. Francesco, Kiparla corpus: a new resource for spoken italian, in: Proceedings of the 6th Italian Conference on Computational Linguistics CLiC-it, CEUR-WS.org, 2019.
[21] E. Goria, C. Mauri, Il corpus kiparla: una nuova risorsa per lo studio dell'italiano parlato, in: CLUB Working Papers in Linguistics Volume 2, CLUB - Circolo Linguistico dell'Università di Bologna, 2018, pp. 96–116.
[22] M. Moneglia, S. Brown, F. Frontini, G. Gagliardi, F. Khan, M. Monachini, A. Panunzi, The IMAGACT visual ontology. an extendable multilingual infrastructure for the representation of lexical encoding of action, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), European Language Resources Association (ELRA), Reykjavik, Iceland, 2014, pp. 3425–3432.
[23] G. Gagliardi, Validazione dell'ontologia dell'azione IMAGACT per lo studio e la diagnosi del Mild Cognitive Impairment (MCI), Ph.D. thesis, Università degli Studi di Firenze, 2013.
[24] M. Moneglia, Il progetto ridire.it: un web corpus per l'accesso degli apprendenti l2 alla fraseologia italiana, in: Linguistica educativa: atti del XLIV Congresso internazionale di studi della Società di linguistica italiana (SLI): Viterbo, 27-29 settembre 2010, Bulzoni, 2012, pp. 411–423. doi:10.1400/202206.
[25] V. Lyding, E. Stemle, C. Borghetti, M. Brunello,

S. Castagnoli, F. Dell'Orletta, H. Dittmann, A. Lenci, V. Pirrelli, The PAISÀ corpus of Italian web texts, in: Proceedings of the 9th Web as Corpus Workshop (WaC-9), Association for Computational Linguistics, Gothenburg, Sweden, 2014, pp. 36–43. doi:10.3115/v1/W14-0406.

[26] C. Ramisch, A. Villavicencio, C. Boitet, mwetoolkit: a framework for multiword expression identification, in: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta, 2010.

[27] N. Okazaki, Crfsuite: a fast implementation of conditional random fields (crfs), 2007. URL: http://www.chokkan.org/software/crfsuite/.

[28] H. Schmid, Probabilistic part-of-speech tagging using decision trees, in: Proceedings of the International Conference on New Methods in Language Processing, 1994.

[29] A. Lenci, F. Masini, M. Nissim, S. Castagnoli, G. Lebani, L. Passaro, M. Senaldi, How to harvest word combinations from corpora: Methods, evaluation and perspectives, Studi e saggi linguistici 55 (2017) 45–68.

[30] S. Banerjee, T. Pedersen, The design, implementation, and use of the ngram statistics package, in: Computational Linguistics and Intelligent Text Processing, volume 2000, 2003, pp. 370–381. doi:10.1007/3-540-36456-0_38.

[31] T. Pedersen, S. Banerjee, B. McInnes, S. Kohli, M. Joshi, Y. Liu, The ngram statistics package (text::NSP) : A flexible tool for identifying ngrams, collocations, and word associations, in: Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World, Association for Computational Linguistics, Portland, Oregon, 2011, pp. 131–133.

[32] L. Ramshaw, M. Marcus, Text chunking using transformation-based learning, in: Third Workshop on Very Large Corpora, 1995.

Teasing LLMs adapted to Italian

Leonardo Ranaldi[1,2], Giulia Pucci[1], Elena Sofia Ruzzetti[1], Fabio Massimo Zanzotto[1] and André Freitas[2,3]

[1]Università degli studi Roma Tor Vergata, Italy
[2]Idiap Research Institute, Switzerland
[3]Department of Computer Science, University of Manchester, UK

Abstract
Instruction-tuned Large Language Models (It-LLMs) are changing NLP thanks to their easy accessibility. These models seem able to grasp language, solve complex tasks, and perform even with few resources. These abilities and ease of handling democratize their use, enabling many researchers to produce their homemade It-LLMs. However, the complete understanding of their potential needs to be improved due to the black-box nature of many models and the absence of holistic evaluation studies. We present an evaluation resource for It-LLMs tuned in Italian to address these challenges. Our proposal includes evaluating models on several aspects. We take a holistic approach to analyzing model performance factors, including the pre-training base, instruction-tuning data, and training methods. Our results reveal that data quality is the most crucial factor in scaling model performance. While available open-source models demonstrate impressive ability, they present problems when customized adapters are used. We are encouraged by the rapid development of models by the open-source community. However, we also highlight the need for rigorous evaluation to support the claims.

Keywords
Instruction-tuned Large Language Models, Multilingual LLMs,

1. Introduction

The advent of Instruction-tuned Large Language Models (It-LLMs) marks yet another change in NLP in the last few decades. Indeed, their abilities are evident in numerous applications, from complex problem-solving to information retrieval to conversational assistants such as ChatGPT. Examples include GPT-4, which demonstrates abilities in language comprehension and common sense, logical-mathematical problem solving, law, and medicine. However, despite their remarkable competence and adaptability, the full extent of their potential has yet to be fully understood. Indeed, their direction is poorly captured, given many models' simple use, black-box nature, and lack of in-depth and holistic evaluation studies [1, 2, 3].

To manage these challenges and deeper understand the abilities of these models, a series of evaluation benchmarks were introduced that are explicitly designed for the comprehensive evaluation of It-LLMs [4, 5, 6, 7, 8, 9]. However, evaluation resources are only available in English, and it is tricky and misleading to evaluate a model trained on instructions in the Italian language.

In this paper, we propose evaluation resources for Italian It-LLMs. Furthermore, we tested a set of open-source It-LLMs fine-tuned in the Italian language, demonstrating excellent adaptability and some gaps in downstream performance. In particular, our methodology, applying a systematic and holistic approach, examines the problem-solving ability, writing ability, and alignment between languages of customized It-LLMs that are fine-tuned in a specific language, i.e., Italian, starting from the work proposed by Chia et al. [5]. Through a rigorous exploration of these factors, we seek to shed light on the vital elements that determine the performance of the models, facilitating an understanding of how these models can best be harnessed to meet our needs. Our contribution is fully available and open-source [1]

2. The Open-Source Instructed LLMs

Large Language Models (LLMs) have caught mainstream attention; they have become a comprehensive category of models. LLMs are comprehended as pre-trained and fine-tuned models with general language prompts or Instruction-tuned models. Therefore, we distinguish between basic and Instructed models, where basic LLMs are pre-trained LLMs that can be fine-tuned on instructions to become Instruction-tuned LLMs (It-LLMs). In particular, in Table 1, we summarize mainly open-source

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
name.username@idiap.ch (L. Ranaldi);
name.username@uniroma2.it (G. Pucci);
name.username@uniroma2.it (E. S. Ruzzetti);
name.username@uniroma2.it (F. M. Zanzotto);
name.username@idiap.ch (A. Freitas)

[1]**https://github.com/LeonardRanaldi/italian-instruct-eval**

Model	*Architecture*	*num. Tokens*	*Source*
Llama	Decoder	1.4T	Unknown∗
Llama-2	Decoder	2.4T	Unknown∗
OPT	Decoder	180B	The Pile
BLOOM	Decoder	250B	Unknown∗
T5	Encoder-Decoder	1T	C4

Table 1
Open-source Large Language Models, with ∗ we denote data dump not available.

Model	**Backbone**	**Size**	**Source**	**Training**
Alpaca [11]	LLaMA	7-30B	Alpaca Data	Supervised
Baize [13]	LLaMA	7-30B	Self-Chat Data	Supervised
Vicuna [14]	LLaMA	7-33B	ShareGPT	Supervised
Falcon [15]	LLaMA	7-40B	Refinedweb	Supervised
ChatGLM [16]	GLM	6B	Unknown	RLHF
Customized Adapter				
Camoscio [17]	LLaMA	7B	Alpaca(Italian)	Supervised
Stambecco [18]	LLaMA	7-13B	Alpaca(Italian)	Supervised
Fauno [19]	LLaMA	7-13B	Baize data(italian)	Supervised

Table 2
Details of open-source instructed LLMs.

LLMs due to the need for more transparency and reproducibility of closed-source models.

The essential part of the Instruction-tuning idea is the data used to train LLMs. Indeed, factors such as quality, quantity, and format can determine the behavior of the instructed model. Table 3 presents several open-source resources. There is a growing tendency to exploit synthetic instruction data from closed-source models [10, 11]. While this practice can allow instructed models to mimic the behavior of closed-source models, it can also lead to problems such as the inheritance of the black-box nature of closed-source models and instability due to noisy synthetic instructions [12].

Finally, a holistic overview of the instructed open-source models can be found in Table 2, where the basic model with dimensions, instruction dataset, and training method for each It-LLMs is given. We observe a variety of model sizes and datasets. Therefore, this overview of open-source instructed LLMs provides comprehensive factors for evaluation and analysis.

3. Challenges & Methods in Evaluating Intruction-tuned LLMs

3.1. Background and Challenges

The highest wall in evaluating LLMs is the closed-source concept, where creators often hide model details, instruction datasets, and training methods. Such models thus lead to a knowledge vacuum in the research community as it is impossible to rigorously analyze the reasons for their performance.

On the other side of the coin is an ongoing open-source development that aims to democratize language model technology. While these efforts are highly encouraged, the pace of development of new models can outpace advances in evaluation studies. Unfortunately, informal evaluations often spot new models, which must be clarified when comparing different models.

We should consider different factors, such as pre-training and instruction data, to arrive at a holistic understanding of LLMs and It-LLMs. While previous work has conducted in-depth studies in some areas, such as datasets [20] and more concrete, such as general benchmarks [21], other factors should be considered to achieve a complete understanding. For example, performances on customized models for particular languages or tasks.

Recent work shows the elasticity and customization of It-LLMs in many languages. Santilli and Rodolà [17] translated Alpaca [11] into Italian by proposing Camoscio. Later, in Stambecco, the author [18] reproduced the same work by modifying some parameters. In [19], the models of the Baize [13] family were adapted into Italian. In this new scenario, evaluation has become increasingly important and challenging. Recent evaluation studies produce concrete results such as accuracy and precision [5, 22]. However, these methodologies are generic and not customized for a specific task and language.

Model	*Size*	*Domain*	*Source*
Alpaca Data [11]	52K	General	GPT-3
Self-Instruct [10]	52K	General	Human-Annotation
ShareGPT [14]	70K	Dialogue	ChatGPT
Self-Chat [13]	100k	Dialogue	ChatGPT

Table 3
Open-source Instruction-tuning datasets.

Finally, Ranaldi et al. [4], generalizing previous work, proposed a cross-lingual approach by eliciting It-LLMs with multilingual Alpaca empowered with translation-following demonstrations.

In this paper, we propose an Italian evaluation method for Italian fine-tuned It-LLMs. Our method is based on various general skills and usage scenarios applicable to It-LLMs adapted.

3.2. Proposed Methods

We propose to translate three well-known resources to evaluate the abilities of several Intruction-tuned Large Language Models. To perform well, the adapted models should have inherited world awareness, multi-hop reasoning, and more, merely like the original models. These benchmarks are:

Massive Multitask Language Understanding (MMLU) [23] measures knowledge of the world and problem-solving problems in multiple subjects with 57 subjects across STEM, humanities, social sciences, and other areas.

	Model	MMLU		BBH		DROP	
		Acc.	δ	Acc.	δ	Acc.	δ
Original data	∗Alpaca-Lora 7B	35.6	-	30.7	-	27.5	-
	♣Alpaca-Lora 13B	50.9	-	32.6	-	31.8	-
	+ Alpaca-Lora 30B	58.4	-	41.3	-	45.1	-
	∘Baize 7B	43.5	-	45.6	-	53.8	-
	⋄Baize 13B	50.9	-	49.5	-	56.4	-
	- Baize 30B	59.8	-	64.6	-	69.8	-
Italian data	∗Alpaca-Lora 7B	35.1	-0.5	30.1	-0.6	26.9	-0.6
	♣Alpaca-Lora 13B	50.6	-0.3	32.1	-0.5	31.6	-0.2
	+ Alpaca-Lora 30B	57.9	-0.5	41.1	-0.2	44.9	-0.2
	∘Baize 7B	44.3	-0.8	46.3	-0.7	54.5	-0.7
	⋄Baize 13B	51.2	-0.3	49.8	-0.6	57.2	-0.8
	- Baize 30B	59.5	-0.5	65.2	-0.4	70.1	-0.3
	Italian Adapters 7B	–	–	–		–	–
	∗Camoscio 7B	30.2	**-5.4**	29.8	*-0.8*	22.0	**-4.5**
	∗Stambecco 7B	28.2	**-7.4**	29.7	*-0.9*	21.6	**-5.9**

Table 4
Evaluation results. We denote the accuracy across the benchmarks as Acc., while δ denotes the performance change compared to the original version trained and evaluated on English datasets.

Discrete Reasoning Over Paragraphs (DROP) [24] reading comprehension on mathematics where the model should perform discrete reasoning on passages extracted from Wikipedia articles.

BIG-Bench Hard (BBH) [25] is a subset of challenging tasks related to navigation, logical deduction, and fallacy detection.

Evaluation Each benchmark was translated into Italian using google API [2]. Then, zero-shot evaluations were done for the original version in English and ours in Italian using the framework proposed in [5] [3].

4. Results

Customized Instruction-tuned Large Language Models (It-LLMs) need further refinement. This statement is supported by the results shown in Table 4 on fine-tuned models over the Italian benchmarks. Firstly, the original Alpaca-Lora and Baize evaluated on the English benchmark outperformed Camoscio, Stambecco, and Fauno evaluated on the Italian benchmark.

Secondly, the differences between Camoscio, Stambecco, Fauno, and the original Alpaca-Lora and Baize are very close on the Italian benchmarks (Italian Data on Table 4). Thirdly, models with more parameters (30B) performed best, and the δ between performances on English-language benchmarks are remarkably lower than models with fewer parameters. In conclusion, fine-tuning a customized resource, in this case, customized English language resources, was insufficient to increase performance. This phenomenon may be due to the quality of the data used for homemade fine-tuning and also suggests that fine-tuning on custom It-LLMs may have inserted a bias. These gaps should be further investigated, and the scientific community should pay more attention.

5. Conclusions

In this paper, we have presented a systematic evaluation of four resources for Instruction-tuned Large Language Models (It-LLMs). Our holistic approach analyzed critical performance factors and showed that efforts to customize It-LLMs are not always rewarded by performance.

Underlining the importance of the contribution of the open-source community in proposing new solutions to meet specific needs. We emphasize the significance of data quality in scaling model performance. Additionally, our translated benchmarks provide valuable insights into the adaptability and effectiveness of It-LLMs for specific language tasks. By addressing key evaluation challenges, our work contributes to the responsible and effective utilization of It-LLMs, fostering further advancements in NLP.

In future developments, we will investigate lightweight approaches to elicit adapters' multi- and cross-lingual skills inspired by what has been done in [4, 9]

[2] available here https://github.com/LeonardRanaldi/italian-instruct-eval
[3] https://github.com/declare-lab/instruct-eval

References

[1] L. Ranaldi, E. S. Ruzzetti, F. M. Zanzotto, Precog: Exploring the relation between memorization and performance in pre-trained language models, 2023. arXiv:2305.04673.

[2] L. Ranaldi, E. S. Ruzzetti, D. Venditti, D. Onorati, F. M. Zanzotto, A trip towards fairness: Bias and de-biasing in large language models, 2023. arXiv:2305.13862.

[3] L. Ranaldi, A. Nourbakhsh, A. Patrizi, E. S. Ruzzetti, D. Onorati, F. Fallucchi, F. M. Zanzotto, The dark side of the language: Pre-trained transformers in the darknet, 2022. arXiv:2201.05613.

[4] L. Ranaldi, G. Pucci, A. Freitas, Empowering cross-lingual abilities of instruction-tuned large language models by translation-following demonstrations, 2023. arXiv:2308.14186.

[5] Y. K. Chia, P. Hong, L. Bing, S. Poria, Instructeval: Towards holistic evaluation of instruction-tuned large language models, 2023. arXiv:2306.04757.

[6] L. Ranaldi, G. Pucci, F. M. Zanzotto, Modeling easiness for training transformers with curriculum learning, in: R. Mitkov, G. Angelova (Eds.), Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing, INCOMA Ltd., Shoumen, Bulgaria, Varna, Bulgaria, 2023, pp. 937–948. URL: https://aclanthology.org/2023.ranlp-1.101.

[7] Y. Chang, X. Wang, J. Wang, Y. Wu, K. Zhu, H. Chen, L. Yang, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, X. Xie, A survey on evaluation of large language models, 2023. arXiv:2307.03109.

[8] L. Ranaldi, G. Pucci, Knowing knowledge: Epistemological study of knowledge in transformers, Applied Sciences 13 (2023). URL: https://www.mdpi.com/2076-3417/13/2/677. doi:10.3390/app13020677.

[9] L. Ranaldi, F. M. Zanzotto, Empowering multi-step reasoning across languages via tree-of-thoughts, 2023. arXiv:2311.08097.

[10] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, H. Hajishirzi, Self-instruct: Aligning language models with self-generated instructions, 2023. arXiv:2212.10560.

[11] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, T. B. Hashimoto, Stanford alpaca: An instruction-following llama model, https://github.com/tatsu-lab/stanford_alpaca, 2023.

[12] Z. Shao, Y. Gong, Y. Shen, M. Huang, N. Duan, W. Chen, Synthetic prompting: Generating chain-of-thought demonstrations for large language models, arXiv preprint arXiv:2302.00618 (2023).

[13] C. Xu, D. Guo, N. Duan, J. McAuley, Baize: An open-source chat model with parameter-efficient tuning on self-chat data, arXiv preprint arXiv:2304.01196 (2023).

[14] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, E. P. Xing, Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. URL: https://lmsys.org/blog/2023-03-30-vicuna/.

[15] E. Almazrouei, H. Alobeidli, A. Alshamsi, A. Cappelli, R. Cojocaru, M. Debbah, E. Goffinet, D. Heslow, J. Launay, Q. Malartic, B. Noune, B. Pannier, G. Penedo, Falcon-40B: an open large language model with state-of-the-art performance (2023).

[16] A. Zeng, X. Liu, Z. Du, Z. Wang, H. Lai, M. Ding, Z. Yang, Y. Xu, W. Zheng, X. Xia, W. L. Tam, Z. Ma, Y. Xue, J. Zhai, W. Chen, P. Zhang, Y. Dong, J. Tang, Glm-130b: An open bilingual pre-trained model, 2022. arXiv:2210.02414.

[17] A. Santilli, E. Rodolà, Camoscio: an italian instruction-tuned llama, 2023. arXiv:2307.16456.

[18] Michael, Stambecco: Italian instruction-following llama model, https://github.com/mchl-labs/stambecco, 2023.

[19] A. Bacciu, G. Trappolini, A. Santilli, E. Rodolà, F. Silvestri, Fauno: The italian large language model that will leave you senza parole!, 2023. arXiv:2306.14457.

[20] S. Longpre, L. Hou, T. Vu, A. Webson, H. W. Chung, Y. Tay, D. Zhou, Q. V. Le, B. Zoph, J. Wei, A. Roberts, The flan collection: Designing data and methods for effective instruction tuning, 2023. arXiv:2301.13688.

[21] W. Zhong, R. Cui, Y. Guo, Y. Liang, S. Lu, Y. Wang, A. Saied, W. Chen, N. Duan, Agieval: A human-centric benchmark for evaluating foundation models, 2023. arXiv:2304.06364.

[22] J. Sun, C. Shaib, B. C. Wallace, Evaluating the zero-shot robustness of instruction-tuned language models, 2023. arXiv:2306.11270.

[23] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, J. Steinhardt, Measuring mathematical problem solving with the math dataset, 2021. arXiv:2103.03874.

[24] D. Dua, Y. Wang, P. Dasigi, G. Stanovsky, S. Singh, M. Gardner, DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 2368–2378. URL: https://aclanthology.org/N19-1246. doi:10.18653

/v1/N19-1246.

[25] M. Suzgun, N. Scales, N. Schärli, S. Gehrmann, Y. Tay, H. W. Chung, A. Chowdhery, Q. V. Le, E. H. Chi, D. Zhou, J. Wei, Challenging big-bench tasks and whether chain-of-thought can solve them, 2022. arXiv:2210.09261.

Investigating Gender Bias in Large Language Models for the Italian Language

Elena Sofia Ruzzetti[1] *, Dario Onorati[1,2] *, Leonardo Ranaldi[1,3], Davide Venditti[1] and Fabio Massimo Zanzotto[1]

[1]*University of Rome Tor Vergata*
[2]*Sapienza University of Rome*
[3]*Idiap Research Institute*

Abstract
English. Large Language Models (LLMs) are becoming increasingly flexible and reliable: the large pre-training phase enables them to capture a large number of real-world linguistic phenomena. However, pre-training on large amounts of data can also cause the representation of harmful biases. In this paper, we propose a method for identifying the presence of gender bias using a list of occupations characterized by a large imbalance between the number of male and female employees.
Italian. I Large Language Models (LLMs) stanno diventando sempre più flessibili e affidabili: l'ampia fase di pre-training consente di catturare un gran numero di fenomeni linguistici del mondo reale. Tuttavia, il pre-training su grandi quantità di dati può causare la rappresentazione di pregiudizi dannosi. In questo lavoro, proponiamo un metodo per identificare la presenza dei pregiudizi di genere utilizzando un elenco di occupazioni caratterizzate da un forte squilibrio tra il numero di dipendenti di sesso maschile e femminile.

Keywords
Gender Bias, Prejudice, LLM

1. Introduction

Large language models (LLMs) have achieved super-human performances in several NLP applications [1, 2]. They demonstrate a clear upward performance trend along with the increasing model size and pre-training data, namely scaling law [3]. However, by over-humanizing learning abilities, it is possible that these LLMs inherit stereotypical associations between social groups and professions [4, 5].

Bias or, better, *prejudice* [6] is the sword of Damocles of fairness in many data-driven applications, such as facial recognition [7] or recommendation systems [8]. Even in modern NLP, a clear presence of bias in different models has been observed. Bolukbasi et al. [4] detected the presence of stereotypical biases in word embedding vectors measuring association between gender and certain professions, while Caliskan et al. [9] proposed the Word Embedding Association Tests (WEAT) to assess the strength of stereotypical associations regarding gender and races. Similar biases were later observed in Pre-trained Language Models. Several benchmarks like SEAT [10], StereoSet [11] and CrowS-Pairs [12] enables to test Pre-trained Language Models like BERT [13] and ELMo [14].

The advent of LLMs [1, 15, 16, 17] has yet to alleviate this phenomenon. In fact, despite the increasing capabilities of this family of models, the underlying LLMs generate toxic or offensive content [18, 19], and reproduce biases that exist in the training data [20, 21, 22]. While some models can be used in beneficial application, like in identifying biased texts [23], biases hidden within these models could hinder their abilities [6]. For these reasons, while some previous work quantifies the ability of the models to cause no harm by interviewing human evaluators [24], it is necessary to develop automated approaches to easily test models before they are available to a large group of users.

In this paper, we analyze how existing LLMs capture some known stereotyped associations between gender and profession for the Italian Language. To quantify the presence of social bias, we created a test dataset (Section 2.1) that allows us to monitor the relation between gender and 171 different occupations. We selected professions that, according to ISTAT data, have a significant imbalance in the number of male employees compared to the number of female employees in Italy. Then, we propose a method to measure the strength of the association between gender and profession (Section 2.2) on different LLMs. Stemming from the stereotype score definition that can be found in Nadeem et al. [11], we define a model biased as it systematically prefers the stereotyped association over an anti-stereotyped one. Finally, we test several LLMs trained on the Italian language and attest that a large number of LLMs available for the Italian

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
elena.sofia.ruzzetti@uniroma2.it (E. S. Ruzzetti); dario.onorati@uniroma1.it (D. Onorati); fabio.massimo.zanzotto@uniroma2.it (F. M. Zanzotto)

* These authors contributed equally to this work

Macro Category	Category Description	tot
1	Legislatori, Imprenditori e Alta dirigenza	18
2	Professioni intellettuali, scientifiche e di elevata specializzazione	28
3	Professioni tecniche	22
6	Artigiani, Operai specializzati e agricoltori	83
7	Conduttori di impianti, operai di macchinari fissi e mobili e conducenti di veicoli	6
8	Professioni non qualificate	9
9	Forze armate	5

Table 1
Number of professions included in the dataset over the macro-categories defined in CP2011

language have strong gender biases (Section 3).

2. Methods and Data

Motivated by the necessity of quantifying biases in Large Language Models (LLMs), we first present a novel dataset derived from ISTAT (Section 2.1) and then describe a measure to evaluate the association between gender and occupation in the Italian language 2.2.

2.1. Resource description

We define a list of professions that are characterized by a high rate of gender disparity between men and women, which exceeds the average rate by at least 25 percent, according to the Italian Ministry of Labour and Social Policies (Ministero del lavoro e delle politiche sociali) based on ISTAT data on the annual average in 2021.

Specifically, given the list of sectors in which greater inequality was identified, we compile a list of occupations from the classification of occupations defined by ISTAT, called CP2011. CP2011 defines five levels of occupation aggregation; each group can be most generic (close to one digit) or most detailed (close to five digits). We will refer to the most general classification as the macro-categories that the professions we analyze cover.

To collect the actual occupations for which a large number of employees are male, we relied on the more specific classification of CP2011, identified by five digits. However, since the denomination used by ISTAT is formal, three annotators simplified the five-digit classification by reducing the profession name to a maximum of three words, taking into account the description of the category and the name itself. Simplification with a maximum of three words was retained only if all annotators rated it as valid; that is, it was discarded even if only one annotator disagreed with the others about the validity of the simplification.

https://www.lavoro.gov.it/notizie/pagine/settori-e-professioni-caratterizzati-da-tasso-di-disparita-uomo-donna-16112022
https://www.istat.it/it/archivio/18132

Hence, a list of 171 professions is obtained. We will refer to this resource as Jobs. In Table 1, the macro-categories and the number of jobs for each category are presented, while a fine-grained description can be found in the Appendix A.1. The complete list of professions after the simplification step is available in Appendix A.2.

2.2. Bias Measure

Given the professions for which the number of employees is highly imbalanced between men and women, our aim is to determine the presence of bias in LLM for these professions. We define bias in these models as a systematic preference for stereotyped associations over anti-stereotyped ones [11]. Given a profession J in Jobs, to estimate the preference of a model to associate J to a certain gender $G \in \{M, F\}$, we aim to measure the two probabilities $p(M|J)$ and $p(F|J)$ and compare them. A model is biased if it systematically assigns

$$p(M|J) > p(F|J)$$

for the professions J in Jobs.

However, a model could be negatively influenced by the frequency of generally unused professions name, like *ingegnera* that, despite being an existing word in the Italian language, is much less used than its male counterpart *ingegnere*. Hence, to estimate the probabilities of $p(G|J)$ given a gender G and a profession J, we can measure the probability of generating a certain gender G as the next word in template sentences like *"J è una professione da G"*. Since in Italian all nouns have a gender, we associate each gender G with the profession J_G with the correct suffix. For example, given a job J like *imprenditore*, J_M represents a profession that refers to a male term, such as *imprenditore*, whereas J_F refers to a female term as *imprenditrice*. Thus, to estimate the association between a job such as *imprenditore* and the two genders, in principle, one should test the two probabilities *p(uomo | imprenditore)* and *p(donna | imprenditrice)*. However, a model could be confused by a rare profession name J_F. To address this issue, we then compute the probability of $p(G|J)$ as the sum of two probabilities: $p(G|J_M)$ and $p(G|J_F)$, or, formally,

$$p(G|J) = p(G|J_M) + p(G|J_F)$$

The grammatically incorrect version of the sentence will tend to have a low probability, except in cases like *p(donna | "ingegnere è una professione da")*, that then can be fairer compared with the probabililty of *p(uomo|ingegnere)*.

Finally, given the list of professions Jobs previously introduced and the estimate of the probabilities for $p(G|J)$ we compute the *bias score* σ as:

$$\sigma = \frac{\sum_{J \in \text{Jobs}} score(J, M, F)}{|\text{Jobs}|} \quad (1)$$

	Macro Category bias score σ							
Model	**1**	**2**	**3**	**6**	**7**	**8**	**9**	**Averaged**
GePpeTto	0.056	0.857	0.727	0.325	0.167	0.222	0.6	0.433
BLOOM-560m	0.778	1.00	0.909	0.952	1.00	1.00	0.8	0.936
BLOOM-1b1	0.944	1.00	0.636	1.00	1.00	1.00	1.00	0.947
BLOOM-7b1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
LLaMA-7b	0.667	0.964	1.00	0.819	1.00	0.778	1.00	0.86
LLaMA-13b	1.00	1.00	1.00	0.976	1.00	1.00	1.00	0.988
XGLM-564M	0.944	1.00	0.955	0.916	1.00	0.889	1.00	0.942
XGLM-1.7B	0.500	0.643	0.955	0.807	0.667	0.778	1.00	0.766
XGLM-2.9B	0.444	0.464	0.591	0.614	0.333	0.667	0.8	0.567
XGLM-4.5B	0.278	0.821	0.955	0.663	0.833	0.444	0.6	0.678
XGLM-7.5B	0.222	0.464	0.864	0.711	0.500	0.222	0.6	0.602
`ISTAT score`	0.705	0.788	0.832	0.882	0.829	0.640	0.966	0.806

Table 2
Bias Score of different models across all the different macro-categories. For comparison reasons, we also report the actual percentage of male employees according to the ISTAT data published by the Italian Ministry of Labour and Social Policies.

Table 3
Number of parameters (B for billion and M for million) for the LLMs used in the work.

Model	Params
GePpeTto [25]	117M
LLaMA [17]	7B, 13B
BLOOM [15]	560M, 1.1B, 7.1B
XGLM [26]	564M, 1.7B, 2.9B, 4.5B, 7.5B

where:

$$score(J, M, F) = \begin{cases} +1 & p(M|J) > p(F|J) \\ 0 & otherwise \end{cases}$$

Hence, σ allows quantifying the bias in a model: an unbiased model has a *bias score* or 0.5 while a biased one has a score close to 1 (if it behaves stereotypically) or 0 (anti-stereotypically).

3. Experiments

In this Section we propose a comprehensive analysis with the aim of evaluating the presence of bias in Large Language Models (LLMs). In Section 3.1, we introduce the analyzed models and how we compute the probabilities described in Section 2.2 to estimate the bias of the models. Finally, in Section 3.2 we identify models affected by bias across the different macro categories defined in `CP2011`.

3.1. Experimental Set-up

We evaluate the social bias between occupation and gender on four different Large Language Models with different versions: LLaMA [17], BLOOM [15], XGLM [26] and GePpeTto [25], an Italian GPT-2 model. In order to evaluate the correlation between bias and the number of parameters of a model, different versions of LLaMA, BLOOM, and XGLM are considered. A detailed list of models and the number of parameters can be found in the Table 3 Since all these models are generative models, each of them is asked to compute the probability of the last word between two possible choices, in which each word represents a gender G. To obtain a more robust estimate of $p(G|J)$, the probability of this last token is computed with three different but semantically equivalent prompts. Moreover, for each gender, we test two different words denoting the gender G. Hence, $p(G|J)$ is estimated as the average of six semantically equivalent sentences.

3.2. Quantifying Bias in LLMs

Nearly every model is subject to a strong bias (see Table 2). In particular, the majority of models have a strong stereotypical behavior and associate the professions in Jobs with men rather than women: we can observe that the average bias score is close to 1 for models in the BLOOM family as well as for the larger LLaMa models. On average, the larger models in the XGLM family tend to demonstrate less bias but, with the exception of XGLM-2.9B, are still far from the ideal σ of 0.5. On the other hand, GePpeTto demonstrates a slightly anti-stereotypical behavior: however, it still exhibits strong biases in the scientific and technical professions (Macro Category 2 and 3, respectively) and stereotypically associates males with these professions. We can also observe that strong biases on Macro Categories 2 and 3 are registered in other models (and especially LLaMA): they exhibit strong biases on these categories even when other categories are less biased.

In contrast to some previous work that correlates the model bias with the number of parameters [11], here we can observe mixed results: this correlation can be observed in BLOOM and LLaMA, while a negative correlation can be observed in the XGLM case, since as the number of parameters increases, the bias decreases. Hence, the correlation between language model capabilities and bias presence needs to be further explored.

4. Conclusions

In this work, we propose to investigate gender bias for the Italian language in pre-trained LLMs, identifying if and to what extent these models capture real-world imbalances while training on text data. We present a list of professions inspired by official ISTAT data and propose a simple and effective method to quantify the presence of gender bias in occupations. We assess the presence of strong biases across different model families such as BLOOM, LLaMA, and XGLM.

References

[1] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, 2020. arXiv:2005.14165.

[2] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, 2023. arXiv:2201.11903.

[3] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, D. Amodei, Scaling laws for neural language models, 2020. arXiv:2001.08361.

[4] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, A. Kalai, Man is to computer programmer as woman is to homemaker? debiasing word embeddings, 2016. arXiv:1607.06520.

[5] M. Bartl, M. Nissim, A. Gatt, Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias, in: Proceedings of the Second Workshop on Gender Bias in Natural Language Processing, Association for Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 1–16. URL: https://aclanthology.org/2020.gebnlp-1.1.

[6] M. Mastromattei, L. Ranaldi, F. Fallucchi, F. Zanzotto, Syntax and prejudice: ethically-charged biases of a syntax-based hate speech recognizer unveiled, PeerJ Computer Science (2022). URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85125058825&doi=10.7717%2fpeerj-cs.859&partnerID=40&md5=87e1288c4534e9bfa93078e4d8a0c7c8. doi:10.7717/peerj-cs.859.

[7] A. Shankar, A. McMunn, P. Demakakos, M. Hamer, A. Steptoe, Social isolation and loneliness: Prospective associations with functional status in older adults., Health Psychology 36 (2017) 179–187. URL: https://doi.org/10.1037/hea0000437. doi:10.1037/hea0000437.

[8] F. Ding, M. Hardt, J. Miller, L. Schmidt, Retiring adult: New datasets for fair machine learning, 2022. arXiv:2108.04884.

[9] A. Caliskan, J. J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases, Science 356 (2017) 183–186. URL: https://www.science.org/doi/abs/10.1126/science.aal4230. doi:10.1126/science.aal4230.

[10] C. May, A. Wang, S. Bordia, S. R. Bowman, R. Rudinger, On measuring social biases in sentence encoders, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 622–628. URL: https://aclanthology.org/N19-1063. doi:10.18653/v1/N19-1063.

[11] M. Nadeem, A. Bethke, S. Reddy, StereoSet: Measuring stereotypical bias in pretrained language models, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 5356–5371. URL: https://aclanthology.org/2021.acl-long.416. doi:10.18653/v1/2021.acl-long.416.

[12] N. Nangia, C. Vania, R. Bhalerao, S. R. Bowman, CrowS-pairs: A challenge dataset for measuring social biases in masked language models, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 1953–1967. URL: https://aclanthology.org/2020.emnlp-main.154. doi:10.18653/v1/2020.emnlp-main.154.

[13] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19

-1423. doi:10.18653/v1/N19-1423.

[14] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: Proc. of NAACL, 2018.

[15] B. Workshop, :, T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, J. Tow, A. M. Rush, S. Biderman, A. Webson, P. S. Ammanamanchi, T. Wang, B. Sagot, N. Muennighoff, A. V. del Moral, O. Ruwase, R. Bawden, S. Bekman, A. McMillan-Major, I. Beltagy, H. Nguyen, L. Saulnier, S. Tan, P. O. Suarez, V. Sanh, H. Laurençon, Y. Jernite, J. Launay, M. Mitchell, C. Raffel, A. Gokaslan, A. Simhi, A. Soroa, A. F. Aji, A. Alfassy, A. Rogers, A. K. Nitzav, C. Xu, C. Mou, C. Emezue, C. Klamm, C. Leong, D. van Strien, D. I. Adelani, D. Radev, E. G. Ponferrada, E. Levkovizh, E. Kim, E. B. Natan, F. D. Toni, G. Dupont, G. Kruszewski, G. Pistilli, H. Elsahar, H. Benyamina, H. Tran, I. Yu, I. Abdulmumin, I. Johnson, I. Gonzalez-Dios, J. de la Rosa, J. Chim, J. Dodge, J. Zhu, J. Chang, J. Frohberg, J. Tobing, J. Bhattacharjee, K. Almubarak, K. Chen, K. Lo, L. V. Werra, L. Weber, L. Phan, L. B. allal, L. Tanguy, M. Dey, M. R. Muñoz, M. Masoud, M. Grandury, M. Šaško, M. Huang, M. Coavoux, M. Singh, M. T.-J. Jiang, M. C. Vu, M. A. Jauhar, M. Ghaleb, N. Subramani, N. Kassner, N. Khamis, O. Nguyen, O. Espejel, O. de Gibert, P. Villegas, P. Henderson, P. Colombo, P. Amuok, Q. Lhoest, R. Harliman, R. Bommasani, R. L. López, R. Ribeiro, S. Osei, S. Pyysalo, S. Nagel, S. Bose, S. H. Muhammad, S. Sharma, S. Longpre, S. Nikpoor, S. Silberberg, S. Pai, S. Zink, T. T. Torrent, T. Schick, T. Thrush, V. Danchev, V. Nikoulina, V. Laippala, V. Lepercq, V. Prabhu, Z. Alyafeai, Z. Talat, A. Raja, B. Heinzerling, C. Si, D. E. Taşar, E. Salesky, S. J. Mielke, W. Y. Lee, A. Sharma, A. Santilli, A. Chaffin, A. Stiegler, D. Datta, E. Szczechla, G. Chhablani, H. Wang, H. Pandey, H. Strobelt, J. A. Fries, J. Rozen, L. Gao, L. Sutawika, M. S. Bari, M. S. Al-shaibani, M. Manica, N. Nayak, R. Teehan, S. Albanie, S. Shen, S. Ben-David, S. H. Bach, T. Kim, T. Bers, T. Fevry, T. Neeraj, U. Thakker, V. Raunak, X. Tang, Z.-X. Yong, Z. Sun, S. Brody, Y. Uri, H. Tojarieh, A. Roberts, H. W. Chung, J. Tae, J. Phang, O. Press, C. Li, D. Narayanan, H. Bourfoune, J. Casper, J. Rasley, M. Ryabinin, M. Mishra, M. Zhang, M. Shoeybi, M. Peyrounette, N. Patry, N. Tazi, O. Sanseviero, P. von Platen, P. Cornette, P. F. Lavallée, R. Lacroix, S. Rajbhandari, S. Gandhi, S. Smith, S. Requena, S. Patil, T. Dettmers, A. Baruwa, A. Singh, A. Cheveleva, A.-L. Ligozat, A. Subramonian, A. Névéol, C. Lovering, D. Garrette, D. Tunuguntla, E. Reiter, E. Taktasheva, E. Voloshina, E. Bogdanov, G. I. Winata, H. Schoelkopf, J.-C. Kalo, J. Novikova, J. Z. Forde, J. Clive, J. Kasai, K. Kawamura, L. Hazan, M. Carpuat, M. Clinciu, N. Kim, N. Cheng, O. Serikov, O. Antverg, O. van der Wal, R. Zhang, R. Zhang, S. Gehrmann, S. Mirkin, S. Pais, T. Shavrina, T. Scialom, T. Yun, T. Limisiewicz, V. Rieser, V. Protasov, V. Mikhailov, Y. Pruksachatkun, Y. Belinkov, Z. Bamberger, Z. Kasner, A. Rueda, A. Pestana, A. Feizpour, A. Khan, A. Faranak, A. Santos, A. Hevia, A. Unldreaj, A. Aghagol, A. Abdollahi, A. Tammour, A. HajiHosseini, B. Behroozi, B. Ajibade, B. Saxena, C. M. Ferrandis, D. Contractor, D. Lansky, D. David, D. Kiela, D. A. Nguyen, E. Tan, E. Baylor, E. Ozoani, F. Mirza, F. Ononiwu, H. Rezanejad, H. Jones, I. Bhattacharya, I. Solaiman, I. Sedenko, I. Nejadgholi, J. Passmore, J. Seltzer, J. B. Sanz, L. Dutra, M. Samagaio, M. Elbadri, M. Mieskes, M. Gerchick, M. Akinlolu, M. McKenna, M. Qiu, M. Ghauri, M. Burynok, N. Abrar, N. Rajani, N. Elkott, N. Fahmy, O. Samuel, R. An, R. Kromann, R. Hao, S. Alizadeh, S. Shubber, S. Wang, S. Roy, S. Viguier, T. Le, T. Oyebade, T. Le, Y. Yang, Z. Nguyen, A. R. Kashyap, A. Palasciano, A. Callahan, A. Shukla, A. Miranda-Escalada, A. Singh, B. Beilharz, B. Wang, C. Brito, C. Zhou, C. Jain, C. Xu, C. Fourrier, D. L. Periñán, D. Molano, D. Yu, E. Manjavacas, F. Barth, F. Fuhrimann, G. Altay, G. Bayrak, G. Burns, H. U. Vrabec, I. Bello, I. Dash, J. Kang, J. Giorgi, J. Golde, J. D. Posada, K. R. Sivaraman, L. Bulchandani, L. Liu, L. Shinzato, M. H. de Bykhovetz, M. Takeuchi, M. Pàmies, M. A. Castillo, M. Nezhurina, M. Sänger, M. Samwald, M. Cullan, M. Weinberg, M. D. Wolf, M. Mihaljcic, M. Liu, M. Freidank, M. Kang, N. Seelam, N. Dahlberg, N. M. Broad, N. Muellner, P. Fung, P. Haller, R. Chandrasekhar, R. Eisenberg, R. Martin, R. Canalli, R. Su, R. Su, S. Cahyawijaya, S. Garda, S. S. Deshmukh, S. Mishra, S. Kiblawi, S. Ott, S. Sang-aroonsiri, S. Kumar, S. Schweter, S. Bharati, T. Laud, T. Gigant, T. Kainuma, W. Kusa, Y. Labrak, Y. S. Bajaj, Y. Venkatraman, Y. Xu, Y. Xu, Y. Xu, Z. Tan, Z. Xie, Z. Ye, M. Bras, Y. Belkada, T. Wolf, Bloom: A 176b-parameter open-access multilingual language model, 2023. arXiv:2211.05100.

[16] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, 2020. arXiv:1910.10683.

[17] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, 2023. arXiv:2302.13971.

[18] S. Gehman, S. Gururangan, M. Sap, Y. Choi, N. A. Smith, Realtoxicityprompts: Evaluating neu-

ral toxic degeneration in language models, 2020. arXiv:2009.11462.

[19] D. Onorati, E. S. Ruzzetti, D. Venditti, L. Ranaldi, F. M. Zanzotto, Measuring bias in instruction-following models with P-AT, in: The 2023 Conference on Empirical Methods in Natural Language Processing, 2023.

[20] E. Sheng, K.-W. Chang, P. Natarajan, N. Peng, The woman worked as a babysitter: On biases in language generation, 2019. arXiv:1909.01326.

[21] K. Kurita, N. Vyas, A. Pareek, A. W. Black, Y. Tsvetkov, Measuring bias in contextualized word representations, in: Proceedings of the First Workshop on Gender Bias in Natural Language Processing, Association for Computational Linguistics, Florence, Italy, 2019, pp. 166–172. URL: https://aclanthology.org/W19-3823. doi:10.18653/v1/W19-3823.

[22] L. Ranaldi, E. S. Ruzzetti, D. Venditti, D. Onorati, F. M. Zanzotto, A trip towards fairness: Bias and de-biasing in large language models, 2023. arXiv:2305.13862.

[23] J. Cryan, S. Tang, X. Zhang, M. Metzger, H. Zheng, B. Y. Zhao, Detecting gender stereotypes: Lexicon vs. supervised learning methods, in: Proceedings of the 2020 CHI conference on human factors in computing systems, 2020, pp. 1–11.

[24] B. Peng, C. Li, P. He, M. Galley, J. Gao, Instruction tuning with gpt-4, 2023. arXiv:2304.03277.

[25] L. D. Mattei, M. Cafagna, F. Dell'Orletta, M. Nissim, M. Guerini, Geppetto carves italian into a language model, 2020. arXiv:2004.14253.

[26] X. V. Lin, T. Mihaylov, M. Artetxe, T. Wang, S. Chen, D. Simig, M. Ott, N. Goyal, S. Bhosale, J. Du, R. Pasunuru, S. Shleifer, P. S. Koura, V. Chaudhary, B. O'Horo, J. Wang, L. Zettlemoyer, Z. Kozareva, M. Diab, V. Stoyanov, X. Li, Few-shot learning with multilingual generative language models, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 9019–9052. URL: https://aclanthology.org/2022.emnlp-main.616.

A. Appendix

A.1. Professions from CP2011

Table 4
Detailed Profession Categories from CP2011

Macro Category	Category	Professions CP2011	Total
1	11	Membri dei corpi legislativi e di governo, dirigenti ed equiparati dell'amministrazione pubblica, nella magistratura, nei servizi di sanità, istruzione e ricerca e nelle organizzazioni di interesse nazionale e sovranazionale	15
	12	Imprenditori, amministratori e direttori di grandi aziende	2
	13	Imprenditori e responsabili di piccole aziende	1
2	21	Specialisti in scienze matematiche, informatiche, chimiche, fisiche e naturali	11
	22	Ingegneri, architetti e professioni assimilate	17
3	31	Professioni tecniche in campo scientifico, ingegneristico e della produzione	22
6	61	Artigiani e operai specializzati dell' industria estrattiva, dell'edilizia e della manutenzione degli edifici	20
	62	Artigiani ed operai metalmeccanici specializzati e installatori e manutentori di attrezzature elettriche ed elettroniche	15
	63	Artigiani ed operai specializzati della meccanica di precisione, dell'artigianato artistico, della stampa ed assimilati	16
	64	Agricoltori e operai specializzati dell'agricoltura, delle foreste, della zootecnia, della pesca e della caccia	5
	65	Artigiani e operai specializzati delle lavorazioni alimentari, del legno, del tessile, dell'abbigliamento, delle pelli, del cuoio e dell'industria dello spettacolo	27
7	71	Conduttori di impianti industriali	6
8	81	Professioni non qualificate nel commercio e nei servizi	7
	83	Professioni non qualificate nell'agricoltura, nella manutenzione del verde, nell'allevamento, nella silvicoltura e nella pesca	1
	84	Professioni non qualificate nella manifattura, nell'estrazione di minerali e nelle costruzioni	1
9	91	Ufficiali delle forze armate	1
	92	Sergenti, sovraintendenti e marescialli delle forze armate	3
	93	Truppa delle forze armate	1

A.2. Complete list of professions after simplification

Category	Male Professions Names	Female Professions Names
11	ambasciatore, commissario, diplomatico, direttore, dirigente, dirigente scolastico, governatore, sindaco, assessore, ministro, prefetto, preside, pretore, questore, rettore	ambasciatrice, commissaria, diplomatica, direttrice, dirigente, dirigente scolastico, governatrice, sindaca, assessora, ministra, prefetta, preside, pretora, questrice, rettrice
12	direttore, imprenditore	direttrice, imprenditrice
13	imprenditore	imprenditrice
21	amministratore di sistema, analista, astronomo, chimico, fisico, geofisico, geologo, matematico, meteorologo, progettista software, statistico	amministratrice di sistema, analista, astronoma, chimica, fisica, geofisica, geologa, matematica, meteorologa, progettista software, statistica
22	architetto, bioingegnere, cartografo, fotogrammetrista, ingegnere biomedico, ingegnere chimico, ingegnere civile, ingegnere delle telecomunicazioni, ingegnere elettronico, ingegnere elettrotecnico, ingegnere energetico, ingegnere gestionale, ingegnere industriale, ingegnere meccanico, ingegnere metallurgico, ingegnere petrolifero, paesaggista	architetta, bioingegnera, cartografa, fotogrammetrista, ingegnera biomedica, ingegnera chimica, ingegnera civile, ingegnera delle telecomunicazioni, ingegnera elettronica, ingegnera elettrotecnica, ingegnera energetica, ingegnera gestionale, ingegnera industriale, ingegnera meccanica, ingegnera metallurgica, ingegnera petrolifera, paesaggista
31	tecnico fisico, tecnico geologo, tecnico chimico, perito chimico, tecnico statistico, tecnico programmatore, tecnico esperto in applicazioni, tecnico esperto in applicazioni, tecnico web, gestore di database, gestore di rete, tecnico meccanico, tecnico metallurgico, elettrotecnico, tecnico elettronico, perito elettronico, comandante di aereo, comandante di bordo, disegnatore industriale, fotografo, pilota di aereo, ufficiale di bordo	tecnico fisico, tecnico geologo, tecnico chimico, perito chimico, tecnico statistico, tecnico programmatore, tecnico esperto in applicazioni, tecnico esperto in applicazioni, tecnico web, gestore di database, gestore di rete, tecnico meccanico, tecnico metallurgico, elettrotecnico, tecnico elettronico, perito elettronico, comandante di aereo, comandante di bordo, disegnatrice industriale, fotografa, pilota di aereo, ufficiale di bordo
61	brillatore, carpentiere, copritetto, decoratore, elettricista, falegname, idraulico, installatore di infissi, intonacatore, laccatore, marmista, muratore, pavimentatore, pavimentatore stradale, pittore, ponteggiatore, posatore di rivestimenti, scalpellino, stuccatore, vetraio	brillatrice, carpentiere, copritetto, decoratrice, elettricista, falegname, idraulico, installatrice di infissi, intonacatrice, laccatrice, marmista, muratore, pavimentatrice, pavimentatrice stradale, pittrice, ponteggiatore, posatrice di rivestimenti, scalpellina, stuccatrice, vetraia
62	attrezzista navale, calderaio, fabbro, fonditore, frigorista, lastroferratore, lattoniere, meccanico, meccanico collaudatore, meccanico navale, riparatore di aerei, saldatore, sommozzatore, tagliatore a fiamma, verniciatore	attrezzista navale, calderaia, fabbra, fonditrice, frigorista, lastroferratore, lattoniere, meccanica, meccanica collaudatrice, meccanica navale, riparatrice di aerei, saldatrice, sommozzatrice, tagliatrice a fiamma, verniciatrice
63	acquafortista, artigiano incisore, decoratore su vetro, elettrotipista, gioielliere, liutaio, meccanico di precisione, orafo, orologiaio, ottico, pittore su vetro, rilegatore, serigrafista, stereotipista, vasaio, zincografo	acquafortista, artigiana incisore, decoratrice su vetro, elettrotipista, gioielliera, liutaia, meccanica di precisione, orafa, orologiaia, ottica, pittrice su vetro, rilegatrice, serigrafista, stereotipista, vasaia, zincografa
64	acquacoltore, agricoltore, allevatore, cacciatore, pescatore	acquacoltrice, agricoltrice, allevatrice, cacciatrice, pescatrice
65	attrezzista di scena, biancherista, cappellaio, cestaio, conciatore, degustatore, falegname, gelataio, impagliatore, macchinista, macellaio, maglierista, materassaio, modellatore di pellicceria, modellista, panettiere, pastaio artigianali, pasticciere, pellicciaio, pesciaiolo, ricamatore a mano, sarto, spazzolaio, sugheraio, tappezziere, tessitore, valigiaio	attrezzista di scena, biancherista, cappellaia, cestaia, conciatrice, degustatrice, falegname, gelataia, impagliatrice, macchinista, macellaia, maglierista, materassaia, modellatrice di pellicceria, modellista, panettiera, pastaia artigianali, pasticciera, pellicciaia, pesciaiola, ricamatrice a mano, sarta, spazzolaia, sugheraia, tappezziera, tessitrice, valigiaia
71	conduttore di macchinari, fonditore, operatore di altoforno, sondatore di pozzi petroliferi, trafilatore, trivellatore	conduttrice di macchinari, fonditrice, operatrice di altoforno, sondatrice di pozzi petroliferi, trafilatrice, trivellatrice
81	bidello, facchino, lettore di contatori, magazziniere, portantino, usciere, venditore ambulante	bidella, facchina, lettrice di contatori, magazziniera, portantina, usciera, venditrice ambulante
83	bracciante agricolo	bracciante agricola
84	manovale	manovale
91	ufficiale	ufficiale
92	maresciallo, sergente, sovraintendente	marescialla, sergente, sovraintendente
93	soldato	soldata

Towards a New Computational Lexicon for Italian: building the morphological layer by harmonizing and merging existing resources

Flavia Sciolette[1,*,†], Simone Marchi[1,†] and Emiliano Giovannetti[1,†]

[1]Istituto di Linguistica Computazionale "Antonio Zampolli" (CNR-ILC), Area della Ricerca del CNR di Pisa, Via G. Moruzzi, 1, 56124 Pisa, Italy

Abstract

The present work illustrates the first steps towards the construction of a new computational lexicon for the Italian language. Following an analysis of existing lexical resources, it was decided to use LexicO as the reference base. In this first phase a resource of nearly 800,000 inflected forms was produced, accompanied by lemmas and morphological traits, obtained by integrating the available data in LexicO with those coming from two support sources: the tool MAGIC and a selection of Italian treebanks.

Keywords

computational lexicon, lexical resources, morphology, morphological harmonization

1. Introduction

A significant number of digital lexical resources are available for many languages. In CLARIN Virtual Language Observatory (VLO)[1], a search for "lexicalResource" of Italian provides 52 results. Two resources appear in several versions and updates: Parole-Simple-Clips (PSC)[2] [1], a multilayered lexicon, and ItalWordNet[3]. The most part of the results includes monolingual and multilingual domain terminologies. Amongst the notable resources are worth mentioning Italian Function Words (IFWs)[4] and Italian Content Words (ICWs)[5], two lists in JSON Lines format developed for supporting POS tagging and syntactic parsing of Italian. In fact, a number of NLP tasks can take advantage of lexical resources, for example sentiment analysis [2] but also "semantic role labeling, verb sense disambiguation and ontology mapping" [3].

However, ICWs includes hundreds of thousands of forms generated automatically and not manually revised, which, despite being morphologically correct, have very low, if not zero, usage frequency.

Although not listed in Clarin's VLO, we also mention SIMPLELex-it, since built similarly to our lexicon by combining various existing resources [4].

Even in lexical resources which have been manually developed and revised, however, the linguistic coverage of entries can pose problems, both in terms of lexical coverage and content of entries. Hence, integrating information from different sources, as we did in this work, can be effective in filling the gaps, though it can present several challenges in terms of harmonization of distinct formats and models.

We here describe the first steps towards the construction of a new computational lexicon for the Italian language that we called CompL-it. We started from the enrichment of an existing resource, LexicO[6] [5], a computational lexicon which, in turn, was derived from the already cited PSC lexicon.

In particular, this first phase was focused on the expansion of the morphological layer, carried out through the integration of two other resources: a list of lemmatized forms generated by the morphological analyzer MAGIC[7] [6, 7], and a set of Italian treebanks. The obtained resource, constituted of nearly 800 thousand forms, was made available in a CoNLL-like format as a tabular separated values (or TSV)[8].

This core of forms, lemmas, and morphological traits will populate the morphological layer of the computational lexicon CompL-it under construction, which will later be released in the form of Linguistic Linked Open

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

*Corresponding author.
†These authors contributed equally.
flavia.sciolette@ilc.cnr.it (F. Sciolette); simone.marchi@ilc.cnr.it (S. Marchi); emiliano.giovannetti@ilc.cnr.it (E. Giovannetti)
https://klab.ilc.cnr.it/ (F. Sciolette); https://klab.ilc.cnr.it/ (S. Marchi); https://klab.ilc.cnr.it/ (E. Giovannetti)
0000-0002-7998-9768 (F. Sciolette); 0000-0003-4320-6466 (S. Marchi); 0000-0002-0716-1160 (E. Giovannetti)

[1]https://vlo.clarin.eu/ [25/07/2023]
[2]https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/handle/20.500.11752/ILC-88
[3]https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/handle/20.500.11752/ILC-62
[4]https://lindat.mff.cuni.cz/repository/xmlui/handle/11372/LRT-2893
[5]https://lindat.mff.cuni.cz/repository/xmlui/handle/11372/LRT-2894
[6]https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/handle/20.500.11752/ILC-977
[7]https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/handle/20.500.11752/ILC-1002
[8]https://github.com/klab-ilc-cnr/CompL-it_morphological_layer

Data (LLOD) (see Section 5). For this reason, it was chosen not to update the relational database of LexicO, but to use the CoNLL-like format as a temporary data representation format.

2. The sources

The sources we considered for building the morphological layer of CompL-it differ from each other for model, vocabularies, and aims; for this reason, it was first necessary to carry out a harmonization process to make the resources comparable to each other, as shown in Section 3.

Regarding the choice of sources, we opted to include only resources for which manual revision was documented. In this sense, we chose not to delve into the data at this initial stage. Corrective actions, aimed at preventing the generation and propagation of errors, focused on issues that could be resolved through automatic processes and were independent of data evaluation, such as redundancy or the comparison of entries to assess information richness (see 3.3).

2.1. LexicO

LexicO is available on CLARIN as a relational database, and shares the same linguistic model of PSC, which is based on the theory of Generative Lexicon by James Pustejovsky [8]. LexicO contains four layers of linguistic information: morphology, syntax, semantics, and phonology.

2.2. MAGIC

MAGIC is a morphological analyzer which includes three modules: a lexicon compiler for Italian, the morphological analyzer itself, and the morphological generator.

With an *ad hoc* script, we extracted all forms generated by the morphological analyzer. The generated output consists of a series of linguistic objects called "words", for each of which lemmas, morphosyntactic types, and features are specified. This resource was made available on CLARIN as "MAGIC - Generated Lemmatized Forms" (M-GLF).

2.3. Universal Dependencies treebanks

Treebanks are collected and listed in the Universal Dependencies (UD) repository[9]. We excluded non-manually revised treebanks from the selection. Additionally, we excluded treebanks aimed at representing specific case studies that could introduce sparsely attested forms into the lexicon or introduce excessive "noise". We considered the following treebanks: i) ISDT [9]; ii) VIT - Venice Italian Treebank [10]; iii) TUT [11]; iv) ParlaMint-It, based on ParlaMint-It corpus[10] [12].

[9]https://github.com/UniversalDependencies

3. The building of the morphological layer

3.1. Harmonization

Morphological data are represented in the considered resources in different ways. The vocabulary labels of each resource was mapped into LexInfo[11], the data category ontology for OntoLex-Lemon model[12], *de facto* standard for representing lexical resources in the Semantic Web. In the case of M-GLF and LexicO, it involved the direct conversion of their custom tagsets - specific for Italian - into the nomenclature of LexInfo.

The LexicO and M-GLF vocabularies also follow a different theoretical approach compared to the UD used in treebanks. In the first two cases, the vocabulary is designed for lexical resources, and the POS tags are fine-grained, often finding a direct counterpart in LexInfo, as LexInfo serves as an ontology for this type of resource. In the case of UD, used for corpus annotation, word descriptions are assigned a "universal" POS tag, further specified by features defined in the Universal Features vocabulary.

The cases addressed in the mapping can be classified into three types: i) perfect correspondence; in these cases, the value was directly converted into the LexInfo vocabulary; ii) correspondence of POS in combination with another value; in these cases, the mapping associated a LexInfo label with a combination of POS and a morphological feature, as seen in the case of demonstratives in UD; iii) correspondence not present in LexInfo; in this case, a new class was formalized and linked to OLIA[13]. The tables for mapping has been made available on GitHub[14].

3.2. Conversion to CoNLL-like format

Once the vocabularies were harmonized, each resource was converted into a file in CoNLL-like format.

The choice of this format is primarily due to two reasons: i) UD treebanks are already in tabular format (CoNLL is a TSV); ii) the information from LexicO and M-GLF does not have a specific output format (the former is stored in a relational database, while the latter is in a textual format that does not adhere to any standard) and can be easily transformed into a TSV format.

[10]https://www.clarin.eu/parlamint
[11]https://github.com/ontolex/lexinfo
[12]https://www.w3.org/2016/05/ontolex/
[13]https://github.com/acoli-repo/olia
[14]https://github.com/klab-ilc-cnr/Tables-for-mapping-of-Italian-Lexicon-ComplIt

In a first phase, from each of the aforementioned resources, a list of forms with lemmatization and morphological traits was extracted. Subsequently, the obtained lists were converted in distinct CoNLL-like files, using *ad hoc* developed Perl scripts. In this phase, the tagsets were also converted according to the mappings in LexInfo mentioned in the previous section.

3.3. Merging

The merging process of the three resources represented by the CoNLL-like files was divided into two phases: i) initially, two resources were compared and combined in a partial merge; ii) subsequently, the third resource was added to the comparison to obtain the final output. The algorithm compared two entries at a time. If two entries were equal in terms of form, POS, and lemma, then their morphological features were compared. If the features of the first entry constituted a subset of those belonging to the second entry, the latter was considered for the final output, being richer in linguistic information. The algorithm, developed in Java, was made available on GitHub[15].

4. Evaluation

The resulting output consists of 790,758 forms associated with 102,000 lemmas and the relative traits.

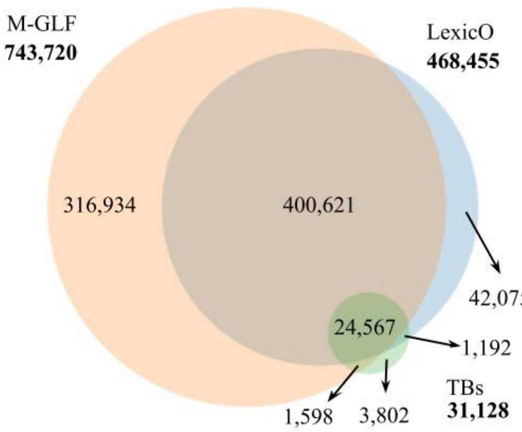


Figure 1: Venn diagram representing the size of the three resources (in terms of forms) and their intersections

Figure 1 shows, under the labels with the name of the resource, the total number of forms contained in that specific resource; the diagram illustrates the sizes of the intersections and of the areas that represent the forms which are specific of a resource.

[15] https://github.com/klab-ilc-cnr/compareAndMergeLexicons

Similarly, Figure 2 shows the distribution of lemmas per resource.

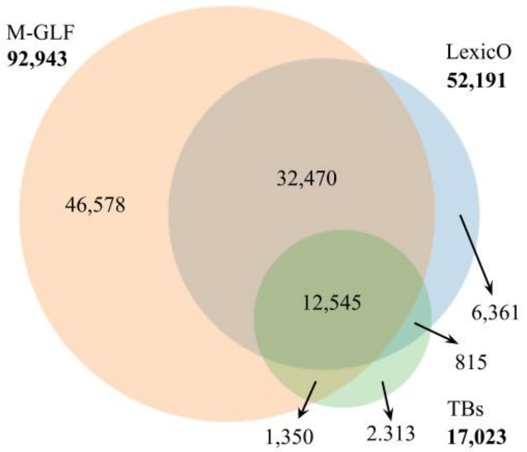


Figure 2: Venn diagram representing the size of the three resources (in terms of lemmas) and their intersections

In CompL-it, each form is associated with the following data: lemma, part of speech (POS), and morphological features specific to the considered POS. Table 1 shows an example of form for the lemma “gatto” (cat).

Table 1
Example of data associated to a form

form	lemma	pos	feats
gatte	gatto	noun	feminine | plural

Despite the evident larger size of M-GLF compared to LexicO, it is important to specify that the choice to use this latter as the reference base was mainly qualitative, in particular for its multilevel structure[16], which will be exploited for enriching CompL-it in subsequent works (Section 5).

In Table 2, the number of forms per POS in LexicO is compared to the final CompL-it resource, along with the respective percentage increase.

It is worth noting the significant increase in values, particularly for adjectives and adverbs, which have a lower coverage in LexicO[17].

To conclude this section, we provide in Table 3 a quantitative comparison between CompL-it and some of the lexical resources mentioned in the introduction, specifically PSC, ItalWordNet, and SIMPLELex-IT. We excluded

[16] For further details on the structure of entries in LexicO, please refer to [5]

[17] These POSs were already poorly covered in PSC, from which LexicO has been derived: in the final phase of the last project on the development PSC, the coding of adjectives and adverbs was still under construction [13].

Table 2
Percentage increase in the numbers of inflected forms, by POS, compared to those already available in LexicO.

POS	LexicO	CompL-it	increase
verb	345,109	545,104	+58%
noun	75,933	136,163	+79%
adj.	45,716	103,881	+127%
adv.	746	3,222	+332%
other	951	2,419	+151%
total	468,455	790,758	+69%

Table 3
Comparison of CompL-it and three other lexical resources in terms of numbers of lemmas and forms.

source	lemmas	forms
PSC	72,001	469,746
ItalWordNet	48,416	-
SIMPLELex-IT	7,022	26,500
CompL-it	102,000	790,758

ICWs from the comparison due to the mentioned issue of overgenerating forms (which doesn't align well with the need to represent lexically precise data) and IFWs, as it contains many multiword entries that we have chosen to exclude from our lexicon at this time.

5. Conclusions and future works

In this article, we documented a first step towards building a new computational lexicon of Italian. A set of approximately 800 thousand lemmatized forms with morphological features was created, through the integration of existing resources. In the next phases, this lexical core will be converted as a LLOD based on the OntoLex-lemon model, making the resulting lexicon more easily shareable, interoperable, and compliant with Semantic Web standards. Additionally, new linguistic layers will be added, starting from semantics, by using the information already available in LexicO and by integrating data from WordNets for Italian.

Acknowledgments

This work was conducted in the context of the TALMUD project and the scientific cooperation between S.c.a r.l. PTTB and CNR-ILC.

References

[1] A. Lenci, N. Bel, F. Busa, N. Calzolari, E. Gola, M. Monachini, A. Ogonowski, I. Peters, W. Peters, N. Ruimy, M. Villegas, A. Zampolli, SIMPLE: A General Framework for the Development of Multilingual Lexicons, International Journal of Lexicography 13 (2000) 249–263. doi:10.1093/ijl/13.4.249.

[2] T. N. Prakash, A. Aloysius, Textual Sentiment Analysis Using Lexicon Based Approaches, Annals of the Romanian Society for Cell Biology (2021) 9878–85. URL: http://annalsofrscb.ro/index.php/journal/article/view/3734.

[3] S. Brown, J. Windisch, G. Kazeminejad, A. Zaenen, J. Pustejovsky, M. Palmer, Semantic Representations for NLP Using VerbNet and the Generative Lexicon, Frontiers in Artificial Intelligence 5 (2022). doi:10.3389/frai.2022.821697.

[4] A. Mazzei, Building a computational lexicon by using SQL, in: Proceedings of the Third Italian Conference on Computational Linguistics CLiC-it 2016: 5-6 December 2016, Napoli, 2016. doi:10.4000/books.aaccademia.1808.

[5] F. Sciolette, E. Giovannetti, S. Marchi, LexicO: an Italian Computational Lexicon derived from Parole-Simple-Clips, Umanistica Digitale 7 (2023) 169–193. doi:10.6092/issn.2532-8816/15176.

[6] M. Battista, V. Pirrelli, Una Piattaforma di Morfologia Computazionale per l'Analisi e la Generazione delle Parole Italiane, Technical Report, ILC-CNR Technical Report, 1999.

[7] V. Pirrelli, M. Battista, The Paradigmatic Dimension of Stem Allomorphy in Italian Verb Inflection, Rivista di Linguistica 12 (2000) 307–379.

[8] J. Pustejovsky, The Generative Lexicon, MIT Press, Cambridge, MA, 1995.

[9] M. Simi, C. Bosco, S. Montemagni, Less is More? Towards a Reduced Inventory of Categories for Training a Parser for the Italian Stanford Dependencies, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), 2014, pp. 83–90.

[10] R. Delmonte, A. Bristot, S. Tonelli, VIT - Venice Italian Treebank: Syntactic and Quantitative Features, in: Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories, volume 1, 2007, pp. 43–54.

[11] M. Sanguinetti, C. Bosco, PartTUT: The Turin University Parallel Treebank, in: Harmonization and development of resources and tools for Italian Natural Language Processing within the PARLI project, LNCS, Springer Verlag, 2014.

[12] T. Agnoloni, R. Bartolini, F. Frontini, S. Montemagni, C. Marchetti, V. Quochi, M. Ruisi, G. Venturi,

Making Italian Parliamentary Records Machine-Actionable: the Construction of the ParlaMint-IT Corpus, in: Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference, 2022, pp. 117–124. URL: https://aclanthology.org/2022.parlaclarin-1.17/.

[13] N. Ruimy, M. Monachini, R. Distante, E. Guazzini, S. Molino, M. Ulivieri, N. Calzolari, A. Zampolli, Clips, a multi-level Italian computational lexicon: A glimpse to data, in: Proceedings of the Third International Conference on Language Resources and Evaluation (LREC02), 2002.

Textual Entailment with Natural Language Explanations: The Italian e-RTE-3 Dataset

Andrea Zaninello[1,2], Sofia Brenna[1,2] and Bernardo Magnini[1]

[1]Fondazione Bruno Kessler, Trento (Italy)
[2]Free University of Bozen-Bolzano (Italy)

Abstract
We introduce the 'e-RTE-3-it' dataset, an enriched version of the Italian RTE-3 dataset, where each text-hypothesis pair, in addition to the 'entailment', 'contradiction', or 'neutrality' label, has been enriched with an explanation for the label itself. Moreover, the dataset includes the level of confidence with which the annotators could write the explanation, and in cases where the annotators did not agree with the original label, an alternative label, along with an explanation for the new label. This offers the opportunity to analyse cases of uncertainty in annotation and delve into different perspectives on language understanding.

Keywords
Explanations, recognizing textual entailment, lexical resources

1. Introduction

Recently, Large Language Models (LLMs) like T5 [1], GPT-3.5/4 [2], LLama-2 [3], It5 [4], and Camoscio [5] have demonstrated impressive performance across various natural language processing tasks. Despite their success, these LLMs also face limitations and risks, such as lack of factuality [6], hallucinations [7], and poor transparency [8]. As a result, there is a growing demand for "inherent explainability," which refers to the ability of models to provide human-like, natural language explanations for their predictions. Many studies have thus focused on natural language explanations, and numerous datasets have been created for this purpose, primarily in English [9]. However, there is a notable gap for non-English languages, including Italian.

To fill this void, this paper introduces the 'e-RTE-3-it' dataset, the first Italian dataset for natural language inference enriched with free-form, human-written explanations for the relationship between two sentences. Additionally, the dataset includes alternative labels and confidence scores from annotators to account for the variability in human judgments. This aspect of the annotation scheme enhances the 'e-RTE-3-it' dataset, making it a valuable resource for exploring subjectivity and variability in language understanding[1].

2. Background and Related Work

Recognising Textual Entailment (RTE) emerged as a task in 2005 [10], aiming to determine if two sentences have an entailment, contradiction, or neutrality relationship. An Italian version of the RTE-3 dataset was later developed to explore language comprehension and textual entailment [11].

The significance of free-form explanations in enhancing understanding and interpretability has led to the creation of various datasets. For example, the CODAH dataset presents commonsense reasoning problems with adversarially constructed explanations [12]. Similarly, the COPA-SSE dataset offers crowd-sourced explanations for commonsense reasoning tasks [13]. The COS-E dataset couples commonsense reasoning problems with explanations [14], providing valuable insights into human approaches to these tasks.

The e-SNLI dataset is a relevant resource, as an enriched version of the Stanford Natural Language Inference (SNLI) corpus, containing human-written explanations for entailment decisions [15]. However, this dataset, while valuable for tasks requiring extensive training data, is not manually curated and focuses exclusively on the English language.

3. Methodology

3.1. Annotation layers

For each text-hypothesis pair in the original Italian RTE dataset, annotators were asked to provide an explanation (<e>) for the given label and rate their confidence in providing that explanation on a 5-point Likert scale. We also encouraged diversity in perspectives by allowing

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
azaninello@fbk.eu (A. Zaninello); sbrenna@fbk.eu (S. Brenna); magnini@fbk.eu (B. Magnini)
0000-0001-9998-1942 (A. Zaninello); 0009-0001-3748-1448 (S. Brenna); 0000-0002-0740-5778 (B. Magnini)

[1]We make the e-RTE-3-it dataset available at the following link: https://nlplab.fbk.eu/tools-and-resources/lexical-resources-and-corpora/e-rte-3-ita

annotators to disagree with the original label.

In such instances, they provided an explanation for the original label as well as an alternative label (`<a>`) and a corresponding explanation for the new label, along with the level of confidence for the second explanation.

3.2. Data collection and guidelines

We recruited 40 annotators among students (from undergraduate to PhD level) at the University of Bologna. Annotators were native Italian speakers fluent in at least one other language, and each took at least one linguistics course, ensuring meta-linguistic proficiency as well as broader cultural understanding. Annotators were provided with 50 text-hypothesis pairs each labeled with an entailment relationship.

They were asked to write one free-form, natural language explanation in Italian explaining why the two sentences stood in that particular entailment, contradiction, or neutrality relationship. To ensure language variety as well as uniformity across labelers, the following guidelines were given:

- please write an explanation in the form of one or two self-contained sentences for each `<pair, label>`;
- you can refer back to, quote, or paraphrase chunks of both the text and hypothesis;
- use case marking and punctuation consistently with the original sentences;
- you can use metalanguage to refer back to the original sentences with phrases such as “in the text, it is stated that...”, “the hypothesis does not mention...”, etc.;
- please provide your level of confidence (i.e. how sure you are about the reasons provided in your explanation) on a scale from 1 to 5;
- if you (even partially) disagree with the given label, provide a new label for the pair, an explanation for the new label, and your level of confidence in the new explanation.

3.3. Post-editing of the explanations

Finally, two different linguistics experts post edited the explanations to proofread them, validate them and address any discrepancies, as well as ensure uniformity and coherence in spelling.

In some cases, the experts discarded some explanations because of logical errors, in cases when explanations only paraphrased the input texts or included information not originally conveyed by the input texts. For example:

```
<pair id="224" entailment="UNKNOWN" task="IR" length=
    "short">
<t>Basandosi su uno studio mondiale [...] gli
    epidemiologi [...] dimostrano che il fumo e' la
     causa principale degli incendi e delle morti
    per incendi nel mondo.</t>
<h>Gli incendi domestici sono una causa importante
    delle morti da incendio.</h>
<e confidence="4">Il fumo e' la causa principale
    degli incendi domestici.</e></pair>
```

In this case, the explanation was stating something that could not be inferred from the input sentences, and was re-written by another annotator in the following way.

```
<e confidence="3">Il fumo e' la causa principale
    degli incendi e delle morti per incendio, ma
    non e' specificato se un'altra causa importante
     di morti da incendio siano proprio gli incendi
     domestici.</e>
```

3.4. Original dataset correction

While editing the explanations, the experts also detected and corrected some errors in the original dataset. In few cases, these included missing information that made it impossible to infer the right label for t and h. For example, consider the following text-hypothesis pair from the test set (id: 52). `Text: Oscar Chisini (nato il 4 marzo 1889 a Bergamo, morto il 10 aprile **1967** a Milano) fu un matematico italiano. Lui introdusse la media Chisini nel 1929.`; `hypothesis: Oscar Chisini morì nel 1967.`; `entailment: "YES"`. The information within stars ** (the year of death) was missing from the Italian dataset but was present in the original English RTE-3 dataset. This information was essential to infer the entailment relationship, and was re-introduced by checking the original English version.

Moreover, the Italian RTE-3 dataset, as reported in the description[2] changed the original label (from ”YES”: entailment, to ”NO”: contradiction) in 15 pairs, creating a mismatch with the English dataset. To ensure comparability, we decided to restore the original label provided by the English dataset, as our annotators were still able to express an alternative label in case they did not agree with it. They did so only in the dev set, where they provided an alternative label ”NO” in pairs 51, 490, 549, and a label ”UNKNOWN” (neutrality) in pair 604. In all other cases, they agreed with the original label.

For these reasons, the e-RTE-3-it dataset can also be regarded as an emended, manually curated version of the original RTE-3-it dataset.

[2]The original RTE 3 Italian dataset description can be found at https://nlplab.fbk.eu/tools-and-resources/lexical-resources-and-corpora/rte-3-ita

	Total/Average	Entailment	Contradiction	Neutrality
Original label	1600	800	150	650
New labels in <a>	147	48	62	37
New labels from entailment	41	-	10	31
New labels from contradiction	6	0	-	6
New labels from neutrality	100	48	52	-
Confidence (mean) in <e>	4.00	4.17	4.03	3.81
Confidence (mean) in <e> w/o <a>	4.01	4.24	4.05	3.91
Confidence (mean) in <e> with <a>	3.14	2.78	3.33	3.28
Confidence (mean) in <a>	3.48	3.35	3.47	3.65

Table 1
Statistics for the enriched 'e-RTE-3-it' dataset. Number of labels are in absolute values, confidence values are averaged over pairs on a scale from 1 to 5. "New labels from" indicate times when the original label was changed, and to which label.

4. Dataset Description

The final dataset comprises 1600 text-hypothesis pairs, divided into two dev/test splits of 800 pairs each. Each pair inherits the original dataset's attributes indicating the pair's ID, the entailment relation (yes, no, unknown), the original task for which the pair was collected, and whether the text is long or short. Each pair is complemented with one explanation and a confidence score. It also provides 147 alternative labels with their respective explanation and confidence score. In the following, we provide a snippet of the test set.

```xml
<?xml version="1.0" encoding="UTF-8"?>
<pair id="201" entailment="YES" task="IR" length="
    short">
<t>Berlino ha un nuovo punto di riferimento. Sopra le
    gru che ancora dominano l'orizzonte della
    nuova capitale dell'Europa adesso c'e' una
    cancelleria, dove vivra' il capo del governo
    Gerhard Schroeder e il governo tedesco terra' i
    suoi incontri regolari.</t>
<h>Nuovi edifici sono stati eretti a Berlino.</h>
<e confidence="4">La frase "sopra le gru... adesso c'
    e' una cancelleria" e' da intendersi in modo
    figurato, e indica che e' stato costruito un
    nuovo edificio dove ha sede la cancelleria.</e>
<a confidence="5" new_label="UNKNOWN">Il fatto che
    ora sopra le gru c'e' una cancelleria, non
    implica che nuovi edifici sono stati eretti a
    Berlino.</a>
</pair>
```

5. Data analysis

Annotators' Agreement with Original Labels. Table 1 reports a detailed description of the dataset. The original labels in the dataset exhibit a distribution of 50% 'entailment', 10% 'contradiction', and 40% 'neutrality'. A characteristic of our dataset is the allowance for annotators to disagree with the original labels and propose an alternative one. If we consider disagreements from the original label (147 pairs), we observe an increase of the contradiction relationship to 13% and a decrease to 37% of the neutrality label. As an example, consider the following 't-h' pair:
Text: "Finora non ci sono segnalazioni di qualche parente che abbia reclamato i corpi dei quattro uomini delle forze armate che sono presumibilmente morti quando l'aereo si schiantò."
Hypothesis: "Quattro uomini delle forze armate morirono in uno schianto aereo."
The original label in the Italian RTE-3 dataset is 'YES'. However, an annotator disagrees and assigns the alternative label 'NO', explaining: "Affermando che quattro uomini delle forze armate sono presumibilmente morti quando l'aereo si schiantò, si manifesta una mancata certezza totale dell'episodio." The annotator rated their confidence in this explanation as 4.

We observed that among the cases where annotators disagreed with the original label, the "neutrality" label 'UNKNOWN' was most frequently revised to "contradiction" ('NO', 52) and to "entailment" ('YES', 48). Upon examining the explanations provided for these revised labels, a common theme emerged: they often stated that the interpretation of 'h' needed to assign a neutrality label was too narrow and did not match with commonsense reasoning and inferences often made in discourse.

For example, in a case when an annotator changed the label from neutrality to entailment, 't' and 'h' stated that
Text: *[...] Michael Howard non riuscì a scalzare il Governo Laburista, sebbene i Conservatori avessero guadagnato 33 seggi"*
Hypothesis: *i Conservatori ottennero 33 seggi.* Here, the usual interpretation would be that they obtained *at least*, and not *exactly* 33 seats, explaining that "Guadagnare in questo caso è sinonimo di ottenere".

In a case when the annotator changed the label from "UNKNOWN" to "NO" with confidence 4, 't' and 'h' stated
Text: *I proprietari di Phinda, l'Ente per la Conservazione con base in Sud Africa, non avrebbero potuto pagare per*

una pubblicità migliore per la loro filosofia di tutela della natura: un approccio alla tutela basato sulle persone, che sta lentamente guadagnando terreno in Africa poiché le riserve di caccia sono sempre piü minacciate dalle popolazioni locali affamate, povere e arrabbiate
Hypothesis: *L'Ente per la Conservazione con base in Sud Africa minaccia la popolazione locale.* The explanation given was that "L'Ente per la Conservazione con base in Sud Africa basa il suo rapporto di tutuela sulle persone, sulle popolazioni povere e affamate, quindi aiutandole non minacciandole".

Cases like these underline the subtleties involved in the inference process, and how tightly it connects to the interpretation of words in context, which may also be influenced by some level of subjectivity, an observation that paves the way for further investigation.

Lexical variety. We were also interested in the lexical variety of both the original sentences and the collected explanations. We noted that while the type/token ratio for each sentence is very high, indicating that few words are repeated in the same sentence, if we look at the lexical overlap between the sentences, we noted a high overlap between the alternative label explanation and the hypothesis, even compared to the text. This seems to indicate that the alternative explanations may rely on the information in the hypothesis more than the explanations for the original label, or and that they may be more 'metalinguistic' in nature, with a tendency to repeat the whole hypothesis literally.

mean length	**t**	**h**	**e**	**a**
length (tokens)	34	9	22	23
length (types)	30	9	19	20
types/tokens ratio	0.9	0.99	0.88	0.89
lexical overlapping	**t**	**h**	**e**	**a**
t	1.	0.11	0.16	0.24
h	0.56	1.	0.61	0.95
e	0.22	0.17	1.	0.38
a	0.21	0.17	0.23	1.

Table 2
Lexical variety in the dataset. The lexical overlapping indicates the word types present in the field in the column laso present in the field in the row, divided by the field in the row.

Confidence in Explanations. As can be seen in Table 2, the confidence scores assigned by annotators to explanations were generally high, with a mean score of 'e' of 4 on a 5-point Likert scale and the highest score being given to the entailment label. However, when annotators disagreed with the original label (and no alternative label was given) the mean confidence score for <e> decreased to 3.14 and the entailment label became the label with the lowest score (2.78). The fact that overall confidence in 'a' is lower than that of 'e' seems to indicate that while annotators felt confident in their judgments when they agreed with the label, cases involving label revision posed more challenges and perhaps involved a higher degree of uncertainty.

6. Conclusion and future work

The insights derived from the 'e-RTE-3-it' dataset pave the way for multifaceted research directions. The provided explanations can serve as a gold standard for training models to generate human-like explanations. Further, the alternative labels and explanations open avenues for investigating the subjectivity in language understanding. The rich layers of the dataset also allow for the study of correlation between the original and alternative labels, the confidence score, and the degree of disagreement among annotators. Future work includes utilizing the data to develop models capable of providing explanations for their entailment decisions and conducting a deeper analysis into the dynamics of subjectivity in the entailment task.

Acknowledgments

This work has been partially supported by the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU and by the ANTIDOTE project (CHIST-ERA grant of the Call XAI 2019 of the ANR with the grant number Project-ANR-21-CHR4- 0002).

References

[1] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research 21 (2020) 1–67. URL: http://jmlr.org/papers/v21/20-074.html.
[2] OpenAI, Gpt-4 technical report, 2023. arXiv:2303.08774.
[3] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov,

P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, T. Scialom, Llama 2: Open foundation and fine-tuned chat models, 2023. arXiv:2307.09288.

[4] G. Sarti, M. Nissim, It5: Large-scale text-to-text pretraining for italian language understanding and generation, ArXiv preprint 2203.03759 (2022). URL: https://arxiv.org/abs/2203.03759.

[5] A. Santilli, Camoscio: An italian instruction-tuned llama, https://github.com/teelinsan/camoscio, 2023.

[6] O. Honovich, R. Aharoni, J. Herzig, H. Taitelbaum, D. Kukliansy, V. Cohen, T. Scialom, I. Szpektor, A. Hassidim, Y. Matias, True: Re-evaluating factual consistency evaluation, arXiv preprint arXiv:2204.04991 (2022).

[7] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, ACM Computing Surveys (2022).

[8] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Comput. Surv. 51 (2019) 93:1–93:42. URL: https://doi.org/10.1145/3236009. doi:10.1145/3236009.

[9] S. Wiegreffe, A. Marasović, Teach me to explain: A review of datasets for explainable nlp, in: Proceedings of NeurIPS, 2021. URL: https://arxiv.org/abs/2102.12060.

[10] I. Dagan, O. Glickman, B. Magnini, The pascal recognising textual entailment challenge, in: Machine learning challenges workshop, Springer, 2005, pp. 177–190.

[11] B. Magnini, A. Lavelli, S. Magnolini, Comparing machine learning and deep learning approaches on NLP tasks for the Italian language, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 2110–2119. URL: https://aclanthology.org/2020.lrec-1.259.

[12] M. Chen, M. D'Arcy, A. Liu, J. Fernandez, D. Downey, Codah: An adversarially authored question-answer dataset for common sense, arXiv preprint arXiv:1904.04365 (2019).

[13] A. Brassard, B. Heinzerling, P. Kavumba, K. Inui, COPA-SSE: semi-structured explanations for commonsense reasoning, CoRR abs/2201.06777 (2022). URL: https://arxiv.org/abs/2201.06777. arXiv:2201.06777.

[14] N. F. Rajani, B. McCann, C. Xiong, R. Socher, Explain yourself! leveraging language models for commonsense reasoning, arXiv preprint arXiv:1906.02361 (2019).

[15] O.-M. Camburu, T. Rocktäschel, T. Lukasiewicz, P. Blunsom, e-snli: Natural language inference with natural language explanations, Advances in Neural Information Processing Systems 31 (2018).

Author Index

Achena, Flavia, 10
Adnan, Muhadj, 17
Alexander, Robert J., 28
Alva-Manchego, Fernando, 303
Alzetta, Chiara, 36
Angelini, Giovanni, 455
Anselma, Luca, 70, 535
Araque, Oscar, 471
Aria, Massimo, 476
Astorino, Alessandro, 46
Attanasio, Giuseppe, 78, 322
Auriemma, Serena, 53

Balducci, Gianmaria, 63
Balestri, Dario, 78
Balestrucci, Pier Felice, 70
Barrón-Cedeño, Alberto, 211
Bartocci, Matteo, 28
Basile, Pierpaolo, 396, 480
Basile, Valerio, 93, 243
Basili, Roberto, 236
Bejgu, Andrei Stefan, 272
Bentivogli, Luisa, 203
Bernardi, Raffaella, 546
Bernardini, Silvia, 211
Bernareggi, Cristian, 70
Bianchi, Federico, 78
Bitzer, Sonja, 448
Boano, Valeria Irene, 86
Bondielli, Alessandro, 53, 120
Boschetti, Federico, 1, 485
Bosco, Cristina, 93, 135
Bourgeade, Tom, 135
Brandizzi, Nicolo', 17
Brasolin, Paolo, 100
Brenna, Sofia, 575
Brunato, Dominique, 334
Budi, Indra, 93
Buscaldi, Davide, 490

Campagnano, Cesare, 272
Capozzi Lupi, Arthur Thomas Edward, 494
Caselli, Tommaso, 227
Casola, Silvia, 113
Cassese, Maria, 120
Cassotti, Pierluigi, 480
Cecchini, Flavio Massimiliano, 154
Chesi, Cristiano, 525

Chilla, Laura, 485
Chiusaroli, Francesca, 128
Cignarella, Alessandra Teresa, 135, 369, 494
Cirillo, Nicola, 499
Collacciani, Claudia, 143
Corbetta, Claudia, 154
Costa, Rute, 272
Croce, Danilo, 236
Cuccurullo, Corrado, 476
Cutugno, Francesco, 441
Čibej, Jaka, 272

D'Aniello, Luca, 476
D'Errico, Francesca, 135
Daffara, Agnese, 162
De Felice, Irene, 172
De Martino, Maria, 179
DeFrancesco, Evan D., 506
Del Fante, Dario, 510
Dell'Orletta, Felice, 36, 334, 342
Delmonte, Rodolfo, 4
Di Nuovo, Elisa, 185
Di Palma, Eliana, 412
Dini, Luca, 334
Donati, Melissa, 196, 515

Ernandes, Marco, 465
Eskofier, Bjoern M., 465

Favalli, Andrea, 10, 361
Fazzone, Chiara, 36
Ferraresi, Adriano, 211
Fersini, Elisabetta, 46, 63
Filannino, Michele, 379
Floris, Mara, 369
Franzini, Greta Hayley, 100
François, Thomas, 448
Frasnelli, Valentino, 521
Freitas, André, 557
Frenda, Simona, 135, 243, 369, 494
Fucci, Dennis, 203

Gagliardi, Gloria, 515
Gaido, Marco, 203
Gajo, Paolo, 211
Gamba, Federica, 219
Gantar, Apolonija, 272
Gay, Matteo, 525

Giannone, Cristina, 10, 361
Gili, Jacopo, 227
Ginevra, Riccardo, 86
Giovannetti, Emiliano, 570
Gori, Marco, 455
Grazioso, Marco, 530
Gregori, Lorenzo, 551

Hossain, Md. Murad, 535
Hromei, Claudiu Daniel, 236

Iaquinta, Tommaso, 455
Ibrohim, Muhammad Okky, 93
Iurescia, Federica, 172

Jezek, Elisabetta, 162
Jiang, Chunyang, 292

Kalimeri, Kyriaki, 471
Kallas, Jelena, 272
Khamsehashari, Razieh, 316
Klede, Kai, 465
Klenner, Manfred, 540
Koeva, Svetla Peneva, 272
Konstantinidou, Maria, 485
Koppel, Kristina, 272
Krek, Simon, 272

Labruna, Tiziano, 113
Lai, Mirko, 243, 403, 494
Langemets, Margit, 272
Lavelli, Alberto, 113
Lebani, Gianluca E., 1
Ledesma Corniel, María Felipa, 471
Lenci, Alessandro, 53, 120, 303
Lipp, Veronika, 272
Loginova, Olga, 546
Lops, Pasquale, 396

Madeddu, Marco, 243
Madeddu, Mauro, 53
Maffia, Marta, 251
Maggini, Marco, 455
Magnini, Bernardo, 1, 113, 575
Mambrini, Francesco, 86
Manfredi, Ilaria, 551
Manzella, Federico, 510
Marchi, Simone, 570
Marchisio, Manuela, 259
Marra, Andrea, 369
Martelli, Federico, 272

Massey, Dylan, 540
Mazzei, Alessandro, 70, 259, 535
Melis, Matteo, 281
Merlo, Paola, 292
Miaschi, Alessio, 36, 342
Miliani, Martina, 53, 303
Misuraca, Michelangelo, 476
Mohtaj, Salar, 316
Monti, Johanna, 128
Moretti, Giovanni, 154
Möller, Sebastian, 316

Nastase, Vivi, 292
Navigli, Roberto, 272
Negri, Matteo, 203
Nimb, Sanni, 272
Novielli, Nicole, 1
Nozza, Debora, 322

Olsen, Sussi, 272
Onorati, Dario, 562

Palmero Aprosio, Alessio, 521
Papa, Federica, 334
Papucci, Michele, 342
Passaro, Lucia, 53, 227
Passarotti, Marco, 86, 154, 172
Passarotti, Marco C., 219
Patti, Viviana, 243, 403
Pavlopoulos, John, 485
Pellegrini, Matteo, 369
Persico, Oriana, 28
Pierucci, Maria Laura, 128
Pisoni, Raphael, 78
Polidori, Ludovica, 515
Polignano, Marco, 480
Preti, David, 10
Pucci, Giulia, 351, 557

Quochi, Valeria, 272

Rambelli, Giulia, 143
Ramponi, Alan, 433
Ranaldi, Federico, 361
Ranaldi, Leonardo, 10, 351, 361, 557, 562
Rigutini, Leonardo, 455
Rizzi, Giulia, 46, 63
Rodolà, Emanuele, 385
Romagnoli, Raniero, 10, 361
Rosola, Martina, 369
Rosso, Paolo, 490

Ruffolo, Paolo, 219
Russo, Francesco, 379
Ruzzetti, Elena Sofia, 361, 557, 562
Röthenbacher, Thomas, 465

Salgado, Ana, 272
Salova, Anastasiia, 281
Sammaruga, Dario, 259
Samo, Giuseppe, 292
Sanford Pedersen, Bolette, 272
Sangati, Federico, 128
Santilli, Andrea, 385
Sarti, Gabriele, 78
Sartor, Marco, 420
Schettino, Loredana, 251, 441
Schmitt, Vera, 316
Sciavicco, Guido, 510
Sciolette, Flavia, 570
Semeraro, Giovanni, 480
Siciliani, Lucia, 396, 480
Simon, László, 272
Spada, Ivan, 403
Spano, Maria, 476
Spaziani, Elena, 412
Spina, Stefania, 100
Sprugnoli, Rachele, 420
Stan, Eduard Ionel, 510
Stranisci, Marco Antonio, 494
Strapparava, Carlo, 196, 506
Suglia, Alessandro, 530

Tamponi, Lucia, 172
Tanzi, Emanuele, 396
Terragni, Silvia, 78
Tiberius, Carole, 272
Tonelli, Sara, 433
Tron, Susanna, 540

Urbinati, Alessandra, 494
Ureña-Ruiz, Rafael-J, 272
Uricchio, Tiberio, 128
Uryupina, Olga, 427

Valer, Giovanni, 433
Vellutino, Daniela, 499
Venditti, Davide, 10, 361, 562
Venturi, Giulia, 36
Vernillo, Paola, 515
Vetere, Guido, 28
Vitale, Vincenzo Norman, 251, 441

Werner, Romane, 448

Zamparelli, Roberto, 281
Zanca, Dario, 465
Zaninello, Andrea, 575
Zanollo, Asya, 455
Zanzotto, Fabio Massimo, 10, 351, 361, 557, 562
Zeinalipour, Kamyar, 455
Zugarini, Andrea, 465