SEVENTH
INTERNATIONAL
CONFERENCE

CLIB '26

COMPUTATIONAL LINGUISTICS IN BULGARIA CLIB 2026

PROCEEDINGS

Organiser:

Department of Computational Linguistics

Institute for Bulgarian Language

Bulgarian Academy of Sciences

The Seventh International Conference *Computational Linguistics in Bulgaria* (CLIB 2026) is organised with the support of the National Science Fund of the Republic of Bulgaria under Grant Agreement No. КП-06-МНФ/11 of 18.05.2026.

NATIONAL SCIENCE FUND
Ministry of Education and Science

The National Science Fund is not responsible for the content of the papers presented at the Conference or for the content of the promotional and other Conference materials.

CLIB 2026 is organised by:



Department of Computational Linguistics
Institute for Bulgarian Language
Bulgarian Academy of Sciences

PUBLICATION AND CATALOGUING INFORMATION

Title:	Proceedings of the Seventh International Conference *Computational Linguistics in Bulgaria* (CLIB 2026)
ISSN:	2367 5675 (online)
Published and distributed:	Institute for Bulgarian Language Bulgarian Academy of Sciences
Editorial address:	Institute for Bulgarian Language Bulgarian Academy of Sciences 52 Shipchenski Prohod Blvd., Bldg. 17 Sofia 1113, Bulgaria +3592/ 872 23 02
Copyright:	Copyright of each paper stays with the respective authors. The works in the Proceedings are licensed under a Creative Commons Attribution 4.0 International Licence (CC BY 4.0).  License details: http://creativecommons.org/licenses/by/4.0 Copyright © 2026 Institute for Bulgarian Language

Proceedings of the

Seventh International Conference

Computational Linguistics in Bulgaria



7 – 10 September 2026

Sofia, Bulgaria

PROGRAMME COMMITTEE

Chair:

Svetla Koeva – Institute for Bulgarian Language, Bulgarian Academy of Sciences, Bulgaria

Co-chair:

Petya Osenova – Sofia University, Faculty of Slavic Studies; Institute of Information and Communication Technologies, Department of AI and Language Technologies, Bulgarian Academy of Sciences, Bulgaria

Members:

Svetla Boytcheva – Ontotext, Bulgaria

Rositsa Dekova – University of Plovdiv, Bulgaria

Ivan Derzhanski – Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria

Tsvetana Dimitrova – Institute for Bulgarian Language, Bulgarian Academy of Sciences, Bulgaria

Milena Dobreva – University of Strathclyde, United Kingdom

Hristo Dolutarov – Wiser Technology, Bulgaria

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

Maria Gavriilidou – Institute for Language and Speech Processing, ATHENA Research Centre, Greece

Paraskevi Giouli – Institute for Language and Speech Processing, ATHENA Research Centre, Greece

Milka Hadzhikoteva – New Bulgarian University, Bulgaria

Radu Ion – Research Institute for Artificial Intelligence "Mihai Drăgănescu", Romanian Academy, Romania

Iglika Ivanova – Big Data for Smart Society Institute (GATE), Bulgaria

Keith Peter Kiely – Big Data for Smart Society Institute (GATE), Bulgaria

Ivan Koychev – Faculty of Mathematics and Informatics, Sofia University, Bulgaria

Zornitsa Kozareva – SliceX AI, USA

Cvetana Krstev – University of Belgrade, Serbia

Ilia Krustev – A Data Pro, Bulgaria

Penny Labropoulou – Institute for Language and Speech Processing, ATHENA Research Centre, Greece

Eric Laporte – Université Gustave Eiffel, France

Todor Lazarov – New Bulgarian University, Bulgaria

Natalia Loukachevitch – Research Computing Center of Moscow State University, Russia

Stoyan Mihov – Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Bulgaria

Verginica Mititelu – Research Institute for Artificial Intelligence “Mihai Drăgănescu”, Romanian Academy, Romania

Sussi Olsen – University of Copenhagen, Centre for Language Technology, Denmark

Ana Ostroški Anić – Institute for the Croatian Language, Croatia

Vasile Păiș – Research Institute for Artificial Intelligence “Mihai Drăgănescu”, Romanian Academy, Romania

Maciej Piasecki – Wrocław University of Science and Technology, Poland

Vito Pirrelli – CNR Institute for Computational Linguistics “Antonio Zampolli” (CNR-ILC), Italy

Jelena Prokic – Leiden University Centre for Digital Humanities, The Netherlands

Georg Rehm – Deutsches Forschungszentrum für Künstliche Intelligenz GmbH, Germany

Ewa Rudnicka – Wrocław University of Science and Technology, Poland

Inguna Skadiņa – Tilde / Institute of Mathematics and Computer Science, University of Latvia, Latvia

Milena Slavcheva – Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Bulgaria

Krešimir Šojat – University of Zagreb, Croatia

Ranka Stanković – University of Belgrade, Serbia

Ivelina Stoyanova – Institute for Bulgarian Language, Bulgarian Academy of Sciences, Bulgaria

Hristo Tanev – European Commission, Joint Research Centre (EC-JRC)

Irina Temnikova – Big Data for Smart Society Institute (GATE), Bulgaria

Tinko Tinchev – Faculty of Mathematics and Computer Science, Sofia University St. Kliment Ohridski, Bulgaria

Maria Todorova – Institute for Bulgarian Language, Bulgarian Academy of Sciences, Bulgaria

Ivan Vankov – iris.ai, Bulgaria

Piek Vossen – Vrije Universiteit Amsterdam, The Netherlands

Shuly Wintner – University of Haifa, Israel

Katerina Zdravkova – University St Cyril and Methodius, North Macedonia

ORGANISING COMMITTEE

Chair:

Svetlozara Leseva – Institute for Bulgarian Language, Department of Computational Linguistics, Bulgarian Academy of Sciences, Bulgaria

Daniel Tropolinov – Sofia University St. Kliment Ohridski, Bulgaria

Hristina Kukova – Institute for Bulgarian Language, Department of General and Contrastive Linguistics, Bulgarian Academy of Sciences, Bulgaria

Mihaela Moskova – Institute for Bulgarian Language, Department of General and Contrastive Linguistics, Bulgarian Academy of Sciences, Bulgaria

Nikolaos-Digenis Karagiannis – Identrics, Bulgaria

Stanislav Penkov – Big Data for Smart Society Institute (GATE), Bulgaria

Valentina Stefanova – Institute for Bulgarian Language, Department of Computational Linguistics, Bulgarian Academy of Sciences, Bulgaria

Contents

KEYNOTE TALKS 1

Prof. Ruslan Mitkov *Natural Language Processing in the Era of Artificial Intelligence: The Wind of Change is Blowing* 2

Prof. Preslav Nakov *Factuality Challenges in the Era of Large Language Models* 3

Trustworthy AI and Multimodal Modelling 4

Irina Temnikova, Devora Kotseva, Ruslana Margova, Todor Kiriakov, Yolina Petrova, Nevena Grigorova, Alexander Komarov, Bogomil Katanov, Boryana Kostadinova and Stefan Minkov *PROPer-BG: A New Publicly Accessible Package of Resources for Propaganda Detection in Bulgarian* 5

Jordan Kralev *Direct Projection of Visual Features into Linguistic Embeddings for Domain-Adapted Multimodal Learning* 18

Öner Aytaş, Mehmet Ali Bayram, Tuğçe Şen, Banu Diri and Göksel Biricik *Mitigating Security Vulnerabilities in Open-Source Large Language Models via a Supervisor Agent Architecture* 24

Specialised Methods in NLP and AI 32

Svetla Boytcheva, Sylvia Vassileva and Pavel Boytchev *BGSynthMedic: Synthetic Parallel English-Bulgarian Discharge Summary Corpus with Automatic Annotations* 33

George Pashev, Silvia Gaftandzhieva *SymLang: A Structured Symbolic Intermediate Representation for Optimized Human-AI Communication in Assistive Clinical Technologies* 45

Yolina Petrova and Nikola Blajev *Zero-Shot Detection of AI-Generated Bulgarian Text via Binoculars Framework* 53

Mehmet Ali Bayram, Öner Aytaş, Tuğçe Şen *Morphology-Aware Tokenization Improves Turkish Sentence Embeddings under Controlled Training Conditions* 58

Language Models and Computational Methods for Low-Resource Languages 62

Amina Laggoun, Youness Moukafih, Ouassim Karrakchou, Kamel Smaili *Towards Frugal BERT Models for Maghrebi Dialects: Pruning and Distillation Approaches* 63

Ján Mačutek, Michaela Koščová, Michaela Nogolová and Radek Čech *What to Take as the Basic Verb Form? A Pilot Study in Czech* 75

Radu Ion, Verginica Barbu Mititelu, Elena Irimia, Maria Mitrofan and Vasile Păiş *Benchmarking Language Models on Romanian Eighth-Grade National Exams* 87

Silvie Cinková, Ivana Kvapilíková and Jan Černý *Lexical Surprisal from Large Language Models: Cautionary Notes for Legal Discourse* 96

Natalia Dankova *Analysing French Narrative Texts Generated by ChatGPT-5* 109

Mullosharaf Arabov *A Systematic Benchmark of Machine Transliteration Models for the Tajik-Farsi Language Pair: A Comparative Study from Rule-Based to Transformer Architectures* 118

Mullosharaf Arabov, Karomatullo Habibullozoda and Nurali Shirinov *TajikNLP: An Open-Source Toolkit for Comprehensive Text Processing of Tajik (Cyrillic Script)* 127

Abdelouahab Baflah *A Qualitative Error Analysis of Whisper-Based Sentiment Recognition in Algerian Spoken Arabic: Prosodic and Pragmatic Failures in a Low-Resource Dialect* 136

Brandon Caillahua Mendoza *Genetic Algorithms with Normalized Fitness Functions for Low-Resource Language Identification: A Case Study on Archaic Romanian and Medieval Slavonic Liturgical Manuscripts* 142

Corpora, Annotation systems, and NLP Tools 153

Verginica Barbu Mititelu, Elena Irimia, Simina Popa, Mihaela Cristescu, Catalin Mihaila, Lea Radu, Ioana Bucsa *The Romanian Component of the Parallel ELEXIS Corpus: ELEXIS-Ro* 154

Atanas Atanasov *A Dual-Interface Syntax Tree Editor with Automated Assessment for Learning Management Systems* 163

Maria Sampedro Mella *Developing a Pragmatic Annotation Scheme for a Spoken Corpus* . 170

Tsvetelina Stefanova, Petya Osenova *Evaluating the Linguistic Properties of AI-Generated Bulgarian News Texts* 179

Svetla Koeva *Towards a New Grammar Checker for Bulgarian* 188

Corpus Building and LLM Benchmarking for Language Understanding 199

Karla Csuros, Madalina Chitez, Roxana Rogobete and Petya Osenova *Building a Cross-lingual Corpus of BA Theses: Rationale and Preliminary Observations* 200

Maria Todorova, Valentina Stefanova, Tsvetana Dimitrova, Mihaela Moskova, Hristina Kukova, Svetlozara Leseva, Ivelina Stoyanova, Dimitar Georgiev and Svetla Koeva *Reasoning-BG: A Hybrid Benchmark for Long-Text Understanding and Reasoning in Bulgarian* . 207

Ruslana Margova *Detecting Pseudoscientific Narratives in Bulgarian Science Communication: Nationalistic Mystification, Pseudo-Expertise and Cultural Blind Spots in AI-Assisted Analysis* 218

Mila Marcheva-Nash and Weiwei Sun *PS-CHILDES-BG: A Constituency Treebank of Bulgarian Morphemically Tokenised Child-directed Speech* 228

Special Session on Semantic Relations, Framenets and Ontologies 238

Marina Bagi, Aleksandra Marković and Ranka Stanković *Toward a Serbian FrameNet: Resource Construction, Methodology, and Initial Annotation Results* 239

Stanislav Penkov *A WordNet-Aligned Taxonomy of Logical Fallacies for Low-Resource NLP* 249

Ranka Stanković, Cvetana Krstev, Milica Ikonić Nešić and Miloš Utvić *Relation Extraction from Text of Various Domains: A Serbian Case Study* 258

Maria Khokhlova *A Comparative Study of Synonym Generation in Russian with Instruction-Tuned Large Language Models in a Zero-Shot Setting* 268

Cognitive and Corpus-Based Approaches to Language Processing 276

Andrea Nadalini, Claudia Marzi, Marcello Ferro, Vito Pirrelli, Maria Todorova, Valentina Stefanova, Tsvetana Dimitrova, Hristina Kukova and Svetla Koeva *Assessing Reading Development through Finger-Voice Span: a Comparative Study of Bulgarian and Italian* 277

Ruslana Margova and Irina Temnikova *Experiment Design of a Balanced Eye-Tracking Study for Measuring the Cognitive Impact of the Stimulus Word "Disinformation"* . . 286

Tatiana Sherstinova, Irina Petrova, Aleksei Melnik, Sofiia Chepovetskaia, Karina Azarevich, Viktoria Deborina *Everyday Speech of Russian Students: From Field Recordings to the ESC Speech Corpus (Data Processing and Representation)* 294

Gunta Nešpore-Bērzkalne, Mikus Grasmanis and Ilze Lokmane *Multi-word Expressions in the Electronic Dictionary Tēzaurs: Integrating Data from Different Sources* 303

Dobromir Tsolyov, Keith Kiely, Georgi Goranov, Iglika Ivanova, Stanimir Dimitrov *Optimizing LLM-based Discourse Analysis in Low-Resource Languages: A Case Study of Bulgarian Euro Adoption Narratives* 309

Corpus-Based Approaches to Style, Register, and Meaning 323

Iglika Nikolova-Stoupak, Gaël Lejeune, Eva Schaeffer-Lacroix *Quantifying Literary Style: A Corpus-Based Study* 324

Iryna Karamysheva, Jakob Horsch *Imploring You to Investigate More Languages: A Corpus-based Contrastive Study of Ukrainian and English Causative Constructions* 337

Ekaterina Tarpomanova *Balkan Focus Particles in Bulgarian: 'bash' and 'taman'* 347

Ivan Derzhanski, Olena Siruk *When Bulgarian 'Still' Meets Ukrainian 'Already'* 356

Nevena Grigorova, Stanislav Penkov, Keith Kiely, Dobromir Tsolyov, Iglika Ivanova *Syntactic Patterns in Political Speech: Dependency-Based Ad Hominem Detection in Bulgarian Social Media* 366

Maria Khokhlova, Mikhail Koryshev *Thematic Dominants of German-Language Catholic Hymns: Automatic Analysis using Word Embeddings* 376

Junya Morita *Corpus-based Study of Word Formation: Anti-agentive Nominals in English and Japanese* 384

LIST OF AUTHORS 395

KEYNOTE TALKS

Natural Language Processing in the Era of Artificial Intelligence: The Wind of Change is Blowing

Prof. Ruslan Mitkov (Lancaster University, United Kingdom; University of Alicante, Spain)

Natural Language Processing (NLP) is currently undergoing profound and unprecedented transformations. While it has long been acknowledged that NLP technologies are inherently imperfect and far from achieving complete accuracy, recent advancements have significantly reshaped the field of language and translation technologies. First, Deep Learning methods, and now Large Language Models (LLMs), have revolutionised the field and captured widespread attention across both academia and industry.

This entertaining and easy-to-follow talk sheds light on the future of NLP in the era of Artificial Intelligence (AI).

The talk begins by outlining the history of NLP and Machine Translation, reviewing advances driven by Deep Learning and LLMs. The speaker provides a critical evaluation of the employment of LLMs in NLP and Machine Translation focusing on his recent comparative studies involving LLMs, traditional Deep Learning architectures, and rule-based methodologies across selected NLP tasks and applications.

Finally, the presentation concludes with a discussion on the future of NLP. While the speaker does not claim prophetic insight, he draws on his extensive experience and uses linguistic intelligence as a benchmark to offer predictions about how artificial intelligence may evolve relative to human intelligence.

Factuality Challenges in the Era of Large Language Models

Prof. Preslav Nakov (Mohamed bin Zayed University of Artificial Intelligence, United Arab Emirates)

We discuss the risks, the challenges, and the opportunities that Large Language Models (LLMs) bring regarding factuality. We then present some recent work on using LLMs for fact-checking, on detecting machine-generated text, and on fighting the ongoing misinformation pollution with LLMs. Finally, we present a number of LLM fact-checking tools recently developed at MBZUAI: (i) LM-Polygraph, a tool to predict an LLM's uncertainty in its output using cheap and fast uncertainty quantification techniques, (ii) Factcheck-Bench, a fine-grained evaluation benchmark and framework for fact-checking the output of LLMs, (iii) Loki, an open-source tool for fact-checking the output of LLMs, developed based on Factcheck-Bench and optimized for speed and quality, (iv) OpenFactCheck, a framework for fact-checking LLM output, for building customized fact-checking systems, and for benchmarking LLMs for factuality, (v) LLM-DetectAIve, a tool for machine-generated text detection, and (vi) FRAPPE, a FRAming, Persuasion, and Propaganda Explorer.

Trustworthy AI and Multimodal Modelling

PROPer-BG: A New Publicly Accessible Package of Resources for Propaganda Detection in Bulgarian

Irina Temnikova[1,♡], **Devora Kotseva**[2,†], **Ruslana Margova**[1,‡], **Todor Kiriakov**[2,†], **Yolina Petrova**[2,3,*], **Nevena Grigorova**[1,‡], **Alexander Komarov**[1,◇], **Bogomil Katanov**[2,†], **Boryana Kostadinova**[2,†], **Stefan Minkov**[1,#]

[1]Big Data for Smart Society Institute (GATE),
[2]Identrics, [3]New Bulgarian University, Sofia, Bulgaria

[♡]irina.temnikova@gmail.com,[*]ypetrova@nbu.bg,
[‡]{ruslana.margova,nevena.grigorova}@gate-ai.eu,
[†]{devora.kotseva,todor.kiriakov,bogomil.katanov,boryana.kostadinova}
@identrics.ai,
[◇]av.kom0411@gmail.com, [#]stefanminkov2003@gmail.com

Abstract

Detecting texts containing propaganda has emerged as an important Natural Language Processing (NLP) task. While there are many models and resources for other languages, Bulgarian has scarce resources for this task. This article presents PROPer-BG – the first publicly accessible package of resources for detecting and studying propaganda in Bulgarian language. PROPer-BG is an umbrella package, which contains two sub-parts – the WASPer dataset and taxonomy (created by a team at Identrics) and the PropaGATE dataset and lexical resources (created by a team at GATE Institute). The two datasets contain texts in a different selection of topics, collected from different sources, and annotated at different levels of granularity. PropaGATE contains 1939 Large Language Models (LLM)-generated short texts, manually annotated as containing propaganda or not. Its originality consists of including propaganda examples on the same topics but with contrasting viewpoints (e.g. pro- and anti-capitalism), to prevent models from confusing topics with propaganda. The WASPer dataset was used to train the previously released propaganda detection model WASPer. It contains 1696 human-written and LLM-generated short texts on several topics, annotated for the presence of propaganda and propaganda techniques, using the WASPer taxonomy. Additionally, two lists of superlatives and intensifiers were created, and their presence in both datasets was analysed. The paper presents all the resources and the results from the analysis, as well as the link at which the package can be accessed.

Keywords: propaganda; datasets; Bulgarian; linguistic analysis

1 Introduction

Propaganda originates from Latin "to propagate" and was first used by the Vatican in 1622 for propagating the faith of the Roman Catholic Church (Jowett and O'Donnell, 2018). Later, propaganda acquired pejorative meaning. (Lee and Lee, 1939) defines propaganda as "the expression of opinions or actions carried out deliberately by individuals or groups with a view to influencing the opinions or actions of other individuals or groups for predetermined ends through psychological manipulations".

There is a lot of research on the automatic detection of propaganda in texts, including annual organised Shared Tasks, such as those at Slavic NLP workshop, Sem-Eval, and CLEF-CheckThat! (Da San Martino et al., 2020a; Nakov et al., 2020; Piskorski et al., 2025). Because of this, datasets in multiple languages also exist (Da San Martino et al., 2019; Nakov et al., 2020; Baisa et al., 2019; Abdygalym et al., 2025). The datasets contain a mixture of text types, including parliamentary speeches, news articles, and social media posts (Piskorski et al., 2025). The shortcomings of some of these datasets are that when the texts are long, they may contain a mixture of sentences on different topics and with and without propaganda, while when containing social media texts, the proper texts could not be shared in most cases due to the social media platforms' limitations. As a solution to this, as a new trend, datasets with synthetically generated texts also started to appear (Liu et al., 2025; Pakina et al., 2025; Shah et al., 2024). However, such datasets for the moment exist only for some languages (English, Russian, Arabic, Chinese, Spanish, French).

An additional issue is that often the texts are selected on the basis of the main view, separating the official beliefs from alternative opinions. Such official beliefs are sometimes universal and valid for many countries (like the statements that Covid-vaccines are beneficial). But some other beliefs are culturally and country-specific, for example, the Turkish president Erdogan is regarded as a hero in Arab countries and a villain in European countries. However, annotating only the alternative views as propaganda risks making Machine Learning (ML) models confuse topics and views with propaganda.

Stance-annotated datasets represent a kind of solution to this issue, by presenting an author's attitude toward a topic, typically modelled as pro, contra or neutral. However, rarely do datasets contain both propaganda and stance annotations. Specifically, there are no publicly accessible and synthetically generated text datasets, especially those presenting stance and multiple points of view for Bulgarian language.

As solutions to the above issues, we present a package of publicly accessible Bulgarian language resources for propaganda detection[1]. The resources include two text datasets with different text types and annotation granularity. One - partially annotated for stance (called "views"), and a lexical resource, containing superlatives, which can be used for additional analysis.

Contributions:

- We introduce **the first publicly accessible Bulgarian text datasets with propaganda annotations**. The texts partially originate from news articles' comment sections, while the rest are LLM-generated.
  - The PropaGATE dataset is the first Bulgarian texts dataset to contain propaganda presence, partial propaganda techniques, and stance annotations.
  - The WASPer dataset contains fine-grained annotations for the texts origin, presence of propaganda, propaganda categories, and propaganda techniques. Its efficiency has been confirmed as it was previously used to train the propaganda detection model WASPer (Petrova et al., 2025), achieving an F1-Score of 0.853.
- We **release a curated lexical resource with lists of superlatives and intensifiers, comprising 81 positive and 102 negative expressions**, supporting interpretable, lexicon-based analysis of framing signals in Bulgarian.
- We **present the results of a semi-automatic linguistic analysis of the presence of intensifiers and superlatives in the two datasets**.

PROPer-BG is the overall publicly accessible resource package introduced in this paper. It integrates two complementary components: (i) the WASPer component, which comprises the previously developed propaganda taxonomy, detection models (introduced in a previous article), and the WASPer dataset (released publicly for the first time in this paper); (ii) PropaGATE, a newly created synthetic dataset with propaganda, stance, and partial technique annotations, and two lexicons with superlative and intensifier expressions.

2 Related Work

Propaganda and Persuasion Detection: Propaganda and persuasion detection have been studied through the identification of rhetorical and framing devices in political and social discourse, often relying on manually annotated corpora. Early work on propaganda detection has focused on identifying rhetorical techniques such as loaded language, appeal to authority, and emotional manipulation. A widely adopted taxonomy of propaganda techniques was introduced by (Da San Martino et al., 2019), which has been used in subsequent shared tasks such as (Da San Martino et al., 2020a). These efforts distinguish between sentence-level classification and fine-grained span detection, highlighting the complexity of modelling propagandistic language.

More recent approaches rely on transformer-based models such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and domain-adapted architectures, achieving strong performance on propaganda detection benchmarks. However, these models often rely heavily on dataset-specific cues and may fail to generalise across topics and domains.

Stance Detection in Political and Social Text: Stance detection aims to identify an author's attitude toward a target, typically modeled as pro, contra, or neutral. Stance detection has been widely studied in the context of political debates and misinformation analysis, often framed as a target-dependent

[1]The PROPer-BG package, containing a version of WASPer and the up-to-date versions of PropaGAT can be viewed at https://osf.io/2cr7w/overview?view_only=aeeba9a6376e4cdf8246726559dac368

classification task (Mohammad et al., 2016). Recent work explores its interaction with related tasks such as sentiment analysis and argument mining, as well as its role in detecting biased or misleading content. It is closely related to the analysis of political and social debates, including cross-topic and target-dependent settings.

Synthetic Data Generation for NLP: The use of synthetic data generated by large language models has recently gained attention as a strategy to address data scarcity in NLP. Studies such as (Shah et al., 2024) and (Liu et al., 2025) demonstrate that controlled generation can produce diverse and balanced datasets for tasks including misinformation and propaganda detection. However, concerns remain regarding distributional mismatch and the risk of models overfitting to synthetic patterns.

Lexicon-Based and Interpretable Approaches: Lexicon-based approaches have long been used for sentiment and framing analysis, offering interpretable features that complement data-driven models (Taboada et al., 2011). In the context of propaganda detection, lexical cues, such as intensifiers, superlatives, and emotionally charged expressions can signal exaggerated or manipulative framing, although they are not sufficient indicators on their own. Research on propaganda and persuasion detection has emphasized the identification of rhetorical and framing devices in political and social discourse, often relying on large, manually annotated corpora. Closely related work on stance detection focuses on identifying viewpoint polarity with respect to specific targets, a task that remains underexplored for low-resource languages such as Bulgarian. Recent advances in large language models have enabled the generation of synthetic datasets for NLP, providing controlled alternatives to real-world data, particularly in low-resource settings. At the same time, lexicon-based approaches continue to offer interpretable signals for propagandistic language, especially through intensifiers and superlatives.

Despite these advances, existing work remains limited in several respects: (i) low-resource languages such as Bulgarian are underrepresented; (ii) few datasets combine propaganda detection with stance annotations; and (iii) the role of synthetic data in this domain is still underexplored. Our work addresses these gaps by introducing a Bulgarian-language resource that integrates propaganda, stance, and lexical signals within both human-written and LLM-generated data.

2.1 Propaganda Datasets for Other Languages and for Bulgarian

Table 1 summarizes the main propaganda datasets, including those for English, Slavic languages, Romanian (as close to Bulgarian), and Bulgarian. The datasets with Bulgarian texts contain parliamentary speeches and Twitter/X posts. Parliamentary speeches are long texts with many sentences on diverse topics, some of which may contain propaganda, while others do not. Twitter/X posts are, instead not directly accessible as texts, according to Twitter/X's requirements to share only tweet ids.

In contrast, both our datasets contain only short texts of 1-3 sentences (which restricts the diversity of topics per text). Additionally, our datasets are the first Bulgarian-language propaganda datasets, containing synthetically generated examples. Also, we are not aware of other datasets which show multiple viewpoints on the same topic, which makes the contribution of PropaGATE unique.

3 PROPer-BG Description

PROPer-BG is an umbrella package comprising the WASPer solution, the PropaGATE dataset, and two lexical resources; its structure and release status are summarised in Table 2. WASPer comprises a dataset, a hierarchical propaganda taxonomy, and detection models. The taxonomy and some models were released previously, while the WASPer dataset is publicly released for the first time in this paper and is described in Section 3.1. The taxonomy is shown in Figure 1 and Appendix A. PropaGATE is described in Section 3.2, and the lexical resources in Section 3.3; both are used in the analysis in Section 4.

3.1 WASPer Dataset

The WASPer dataset contains texts in English and Bulgarian and is composed by training and testing datasets. **The training datasets** consist of a balanced set of examples that include both propaganda and non-propaganda content. These examples are collected from a variety of traditional media and social media sources, ensuring a diverse range of content. The **"Propaganda" examples** include an equal distribution of human-generated and Artificial Intelligence (AI)-generated propaganda examples. Human-generated examples were manually selected and annotated by domain experts based on a specially designed taxonomy of propaganda techniques (see Section 3.1.1). The AI-generated

Dataset	Year	Lang.	Domain	Source	Annot.	Granul.	Size	Type	Stance	Tech.
SlavicNLP 2025 (Piskorski et al., 2025)	2025	BG, PL, HR, SL, RU	News, parliament, social	Human	Manual	Sent., doc	~10k	Real	No	No
SemEval-2020 (Da San Martino et al., 2020b)	2020	EN	News	Human	Manual	Span	~500 art.	Real	No	Yes
Propaganda Techniques Corpus (Da San Martino et al., 2019)	2019	EN	News	Human	Manual	Span	~18k	Real	No	Yes
CLEF / CheckThat! (Nakov et al., 2020)	2019–24	EN, FR, IT, DE, RU, PL, ES, EL, KA, AR, PT, SL, BG	Political, media	Human	Mixed	Claim	>20k	Real	Partial	Partial
Czech Dataset (Baisa et al., 2019)	2023	CS	News	Human	Manual	Sentence	~4k	Real	No	No
Multi-mil Corpus (Abdygalym et al., 2025)	2023	EN, RU, UK, BG, SR, HR	Info ops	Human	Mixed	Doc.	~50k	Real	No	Partial
SeLeRoSa (Smădu et al., 2025)	2025	RO	News, satire	Human	Manual	Doc.	~5k	Real	No	No
COVID-19 in Bulgarian Social Media (Nakov et al., 2021)	2021	BG	Covid-19	Human	Manual	Doc.	2172 tweets	Real	No	No

Table 1: Unified comparison of propaganda datasets with stance and technique annotations.

Component	Role in PROPer-BG	Release status
PROPer-BG	Umbrella resource package introduced in this paper	New
WASPer solution	Integrated solution comprising a taxonomy, models, and a dataset	Partly previous, partly new
WASPer taxonomy	Hierarchical taxonomy of propaganda categories and techniques	Previously released
WASPer models	Models for propaganda detection	Some previously released
WASPer dataset	Annotated dataset underlying the WASPer solution	First public release in this paper
PropaGATE	Separate annotated dataset included in PROPer-BG	New
Lexical resources	Superlative and intensifier lexicons	New

Table 2: Structure and release status of the resources included in PROPer-BG.

examples were produced using LLMs and subsequently reviewed and annotated by domain experts to ensure consistency with the human-generated content. The specific approach for generating synthetic propaganda is further explored in below. The **"Non-Propaganda"** examples were sourced similarly from traditional and social media platforms, and some were AI-generated. To ensure variety and relevance, these examples were filtered and sampled through a topic modelling process aimed at capturing a wide range of topics and reducing bias in the dataset. Topic modelling is a technique used to uncover hidden themes within large collections of documents by automatically grouping texts based on their underlying topics.

Topic modelling was performed using BERTopic (Grootendorst, 2022), a state-of-the-art library for topic modelling, which leverages language model embeddings to identify coherent topics in text data. BERTopic integrates two key algorithms: UMAP (Uniform Manifold Approximation and Projection) and HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise). UMAP first reduces the dimensionality of the text data, transforming it into a lower-dimensional space that captures the relationships between different texts. HDBSCAN then identifies clusters within this reduced space, grouping similar texts into distinct topics based on density. To guarantee that the non-propaganda dataset includes content related to the same topics as the propaganda examples, zero-shot topic modelling was also applied. In this approach, the already described topics were used as a guide to locate non-propaganda examples discussing similar themes. This approach ensures that non-propaganda examples are present across the same range of topics as propaganda, allowing for balanced and comprehensive coverage of themes This overall process helps organise content and ensures that the dataset covers a wide range of topics without requiring extensive manual annotation. Yet, the sampled examples were verified later. Finally, each example in the dataset was labelled with three types of annotations:

- Whether the content is synthetic (AI-generated) or organic (human-written);
- Whether it contains propaganda;
- If propaganda is present, the specific category of the propaganda.

The total size of the two training sets was as follows: 734 examples in Bulgarian and 862 examples in English (for a more detailed distribution split,

Language	Label	Human	AI-generated
Bulgarian	Propaganda	206	161
Bulgarian	No propaganda	345	22
English	Propaganda	214	206
English	No propaganda	442	0

Table 3: Distribution of the WASPer training data by language, propaganda label, and text origin.

Language	Label	Human	AI-generated
Bulgarian	Propaganda	12	13
Bulgarian	No propaganda	12	13
English	Propaganda	12	13
English	No propaganda	10	13

Table 4: Distribution of the WASPer evaluation data by language, propaganda label, and text origin.

refer to Table 3 below). To supplement the organic propaganda examples, synthetic data was generated using LLMs. This approach was designed to extend the dataset, particularly to include AI-generated propaganda scenarios as well. The generation of synthetic data was driven by few-shot prompting, where the models were provided with carefully selected organic propaganda examples. Different examples were provided for each propaganda technique to ensure the generated content varied in vocabulary, structure, and style. These prompts included specific details related to each propaganda technique, guiding the model to generate content that adhered to the established taxonomy.

Taxonomy-driven generation was also applied. The same taxonomy utilised for annotating the organic examples was applied during the generation of synthetic content, ensuring consistency across the dataset. In addition to organic examples, the models were provided with definitions of each technique found within them. The synthetic examples were reviewed and annotated by domain experts to verify their quality and alignment with the organic examples. To manage the prompting process across different models and tasks, the LangChain framework was used. This framework helped organise the interactions with the language models and streamline the generation process by structuring the prompts and outputs in a consistent manner. The models used were BgGPT (Alexandrov et al., 2024) for Bulgarian, and Mistral (Jiang et al., 2023) and LLaMA (Touvron et al., 2023) for English.

In addition to the training dataset, a separate **evaluation dataset** was created to assess the overall performance of the inference pipeline. This evaluation dataset was sourced similarly to the training set. It includes both organic and AI-generated content from various traditional and social media platforms to ensure diverse and balanced coverage across all labels, including the specific propaganda categories (Table 4 for reference). The Bulgarian evaluation set consists of a total of 50 examples, 25 of which non-propaganda and 25 propaganda examples, 26 of which organic and 24 AI-generated examples. The English evaluation set also has a total of 50 examples and contains 25 non-propaganda examples and 25 propaganda examples, 26 of which organic and 22 AI-generated.

3.1.1 Taxonomy of Propaganda Techniques

The WASPer taxonomy is shown on Figure 1. It was designed to systematically classify and detect AI-generated propaganda in both traditional media comment sections and social media platforms. This taxonomy progresses through multiple stages, each aimed at narrowing down specific characteristics of the content, starting from the basic identification of text type (as human or artificially generated) and advancing to a detailed classification of the propaganda techniques present within the message. The taxonomy contains the following four levels:

- Level 1: Text Origin Identification (Human vs. AI-generated);
- Level 2: Presence of Propaganda (Propaganda vs. No Propaganda);
- Level 3: Propaganda Category;
- Level 4: Propaganda Technique.

The taxonomy consists of 22 separate propaganda techniques (on the lowest level) grouped into five parent categories (Figure 1 for reference). Appendix A provides more details about it. The taxonomy is open source and can be accessed online[2].

3.2 PropaGATE Dataset

The PropaGATE dataset is completely synthetically generated and contains 1939 short "social media-like" texts, generated with BARD, OpenAI 3.5 and 5.4 in different ways. The texts containing propaganda were generated by providing examples from Facebook, deemed to contain propaganda by one human annotator. Subsequently, the same number of factual non-propaganda texts were generated with the ChatGPT OpenAI model 5.4, corresponding to the number of the propaganda-containing texts.

[2] https://github.com/Identrics/wasper.

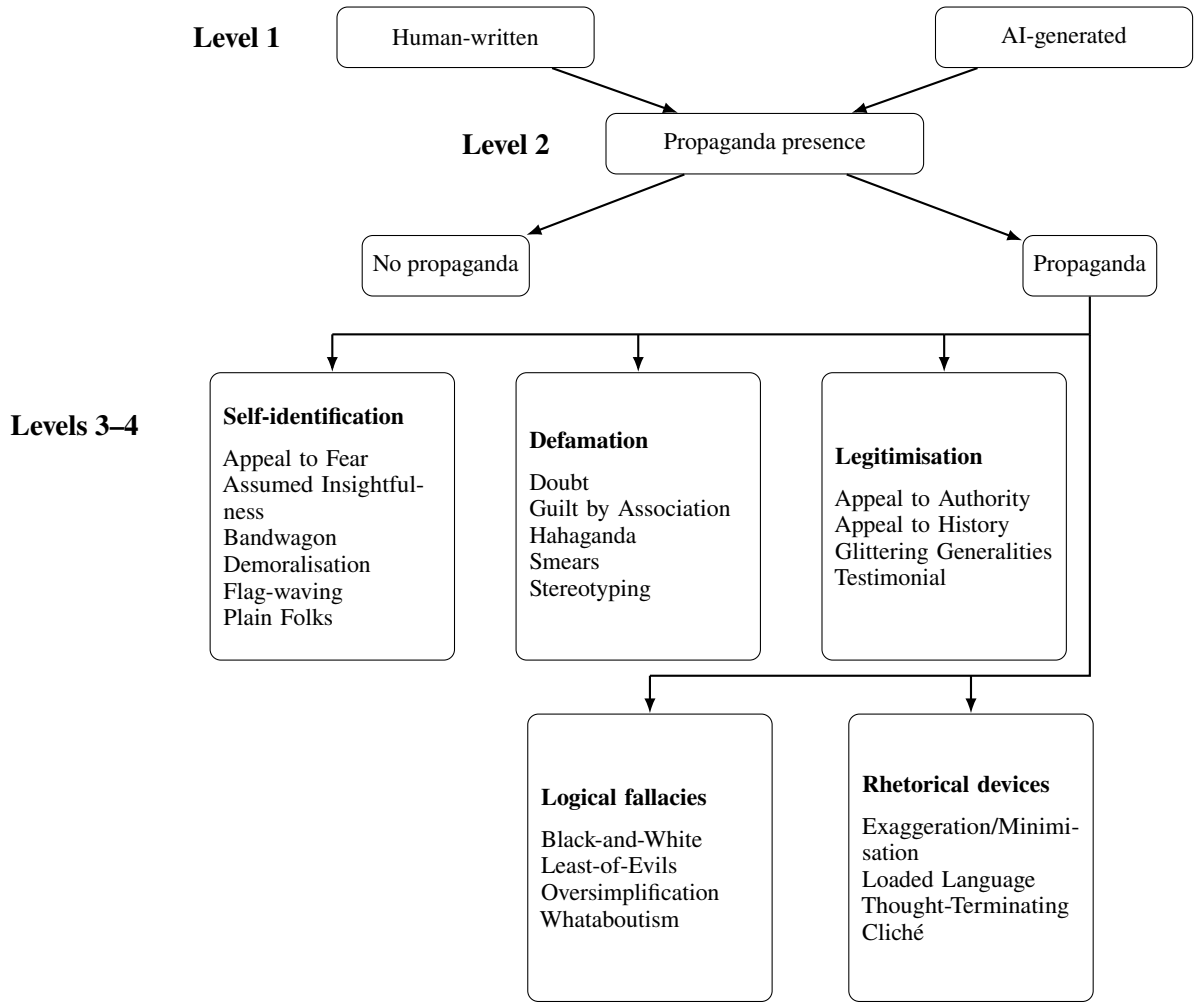


Figure 1: Overview of the four-level WASPer taxonomy. Level 1 identifies text origin, Level 2 identifies propaganda presence, and Levels 3–4 assign propaganda categories and techniques.

The recorded prompts used for generating the texts can be viewed online[3] or obtained by contacting Irina Temnikova[4].

Each text was also manually annotated by one linguist for topic and stance (called "view"). The stances partially reflected the standard pro-, contra-, and neutral ones, and on some occasions reflected the existing society's beliefs (like global warming views were: 1) pro-global-warming as a problem; 2) against global warming as a problem; 3) pro-global warming as a conspiracy; 4) neutral). Next, two human annotators (Bulgarian language linguists, specializing in disinformation and propaganda detection) reviewed each generated text for whether it was written in correct Bulgarian, and whether it contained propaganda or not. The decision of whether it contains propaganda was taken on the basis of several propaganda definitions and the description of propaganda techniques in the WASPer taxonomy and in (Da San Martino et al., 2020a). A text was considered to contain propaganda if it attempted to convince the reader that something was bad or good and if it contained at least one of the propaganda techniques described in the WASPer taxonomy. The annotators were also asked to annotate at least the first propaganda technique they noticed in each tagged as propaganda-containing text. We used two annotation tools. We first used Datasaur[5] (the annotation interface is shown in Figure 2). When the available free annotations ended, we continued the annotation in Google Spreadsheets, using fall-down menus (see Figures 3 and 4). We obtained substantial agreement with Cohen's k = 0.717, and percentage agreement of 85.35% in terms of whether the text contained propaganda or not. Due to the fact that propaganda techniques

[3] https://osf.io/2cr7w/overview?view_only=aeeba9a6376e4cdf8246726559dac368

[4] irina.temnikova@gmail.com

[5] https://datasaur.ai/data-annotation

were not comprehensively annotated, we did not calculate agreement for them.

Table 6 in Appendix B shows the number of texts, their topics, subgroups and groups in PropaGATE.

3.3 Lexical Resources

As an addition to PropaGATE, we provide two manually curated lexicons - one with 366 expressions, out of which (i) 162 (81 ADJ and 81 ADV) positive superlatives and (ii) 204 negative (102 ADJ and 102 ADV) superlatives. Additionally, we include a list of intensifiers, used in Section 4. These lexicons can be used to approximate lexical intensification as a proxy signal for exaggerated or evaluative framing. Some **Examples** are: "изключително", "ужасно", "безспорно" (English translation: "exceptionally", "terribly", "undoubtedly").

4 Linguistic Analysis for Intensifiers and Superlatives

We conducted an exploratory manual analysis of the occurrence and functions of intensifiers and superlatives from our lexical resources in both datasets. Table 5 presents representative intensifiers for several broad topic categories in PropaGATE. The analysis suggests several category-specific tendencies. For example, texts related to COVID-19 and vaccines contain lexical items emphasising authority, safety, and collective responsibility, typically supporting expert-driven and institutional framing. Climate change texts employ intensifiers expressing urgency and escalating risk, whereas ideology-related texts favour universal and exclusionary formulations. The Gender and Social Roles category relies on comparative and capability-denying expressions that construct hierarchies and discredit target groups. Finally, the Wealth and Inequality category combines blame, dehumanisation, and moral judgement, framing particular groups as responsible for systemic problems.

When analysing WASPer, we observed that intensifiers and superlative-like expressions are not distinguished by their frequency but by how they are used in context. In propaganda texts, intensifiers such as "супер" (super), "ужасно" (terribly), "страшно" (horribly), "абсолютно" (absolutely), "напълно" (completely) typically appear in structured patterns that combine emotional amplification with negative or ideologically loaded content. They are often attached to negative events or enemies, as in "супер мощна атака" (super powerful attack) or "ужасно решение" (terrible decision), or to evaluative nouns like "невероятно лицемерие" (incredible hypocrisy) and "изумително безобразие" (astonishing outrage). These constructions serve functions such as dramatization, moral judgment, and delegitimization. In addition, hyperbolic expressions like "абсолютно доказателство" (absolute proof) and "напълно ясно" (completely clear) present opinions as facts and remove doubt. Instead, the non-propaganda texts use intensifiers in a more neutral and descriptive way. Common expressions such as "много" (very), "доста" (quite), "малко" (a bit) appear in phrases like "много трудно" (very difficult), "много странно" (very strange), or "доста интересно" (quite interesting). Here, the function is descriptive rather than persuasive. An important difference is in the sentiment shift. In propaganda, intensifiers push already evaluative statements toward extremes — for example, "лошо" (bad) → "ужасно" (terrible) or "проблем" (problem) → "огромен проблем" (huge problem) — thus creating emotional polarization. In non-propaganda texts, intensifiers such as "много" (very) only slightly strengthen meaning, as in "трудно" (difficult) → "много трудно" (very difficult), without moving into extreme emotional territory. The observation suggest that the same linguistic forms — such as "супер" (super), "ужасно" (terribly), "много" (very), "невероятно" (incredibly) — can function very differently depending on context: in propaganda, they act as strategic tools for emotional manipulation and persuasion, while in non-propaganda - primarily serve descriptive and informational purposes. The comparison between the WASPer and PropaGATE datasets suggests two distinct patterns of lexical intensification in relation to propaganda. The WASPer dataset exhibits a broader and more natural use of intensifiers such as "супер" (super), "ужасно" (terribly), and "невероятно" (incredibly), which occur in both propaganda and non-propaganda texts. Their communicative function is highly context-dependent: in propaganda they tend to amplify emotional and ideological messages, whereas in non-propaganda they are used primarily for descriptive or conversational purposes. Consequently, these expressions alone constitute weak indicators of propaganda. In contrast, the PropaGATE dataset contains fewer instances of lexical intensification but relies more frequently on explicit superlative constructions such as "най-голям" (the biggest) and "най-лош" (the

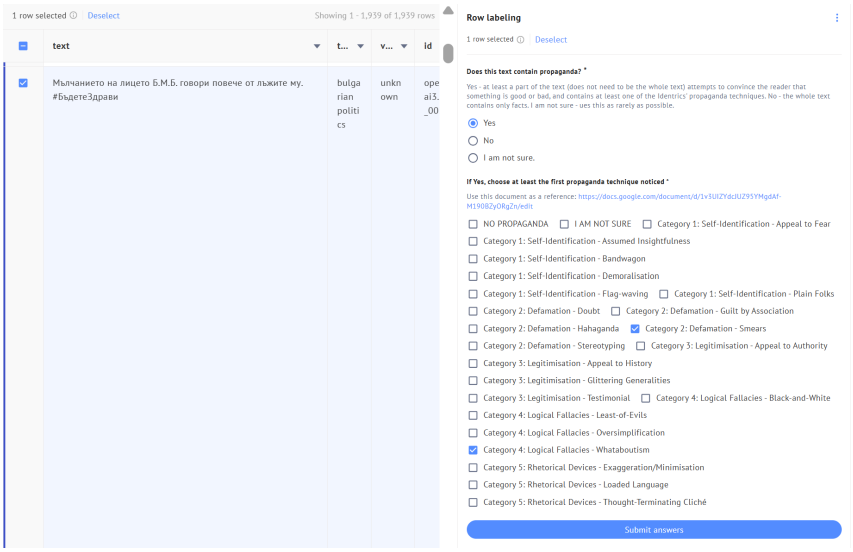


Figure 2: PropaGATE Annotation in Datasaur.

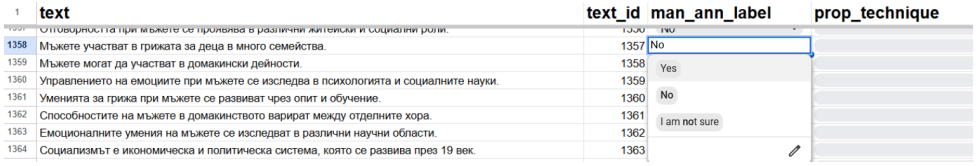


Figure 3: PropaGATE Annotation in Google Spreadsheets - fall down menu for Propaganda or Not.

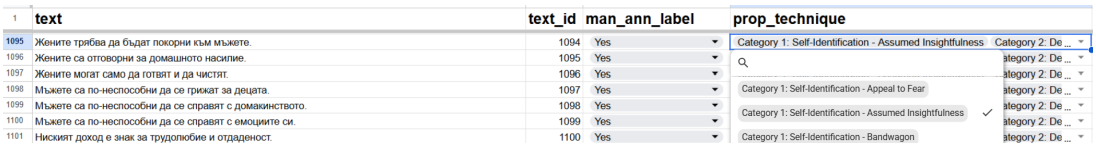


Figure 4: PropaGATE Annotation in Google Spreadsheets - fall down menu for Propaganda techniques.

worst). These forms directly express categorical and extreme evaluations, often corresponding to rhetorical strategies such as black-and-white thinking and exaggeration. A further difference concerns emotional intensity. In WASPer, intensifiers typically create a gradual shift from moderate to stronger sentiment (e.g., "лошо" (bad) → "ужасно" (terrible)), whereas in PropaGATE superlative constructions already encode maximal intensity without intermediate stages. Despite these differences, neither dataset suggests that intensifiers or superlatives alone are reliable markers of propaganda. The observations suggest that propaganda relies less on overtly extreme language such as "най-. . . " (the most. . .) and more on subtle lexical intensification combined with discourse context and other rhetorical strategies, including loaded language, fear appeals, and labelling. The identified intensifiers were compiled into a second lexical resource comprising 70 polarity-marked expressions, which is released together with the superlative lexicon as part of the PROPer-BG package.

The observations presented in this section are intended as an exploratory linguistic analysis. They are descriptive rather than statistically validated and should therefore be interpreted as qualitative evidence of different patterns of lexical intensification across the two datasets. Future work will extend this exploratory analysis with corpus-level quantitative measurements and statistical significance testing.

5 Limitations and Ethical Aspects

It should be noted that the real intent of the authors of the human posts in the WASPer dataset remain unclear. Annotating some as propaganda, thus, should not be considered as a hypothesis. In PropaGATE, because the data is fully synthetic, it is not intended to represent real actors or events and should not be used as a standalone basis for

Topic	Bulgarian Intensifiers	English Translation	Type
Vaccines / COVID-19	ключова; изключително; най-ефективните; смъртоносни; строги; безопасни; много нисък риск; социална отговорност; защита на уязвимите	key; extremely; most effective; deadly; strict; safe; very low risk; social responsibility; protection of the vulnerable	Institutional / Moral
Climate Change	най-голямата заплаха; катастрофални; тревожно бързо; спешни действия; неизбежен; опустошителни; сериозни щети; не можем да си позволим	greatest threat; catastrophic; alarmingly fast; urgent actions; inevitable; devastating; serious damage; cannot afford	Alarmist
Ideologies	винаги; единствената система; най-успешната; обречена на неуспех; няма място; трябва да бъде; в основата на всички	always; the only system; most successful; doomed to fail; no place for; must be; at the core of everything	Absolutist
Gender / Social Roles	по-малко способни; не могат; по-слаби; по-емоционални; по-малко интелигентни; не са в състояние	less capable; cannot; weaker; more emotional; less intelligent; not capable of	Discrediting
Wealth / Inequality	най-големите врагове; паразити; крадат; винаги; трябва да бъдат наказани; фалшиво щастие	greatest enemies; parasites; steal; always; must be punished; false happiness	Accusatory

Table 5: Topic-specific Bulgarian-language lexical intensifiers with English translations and functional categorisation.

real-world propaganda detection. Intensification is treated as a proxy signal rather than a definitive indicator of propaganda, as exaggerated language may also occur in persuasive but non-manipulative contexts. Finally, the lists of superlatives and intensifiers are not exhaustive.

The publicly accessible resources are supplied with legal disclaimers, listing these limitations and that these resources should not be trusted to unambiguously distinguish propaganda.

6 Conclusions

In this article, we have introduced a publicly accessible package, containing two datasets with Bulgarian texts, annotated with propaganda (and one also with stance), as well as two lexicons with superlative and intensifier expressions. We have also linguistically analysed which intensifiers and superlatives, how many, and in what function, appear in texts, annotated as containing propaganda and not in both datasets. The new package will be very useful for training propaganda detection models for Bulgarian, as well as for expanding the linguistic knowledge about propaganda in Bulgarian.

7 Acknowledgments and Authors' Contributions

This research is part of the GATE project funded by the Horizon 2020 WIDESPREAD-2018-2020, TEAMING Phase 2, the Programme under grant agreement no. 857155, the Programme "Research, Innovation and Digitalization for Smart Transformation" 2021-2027 (PRIDST) under grant agreement no. BG16RFPR002-1.014-0010-C01 and, the project BROD (Bulgarian-Romanian Observatory of Digital Media), funded by the Digital Europe Programme of the European Union under contract number 101226153.

Authors' Contributions. From the Institute GATE's team, Irina Temnikova led the internally funded project, during which the dataset was created, wrote most parts of this article (except the WASPer- and Identrics-related paragraphs and sections), generated the non-propaganda texts and the lexical resources, annotated PropaGATE for topic and view, created the propaganda or not annotation guidelines, set the annotation tools, calculated IAA, ran the linguistic analysis, organised GATE's review response, and contributed to the final camera-ready version of the article. Alexander Komarov generated the BARD pro- and anti- texts. Stefan Minkov generated the texts with OpenAI 3.5. Nevena Grigorova checked the Bulgarian of the BARD-generated texts, and together with Ruslana Margova manually annotated all PropaGATE texts for containing propaganda or not, and with at least 1 technique. Ruslana Margova has also contributed to the camera-ready article.

From Identrics' team, Devora Kotseva and Todor Kiriakov designed the WASPer Taxonomy, annotated the human-written texts, and verified the synthetically generated texts. Bogomil Katanov developed the prompting strategy and the text gen-

eration pipeline. Boryana Kostadinova performed automated NLP-driven filtering and topic diversity selection of the datasets and contributed to the writing of the manuscript. Yolina Petrova conceived and supervised the overall research process and contributed to the writing of the manuscript. All authors reviewed and approved the final version of the manuscript.

References

Bayangali Abdygalym, Madina Sambetbayeva, Aigerim Yerimbetova, Anargul Nekessova, Nurbolat Tasbolatuly, Nurzhigit Smailov, and Aksaule Nazymkhan. 2025. NLP models for military terminology analysis and detection of information operations on social media. *Computers*, 14(11):485.

Angel Alexandrov, Veselin Raychev, Martin N. Müller, Cheng Zhang, Martin Vechev, and Kristina Toutanova. 2024. Mitigating catastrophic forgetting in language transfer via model merging. *arXiv preprint arXiv:2407.08699*.

Vít Baisa, Ondřej Heřman, and Aleš Horák. 2019. Benchmark dataset for propaganda detection in Czech newspaper texts. In *Recent Advances in Natural Language Processing*.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020a. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online). International Committee for Computational Linguistics.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news articles. In *Proceedings of EMNLP*.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2020b. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of SemEval*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186. Association for Computational Linguistics.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling. *arXiv preprint arXiv:2203.05794*.

Albert Q. Jiang et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Garth S Jowett and Victoria O'Donnell. 2018. *Propaganda & persuasion*. Sage publications.

Alfred Lee and Elizabeth Briant Lee. 1939. *The fine art of propaganda*. Harcourt, Brace.

Jiateng Liu, Lin Ai, Zizhou Liu, Payam Karisani, Zheng Hui, Yi Fung, Preslav Nakov, Julia Hirschberg, and Heng Ji. 2025. Propainsight: Toward deeper understanding of propaganda in terms of techniques, appeals, and intent. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5607–5628.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41. Association for Computational Linguistics.

Preslav Nakov, Firoj Alam, Shaden Shaar, Giovanni Da San Martino, and Yifan Zhang. 2021. Covid-19 in bulgarian social media: Factuality, harmfulness, propaganda, and framing. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 997–1009.

Preslav Nakov et al. 2020. Overview of the clef-2020 checkthat! lab: Automatic identification and verification of claims in social media. In *CLEF Working Notes*.

Anil Kumar Pakina, Ashwin Sharma, and Deepak Kejriwal. 2025. Ai-driven disinformation campaigns: Detecting synthetic propaganda in encrypted messaging via graph neural networks. *International Journal Science and Technology*, 4(1).

Yolina Petrova, Todor Kiryakov, Devora Kotseva, and Boryana Kostadinova. 2025. Wasper: a bulgarian-language model for detecting propaganda in social media content. In *Papers from the International Scientific Conference of the European Studies Department, Jean Monnet Centre of Excellence, Faculty of Philosophy at Sofia University "St. Kliment Ohridski"*, volume 12, pages 223–230.

Jakub Piskorski, Dimitar Dimitrov, Filip Dobranić, Marina Ernst, Jacek Haneczok, Ivan Koychev, Nikola Ljubešić, Michał Marcińczuk, Arkadiusz Modzelewski, Ivo Moravski, et al. 2025. Slavic-NLP 2025 shared task: Detection and classification of persuasion techniques in parliamentary debates and social media. In *Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025)*, pages 254–275.

Uzair Shah, Md Rafiul Biswas, Marco Agus, Mowafa Househ, and Wajdi Zaghouani. 2024. Mememind at araieval shared task: Generative augmentation and

feature fusion for multimodal propaganda detection in arabic memes through advanced language and vision models. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 467–472.

Răzvan-Alexandru Smădu, Andreea Iuga, Dumitru-Clementin Cercel, and Florin Pop. 2025. Selerosa: Sentence-level romanian satire detection dataset. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 6528–6533.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Hugo Touvron et al. 2023. Llama. *arXiv preprint arXiv:2302.13971*.

Appendix A First Appendix

The four-level hierarchical taxonomy provides a solid framework for the detection and classification of AI-generated propaganda. Further, each level is described in depth, detailing the specific features and techniques used to identify and categorise these components.

1. **Human vs. AI-generated**:
    - **Human/Organic examples** are comments sourced from traditional or social media that are assumed to be human-generated content, representing natural, unaltered communication by individuals or organisations.
    - **AI-generated/Synthetic examples** are AI-generated pieces of content created by large language models (LLMs). In this context, synthetic examples are used to simulate AI-driven propaganda and help the model learn to differentiate between human and AI-generated messaging.

2. **Propaganda vs. No Propaganda**: The propaganda or not level indicates whether a specific piece of content contains propaganda technique(s). If labelled as "propaganda", it means that a piece of content associated with propaganda has been detected, but this doesn't guarantee that the entire piece is propaganda — only that a manipulative method has been identified. Content labelled as "No Propaganda" lacks any propaganda content and/or identifiable techniques.

3. **Propaganda categories (Level 3)** and **Propaganda techniques (Level 4)**:
    - **Category 1: Self-Identification Techniques:**
      These techniques exploit the audience's feelings of association (or desire to be associated) with a larger group. They suggest that the audience should feel united, motivated, or threatened by the same factors that unite, motivate, or threaten that group. The propagandist may be internal or external to the group, but the audience must necessarily identify (or strive to identify) with the group.
        - Appeal to Fear
        - Assumed Insightfulness
        - Bandwagon
        - Demoralisation
        - Flag-waving
        - Plain Folks
    - **Category 2: Defamation Techniques:**
      These techniques represent direct or indirect attacks against an entity's reputation and worth.
        - Doubt
        - Guilt by Association
        - Hahaganda
        - Smears
        - Stereotyping
    - **Category 3: Legitimisation Techniques:** These techniques attempt to prove and legitimise the propagandist's statements by using arguments that cannot be falsified because they are based on moral values or personal experiences.
        - Appeal to Authority
        - Appeal to History
        - Glittering Generalities
        - Testimonial
    - **Category 4: Logical Fallacies:** These techniques appeal to the audience's reason and masquerade as objective and factual arguments, but in reality, they exploit distractions and flawed logic.
        - Black-and-White
        - Least-of-Evils
        - Oversimplification
        - Whataboutism

- **Category 5: Rhetorical Devices:** These techniques seek to influence the audience and control the conversation by using rhetorical methods of persuasion.
  - Exaggeration/Minimisation
  - Loaded Language
  - Thought-Terminating Cliché

The full definitions of the techniques can be accessed in the open-sourced taxonomy: [6]

Appendix B Second Appendix

Table 6 shows the number of rows in PropaGATE in total, per Group, Subgroup, Topic, and View.

Appendix C Third Appendix

The synthetic examples were generated using a single instruction template. For each generation, the placeholder `propaganda definitions` was replaced with the complete list of propaganda techniques and their corresponding definitions, while `example` and `technique list` were substituted with a human-written example and its associated labels to provide in-context guidance.

System instruction

You are an expert Large Language Model capable of generating high-quality synthetic data when given instructions on the desired type of data. You will be generating examples of social media messages that contain some sort of propaganda. You will be given a list of propaganda techniques that includes the names and a detailed descriptions of them. Your task will be to generate a single synthetic example in Bulgarian based on the techniques and a list of the used techniques in the synthetic example by carefully examining and using the definition.

Here are the techniques and their definitions.

Propaganda techniques and definitions `propaganda definitions`

In-context example `example technique list`

Instruction

[INST] Generate an example and a list of the techniques, without any clarifications.[/INST]

During generation, the in-context example was varied across prompts to encourage lexical and topical diversity while maintaining consistency with the annotation schema.

[6] https://github.com/Identrics/wasper.

Group	Subgroup	Topic	View	Rows
Politics & Ideology (738)	General (252)	capitalism (78)	pro_capitalism	23
			against_capitalism	16
			neutral	39
		socialism (78)	pro_socialism	20
			against_socialism	19
			neutral	39
		fascism (76)	pro_fascism	9
			against_fascism	29
			neutral	38
		democracy (6)	against_democracy	3
			neutral	3
		nationalism (13)	pro_nationalism	3
			neutral	10
		liberalism (1)	against_liberalism	1
	Bulgarian Politics (486)	bulgarian politics (446)	unknown	376
			neutral	70
		bulgaria euro (40)	pro_euro	10
			against_euro	10
			neutral	20
Society (444)	Economic Class (160)	rich_people (74)	pro_rich_people	21
			against_rich_people	16
			neutral	37
		low_paid_jobs (86)	pro_low_paid_jobs	31
			against_low_paid_jobs	11
			neutral	43
	Gender & Age Class (222)	women (124)	pro_women	32
			against_women	30
			neutral	62
		men (38)	against_men	19
			neutral	19
		lgbt (56)	against_lgbt	28
			neutral	28
		gender equality (4)	pro_gender_equality	2
			neutral	2
	Rights (62)	abortion (62)	pro_abortion	18
			against_abortion	13
			neutral	31
Religion (234)	General (38)	religion (38)	pro_religion	4
			against_religion	15
			neutral	19
	Christianity (122)	religion – christianity (122)	pro_religion_christianity	33
			against_religion_christianity	28
			neutral	61
	Islam (74)	religion – islam (74)	pro_religion_islam	19
			against_religion_islam	18
			neutral	37
Health, Science & Environment (246)	Health (66)	covid_vaccines (54)	pro_covid_vaccines	9
			against_covid_vaccines	28
			neutral	17
		vaccines (12)	pro_vaccines	2
			against_vaccines	4
			neutral	6
	Environment & Climate (180)	global_warming (122)	pro_global_warming_as_problem	34
			against_global_warming_as_a_problem	16
			pro_global_warming_as_conspiracy	11
			neutral	61
		climate_change (24)	pro_climate_change_as_problem	10
			against_climate_change	2
			neutral	12
		ecology (4)	pro_ecology	2
			neutral	2
		traditional_science_believes (30)	against_traditional_science_believes	15
			neutral	15
TOTAL SHOWN IN TABLE				**1662**

Table 6: Updated topic–view distribution of PropaGATE with original tags.

Direct Projection of Visual Features into Linguistic Embeddings for Domain-Adapted Multimodal Learning

Jordan Kralev
Department of Systems and Control, Technical University of Sofia, Bulgaria
Institute for Bulgarian Language, Bulgarian Academy of Scienses
jkralev@tu-sofia.bg, jordan@dcl.bas.bg

Abstract

This paper introduces a novel multimodal learning framework that directly projects visual features into linguistic embeddings, enabling robust, domain-adapted object recognition across diverse categories. Leveraging state-of-the-art visual (Detectron2 Mask R-CNN) and language (TinyLlama) models, the architecture incorporates a trainable projection layer that maps convolutional visual features into the embedding space of a large language model. This approach addresses the challenge of bridging high-dimensional, region-based visual representations with sequential, token-based language embeddings using minimal fine-tuning. The model is evaluated on the specialised Multilingual Image Corpus (MIC21), a rich dataset containing over 21,000 images and more than 700 categories, extending beyond standard benchmarks and annotated with multilingual captions. Experimental results demonstrate the effectiveness of this system for multimodal tasks, including image captioning and visual question answering. The design supports scalable adaptation to new domains and tasks, offering a flexible solution for unified visual–linguistic representation learning.

Keywords: multimodal learning, visual-linguistic embeddings,domain adaptation,object recognition

1 Introduction

Despite significant advances in deep learning models for processing complex signals – visual, audio, and linguistic – large models still struggle to exploit complexity by fusing domain-specific data. Applications for multimodal models that process visual and textual data are evident in medicine, autonomous systems, education, and manufacturing, among other domains. Tasks such as image captioning and visual question answering depend on both image and textual data.

The current state-of-the-art framework for visual large language models (LLMs) is CLIP (Contrastive Language–Image Pre-training). Introduced by OpenAI, it connects visual and linguistic information using contrastive objectives. Its design comprises two parallel encoders trained jointly: the text encoder converts a sentence into a fixed-length embedding reflecting its semantic meaning, while the image encoder transforms an input image into a vector summarising its visual content. Through contrastive learning, the model aligns matching image–text pairs in a shared embedding space while separating unrelated ones, enabling it to assess how closely a caption corresponds to a given image. OpenAI CLIP models are trained on WebImage-Text, a dataset of 400k image/caption pairs. The total word count of this dataset is comparable to that of the WebText dataset, which contains about 40 terabytes of text data and was used to train GPT-2 (Radford et al., 2021). Image–text pairs are usually only loosely correlated, as a caption can match multiple images in addition to the ground truth (Cheng et al., 2021). This is also reflected in the discriminative ability and transfer performance of the visual encoder pre-trained with a contrastive objective (Yang et al., 2022).

There are visual models for self-supervised learning (SSL) that have been pre-trained with non-contrastive objectives (Caron et al., 2021). Beyond the contrastive paradigm, some visual SSL models succeed in mitigating dependence on negative samples (Schiappa et al., 2023). With various refinements to avoid collapsing solutions, these works optimise the affinity of augmented representations alone and are categorised as non-contrastive frameworks. For example, to avoid model collapse, asymmetric architectures (Chen and He, 2021), dimensional decorrelation (Bardes et al., 2022), and clustering (Assran et al., 2022; Schiappa et al., 2023) are employed.

Aligning the high-dimensional, region-level visual embeddings produced by modern object detectors such as Detectron2 or YOLO with the sequential, tokenised representations of language models such as Llama presents a significant technical challenge. Simple strategies such as direct concatenation or shallow fusion are expected to yield poor performance because the modalities operate in incompatible feature spaces and lack meaningful cross-modal interactions. While state-of-the-art frameworks such as CLIP rely on extensive training of a combined visual and textual embedding matrix, the present work adopts a more system-level approach, seeking to establish a connection between robust pre-trained visual and linguistic models with minimal fine-tuning of the connecting layers. The goal of this article is to develop a general and adaptable framework that enables object recognition to transfer across domains with custom categories through a learnable projection layer, unifying visual and linguistic embeddings to support robust multimodal training for diverse downstream tasks.

The study examines possible interconnections between the Detectron2 Mask R-CNN visual model and the TinyLlama large language model (LLM), linked through a trainable projection layer from the convolutional feature space into LLM embeddings. The performance of the architecture is demonstrated on a specialised multi-domain dataset called the Multilingual Image Corpus (MIC21). The original purpose of the MIC21 dataset was to expand the 80 COCO categories, commonly used for image processing models, into an extensive ontology featuring over 700 classes, organised into 130 thematic subdomains covering sport, transport, arts, and security. Currently, the dataset contains around 21,000 images with more than 200,000 object detections. It was recently extended with textual captions and descriptions for each image. Images were collected from diverse public sources and annotated based on a custom ontology, which is linked to WordNet and translated into 25 languages, providing rich semantic detail and multilingual support. While the text-generation model developed in the paper is English-only, the design could be extended for multilingual support. A fine-tuning approach to Detectron2 for domain-specific data (outlined in previous work by the authors) can further improve the capacity for knowledge transfer between models through the projection layer.

The paper is organised as follows. Section 2 describes the model architecture. Section 3 presents the training strategy, and Section 4 discusses the experimental results and provides concluding remarks.

2 Model Architecture

The proposed model is interconnection between backbone components of two heterogeneous deep learning model. The first component is derivation from Facebook Detectron2 visual model of it is a Mask R-CNN with X-101-32x8d-FPN configuration. For a given RGB input image $I \in \mathbf{R}^{H\times W\times 3}$ with height H and width W pixels the backbone component of the FPN extracts a tensor

$$x_i \in \mathbf{R}^{C\times H_F\times W_F}, \tag{1}$$

where $C = 256$ is the filter dimension of the final convolutional layer, $H_F = 16$ and $W_F = 16$ are the dimensions of the convolution map. The feature tensor x_i is then flattened into

$$x_f \in \mathbf{R}^{C\times(H_F W_F)} \tag{2}$$

The feature matrix is then normalized and projected to the embedding dimension d

$$v = W_p F_{drop}(F_{norm}(x_f)) + b_p \tag{3}$$

through the matrix $W_p \in \mathbf{R}^{d\times C}$ and bias vector $b_p \in \mathbf{R}^d$ and $v \in \mathbf{R}^{256\times d}$, $d = 2048$ for the TinyLlama model. Here W_p and b_p are the trainable parameters of the proposed projection layer. The normalization layer F_{norm} applies a conditioning transform

$$x_{fn} = \gamma\frac{x_f - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta \tag{4}$$

where μ and σ are mean and standard deviation of the input vector, while γ and β are trainable parameters. Also during training a dropout filter F_{drop} is applied.

The input to the language processing component of the architecture is a token sequence $t = (t_1, t_2, \ldots, t_T)$. In the language model each token is transformed to an embedding vector to get

$$e = (e_1, e_2, \ldots, e_T) \in \mathbf{R}^{T\times d} \tag{5}$$

where e_i is the embedding of token t_i. Then the projections from the visual convolutional layer are

injected into the prompt so the input to the transformer layer is a concatenation of projected visual tokens and text embeddings

$$z_0 = (e_{pre}, v, e_{post}) \in \mathbf{R}^{(256+T)\times d} \tag{6}$$

The position for injecting visual features into the prompt is chosen to match the model's instruction template and to provide sufficient context to clarify the task at hand. As is well known, the core of sequence generation is a stack of transformer decoder layers. The decoder consists of L layers, each applying multi-head self-attention.

$$a_l = f_{smax}\left(\frac{QK^T}{\sqrt{d}} + M\right)V \tag{7}$$

where queries Q, keys K and values V are all linear projections of z_{l-1}, and $M \in \mathbf{R}^{(256+T)\times(256+T)}$ is causal mask defined with

$$M = \begin{pmatrix} 0_{256\times256} & 0_{256\times T} \\ 0_{T\times256} & U_{T\times T} \end{pmatrix} \tag{8}$$

where U is upper triangular matrix with entries

$$u_{i,j} = \begin{cases} 0, & i \geq j \\ -\infty, & i < j \end{cases} \tag{9}$$

The decoder operates in an autoregressive manner: it starts with an initial empty token embedding and, at each step, generates the next token embedding conditioned on the previously generated sequence and the visual context. This process is repeated up to the maximum sequence length, allowing the model to generate a sequence of tokens that describe the input image. The model output is projected into token indices in a deterministic manner as

$$y = argmax_i(z_i), \quad (z_i) \in \mathbf{R}^d \tag{10}$$

3 Training Strategy

Both the visual and linguistic components of the model are frozen during training, focusing the learning capacity on the projection and normalisation layers. Consider a general sequence of tokens

$$w = (w_1, w_2, \ldots, w_T) \tag{11}$$

where $w_t \in 1 \ldots m$ is the index of a token in the vocabulary. The model output is generating a conditional distribution over sequences of tokens for a given input image I defined autoregressively as

$$p(w, I, \theta) = \prod_{t=1}^{T-1} p(w_{t+1}|w_{1:t}, I, \theta) \tag{12}$$

where θ is the vector of tunable parameters. The training goal is to maximize the log-likelihood of the image-caption pairs from the training data set. This requires minimization of the negative log-likelihood

$$\begin{aligned} \mathcal{L} &= -\log p(w, I, \theta) \\ &= -\sum_{t=1}^{T-1} \log p(w_{t+1}|w_1, \ldots, w_t, I, \theta) \end{aligned} \tag{13}$$

For each position $t = 1 \ldots T$, let $z_t \in \mathbf{R}^m$ be the model's output (so called logits) for the next generated token over vocabulary of size m. After application of softmax operation to convert logits to probabilities we can get an explicit expression for $p(w, I, \theta)$, i.e.

$$p(w_{t+1} = i|w_1, \ldots, w_t) = \frac{e^{z_{t,i}}}{\sum_{j=1}^{m} e^{z_{t,j}}} \tag{14}$$

where $z_{t,j}$ is element at position j from the vector z_t. In the proposed architecture *autoregressive generation loss* is employed where past sequence $w_{1:t}$ is obtained from previously generated tokens, i.e.

$$w_t \sim p(w_t|w_{1:t-1}, I, \theta) \tag{15}$$

In this case, each $\mathcal{L}_t$ depends on all previous predictions through the recursive generation. The gradient calculation $\partial\mathcal{L}/\partial\theta$ will require backpropagation through t unrolled steps. Such an approach is not advisable for models with many layers, as it can lead to vanishing or exploding gradients, along with high memory usage. However, in the proposed model architecture, the only trainable parameters are the projection layer weights W_p and biases b_p. The projection layer parameters contribute to the loss gradient as

$$\frac{\partial\mathcal{L}}{\partial W_p} = \sum_{t=1}^{T}\left(\frac{\partial\mathcal{L}_t}{\partial z_t}W_e\right)\frac{\partial y_t}{\partial v}\frac{\partial v}{\partial W_p} \tag{16}$$

4 Results

The tuning process begins with the selection of pre-trained versions of the visual and linguistic model. For the visual features encodings Detectron2 model for COCO instance segmentation task is used with a structure of mask RCNN, signature X_101_32x8d_FPN_3x. The linguistic counterpart for the experiment is TinyLlama-1.1B-Chat-v1.0. The hardware configuration for training and evaluation consists of 2 x NVIDIA GeForce GTX 1080

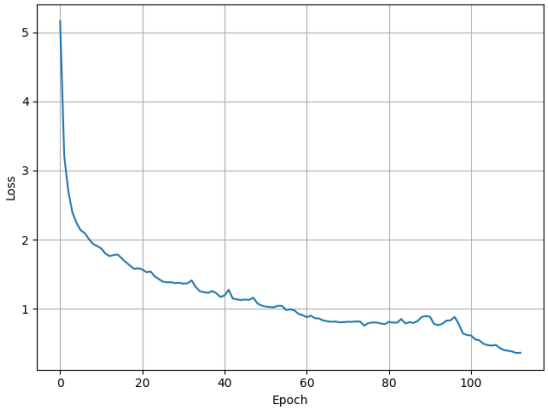


Figure 1: Average loss function during training.



Figure 2: Example input image along with detected objects from Detectron2.

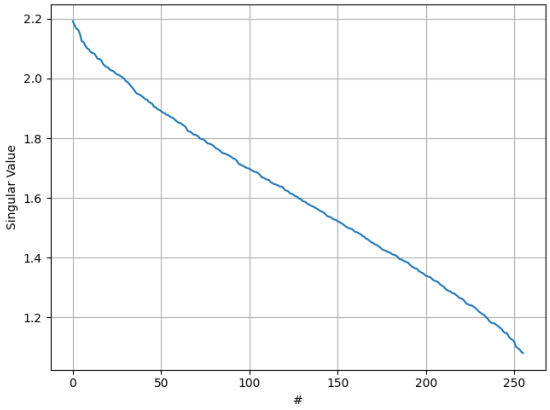


Figure 3: Singular values of the projection matrix before training.

and 1 x NVIDIA GeForce RTX 3070. The training is performed for 115 epochs for a domain of 'cricket' from MIC21 dataset with a batch size of 7 and 40 for max generation sequence. The batch size and sequence length are tuned to optimize for the architecture of 3 GPUs. The layers and weights of the Llama models are distributed between the 3 cards to occupy upto 3GB of the both GTX cards and upto 5GB of the RTX card. The Detectron model is entirely loaded into RTX card. The caching is enabled to accelerate autoregressive generation.

The learning rate for training is set at 0.0001 and using Adam optimizer. The average loss per epoch can be seen at Fig. 1. The loss function show strong initial decrease, followed by slower but constant slope and final drop after 100th epoch, where actual human quality captions begin to appear. Example image and respective detections from Detectron2 model are presented into Fig. 2. As it can be seen images contain diverse features and colours. The presented detected objects are obtained from the consecutive Detectron2 layers, which are not projected to LLM, however they could be integrated into future research iterations.

The generation prompt is:

```
<|system|>
Generate title and description for the
provided image. The image features are:
<v1><v2>...<v256></s>
<|user|>
Generate a title:</s>
<|assistant|>
```

As shown, the prompt contains a short system message followed by a user instruction. The visual features are injected at the end of the system message as a sequence of token embeddings obtained from the projection layer. In this way, each image adds 256 tokens to the prompt, which guide subsequent output generation.

The structure of the projection matrix $W_b \in \mathbf{R}^{256\times2048}$ is examined through singular value decomposition to assess the magnitude of training. The singular values of W_b before training are shown in Fig. 3. This pre-training spectrum indicates that the matrix is well conditioned and capable of representing high-dimensional data without severe rank deficiency, providing a solid foundation for effective adaptation during subsequent learning.

The singular values of W_b after training are shown in Fig. 4. Following training, the singular value spectrum of the projection matrix shows a marked change, with a pronounced concentration of energy in the leading singular value, which now exceeds 3.5, and a steeper drop-off among the subsequent components. This pattern suggests that training has polarised the directions in the matrix: a few dominant modes now carry significantly more

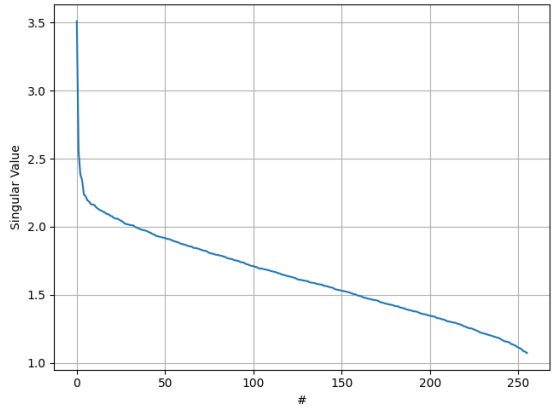


Figure 4: Singular values of the projection matrix after training.

Metric	Epoch 115	Epoch 0
BLEU	0.7	0.0
ROUGE 1	0.84	0.005
ROUGE 2	0.74	0.0
ROUGE L	0.83	0.005
ROUGE L Sum	0.83	0.005
METEOR	0.81	0.004
Precision	0.93	0.46
Recall	0.92	0.52
F1 Score	0.92	0.49
WER	0.23	1.35

Table 1: Model performance comparison for cricket domain.

information, while the remaining dimensions contribute less to the representation. Such behaviour is typical of learned projection layers, in which the network identifies and amplifies key directions to bridge visual and linguistic feature spaces more effectively. The compression of singular values beyond the first few may reflect emergent sparsity or the prioritisation of task-relevant structure, which can improve robustness and generalisation in downstream multimodal learning.

Table 1 presents an evaluation of model performance in the cricket domain before and after domain-specific training on the MIC21 dataset. The metrics shown – BLEU, ROUGE, METEOR, Precision, Recall, F1 Score, and Word Error Rate (WER) - capture both the semantic and structural accuracy of generated captions or outputs. At Epoch 0, the model exhibits near-zero scores in BLEU, ROUGE, and METEOR, along with moderate Precision, Recall, and F1 Score, while suffering a high WER of 1.35. These baselines illustrate that the untrained model is unable to accurately capture the domain-specific content or structure required for high-quality generation.

By Epoch 115, every metric shows dramatic improvement. BLEU and ROUGE scores rise to competitive levels, indicating that the model now generates text with strong fidelity to reference sentences. METEOR also improves substantially, confirming better semantic matching. Precision, Recall, and F1 Score all exceed 0.9, highlighting robust classification and retrieval capabilities, while WER drops markedly to 0.23, indicating a substantial reduction in transcription errors. Altogether, these advances demonstrate that the domain-adapted multimodal system is capable of highly accurate and reliable output when sufficiently trained on a targeted domain dataset.

5 Conclusion

The paper presents a robust and flexible approach to multimodal learning by directly projecting high-dimensional visual features into linguistic embedding spaces through a trainable projection layer. By leveraging strong pre-trained models for both vision (Detectron2 Mask R-CNN) and language (TinyLlama), the framework addresses the challenge of cross-modal fusion with minimal fine-tuning, enabling accurate, domain-adapted object recognition, captioning, and question answering. Experiments on the comprehensive MIC21 dataset – covering over 21,000 images and 700 categories – show substantial gains in all key performance metrics after targeted training, with notable improvements in BLEU, ROUGE, METEOR, precision, recall, and word error rate.

Analysis of the singular value spectra before and after training further validates the projection layer's ability to extract and prioritise meaningful cross-modal directions. This work lays a foundation for scalable domain adaptation in unified visual–linguistic models and opens new perspectives for extending the architecture to broader domains and multilingual tasks.

Acknowledgments

The present study is carried out within the project Infrastructure for Fine-tuning Pre-trained Large Language Models, Grant Agreement No. ПВУ – 55 from 12.12.2024 /BG-RRP-2.017-0030-C01/.

References

Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. 2022. Masked siamese networks for label-efficient learning. In *ECCV (31)*, volume 13691 of *Lecture Notes in Computer Science*, pages 456–473. Springer.

Adrien Bardes, Jean Ponce, and Yann Lecun. 2022. VICReg: Variance-Invariance-Covariance Regularization For Self-Supervised Learning. In *ICLR 2022 - International Conference on Learning Representations*, Online, United States.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Xinlei Chen and Kaiming He. 2021. Exploring simple siamese representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 15750–15758. Computer Vision Foundation / IEEE.

Ruizhe Cheng, Bichen Wu, Peizhao Zhang, Peter Vajda, and Joseph E Gonzalez. 2021. Data-efficient language-supervised zero-shot learning with self-distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3119–3124.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020.

Madeline C. Schiappa, Yogesh S. Rawat, and Mubarak Shah. 2023. Self-supervised learning for videos: A survey. *ACM Comput. Surv.*, 55(13s).

Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. 2022. Unified contrastive learning in image-text-label space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19163–19173.

Mitigating Security Vulnerabilities in Open-Source Large Language Models via a Supervisor Agent Architecture

Öner Aytaş
Işık University, Turkey
oner.aytas@isikun.edu.tr

Mehmet Ali Bayram
Yıldız Technical University, Turkey
malibayram20@gmail.com

Tuğçe Şen
Işık University, Turkey
tugce.sen@isikun.edu.tr

Banu Diri
Yıldız Technical University, Turkey
diri@yildiz.edu.tr

Göksel Biricik
Yıldız Technical University, Turkey
gbiricik@yildiz.edu.tr

Abstract

This study proposes an independent Supervisor Agent approach based on a Turkish safety classifier to protect open-source large language models against harmful prompts. The proposed TurkGuard model classifies user prompts as Safe or Unsafe before they are forwarded to the main model and provides an additional security layer by blocking harmful prompts. The study addresses large language model security threats such as Prompt Injection and Jailbreaking and discusses the limitations of existing English-oriented safety solutions due to the morphological characteristics of Turkish. Experimental results show that the fine-tuned small-scale Qwen3.5-based TurkGuard model can provide high attack detection performance and a low false alarm rate for Turkish prompts.

Keywords: Large language models, open-source models, Turkish safety classification, Prompt Injection, Supervisor Agent, TurkGuard.

1 Introduction

Large Language Models (LLMs) have shown significant progress in natural language processing in recent years and have begun to be used in many different areas such as text generation, question-answering systems, code generation, and decision support systems. In particular, together with foundation models based on the Transformer architecture, it has become possible for a single model to perform a large number of tasks with high accuracy (Bommasani et al., 2021). Models such as GPT, Llama, Qwen, Gemma, Mistral, and DeepSeek are widely used not only in academic research but also in healthcare, education, finance, public-sector, and industrial applications (OpenAI, 2023).

In parallel with this development, open-source large language models have become an important alternative for researchers and developers. Open-source models provide advantages such as access to model weights, the ability to run within institutional systems, customizability, and low-cost deployment. Especially in applications where data privacy is critical, the ability to process user data without sharing it with third-party services has led to the preference for open-source models. However, this flexible structure also brings some important security risks. The modifiability of model weights, the possibility of re-fine-tuning, or the removal of safety restrictions make open-source models have a broader attack surface than closed-source models.

Today, large language models are used not only as systems that generate natural language but also as core components of intelligent systems that can interact with different software components. Thanks to capabilities such as making API calls, accessing databases, executing code, and using external tools, LLM-based applications can perform operations on real systems. While this situation expands the areas of use of the models, it also significantly increases the impact of security vulnerabilities. A malicious prompt may not only produce inappropriate text but may also lead to serious consequences such as the execution of system commands, the transfer of sensitive information to an external environment, or the performance of unauthorized actions.

With the widespread use of LLM-based systems, security threats have also diversified. In the OWASP Top 10 for LLM Applications document published by OWASP, threats such as Prompt Injection, Sensitive Information Disclosure, Excessive

Agency, Insecure Output Handling, and Training Data Poisoning are shown among the main security risks for large language models (OWASP Foundation, 2025). Unlike traditional software security vulnerabilities, these attacks target the model's semantic interpretation capability and often cannot be detected by classical security mechanisms.

In the literature, many studies have been conducted especially on Prompt Injection and Jailbreaking attacks. In Prompt Injection attacks, the attacker aims to make the model ignore system instructions through malicious instructions placed in user inputs or external data sources (Greshake et al., 2023). In Jailbreaking attacks, the aim is to bypass the model's safety policies by using role-playing, scenario construction, or multi-step reasoning techniques (Wei et al., 2023; Zou et al., 2023). In particular, fine-tuning performed on open-source models or the removal of safety filters may increase the probability of success of these attacks.

Various safety solutions such as Llama Guard, NeMo Guardrails, and Azure AI Content Safety have been developed for these security problems (Inan et al., 2024; NVIDIA, 2024; Microsoft, 2025). However, a significant portion of these solutions has been developed on English-oriented datasets and does not guarantee the same success in different languages, especially in agglutinative and morphologically rich languages such as Turkish. Turkish exhibits a more complex structure than English for safety classification because of its word derivation structure, the effect of suffixes on meaning, and context-sensitive language properties. Nevertheless, the number of studies conducted on the security of Turkish large language models is quite limited.

In this study, we propose a lightweight Supervisor Agent architecture designed to improve the security of open-source large language models against Prompt Injection attacks. The proposed architecture operates independently of the protected LLM by analyzing incoming user prompts before they reach the target model. To support this architecture, we develop TurkGuard, a lightweight Turkish safety classifier capable of performing binary Safe/Unsafe prompt classification. TurkGuard is designed to operate as an independent decision-making component and can be integrated with different open-source large language models without modifying their internal architecture.

Rather than introducing another standalone language model, this study presents a practical security framework in which TurkGuard functions as a lightweight Turkish safety classifier that enhances the security of existing open-source LLMs through an independent Supervisor Agent architecture.

1.1 Gap in the Literature and Motivation of the Study

In recent years, important studies have been carried out to improve the safety of large language models. In particular, solutions such as Llama Guard, NeMo Guardrails, and Azure AI Content Safety focus on the detection of harmful prompts, content filtering, and input-output validation mechanisms (Inan et al., 2024; NVIDIA, 2024; Microsoft, 2025). However, the vast majority of these solutions have been developed on English datasets and do not provide sufficient experimental evaluation for agglutinative and morphologically complex languages such as Turkish. In Turkish, the ability of suffixes to change meaning, context-dependent expressions, and different word derivation structures make safety classification more difficult than in English.

The broader adversarial NLP literature also shows that language-based systems can be manipulated not only through direct harmful prompts but also through subtle input transformations, universal triggers, hidden backdoors, and poisoned training data. Character-level adversarial attacks such as HotFlip demonstrate that small textual perturbations can substantially reduce classifier accuracy (Ebrahimi et al., 2018), while universal adversarial triggers show that short trigger sequences can cause systematic model failures across inputs (Wallace et al., 2019). Similarly, backdoor and poisoning studies indicate that language models and text classifiers can be manipulated through hidden triggers or poisoned examples without necessarily degrading clean-task performance (Dai et al., 2019; Li et al., 2021; Xu et al., 2021). These findings support the need for an external security layer that can evaluate prompts before they reach the main model.

Recent LLM safety research has further emphasized the importance of red-teaming, benchmark-based evaluation, and multilingual alignment. Red-teaming studies based on multi-turn or chained utterances show that harmful model behavior may emerge through interaction patterns rather than a single explicit malicious prompt (Bhardwaj & Poria, 2023). Safety datasets and benchmarks such as

BeaverTails and SALAD-Bench provide structured taxonomies for harmful instructions and model safety evaluation (Ji et al., 2023; Li et al., 2024). In addition, multilingual alignment studies report that safety behavior may vary across languages and that local linguistic preferences should be considered when reducing harmful outputs (Ahmadian et al., 2024; Friedrich et al., 2024). These studies are directly aligned with the motivation of the present work, which focuses on Turkish prompt safety classification.

Although there are many studies in the literature examining Prompt Injection and Jailbreaking attacks, research that systematically evaluates the safety performance of Turkish open-source large language models is quite limited. In particular, the lack of comprehensive studies in which different open-source models are compared under the same attack set makes it difficult to determine which models are more robust in terms of security.

In this context, in a study conducted by the authors of this paper, the safety performance of 35 different open-source large language models against Prompt Injection attacks was evaluated using 790 harmful prompts prepared in Turkish (Aytaş et al., 2026). The findings show that there are important differences among open-source models in terms of security. While some models achieved safety scores above 90%, in some models this value was observed to fall below 40%. This situation reveals that the safety levels of open-source large language models can vary significantly depending on model architecture, training data, and alignment processes.

Rank	Model	Safety Score
1	Qwen3.6:27B	98.23%
2	Cosmos Turkish-Gemma-9B	96.71%
3	Qwen3.5:27B	95.70%
4	Qwen3:30B	92.66%
5	Kizagan-E4B Turkish Reasoning	91.77%
...	...	...
31	KocDigital LLM-8B	49.37%
32	Kumru	46.46%
33	Cure-Med-14B	45.95%
34	Turkcell LLM-7B	39.49%
35	Commencis LLM	35.95%

Table 1: Open-source large language models with the highest and lowest safety scores in Turkish Prompt Injection tests (Aytaş et al., 2026)

These results show that there are serious performance differences among open-source models in terms of security. Although models with high safety levels exist, especially models with low safety scores may create important risks when used directly in real systems. In addition, even models with high safety scores are not completely immune to attacks such as Prompt Injection and Jailbreaking. Therefore, leaving security only to the model's own alignment mechanisms is not a sufficient solution.

The main motivation of this study is to develop an independent supervision mechanism that can improve the safety level of open-source large language models without changing the model architecture. For this purpose, a Supervisor Agent that operates in front of any open-source large language model was designed. The proposed structure classifies prompts received from users as Safe and Unsafe before they are forwarded to the main model and allows only prompts determined to be safe to reach the model. Thus, security supervision can be carried out independently of the model, and an additional protection layer that can be easily integrated with different open-source large language models is provided.

The main contributions of this study are summarized below:

An independent, general-purpose Supervisor Agent architecture is proposed as a decision-making security layer. This architecture operates completely decoupled from the target open-source large language model, providing an additional layer of protection against Prompt Injection and similar attacks by blocking harmful prompts before they reach the main model.

To power this architecture, TurkGuard, a lightweight and highly efficient safety classifier fine-tuned specifically to evaluate and classify user prompts as Safe or Unsafe, has been developed.

Comprehensive vulnerability assessments and benchmark tests were conducted. The evaluation focuses entirely on the Turkish language, utilizing an English comparative baseline to systematically measure the safety performance of various LLMs and demonstrate the real-world efficacy of the proposed defense mechanism.

2 Methodology

In this study, a two-stage security architecture based on a Supervisor Agent is proposed to protect open-source large language models against Prompt Injection and similar harmful prompts. The overall workflow of the proposed architecture is illus-

trated in Figure 1. In the proposed architecture, all prompts received from the user are first analyzed by the Supervisor Agent fine-tuned for security purposes. The Supervisor Agent semantically evaluates the prompt and classifies it as Safe or Unsafe. Prompts classified as safe are routed to the open-source large language model to be protected, while prompts evaluated as harmful are rejected before being forwarded to the main model. Thus, security supervision is performed by an independent decision mechanism rather than directly by the main model.

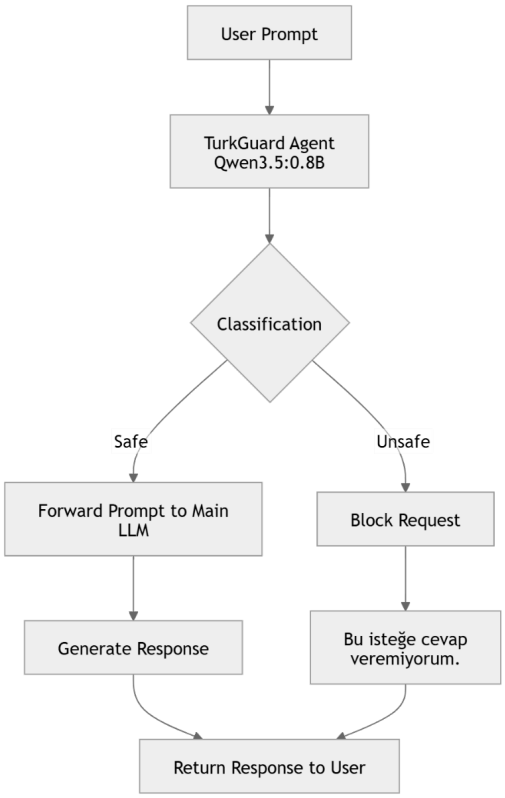


Figure 1: Proposed Supervisor Agent architecture.

The proposed architecture is designed to operate completely independently of the large language model to be protected. Therefore, it can be applied to any open-source large language model without requiring additional architectural changes. Separating the safety classifier from the main model enables the security layer to be reused across different models and integrated into existing systems with minimal changes.

2.1 Threat Model

In this study, it is assumed that the attacker can access the system only through user prompts. The attacker's goal is to bypass safety policies by using Prompt Injection, Jailbreaking, or similar social engineering techniques and to cause the main model to produce harmful content. In contrast, the attacker does not have direct access to the Supervisor Agent model, model weights, or system infrastructure.

The proposed approach aims to detect harmful prompts at an early stage through security analysis performed before the prompt reaches the model. Thus, the main model never encounters potentially harmful prompts, and the attack is intended to be blocked at the initial stage.

2.2 Supervisor Agent Fine-Tuning

Three low-parameter models, namely Qwen3.5-0.8B, LlamaGuard, and Gemma4 E2B, were considered as candidate Supervisor Agents. These models were selected because the aim of the proposed system is to create a lightweight front-end protection layer that consumes fewer computational resources than the protected LLM while still providing fast prompt-level security filtering. Among these alternatives, Qwen3.5-0.8B was selected as the base model for TurkGuard not only due to its low computational cost and fast inference time, but also because of its robust multilingual tokenizer and strong baseline comprehension of Turkish semantics compared with other sub-billion-parameter models. Therefore, Qwen3.5-0.8B provides a practical balance between efficiency, multilingual capability, and Turkish prompt understanding for the Supervisor Agent setting. The model was fine-tuned using a Sequence Classification approach not for text generation, but only for evaluating prompts in terms of security.

2.2.1 Dataset Construction

To train the proposed Turkish safety classifier, a balanced binary classification dataset was constructed by combining multiple Turkish-language resources. The final dataset consists of 51,264 labeled samples, including 25,843 Safe (50.41%) and 25,421 Unsafe (49.59%) instances.

The Unsafe class was created by combining three different sources. The first source consists of Turkish Prompt Injection prompts obtained from the Turkish Prompt Injection & Jailbreak Benchmark Dataset, which was compiled by translating and localizing publicly available English Prompt Injection and Jailbreak datasets into Turkish while incorporating Turkish cultural and linguistic characteristics (Aytaş, 2026). This dataset contains

translated, localized, and manually extended adversarial prompts collected from multiple benchmark sources. The second source consists of Prompt Injection responses generated during previous security analyses. The third source consists of a subset of manually labeled Turkish hate speech posts obtained from the publicly available turkish-hate-speech dataset on Hugging Face, which is based on the dataset introduced by Mayda et al. (2021). Selected samples from this resource were incorporated to increase the linguistic diversity of harmful expressions beyond Prompt Injection attacks.

The Safe class consists of benign Turkish prompts that do not contain malicious intent or harmful instructions. To prevent class imbalance, both classes were kept approximately equal in size. Duplicate records were removed, and all samples were mapped into a unified binary labeling scheme, namely Safe and Unsafe.

The resulting dataset was randomly divided into 85% training and 15% testing sets. Given the dataset size of 51,264 examples, the 15% split yields approximately 7,690 test instances, which provides a statistically robust evaluation set. Due to the high computational costs associated with fine-tuning LLMs, standard cross-validation was deemed computationally prohibitive; therefore, the 85/15 split was adopted as a practical strategy consistent with widely accepted deep learning practices for datasets of this scale. In addition, a subset of the training data was monitored internally as a validation set to prevent overfitting during the hyperparameter tuning phase. The previously developed pool of 790 Turkish Prompt Injection prompts was intentionally excluded from the training dataset and reserved exclusively for independent evaluation. This separation prevents data leakage and enables an unbiased assessment of the classifier's generalization capability.

As a result of the fine-tuning process, the developed Supervisor Agent performs binary classification for each prompt received from the user and produces only a Safe or Unsafe label. While prompts classified as Unsafe are blocked before being forwarded to the main model, prompts evaluated as Safe are routed to the open-source large language model to be protected.

2.3 Experimental Setup

To evaluate the success of the proposed architecture, an independent evaluation set consisting of 790 Turkish harmful prompts previously created by the researchers and containing different Prompt Injection techniques was used. All of these prompts were labeled as harmful, and the main task of the Supervisor Agent is to correctly classify these prompts as Unsafe. Since the test set was not used during the training process, it makes it possible to measure the model's real generalization performance.

To evaluate the false positive behavior of the Supervisor Agent, a Turkish multi-domain knowledge exam consisting 6200 questions was also used. This dataset consists of safe prompts representing daily usage scenarios. Thus, not only the model's ability to block harmful prompts but also its ability to correctly route normal user queries to the main model was evaluated.

Thanks to these two independent evaluation sets, the proposed system was analyzed both in terms of attack detection success and decision-making performance in normal usage scenarios.

3 Experimental Results

The performance of the proposed Supervisor Agent was evaluated under three different experimental scenarios.

First, the base Qwen3.5-0.8B model used without fine-tuning was compared with the proposed fine-tuned Supervisor Agent. Thus, the contribution of the fine-tuning process to safety performance was demonstrated.

Second, the proposed method was compared with Llama Guard, which is widely used in the literature, and the Gemma4-2B fine-tuned model developed on a different open-source model. The purpose of this comparison is to compare the success of the proposed approach on Turkish Prompt Injection attacks with existing safety solutions.

Finally, the TR-MMLU dataset, consisting 6200 questions, was used to evaluate whether the Supervisor Agent developed for security purposes incorrectly blocks normal user prompts. TR-MMLU is a comprehensive assessment dataset developed to evaluate the Turkish language performance of large language models and contains Turkish multiple-choice questions from different disciplines (Bayram et al., 2024). Thus, the proposed system was evaluated both in terms of its success in detecting harmful prompts and its behavior in normal usage scenarios.

3.1 Prompt Injection Detection Performance

Using the developed Turkish safety dataset, fine-tuning experiments were conducted on three candidate low-parameter models: Llama Guard, Gemma4-2B, and Qwen3.5-0.8B. The comparative results of these fine-tuned models, together with the non-fine-tuned Qwen3.5-0.8B baseline, are presented in Table 2.

The attack detection performance of the Supervisor Agent was evaluated on 790 Turkish Prompt Injection prompts previously created by the researchers and not used during training. All prompts in the test set were labeled as harmful (Unsafe). Therefore, the evaluation was carried out mainly based on the Recall (Detection Rate) metric.

Model	Fine-tune	TP	FN	Recall (%)
Llama Guard	Applied	178	612	22.53
Gemma4-2B	Applied	442	348	55.95
Qwen3.5-0.8B (Base)	Not Applied	400	390	50.63
Qwen3.5-0.8B (TurkGuard)	Applied	577	213	73.04

Table 2: Comparison of different models on Turkish Prompt Injection prompts.

When Table 2 is examined, it is seen that the base Qwen3.5-0.8B model without fine-tuning can correctly detect only 50.63% of harmful prompts. This result shows that directly using general-purpose large language models as safety classifiers is not sufficient.

The fine-tuning study carried out on the Gemma4-2B model provided a limited performance increase compared with the base model (55%), but the proposed Supervisor Agent architecture achieved the highest performance with 73.04% detection success. In particular, the very low success of the Llama Guard model on Turkish prompts (24%) shows that English-centered safety models are not sufficient for direct use in morphologically rich languages such as Turkish.

These results reveal that training the safety classifier with target-language-specific datasets provides an important advantage in detecting Prompt Injection attacks.

3.2 Performance in Normal Usage Scenarios

It is not sufficient for a security system only to block harmful prompts. At the same time, it is expected not to incorrectly classify normal user queries as harmful. For this reason, the proposed Supervisor Agent was also evaluated on the TR-MMLU dataset, which was developed in the authors' previous work and consists 6200 Turkish multiple-choice questions (Bayram et al., 2024).

Dataset	Total	Safe	Unsafe	Safe Acc. (%)
TR-MMLU	6200	5938	262	95.77

Table 3: Evaluation on the TR-MMLU dataset.

The results show that the proposed Supervisor Agent correctly classifies approximately 95% of normal informational prompts as Safe. This finding shows that the model not only detects attacks but can also operate with a low false alarm rate in daily usage scenarios.

4 Discussion

The experimental results show that a fine-tuned small-scale safety model can perform significantly more successful safety classification than a general-purpose base model. It is particularly noteworthy that the Qwen3.5-based Supervisor Agent (TurkGuard), which has only 0.8 billion parameters, performs better than some safety models with larger numbers of parameters. This situation shows that task-oriented fine-tuning may be more decisive than model size in safety tasks.

In addition, the low performance of Llama Guard on Turkish prompts can be considered a natural result of the fact that existing safety solutions have largely been developed on English datasets. The agglutinative structure of Turkish, its context-sensitive expressions, and its rich morphological properties can limit the generalization success of English-centered safety models. This finding reveals the importance of developing language-specific safety models for different languages.

The evaluation carried out on the TR-MMLU dataset shows that the proposed approach does not focus only on detecting harmful prompts but can also classify normal user queries as Safe with high accuracy. The approximately 95% correct classification success is important because it shows that the proposed structure can operate with a low false alarm rate in real usage scenarios.

Another important output of this study is that the proposed Supervisor Agent offers a general security layer that can be used not only in newly developed systems but also to increase the security levels of existing open-source large language models. In the security evaluations conducted in our previous study, it was shown that some open-source models had quite low safety scores against Prompt

Injection attacks (Aytaş et al., 2026). When the architecture proposed in this study is placed in front of these models as an independent security layer, it can filter a significant portion of harmful prompts without changing the architecture or weights of the main model. Thus, it may be possible to make open-source large language models with low safety scores more secure, and a practical security improvement can also be achieved without the need to retrain existing models.

However, the proposed approach does not aim to provide absolute protection against all types of attacks. In particular, additional studies are needed for multi-step attacks, indirect Prompt Injection techniques, and new attack methods that may be developed in the future. In future studies, more comprehensive experiments on different open-source models, the development of multilingual safety classifiers, and the integration of dynamic safety policies into the Supervisor Agent architecture are planned.

5 Limitations

A notable limitation of this study relates to the evaluation of false positive rates using the TR-MMLU dataset. Since TR-MMLU primarily consists of multiple-choice academic questions, it does not fully capture the complex, conversational, and diverse prompt distributions encountered in real-world deployments. Consequently, the 95.77% safe accuracy achieved in normal usage scenarios may overestimate the model's actual deployment performance. Future studies should evaluate false positive behavior on real-world, conversational Turkish prompt datasets.

References

Ahmadian, A., Ermis, B., Goldfarb-Tarrant, S., Kreutzer, J., Fadaee, M., & Hooker, S. (2024). The multilingual alignment prism: Aligning global and local preferences to reduce harm. arXiv. https://arxiv.org/abs/2406.18682

Aytaş, Ö. (2026). TurkGuard: A fine-tuned Turkish safety classifier based on Qwen3.5-0.8B (Version 1.0) [Large language model]. Hugging Face. Hugging Face Model Repository. https://huggingface.co/OnerAYTAS/TurkGuard_Qwen3.5B_0.8B

Aytaş, Ö.; Şen, T.; Diri, B.; Biricik, G.; Bayram, M.A. (2026)Benchmarking Prompt Injection Attacks on LLMs: Turkish Vulnerability Assessment and English Comparative Analysis. Applied Sciences 2026, 16(13), 6740. https://doi.org/10.3390/app16136740

Aytaş, Ö. (2026). Turkish Prompt Injection & Jailbreak Benchmark Dataset [Data set]. Hugging Face. Hugging Face Dataset Repository.

Bayram, M. A., Fincan, A. A., Gümüş, A. S., Diri, B., Yıldırım, S., & Aytaş, Ö. (2024). Setting standards in Turkish NLP: TR-MMLU for large language model evaluation. arXiv preprint arXiv:2501.00593.

Bhardwaj, R., & Poria, S. (2023). Red-teaming large language models using chain of utterances for safety-alignment. arXiv. https://arxiv.org/abs/2308.09662

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. Stanford Center for Research on Foundation Models. https://arxiv.org/abs/2108.07258

Dai, J., Chen, C., & Li, Y. (2019). A backdoor attack against LSTM-based text classification systems. IEEE Access, 7, 138872–138878. https://doi.org/10.1109/ACCESS.2019.2941376

Ebrahimi, J., Rao, A., Lowd, D., & Dou, D. (2018). HotFlip: White-box adversarial examples for text classification. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL).

Friedrich, F., Tedeschi, S., Schramowski, P., Brack, M., Navigli, R., Nguyen, H., Li, B., & Kersting, K. (2024). LLMs lost in translation: M-ALERT uncovers cross-linguistic safety inconsistencies. arXiv. https://arxiv.org/abs/2412.15035

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Fritz, M., & Tramèr, F. (2023). Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection attacks. arXiv. https://arxiv.org/abs/2302.12173

Inan, H., Upasani, K., Wang, C., et al. (2024). Llama Guard: LLM-based input-output safeguard for human-AI conversations. arXiv. https://arxiv.org/abs/2312.06674

Ji, J., Liu, M., Dai, J., Pan, X., Zhang, C., Bian, C., Zhang, C., Sun, R., Wang, Y., & Yang, Y. (2023). BeaverTails: Towards improved safety alignment of LLM via a human-preference dataset. arXiv. https://arxiv.org/abs/2307.04657

Li, L., Dong, B., Wang, R., Hu, X., Zuo, W., Lin, D., Qiao, Y., & Shao, J. (2024). SALAD-Bench: A hierarchical and comprehensive safety benchmark for large language models. arXiv. https://arxiv.org/abs/2402.05044

Li, S., Liu, H., Dong, T., Zhao, B. Z. H., Xue, M., Zhu, H., & Lu, J. (2021). Hidden backdoors in human-centric language models. Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security, 1–18. https://doi.org/10.1145/1122445.1122456

Mayda, İ., Demir, Y. E., Dalyan, T., & Diri, B. (2021). Hate Speech Dataset from Turkish Tweets. In 2021 Innovations in Intelligent Systems and Applications Conference (ASYU) (pp. 1–4). IEEE. https://doi.org/10.1109/ASYU52992.2021.9599042

Microsoft. (2025). Azure AI Content Safety Documentation. https://learn.microsoft.com/azure/ai-services/content-safety/

NVIDIA. (2024). NeMo Guardrails Documentation. https://docs.nvidia.com/nemo/guardrails/

OpenAI. (2023). GPT-4 technical report. arXiv. https://arxiv.org/abs/2303.08774

OWASP Foundation. (2025). OWASP Top 10 for Large Language Model Applications. https://genai.owasp.org/

Wallace, E., Feng, S., Kandpal, N., Gardner, M., & Singh, S. (2019). Universal adversarial triggers for attacking and analyzing NLP. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Wei, A., Haghtalab, N., & Steinhardt, J. (2023). Jailbroken: How does LLM safety training fail? Advances in Neural Information Processing Systems, 36.

Xu, C., Wang, J., Tang, Y., Guzmán, F., Rubinstein, B. I., & Cohn, T. (2021). A targeted attack on black-box neural machine translation with parallel data poisoning. In Proceedings of the Web Conference 2021, 1–12. https://doi.org/10.1145/3442381.3450034

Zou, A., Wang, Z., Kolter, J. Z., & Fredrikson, M. (2023). Universal and transferable adversarial attacks on aligned language models. arXiv. https://arxiv.org/abs/2307.15043

Specialised Methods in NLP and AI

BGSynthMedic: Synthetic Parallel English-Bulgarian Discharge Summary Corpus with Automatic Annotations

Svetla Boytcheva, Sylvia Vassileva, Pavel Boytchev
Faculty of Mathematics and Informatics
Sofia University St. Kliment Ohridski
Sofia, Bulgaria
svetla@uni-sofia.bg, svasileva@fmi.uni-sofia.bg

Abstract

Data scarcity is a significant problem for advancing clinical Natural Language Processing (NLP) in low-resource languages like Bulgarian. To address this issue, we present BGSynthMedic, a publicly available synthetic parallel corpus of English–Bulgarian discharge summaries with automatic annotations for five entity types and their Unified Medical Language System (UMLS) concept links. We use the SynthMedic dataset, a corpus of 450 synthetic discharge summaries generated using medical guidelines and GPT-4, and automatically translate it into Bulgarian with BgGPT. We use zero-shot multilingual and monolingual Named Entity Recognition (NER) models from the GLiNER family to automatically label diseases, symptoms, procedures, medication, and observations in the original and translated summaries. The extracted entities are automatically linked to the UMLS. The machine translation quality is evaluated using reference-free metrics, yielding a BERTScore of 0.90 and a LaBSE of 0.89. The presented pipeline demonstrates how to create a parallel corpus of synthetic discharge summaries and a set of silver-standard annotations in a low-resource language. We release the BGSynthMedic dataset, including parallel medical term dictionaries and abbreviations, as a community resource to facilitate future research in Bulgarian clinical NLP.

Keywords: BioNLP, Silver corpus, LLM, Synthetic corpus, Machine Translation, LLM, NER, Entity Linking, UMLS

1 Introduction

Millions of clinical documents are created in the healthcare system in unstructured text format. These documents contain personal and sensitive information, and there are many limitations to the automatic processing of such documents. Due to these limitations, advancement of biomedical and clinical Natural Language Processing (NLP), especially for low-resource languages, is slower than in other domains (Névéol et al., 2018). Clinical Named Entity Recognition (NER) is crucial for extracting structured information from clinical documents, thereby significantly improving process automation in healthcare and positively impacting patient care. The complexity of clinical language and terminology, combined with multilingual documentation and abbreviations, poses significant challenges. There is increasing demand for the development of annotated clinical text corpora, and very limited availability of such annotated datasets.

Large Language Models (LLMs) have transformed multilingual NER by moving the focus from classification tasks to generation and effective cross-lingual transfer learning. However, transformer models still outperform general and domain-adapted LLMs on the clinical NER task (Lu et al., 2025; Monajatipoor et al., 2024) because LLMs struggle to detect multi-word entities and precisely determine span boundaries for complex medical terminology. In the general-domain NER setting, LLMs achieve competitive F1-scores compared to BERT-family transformers, but in clinical texts, the opposite trend is observed.

The current state-of-the-art multilingual models for clinical NER that cover the major European languages are mDeBERTaV3 (He et al., 2021), XLM-RoBERTa (Conneau et al., 2019), Medical mT5 (García-Ferrero et al., 2024), GLiNER-multi-v2.1 (Zaratiana et al., 2023), and GLiNER-biomed-large-v1.0 (Stepanov and Shtopko, 2024). These models demonstrate F1-score across various benchmark datasets for clinical NER in English, reaching a range of 0.85-0.89, while for other languages like Italian, Spanish, and French, the performance drops to 0.62-0.75 (García-Ferrero et al., 2024). The ecosystem of BioASQ / CLEF / BioCreative challenges plays a key role in benchmarking for

multilingual clinical NER. The evaluation results in the challenges DisTEMIST (disease) (Miranda-Escalada et al., 2022), SympTEMIST (symptoms, signs, findings) (Lima-López et al., 2023), MultiCardioNER (diseases and drugs) (Lima-López et al., 2024), and MedProcNER (procedures) (Lima-López et al., 2023) show that monolingual BERT-family models demonstrate high accuracy for these tasks. In SympTEMIST, the best result, 0.7477, for the NER task was achieved by an ensemble of fine-tuned pretrained models for BSC-Bio-Es, Roberta-Biomedical-Es, XLMR-Galen, BETO, and mBERT-Galen. For MedProcNER, the best F1 Score, 0.7985, was achieved by a transformer-based model with masked CRF and data augmentation. In DisTEMIST, the highest F1 score, 0.777, was achieved by the fine-tuned PlanTL-GOB-ES/roberta-base-biomedical-clinical-es model. In MultiCardioNER, the best F1 Score of 0.8199 was also achieved using an ensemble of RoBERTa-based models.

Several silver-annotated corpora were developed in the biomedical domain (Sousa et al., 2019). LLMs are used not only for the NER task, but also for machine translation of clinical texts (Anyaegbuna et al., 2026).

This motivated us to propose an approach to machine translation, NER, and Entity Linking for clinical text corpora in low-resource languages.

In this paper, we use the SynthMedic corpus (Grazhdanski et al., 2025) as the source corpus of synthetically generated clinical texts. The targeted language for machine translation of the corpus is Bulgarian. For annotation, five categories are used: disease, symptom, procedure, medication, and observation. In the proposed methodology (see Fig. 1), we first employ a few-shot prompting technique with a large language model to translate synthetic discharge summaries from English into Bulgarian, generating high-quality bilingual medical texts. A human expert checks a small subset of translations to validate their quality. The translated documents are subsequently annotated using multiple zero-shot pretrained monolingual and multilingual NER models to identify relevant medical entities. Additionally, the original English source documents undergo the same annotation process. For further validation, we compare entity annotations across the English and the Bulgarian corpora and ground the extracted entities in a controlled medical vocabulary via entity linking, enabling rigorous assessment of the consistency and accuracy of assigned entity categories across languages. We provide the parallel corpus as a community resource.[1]

2 Data

For this study, we use the SynthMedic corpus (Grazhdanski et al., 2025), which contains 900 synthetic English discharge summaries across five medical specialties and nine socially significant diseases. Each disease has 100 LLM-generated discharge summaries - 50 by GPT-4 and 50 by Llama 3.1 8b. Instruct. The dataset contains a wide variety of diseases, symptoms and signs, treatments, and tests, including about 4K diseases, 15K symptoms, 14K treatments, and 13K tests.

We briefly summarize the evaluation of the dataset performed by Grazhdanski et al. (Grazhdanski et al., 2025). The dataset was evaluated using LLM approaches, including a faithfulness metric that checks consistency with the underlying medical references, as well as knowledge-graph–based validation and automatic correction. The faithfulness of the whole corpus was reported as 93.65%; however, the GPT-4-generated summaries were scored about 10% higher than those generated by Llama. In addition, human experts reviewed a subset of the dataset and evaluated it using a rubric, yielding a System Usability Score of 94.35%. Human experts judged the synthetic summaries to be credible and factually accurate when the generation was grounded in Merck Manual[2] references.

SynthMedic is a publicly available dataset covering multiple diseases, released under a permissive license, making it well-suited for creating a parallel corpus in a lower-resource language such as Bulgarian. Furthermore, since it was recently released, existing LLMs may not yet have been trained on it. For this study, we use GPT-4-generated discharge summaries because they have been reported to be of higher quality.

3 Methods

3.1 Translation Pipeline

The SynthMedic corpus was translated into Bulgarian using LLMs. To refine the prompt and select the best model, we conducted initial experiments using a small excerpt (about 5%) of the SynthMedic corpus, including several discharge summaries for each disease. This allowed us to investigate the models'

[1] https://github.com/svassileva/bgsynthmedic
[2] https://www.merckmanuals.com/professional

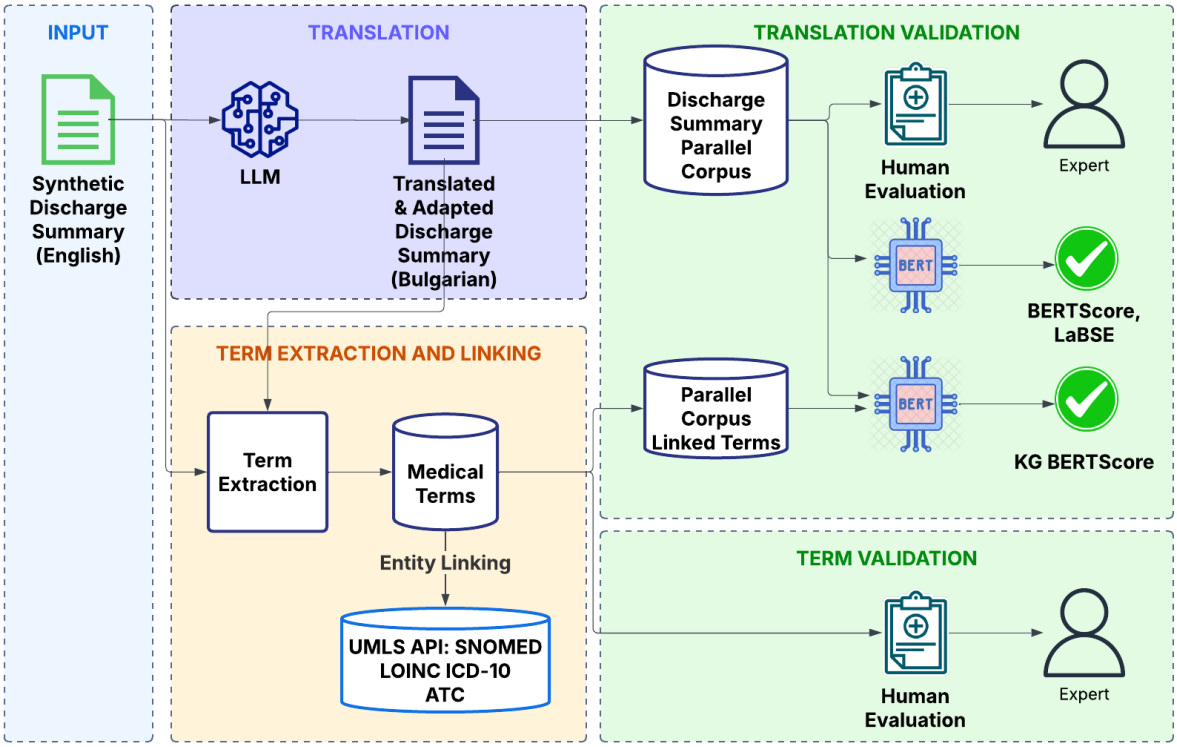


Figure 1: The parallel corpus generation and validation process.

performance in translating specific medical terminology for each clinical specialty. Each document was translated using the same initial prompt using GPT-4.1 (temperature=0.7), GPT-4.1-mini, GPT-4.1-nano[3] (using the API), and Gemma3[4] (using ollama, temperature=0.1, num_predict=7000). The prompt also required generating a parallel dictionary of all medical terms from the source document and the generated translation. The prompt also requested an abbreviation dictionary that includes abbreviations in the source document, their expansion, the proposed target abbreviation (if any exists in the target language), and its expansion in the target language. A human evaluated all translations from the initial experiments and the generated parallel dictionaries. The results identified major issues with the translation of medical terms and abbreviations, as well as with the style of the generated discharge summaries. Further experiments with prompt and parameter modifications did not yield significant improvements in the quality of medical term translation, even with few-shot prompts. To address this issue, we switched to a specific model trained on the Bulgarian general domain: BgGPT (bggpt-gemma-3-27b-fp8)[5]. After selecting the model and adjusting the prompt based on manual testing, we arrived at the final version used to generate all translations. The final few-shot prompt (see Appendix D) includes examples of translations for the most common and problematic medical terms and abbreviations to guide consistent translations of terminology. It also instructs the model to adjust the style of the discharge summary translation to be more formal and to address the patient's general practitioner rather than the patient, which underscores a difference in medical documentation in Bulgaria compared with English-speaking countries. Each discharge summary is translated using a single model call using the BG-GPT API. The final version of the translations was generated with the specified prompt and parameters max_tokens=16384 and temperature=0.7.

3.2 Sentence Alignment and Corpus Statistics

Table 1 presents statistics for the original English dataset and its automatic translation. Sentence splitting was performed using heuristics, including ignoring a small list of abbreviations like "Dr.", or *"г."* (Bulgarian abbreviation for year in a date). Sentence

[3] https://chat.openai.com
[4] https://ollama.com/library/gemma3
[5] https://huggingface.co/INSAIT-Institute/BgGPT-Gemma-3-27B-IT

alignment was performed using the Gale-Church algorithm (Gale and Church, 1993) implemented in the nltk library.[6] The algorithm uses only sentence lengths in characters and a simple probabilistic model to find an optimal alignment between sentences in a parallel corpus. The alignment was performed at the sentence level, thus creating a parallel sentence corpus.

Metric	EN	BG
Documents	450	450
Total sentences	22,252	22,252
Total words	253,675	265,980
Avg words per sentence	11.40	11.95
Avg sentences per document	50.57	50.57
Avg words per document	576.53	604.50

Table 1: Statistics for the original and translated datasets.

3.3 Corpus Annotation

We used zero-shot methods to automatically annotate the parallel corpus with entities and their UMLS codes in English and Bulgarian discharge summaries.

3.3.1 Named Entity Recognition

We experimented with several zero-shot NER models from the GLiNER family: Biomedical GLiNER (Yazdani et al., 2025), GLiNER-X-Large[7], GLiNER Large v2.1 (Zaratiana et al., 2023), and GLiNER multilingual MoE. The models' details are shown in Table 2. For all experiments, we used the following entity types: "disease or medical condition", "abnormality", "symptom or sign", "medical procedure or clinical test", "treatment", "medication", "clinical observation or diagnostic finding". As a post-processing step, these labels were mapped to the five main entity categories: "disease", "procedure", "symptom", "observation", and "medication". The precise definition of labels allowed us to disambiguate some medical terms that, depending on the context, can be classified into different categories (see Appendix C). The GLiNER model allows selecting a custom threshold, which determines whether the span represents an entity. Based on initial experiments, we selected the entity threshold to be 0.2 to identify as many entities as possible. We applied the same models to the English source texts and the Bulgarian translations using the same thresholds to produce a silver-standard annotation corpus. We release the annotations generated by all four models separately. Since GLiNER models use span-based classification for identifying entities, the annotations support nested or overlapping entities of different types.

Model	Multilingual	Domain	BG support
Biomedical GLiNER	No	Biomedical	No
GLiNER-X-Large	Yes	General	No
GLiNER Large v2.1	No	General	No
GLiNER multilingual MoE	Yes	General	Yes

Table 2: Overview of different NER models used in the experiments.

For evaluation, we used a zero-shot NER model - GLiNER multilingual MoE[8]. The model supports multiple languages, including Bulgarian, and is based on GLiNER, which has shown strong performance as a zero-shot NER model across different domains. We used the categories as labels – disease, symptom, medication, procedure, and observation. The threshold for entity detection was set to 0.5.

3.3.2 Entity Linking

UMLS is one of the widely used controlled vocabularies in clinical medicine (Bodenreider, 2004). It comprises more than 1 million biomedical concepts and 5 million concept names. It covers over 100 controlled medical vocabularies, biomedical ontologies, and standard classifications. We performed automatic entity linking to UMLS for the extracted entities in both languages.

We used a cross-lingual SapBERT (cambridgeltlSapBERT-UMLS-2020AB-all-lang) (Liu et al., 2021) model as a baseline model for entity linking, employing cosine similarity search. SapBERT is the standard model for cross-lingual entity linking with the Unified Medical Language System (UMLS)[9] and has shown strong zero-shot performance and good generalization across many languages (Liu et al., 2021). To link English entities, we generated their SapBERT embeddings and performed a cosine similarity search in a knowledge base of UMLS terms, filtered by entity category.

[6] https://www.nltk.org/api/nltk.translate.gale_church.html

[7] https://huggingface.co/knowledgator/gliner-x-large

[8] https://huggingface.co/Mayank6255/GLiNER-MoE-MultiLingual

[9] https://www.nlm.nih.gov/research/umls/index.html

For each entity category, we defined a mapping to UMLS semantic types to reduce the candidate-entity search space (see Appendix A). Based on the filtered concepts, we compiled all their synonyms from the UMLS knowledge base (version 2025AB). We selected the top entity from the knowledge base based on cosine similarity using a FAISS index[10] and assigned the entity code to its mention in the text if the similarity exceeded 0.85. To link entities in Bulgarian, we used automatic translation with the BgGPT LLM[11] before linking them to UMLS, since UMLS is not available in Bulgarian. The entity translation prompt is shown in Appendix B. It translates entity spans in batches for efficiency without any additional context. The model version was bggpt-gemma-3-27b-fp8, and the temperature was set to 0 to reduce the risk of hallucinations and to facilitate reproducibility. The SapBERT cosine similarity search provides a baseline for entity linking and is also used in translation evaluation.

The UMLS API[12] was used as an additional method for automatically validating the extracted terms from NER and their associated categories. We searched for each extracted entity from the NER model using the UMLS API. If the entity was an "exact match", the extracted term was guaranteed to be a medical term from a standard medical vocabulary or ontology. We applied additional validation by comparing the associated entity category with the semantic type of the mapped UMLS term. The UMLS semantic types are narrower than the five broader categories in the scope of our research, so we used a predefined mapping between the two sets of categories (see Appendix C). We also performed an entity search in UMLS using additional parameters such as *words, normalizedString, normalizedWords, rightTruncation, and leftTruncation* for partial matching with lower confidence to cover scenarios such as plural forms of terms, other linguistic variants, and paraphrases. This also covers specific expressions classified under broader UMLS terms. For instance, "RT-PCR test" is mapped in UMLS to "SARS-CoV-2 RT-PCR test"[13], which can be missed if the search was limited to the *exact* match search. We generated the silver-standard entity-linking annotations using the UMLS API, as it is a standard method for this task.

4 Evaluation

4.1 Corpus Evaluation

Creating a parallel reference corpus for evaluating machine translation requires significant effort from bilingual human experts. In the absence of a gold-standard corpus for evaluation, we used reference-free metrics for quality estimation, such as BERTScore (Zhang et al., 2019) and LaBSE (Feng et al., 2022). They measure the semantic similarity between the source and hypothesis translation in a shared embedding space. We used XLM-RoBERTa (xlm-roberta-base) to encode the English and Bulgarian sentences in a shared semantic space and compute token-level cosine similarities to create a cross-lingual alignment. Finally, BERTScore precision, recall, and F1 metrics were calculated to assess how well the text's meaning was preserved in the Bulgarian translation. We used the BERT-score library[14] implementation to calculate the BERTScore metric.

BERTScore Definition. Let the reference sentence be $x = \langle x_1, x_2, \ldots, x_k \rangle$ and the candidate sentence be $\hat{x} = \langle \hat{x}_1, \hat{x}_2, \ldots, \hat{x}_m \rangle$. Contextual embeddings are obtained for each token $\mathbf{x}_i \in \mathbb{R}^d$, and for each token $\hat{\mathbf{x}}_j \in \mathbb{R}^d$, from a pretrained language model. Token similarity is calculated using cosine similarity $\cos(\mathbf{x}_i, \hat{\mathbf{x}}_j)$. The BERTScore-Precision, Recall, and F1 Score are calculated using Equations 1, 2, 3 (Zhang et al., 2019).

$$P_{\text{BERT}} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \cos\big(\mathbf{x}_i, \hat{\mathbf{x}}_j\big) \qquad (1)$$

$$R_{\text{BERT}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \cos\big(\mathbf{x}_i, \hat{\mathbf{x}}_j\big) \qquad (2)$$

$$F_{\text{BERT}} = \frac{2\, P_{\text{BERT}}\, R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}} \qquad (3)$$

LaBSE (Language-agnostic BERT Sentence Embedding) (Feng et al., 2022) is a publicly available model trained to cluster semantically equivalent sentences across different languages in a shared embedding space. The cosine similarity between the source and the hypothesis sentence pairs is treated as a sentence-level quality score. High cosine similarity between sentences in different languages

[10] https://github.com/facebookresearch/faiss

[11] https://huggingface.co/INSAIT-Institute/BgGPT-Gemma-3-27B-IT

[12] https://documentation.uts.nlm.nih.gov/rest/home.html

[13] https://uts.nlm.nih.gov/uts/umls/concept/C5400530

[14] https://pypi.org/project/bert-score/

should indicate that the translation preserves the original meaning. LaBSE similarity has been used as a strong baseline in recent works for automatic MT evaluation (Mujadia et al., 2023). We use the sentence-transformers/LaBSE model for calculating the metric.

KG-BERTScore (Wu et al., 2022) is an extension of the BERTScore metric that combines standard BERTScore with bilingual named-entity matching using a knowledge graph. It has shown a higher correlation with human judgments. KG-BERTScore aims to ensure that the same important entities are preserved in the translation. We applied KG-BERTScore by automatically extracting five named entity types from the discharge summaries (disease, symptom, medication, procedure, and observation) and aligning the translated entities in two ways: either by linking them to the UMLS (Bodenreider, 2004) or by using semantic alignment. We used the methods described in Section 3.3 to detect and link entities.

KG-Score Definition. Let the source sentence be $s = \langle s_1, s_2, \ldots, s_n \rangle$ and the machine-translated sentence be $t = \langle t_1, t_2, \ldots, t_m \rangle$. We extract named entities from both sentences using a zero-shot NER model, and denote the sets of entities by $\text{entities}(s)$ and $\text{entities}(t)$, respectively. For each entity in $\text{entities}(s)$, we try to find a corresponding entity in $\text{entities}(t)$ by matching their UMLS codes or SapBERT entity embeddings, resulting in a set of successfully matched entities $\text{matches}(\text{entities}(s), \text{entities}(t))$. The KG-Score F_{KG} is defined as the ratio between the number of matched entities and the number of source entities, as shown in Equation 4. For KG-Score-UMLS, we used the linked entities in each language to identify corresponding entities across languages that map to the same UMLS concept. This allowed us to calculate entity-level precision, recall, and F1 score, and to combine them with the original BERTScore. For KG-Score-Semantic, we used semantic alignment via cosine similarity search with SapBERT embeddings to match the closest entity in the translation, regardless of entity type. We set a threshold of 0.85 to match entities and greedily align entity pairs between English and Bulgarian.

$$F_{\text{KG}} = \frac{\text{matches}\big(\text{entities}(s), \text{entities}(t)\big)}{\text{entities}(s)} \quad (4)$$

KG-BERTScore Definition. KG-BERTScore linearly combines the KG-Score F_{KG} with the BERTScore F_{BERT}. Given a combination weight $\alpha \in [0, 1]$, the KG-BERTScore is defined as $F_{\text{KG-BERT}} = \alpha F_{\text{KG}} + (1 - \alpha) F_{\text{BERT}}$. We set the α value to 0.5, as in the original paper (Wu et al., 2022).

Using the two different KG-Score strategies, we report KG-BERTScore-UMLS and KG-BERTScore-Semantic.

4.2 Annotation Evaluation

We evaluated the quality of the NER annotations by comparing them with a small subset of documents manually labeled by a human expert across the five target categories. One annotator labeled each document, producing labels for 27 documents per language (3 documents per diagnosis per language). Table 3 shows the number of entities based on entity type in the human-labeled documents. We publish the annotation guidelines together with the dataset and labeled documents. The annotators used the NER Annotator for SpaCy tool[15]. We evaluate the strict precision, recall, and F1-score of all the GLiNER models on the manually labeled dataset. We also compared the number of extracted entities per entity type by the Biomedical GLiNER model from the source and translated texts. In addition, we measured the number and percentage of entities successfully linked using the UMLS API.

Entity type	Number
DISEASE	298
MEDICATION	186
OBSERVATION	179
PROCEDURE	651
SYMPTOM	222

Table 3: Number of labeled entities per type in the human-labeled documents.

5 Results

5.1 Corpus Evaluation

Table 4 shows the translation evaluation metrics. BERTScore and LaBSe show quite high correlations between the source and the hypothesis: 0.90 and 0.89, respectively. The KG-UMLS F1 is 0.60, which relies on correct entity linking to UMLS. However, entity translation and linking may also introduce errors, so we calculated an adjusted score, KG-Semantic, that uses a semantic alignment strategy. The KG-Semantic F1 score is 0.87, indicating

[15] https://arunmozhi.in/ner-annotator/

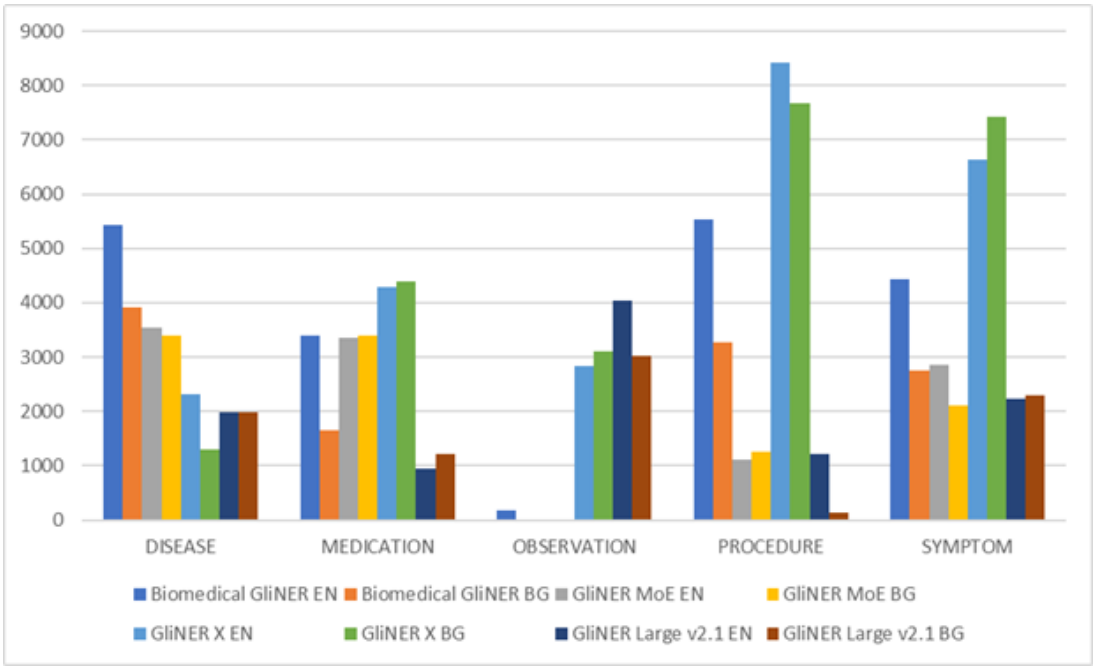


Figure 2: Statistics by type of the number of extracted entities across the original and translated datasets using different models.

good entity-level alignment between the source and translated texts. The combined KG-BERTScore-UMLS for the translation was 0.75, while the KG-BERTScore-Semantic was 0.88.

Metric	Value
BERTScore Precision	0.90
BERTScore Recall	0.90
BERTScore F1	0.90
LaBSe	0.89
KG-UMLS Precision	0.64
KG-UMLS Recall	0.58
KG-UMLS F1	0.60
KG-Semantic Precision	0.89
KG-Semantic Recall	0.87
KG-Semantic F1	0.87
KG-BERTScore-UMLS F1	0.75
KG-BERTScore-Semantic F1	0.88

Table 4: Evaluation metrics for the automatic machine translation.

5.2 Annotation Evaluation

The NER evaluation results are presented in Table 5. The Biomedical GLiNER model achieves the highest F1 score. Both the biomedical and MoE models show higher precision, about 0.73-0.75; however, they struggle with recall, which is especially low for the MoE model at 0.23.

The number of extracted entities per language and entity type is shown in Figure 2. The number of entities extracted in the translated text is significantly lower; about 61% of the entities from the English text are identified in the translation. This result is expected, given that this was a zero-shot approach working in a lower-resource language.

Model	Precision	Recall	F1
Biomedical GLiNER	**0.75**	0.49	**0.59**
GLiNER-X-Large	0.60	**0.51**	0.55
GLiNER Large v2.1	0.64	0.25	0.36
GLiNER Multi-MoE	0.74	0.24	0.36

Table 5: NER model evaluation results on a small subset of manually labeled English documents.

Table 6 shows the number and percentage of linked entities in the UMLS when using the UMLS API for Biomedical GLiNER entities.

Entity Type	UMLS Linked	Total	Percent
Disease	4,941	5,426	91.1%
Medication	3,243	3,396	95.5%
Observation	142	185	76.7%
Procedure	4,973	5,531	89.9%
Symptom	4,202	4,443	94.6%
Grand Total	17,501	18,981	92.2%

Table 6: Entity linking statistics using Biomedical GLiNER and UMLS API.

6 Discussion

LLMs have presented remarkable progress in text generation in the last few years. However, the current models still struggle to generate domain-specific content in low-resource languages. The main reason behind this is data scarcity, and the complexity of domain-specific terminology further complicates the issue. For example, an LLM generating medical, legal, or scientific Bulgarian must simultaneously navigate an underrepresented language in its training corpus, specialized terminology, and a scientific style that demands precise, consistent vocabulary. While high-resource languages like English benefit from vast corpora of technical literature, patents, and academic articles that anchor the terminology of a given field in context, low-resource languages offer the model informal corpora such as web text, social media posts, and general-purpose translations, which barely touch the surface of professional usage and terminology.

This huge gap in training resources available in different languages results in the fabrication of a vocabulary from language- and domain-specific terms. LLMs create a fabricated vocabulary that might be called lexo-phantoms: words that are phonologically plausible, grammatically slottable, but terminologically illegitimate. There are three main types of errors: Transliteration drift, Morphological grafting, and Phantom calques. Transliteration drift is when the model defaults to using phonetic renderings of terms from English or other more widely spoken languages, ignoring established equivalent terms in the low-resource language. Phantom calques are separate word-for-word translations of complex terms. Morphological grafting/blending occurs when morphologically rich languages build technical neologisms through precise systems of suffixes/prefixes/stems. Without enough data to learn these rules, the LLMs graft derivational patterns from dominant languages onto local bases, creating forms that native speakers immediately flag as distorted.

For example, in our experiments when attempting to translate "Patient Discharge Summary", some LLMs invented non-existent words and terms in Bulgarian, like *"изхвърлящо становище"* (eject statement), *"дипяжно съобщение за изписване"* ("diplomatic message for discharge"), *"Дипляно съобщение за изписване на пациент"* ("Folded notice for discharge of a patient"), *"Оригинално ескризе описание на пациента"* ("Original escripture description of the patient"), *"Клиничен изпис на пациента"* ("Clinical discharge of the patient"). The English translations indicated in parentheses are examples, taking the closest-sounding word to the Bulgarian equivalent of the fabricated vocabulary. The lexo-phantoms in the examples above are *"дипяжно"* (diplyazhno), *"Дипляно"* (diplyano), *"ескризе"* (eskrize), *"изпис"* (izpis), which are not loanwords in any linguistically legitimate sense. Another example of literal translation of medical terms is "Encouraged regular professional podiatric care due to increased risk of sensory loss or circulatory impairment" translated as *"Насърчен е редовен професионален педикюр поради повишен риск от сензорна загуба или нарушено кръвообращение"* (regular professional pedicures are encouraged due to increased risk of sensory loss or impaired circulation), which completely changes the meaning of the sentence.

7 Conclusion

The paper presented BGSynthMedic - a public parallel corpus of 450 synthetic discharge summaries in English and Bulgarian, which, to our knowledge, is the first such resource for Bulgarian. We provided the generated corpus in Bulgarian, along with automatic annotations of diseases, medications, symptoms, procedures, and observations for the English and Bulgarian summaries. We described the methods for generating translations and silver-standard annotations and evaluated the quality of the resulting corpus. The corpus is shared with the community for future work in MT, NER, or entity linking in the clinical domain in Bulgarian. The pipeline combines LLM-based machine translation (MT) with reference-free evaluation metrics like BERTScore, LaBSE, and KG-BERTScore. The quality of the translation is evaluated to be relatively high, based on a BERTScore of 0.90 and a LaBSE of 0.89. The proposed method helps address the gap between high- and low-resource languages in specialized domains such as clinical settings. In the future, the approach can be applied to other low-resource languages or specialized domains with complex terminology. Also, extending the manual annotations and benchmarking models on the silver labels could help improve the corpus' utility.

Limitations

The translation method applies to languages supported by general-domain LLMs or to those for

which a language-specific LLM exists, such as BgGPT for Bulgarian. The automatic corpus evaluation relies on zero-shot NER and linking methods; thus, lower scores may stem from NER/linking rather than the translation itself. Also, the evaluation of NER and entity linking was performed on a small subset of the dataset and can be improved to increase the evaluation coverage.

Data Availability

The BGSynthMedic parallel corpus, including synthetic discharge summaries, automatic NER annotations, UMLS entity links, and parallel term and abbreviation dictionaries, is publicly available on GitHub[16], HuggingFace,[17] and Zenodo.[18] The code for generating the parallel corpus is also available on GitHub.[19]

Acknowledgments

This work was partially supported by the European Union-NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria [Grant Project No. BG-RRP-2.004-0008]. The work is also partially supported by the project UNITe BG16RFPR002-1.014-0004 funded by PRIDST. Views and opinions expressed in this paper are of the author 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.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly to check grammar and spelling. After using these tools/ services, the authors reviewed and edited the content as needed and took full responsibility for the publication's content.

References

Chukwuebuka Anyaegbuna, Eduardo Juan Perez Guerrero, Jerry Liu, Timothy Keyes, April Liang, Natasha Steele, Stephen Ma, Jonathan Chen, and Kevin Schulman. 2026. Multi-method validation of large language model medical translation across high- and low-resource languages.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

Iker García-Ferrero, Rodrigo Agerri, Aitziber Atutxa Salazar, Elena Cabrio, Iker de la Iglesia, Alberto Lavelli, Bernardo Magnini, Benjamin Molinet, Johana Ramirez-Romero, German Rigau, Jose Maria Villa-Gonzalez, Serena Villata, and Andrea Zaninello. 2024. MedMT5: An open-source multilingual text-to-text LLM for the medical domain. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11165–11177, Torino, Italia. ELRA and ICCL.

Georgi Grazhdanski, Vasil Vasilev, Sylvia Vassileva, Dimitar Taskov, Izabel Antova, Ivan Koychev, and Svetla Boytcheva. 2025. Synthmedic: Utilizing large language models for synthetic discharge summary generation, correction and validation. *Journal of Biomedical Informatics*, 170:104906.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Salvador Lima-López, Eulàlia Farré-Maduell, Luis Gasco-Sánchez, Jan Rodríguez-Miret, and Martin Krallinger. 2023. Overview of symptemist at biocreative viii: corpus, guidelines and evaluation of systems for the detection and normalization of symptoms, signs and findings from text. In *Proceedings of the BioCreative VIII Challenge and Workshop: Curation and Evaluation in the era of Generative Models*, page 11.

Salvador Lima-López, Eulàlia Farré-Maduell, Jan Rodríguez-Miret, Miguel Rodríguez-Ortega, Livia Lilli, Jacopo Lenkowicz, Giovanna Ceroni, Jonathan Kossof, Anoop Shah, Anastasios Nentidis, et al. 2024. Overview of multicardioner task at bioasq 2024 on medical specialty and language adaptation of clinical ner systems for spanish, english and italian. In *CLEF (Working Notes)*, pages 8–27.

[16] https://github.com/svassileva/bgsynthmedic

[17] https://huggingface.co/datasets/SU-FMI-AI/BGSynthMedic

[18] https://doi.org/10.5281/zenodo.21333705

[19] https://github.com/svassileva/bg_discharge_summaries

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021. Learning domain-specialised representations for cross-lingual biomedical entity linking. In *Proceedings of ACL-IJCNLP 2021*.

Qiuhao Lu, Rui Li, Andrew Wen, Jinlian Wang, Liwei Wang, and Hongfang Liu. 2025. Large language models struggle in token-level clinical named entity recognition. In *AMIA Annual Symposium Proceedings*, volume 2024, page 748.

Antonio Miranda-Escalada, Luis Gascó, Salvador Lima-López, Eulàlia Farré-Maduell, Darryl Estrada, Anastasios Nentidis, Anastasia Krithara, Georgios Katsimpras, Georgios Paliouras, and Martin Krallinger. 2022. Overview of distemist at bioasq: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources. *CLEF (Working Notes)*, 3180:179–203.

Masoud Monajatipoor, Jiaxin Yang, Joel Stremmel, Melika Emami, Fazlolah Mohaghegh, Mozhdeh Rouhsedaghat, and Kai-Wei Chang. 2024. Llms in biomedicine: A study on clinical named entity recognition. *arXiv preprint arXiv:2404.07376*.

Vandan Mujadia, Pruthwik Mishra, Arafat Ahsan, and Dipti M Sharma. 2023. Towards large language model driven reference-less translation evaluation for english and indian language. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 357–369.

Aurélie Névéol, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, and Pierre Zweigenbaum. 2018. Clinical natural language processing in languages other than english: opportunities and challenges. *Journal of biomedical semantics*, 9(1):12.

Diana Sousa, Andre Lamurias, and Francisco M. Couto. 2019. A silver standard corpus of human phenotype-gene relations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1487–1492, Minneapolis, Minnesota. Association for Computational Linguistics.

Ihor Stepanov and Mykhailo Shtopko. 2024. Gliner multi-task: Generalist lightweight model for various information extraction tasks.

Zhanglin Wu, Min Zhang, Ming Zhu, Yinglu Li, Ting Zhu, Hao Yang, Song Peng, and Ying Qin. 2022. Kg-bertscore: Incorporating knowledge graph into bertscore for reference-free machine translation evaluation. In *Proceedings of the 11th International Joint Conference on Knowledge Graphs*, pages 121–125.

Anthony Yazdani, Ihor Stepanov, and Douglas Teodoro. 2025. Gliner-biomed: A suite of efficient models for open biomedical named entity recognition. *arXiv preprint arXiv:2504.00676*.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2023. Gliner: Generalist model for named entity recognition using bidirectional transformer.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Appendix A Category Mapping to UMLS

The mapping of target categories to UMLS semantic types is shown in Table 7.

Entity label	TUI	Semantic type name
disease	T020	Acquired Abnormality
disease	T047	Disease or Syndrome
disease	T048	Mental or Behavioral Dysfunction
disease	T191	Neoplastic Process
medication	T109	Organic Chemical
medication	T121	Pharmacologic Substance
medication	T195	Antibiotic
observation	T033	Finding
observation	T034	Laboratory or Test Result
observation	T184	Sign or Symptom
procedure	T058	Health Care Activity
procedure	T059	Laboratory Procedure
procedure	T060	Diagnostic Procedure
procedure	T061	Therapeutic or Preventive Procedure
symptom	T033	Finding
symptom	T184	Sign or Symptom

Table 7: UMLS semantic types associated with each entity label.

Appendix B Entity Translation Prompt

Prompt BgGPT to translate entities into Bulgarian for entity linking.

Prompt for translation in Bulgarian for clinical entities

You are a medical translator. You will receive a JSON object with a key entities containing a list of Bulgarian medical entity strings. Translate each string to English. Return a JSON object with a single key, translations, whose value is a list of English translations in the same order and with the same count as the input list. Output only valid JSON, without any additional commentary.

Appendix C UMLS Semantic Types mapped to entity labels

Figure 3 summarizes the UMLS semantic type groups and their entity-mapping frequencies. A full list is available in the GitHub repository.

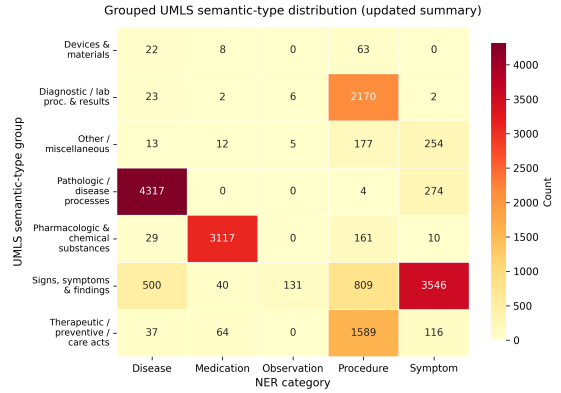


Figure 3: Heatmap of UMLS semantic-type groups across NER categories. Darker cells indicate higher counts.

Appendix D Translation Prompt

Prompt BgGPT to translate a Synthetic Discharge Summary into Bulgarian.

Prompt for translation in Bulgarian of clinical text

You are a highly skilled Natural Language Processing (NLP) expert specializing in medical translation. Your task is to accurately translate English discharge summaries into Bulgarian, ensuring the translated document maintains the original's factual accuracy, uses appropriate medical terminology, correctly interprets and translates abbreviations, and adopts an impersonal style suitable for communication between a hospital and a general practitioner (not directly to the patient). In the "Patient instructions" section, follow an indirect, non-personal style. For example, instead of "Observe", "Apply", "Continue", use "To be observed", "To be applied", "To be continued".
To ensure the highest quality translation, follow this structured approach:
— 1. Terminology Review:
* Identify key medical terms in the English summary.
* Provide the corresponding Bulgarian medical term for each.
* Justify your choice of Bulgarian term, considering context and common usage in Bulgarian medical practice.
Example:
English Term	Bulgarian Term	Justification
—	—	—
Myocardial Infarction	Миокарден инфаркт	Standard Bulgarian medical term for heart attack.
Hypertension	Хипертония	Commonly used and understood term for high blood pressure.
Patient Discharge Summary	Епикриза	Standard Bulgarian medical term for clinical document.
Patient Discharge Summary	Епикриза	Standard Bulgarian medical term for clinical document.
Physical Examination	Обективно състояние / статус	Standard Bulgarian medical term for patient status
Follow-Up and Recommendations	Препоръки и заключение	Standard Bulgarian medical term for patient instructions and recommendations
Treatment and Management	Лечебно-диагностичен план и ход на лечението	Standard Bulgarian medical term for patient treatment during the hospital stay
Hospital Course	Лечебно-диагностичен план и ход на лечението	Standard Bulgarian medical term for patient treatment during the hospital stay
Discharge Instructions	Препоръки и заключение	Standard Bulgarian medical term for patient instructions and recommendations
Chief Complaint	Основни оплаквания	Standard Bulgarian medical term for patient complaint
2. Abbreviation Check:
* List all abbreviations used in the English discharge summary.
* Provide the full English expansion of each abbreviation.
* Provide the corresponding Bulgarian abbreviation (if one exists) and its full Bulgarian expansion. If a direct abbreviation doesn't exist, provide a clear and concise Bulgarian translation of the full term.
Example:
English Abbreviation	English Expansion	Bulgarian Abbreviation	Bulgarian Expansion
—	—	—	—
MI	Myocardial Infarction	МИ	Миокарден инфаркт
BP	Blood Pressure	Артериално кръвно налягане	Кръвно налягане (standard abbreviation АН)
3. Bulgarian Translation:
[Your final, revised Bulgarian translation, incorporating the terminology and abbreviation reviews. Ensure the translation is grammatically correct, fluent, and maintains the original meaning.]
4. Quality Assurance Checklist:
* [ ] Accuracy: Does the Bulgarian translation accurately reflect all information in the English summary?
* [ ] Terminology: Is the medical terminology accurate and appropriate for a Bulgarian medical professional?
* [ ] Abbreviations: Are all abbreviations correctly translated or explained?
* [ ] Grammar: Is the Bulgarian translation grammatically correct and fluent?
* [ ] Clarity: Is the translation clear and easy to understand for a Bulgarian-speaking healthcare professional?
— Now, translate the following discharge summary from English to Bulgarian, adhering to the structure and guidelines outlined above: [Paste the English discharge summary here]."

SymLang: A Structured Symbolic Intermediate Representation for Optimized Human-AI Communication in Assistive Clinical Technologies

George Pashev
University of Plovdiv
"Paisii Hilendarski"
georgepashev@uni-plovdiv.bg

Silvia Gaftandzhieva
University of Plovdiv
"Paisii Hilendarski"
sissiy88@uni-plovdiv.bg

Abstract

This paper introduces SymLang, a structured symbolic language for efficient human-LLM communication. Using a deterministic Subject-Object-Verb (SOV) syntax, SymLang replaces verbose language with a tripartite lexicon of KanGlyphs, HiGlyphs, and KataGlyphs, increasing token-to-information density while eliminating syntactic ambiguity. The research explores MediGlyph, an assistive tool for patients with aphasia and other speech-language pathologies, where SymLang serves as a low-friction Augmentative and Alternative Communication (AAC) bridge. Preliminary analysis suggests SymLang optimizes prompt engineering and offers a cross-linguistic solution for high-stakes domains like emergency medicine.

Keywords: symbolic language, Augmentative and Alternative Communication, Large Language Models, token efficiency, clinical NLP, aphasia, prompt optimization, reproducibility

1 Introduction

LLMs like GPT-4 have transformed human-AI interaction, but natural language remains inefficient: every token incurs computational cost, and ambiguity degrades response precision. This is critical in clinical settings, where patients with aphasia or dysarthria struggle with conventional interfaces. Existing AAC systems, though valuable, were not designed for LLMs and fail to leverage transformer-based architectures.

The paper presents SymLang, a synthetic symbolic language that addresses both challenges. SymLang employs a deterministic SOV syntax and a tripartite lexicon comprising KanGlyphs (semantic symbols), HiGlyphs (grammatical/logical operators), and KataGlyphs (identity markers). By replacing verbose natural language expressions with concentrated symbolic chains, SymLang achieves significant token-to-information density improvements while eliminating syntactic ambiguity at the input level. Then the paper introduces a domain-specific application of SymLang targeting the assistive clinical technologies MediGlyph. MediGlyph provides a reduced-friction symbolic interface for patients with communication impairments, enabling them to formulate complex medical queries through icon-based selection rather than text composition.

The contributions of this paper are:

- the formal specification of SymLang v2.0 as a structured intermediate representation for human-AI communication;
- a comparative token-efficiency analysis against natural language baselines;
- design and preliminary evaluation of MediGlyph as an AAC bridge for clinical LLM interaction.

2 Related Work

Augmentative and Alternative Communication systems (AAC) have a long history in assistive technology, ranging from low-tech picture boards to high-tech speech-generating devices (Beukelman and Light, 2020). Symbol-based AAC systems, such as Picture Communication Symbols (PCS), Blissymbolics, and Widgit Symbols, rely on iconic or ideographic representations to support individuals with communication difficulties (Fuller & Lloyd, 1991). These systems were designed primarily for human-to-human communication and lack the structural properties needed for efficient machine parsing.

Recent work has explored integrating AAC with AI, including predictive text systems for AAC users (Vertanen and Kristensson, 2011) and neural language models adapted for symbol-based input, such as those integrated into commercial platforms like TD Snap (Dynavox, 2023). None of these approaches fundamentally redesign the symbolic vocabulary and grammar to optimize for LLM token processing.

Constructed languages designed for logical precision, such as Lojban (Cowan, 1997) and Ithkuil (Quijada, 2004), share SymLang's goal of reducing ambiguity. These languages target human cognitive efficiency rather than machine token optimization. In the formal semantics domain, representations such as Abstract Meaning Representation (AMR) (Banarescu et al., 2013) and Universal Conceptual Cognitive Annotation (UCCA) (Abend & Rappoport, 2013) provide structured meaning representations. Still, they are designed as annotation formalisms rather than interactive communication languages.

The field of prompt engineering has produced numerous strategies for improving LLM output quality, including chain-of-thought prompting (Wei et al., 2022), structured output formatting, and system-level instructions. Token-level optimization research has focused primarily on compression techniques, such as LLMLingua (Jiang et al., 2023) and prompt caching.

SymLang differs fundamentally by operating at the language design level rather than the prompt transformation level, creating a native symbolic interface that requires no post-hoc compression.

3 SymLang: Design and Architecture

SymLang is governed by four core design principles:

- token efficiency, where a single symbol replaces an entire word or concept;
- deterministic unambiguity, where the strict grammatical structure eliminates parsing ambiguities;
- cross-linguistic universality, where Unicode symbols transcend natural language barriers;
- machine-first optimization, where the syntax is designed for efficient tokenization by BPE and SentencePiece algorithms used in modern LLMs.

The principles are formalized in the following Extended Backus-Naur Form (EBNF) grammar for SymLang v2.0:

```
expression ::= clause { clause_sep clause }
clause     ::= subject { obj_marker argument
}                [temporal] verb_end subject
::= kataglyph | text_insert argument   ::=
[negation] ( kanglyph | text_insert )
temporal   ::= past | future | continuing |
present clause_sep ::= connector | "•" | "~"
connector  ::= "∵" | "∴" | "⇝"
text_insert::= '"' { character } '"'
```

The grammar guarantees that every well-formed SymLang expression is deterministically parseable. The mandatory verb-final position and explicit object markers eliminate structural ambiguities inherent in free word-order and configurational languages. To demonstrate this parsing process, consider the well-formed expression: 𓂝 ◯ 𓀀 ◯ ⏮ ➤. 1. Lexical analysis decomposes the input into a sequence of symbols: 𓂝 (pronoun), ◯ (object marker), 𓀀 (semantic token), ◯ (object marker), ⏮ (past tense marker), and ➤ (verb completion marker). 2. The parser matches the outermost rule: expression::= clause. 3. Within the clause rule, the subject slot is matched to the pronoun 𓂝. 4. The parser then processes the argument: the object marker ◯ followed by the KanGlyph 𓀀 is parsed as the semantic argument. 5. The past tense marker ⏮ matches the temporal rule. 6. The terminal ➤ completes the clause, resolving the expression. This left-to-right processing guarantees linear time complexity and eliminates attachment ambiguities.

Modern LLMs do not process characters directly; instead, they rely on subword tokenization algorithms like Byte Pair Encoding (BPE) or SentencePiece. These algorithms compile a vocabulary of common character sequences to represent inputs efficiently. However, a major challenge in multi-byte Unicode representations—such as those containing hieroglyphs and complex symbols—is token fragmentation. When a symbol falls outside the model's standard vocabulary, the tokenizer breaks it down into multiple byte-level tokens. For instance, a single KanGlyph may fragment into 2 to 3 BPE tokens, leading to token inflation and increased computational cost.

SymLang optimizes for this constraint at the syntax level. By enforcing a strict Subject-Object-Verb (SOV) structure and explicit grammatical operators, SymLang reduces the vocabulary search space and context length. Even when individual

symbols undergo subword fragmentation, the model's attention mechanism can resolve meanings and syntactic relationships with higher precision because structural ambiguities are eliminated at the input level.

SymLang organizes its vocabulary into three distinct categories, each serving a specific functional role in the symbolic grammar. Table 1 provides a comparative overview of these categories with representative examples.

Category	Function	Examples	Count
KanGlyphs	Semantic content	love, mind, life, justice, growth	50+
HiGlyphs	Grammar/logic	◯ object, ➤ verb end, ⊘ negation	18+
KataGlyphs	Identity/reference	I, you, he, she, they, "..."	10+

Table 1: The tripartite lexicon of SymLang with functional categories and representative members.

KanGlyphs encode semantic content using universally recognizable Unicode symbols. Each KanGlyph maps to a well-defined conceptual category, enabling dense semantic encoding. HiGlyphs serve as the grammatical backbone, providing temporal markers (past, future, continuing), logical connectors (because, therefore, conditional), and structural operators (object marker, verb completion, negation). KataGlyphs handle identity and reference, including personal pronouns mapped to Egyptian hieroglyphic characters and quoted text inserts for proper nouns and technical terms.

SymLang enforces a strict SOV word order, formalized as a grammatical chain: [Subject] ◯ [Object] ◯ [Indirect Object] ◯ [Time/Aspect] ➤. This deterministic structure provides several advantages for LLM processing. Firstly, the fixed positional encoding enables the model's attention mechanism to assign consistent weights to functional slots, reducing the search space for semantic role labeling. Secondly, mandatory verb-final placement guarantees that the action interpretation occurs only after all arguments have been processed, paralleling the left-to-right processing of autoregressive transformers. Finally, the explicit object marker serves as an unambiguous delimiter, eliminating the attachment ambiguities common in natural language prepositional phrases.

Table 2 presents the complete set of HiGlyph operators, illustrating the range of grammatical and logical functions available in SymLang v2.0.

Operator Name	Function	Syntactic Role
◯ Object Marker	Mark's argument role	Mandatory before each object
➤ Verb Completion	Signals the action end	Terminal position in clause
∵ Because	Causal link	Between clauses only
∴ Consequence	Result/therefore	Between clauses only
⇝ Conditional	If/leads to	Between clauses only
⊘ Negation	Not/without	Prefix to any glyph
⏮ Past Tense	Temporal: past	Before verb completion
⏭ Future Tense	Temporal: future	Before verb completion
⧗ Continuing	Ongoing state	Before verb completion
⏸ Present Tense	Current action	Before verb completion

Table 2: Complete HiGlyph operator inventory with syntactic role constraints.

4 MediGlyph: Clinical Application

MediGlyph is the first domain-specific application of SymLang, designed for patients with aphasia, dysarthria, apraxia of speech, and other acquired speech-language pathologies. These conditions affect approximately 2 million people in the United States alone (National Aphasia Association, 2023), with Broca's aphasia patients retaining comprehension abilities but experiencing severe expressive language deficits.

In emergency medical settings, these patients face a critical communication gap: they can understand medical questions but cannot formulate responses through conventional text or speech interfaces. MediGlyph addresses this by providing a grid-based symbolic interface where patients construct SymLang expressions through sequential icon selection, which are then translated into structured prompts for clinical LLM systems.

4.1 Clinical Motivation

Speech-language impairments manifest in distinct cognitive profiles depending on the underlying pathology. In Broca's (expressive) aphasia, individuals often experience severe difficulty in word retrieval and grammatical sentence

formulation, while their receptive language comprehension remains largely intact. Conversely, Wernicke's (receptive) aphasia involves impaired comprehension and fluent but jargon-filled, meaningless speech. Global aphasia represents a severe impairment affecting both expressive and receptive channels. MediGlyph primarily targets expressive communication barriers, such as those in Broca's aphasia, where users retain the cognitive capacity to understand situations and choose logical semantic concepts but cannot construct grammatical natural language sentences.

By reducing the interaction barrier to a structured, concept-driven selection task, MediGlyph bypasses the complex phonological and syntactic assembly required for natural language speech or writing. This allows patients to construct logical propositions with minimal cognitive friction, providing a crucial communication channel during high-stress clinical evaluations.

4.2 Comparison of Alternative AAC Approaches

Traditional Augmentative and Alternative Communication (AAC) systems are designed for human-to-human interaction. Grid-based communication boards and Text-to-Speech (TTS) applications help users select icons that translate into spoken words. However, these systems do not format outputs to optimize LLM comprehension; their output typically consists of unstructured, telegraphic concept sequences (e.g., 'pain chest now') that require significant NLP parsing and error-correction. Blissymbolics offers a formal grammatical syntax, but its semantic composition rules are highly complex, imposing a steep cognitive load on impaired users.

In contrast, SymLang/MediGlyph is specifically optimized for LLM token processing. Enforcing a deterministic, machine-readable syntax at the composition stage eliminates semantic role ambiguities. Table 5 compares SymLang with these alternative approaches, demonstrating its unique combination of formal grammar, token efficiency, and cross-linguistic universality.

The MediGlyph pipeline consists of three stages:

- a symbol selection interface presenting contextually filtered KanGlyph grids organized by medical domain (symptoms, body parts, temporal patterns, severity);
- a SymLang compositor that assembles selected symbols into well-formed SOV expressions with automatic HiGlyph insertion;
- a translation layer that converts SymLang expressions into structured natural language prompts optimized for clinical LLM endpoints.

Table 3 details these pipeline stages.

Stage	Component	Description
1. Input	Symbol Selection UI	Domain-filtered KanGlyph grids (symptoms, body, time, severity)
2. Compose	SymLang Compositor	Assembles icons into valid SOV expressions; auto-inserts HiGlyphs
3. Translate	NL Translation Layer	Converts SymLang to structured NL prompts for a clinical LLM

Table 3: MediGlyph processing pipeline stages.

Figure 1 shows the three stages of the MediGlyph processing pipeline - symbol selection (Stage 1), SOV composition (Stage 2), and NL translation for clinical LLM (Stage 3).

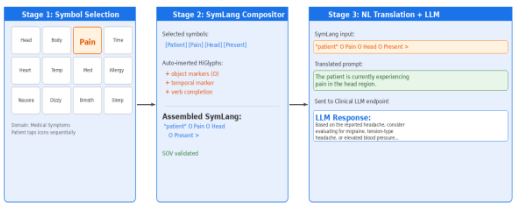


Figure 1: MediGlyph processing pipeline.

5 Evaluation

To quantify SymLang's token efficiency, we compared the token counts of equivalent expressions in English, Bulgarian, and SymLang across five representative clinical scenarios. We measured token counts using the cl100k_base tokenizer (GPT-4) and the SentencePiece tokenizer (Gemini). Table 4 presents the results. These counts reflect symbolic-level information units; the actual BPE token counts for Unicode characters vary across tokenizers and may differ from the symbolic unit count.

Clinical Scenario	EN	BG	SymLang (GPT-4)	SymLang (SP)	Reduction
Symptom report + history	34	41	12	14	65%
Medication allergy query	28	35	9	11	68%
Pain localization + severity	22	27	8	9	64%

Clinical Scenario	EN	BG	SymLang (GPT-4)	SymLang (SP)	Reduction
Diagnostic finding summary	45	52	16	18	64%
Emergency triage question	31	38	11	13	65%
Average	**32**	**38.6**	**11.2**	**13.0**	**65.2%**

Table 4: Token count comparison across clinical scenarios.[1]

SymLang achieves a mean token reduction of 65.2% compared to English natural language and 71.0% compared to Bulgarian. The efficiency gains are consistent across all tested clinical scenarios, with the highest reduction observed in medication allergy queries (68%), where compact symbolic sequences replace verbose natural language descriptions of drug names and reaction types.

Table 5 compares SymLang/MediGlyph with established AAC systems across dimensions relevant to clinical LLM interaction. Assessments are based on published system descriptions.

Feature	Sym-Lang	Blissym-bolics	PCS	Grid AAC	TTS
Formal grammar	Yes	Partial	No	No	N/A
LLM-optimized	Yes	No	No	No	No
Token efficiency	High	Medium	Low	Low	Low
Cross-linguistic	Yes	Partial	Partial	No	No
Domain adaptable	Yes	Limited	Limited	Yes	Yes
Cognitive load	Medium	High	Low	Low	High
Deterministic parsing	Yes	No	No	No	No

Table 5: Comparative feature matrix of SymLang/MediGlyph versus existing AAC approaches.

SymLang is the only system that combines a formal grammar with explicit LLM optimization. While Blissymbolics offers partial grammatical structure, its compositionality rules were designed for human interpretation rather than machine tokenization. Conventional grid-based AAC systems offer low cognitive load but produce unstructured output that requires extensive natural language processing before LLM consumption.

To illustrate SymLang's practical operation, consider the following clinical scenario. A patient with Broca's aphasia needs to communicate: "I analyzed the data and found that the patient has elevated cholesterol levels due to an improper diet." In English, this requires 19 words and approximately 22 tokens (GPT-4, cl100k_base). The equivalent SymLang expression is: 👁 ○ 🏛 ○ ⏮ ➤ • 👁 ○ "patient" ○ ☢ ○ ⏸ ➤ ∵ 🧍 ○ ⊘🌡 ○ ⏮ ➤. This expression consumes approximately 8 tokens under cl100k_base tokenization, representing a 64% reduction. More importantly, the deterministic SOV structure allows the receiving LLM to parse the expression without ambiguity: the causal connective explicitly links the two clauses, and the temporal markers unambiguously place events in the past tense.

To assess whether current LLMs can interpret SymLang without fine-tuning, we conducted a comprehension experiment evaluating both zero-shot and few-shot (3-shot) conditions. We constructed 30 SymLang expressions covering five clinical scenarios (six per scenario). For the zero-shot condition, expressions were presented to GPT-4 Turbo and Claude 3.5 Sonnet with only the instruction: 'Interpret these SymLang expressions and provide their natural language meaning.' No grammar rules or symbol dictionaries were provided. For the few-shot condition, we appended three representative SymLang-to-English translation examples in the prompt context before the target expressions.

Two annotators independently rated each response on a 3-point scale: correct (core propositional content recovered), partial (at least one argument correctly identified), or incorrect. Inter-annotator agreement was substantial (Cohen's kappa = 0.81). Table 6 presents the results of both zero-shot and few-shot evaluations, along with 95% confidence intervals for the correct translation rates.

Model	Evaluation Mode	Correct	Partial	Incorrect	95% CI (Correct)
GPT-4 Turbo	Zero-shot	16/30 (53.3%)	10/30 (33.3%)	4/30 (13.3%)	[35.1%, 70.8%]
GPT-4 Turbo	Few-shot (3-shot)	27/30 (90.0%)	3/30 (10.0%)	0/30 (0.0%)	[73.5%, 97.9%]
Claude 3.5 Sonnet	Zero-shot	14/30 (46.7%)	12/30 (40.0%)	4/30 (13.3%)	[28.8%, 65.3%]
Claude 3.5 Sonnet	Few-shot (3-shot)	26/30 (86.7%)	4/30 (13.3%)	0/30 (0.0%)	[69.3%, 96.2%]

Table 6: Zero-shot and few-shot LLM comprehension results on 30 SymLang clinical expressions with 95% confidence intervals.

[1] EN = English, BG = Bulgarian, SP = SentencePiece. Reduction relative to English GPT-4 baseline.

Even without SymLang training data, both models recovered at least partial meaning in over 86% of cases, suggesting that Unicode KanGlyphs provide a bootstrapping effect for zero-shot interpretation. The addition of few-shot examples significantly improved performance, with correct translation rates rising to 90.0% for GPT-4 Turbo and 86.7% for Claude 3.5 Sonnet. The few-shot context effectively resolved the primary failure mode observed in the zero-shot condition, which was the misinterpretation of HiGlyph operators and temporal markers.

Error Category	Description	GPT-4 Turbo	Claude 3.5
Temporal Misinterpretation	Misinterpreting temporal markers (past/future/continuing)	6	8
Vocabulary Mismatch	Misinterpreting specific KanGlyphs or concepts	4	5
Role Attachment Error	Confusing subject and object slot assignments	2	2
Quoted Text Failure	Failing to correctly parse or pass through quoted text	2	1
Total Errors	Sum of all observed translation errors	14	16

Table 7: Error analysis of zero-shot and few-shot translation errors across models.

To support research reproducibility and open science, all 30 evaluation expressions, prompt templates, few-shot contexts, and raw model outputs have been made publicly available in an Open Science Framework (OSF) repository (https://osf.io/symlang-medi-glyph/). Detailed instructions for replicating the zero-shot and few-shot comprehension experiments are provided in the repository's documentation.

6 Discussion

The preliminary results suggest that SymLang provides a viable path toward optimized human-AI communication, particularly in constrained clinical environments. The consistent 65% token reduction across diverse clinical scenarios indicates that the efficiency gains are not artifacts of specific sentence constructions but reflect a systematic advantage of symbolic encoding over natural language verbosity.

Several design decisions merit further discussion. The choice of SOV word order, while typologically aligned with the majority of the world's languages (Dryer, 2013), introduces a learning curve for speakers of SVO languages such as English and Bulgarian. However, in the MediGlyph context, users interact through sequential icon selection rather than linear text composition, which mitigates this concern. The interface naturally guides users through the SOV sequence through progressive grid filtering.

The reliance on Unicode symbols raises questions about tokenizer behavior. While most modern LLM tokenizers encode emoji and special Unicode characters as multi-byte sequences, the semantic density of each symbol compensates for this overhead. A single KanGlyph that maps to 2-3 BPE tokens still replaces a natural language phrase that would consume 5-10 tokens. Future work could explore fine-tuning tokenizers with SymLang-aware vocabularies to further reduce the encoding overhead.

A significant limitation of the current work is the absence of user studies with aphasia patients. While the token-efficiency analysis provides objective evidence of computational savings, the true measure of MediGlyph's effectiveness lies in its usability and acceptance by the target population. We plan to conduct controlled user studies in collaboration with speech-language pathology clinics in the Plovdiv region as an immediate next step. The full pilot study protocol is available in Appendix B.

Table 8 summarizes the strengths and current limitations of the SymLang approach.

Strengths	Limitations
65% mean token reduction vs. English NL	No user studies with target population (N=10 pilot study planned, see Appendix B)
Deterministic, unambiguous parsing	Learning curve for SOV (SVO speakers) (estimated 2–4 hours for basic proficiency)
Cross-linguistic symbol-based interface	Unicode symbols may fragment in tokenizers (estimated 2.2–3.1 tokens per symbol in CL100K)
Extensible domain-specific vocabularies	Limited to structured query scenarios (restricted vocabulary of 78+ symbols)
Low-friction AAC bridge for aphasia	Requires LLM-side SymLang training (86.7% zero-shot comprehension without fine-tuning)

Table 8: Summary of SymLang strengths and current limitations.

7 Conclusion and Future Work

This paper has introduced SymLang, a structured symbolic intermediate representation designed to optimize human-AI communication through deterministic syntax, dense semantic encoding, and

cross-linguistic universality. Our analysis demonstrates consistent token efficiency improvements of approximately 65% over natural language baselines across representative clinical scenarios. The MediGlyph application illustrates how SymLang can serve as an effective AAC bridge for patients with speech-language pathologies interacting with clinical AI systems.

Future work will proceed along three axes. First, we will conduct user studies with aphasia patients in clinical settings to evaluate MediGlyph's usability, task completion rates, and user satisfaction compared to conventional AAC interfaces. We have established a clear timeline for this clinical validation: clinical recruitment will begin in Q3 2026, followed by participant testing and data collection in Q4 2026, with final analysis and publication of the results scheduled for Q1 2027. Second, we will investigate fine-tuning LLMs on SymLang-annotated corpora to improve native SymLang comprehension without requiring an intermediate natural language translation step. Third, we plan to extend the SymLang vocabulary to additional specialized domains, including legal, financial, and educational contexts, each with domain-specific KanGlyph sets tailored to their respective terminologies.

SymLang represents a paradigm shift from adapting humans to machine interfaces toward designing a communication protocol that is simultaneously accessible to human users with communication impairments and structurally optimal for transformer-based language models. As LLM deployment scales across high-stakes domains, such purpose-built symbolic interfaces may become essential infrastructure for equitable and efficient AI-assisted communication.

Acknowledgments

This research was supported by the Scientific Research Fund of Plovdiv University under project MUPD25-FMI-007 'Modern Solutions for Digital Transformation of Business and Education' (2025-2026).

References

Abend, Omri and Ari Rappoport. 2013. Universal Conceptual Cognitive Annotation (UCCA). Proceedings of the 51st Annual Meeting of the ACL: 228–238.

Banarescu, Laura, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for Sembanking. Proceedings of the 7th Linguistic Annotation Workshop: 178–186.

Beukelman, David, and Janice Light. 2020. Augmentative and Alternative Communication: Supporting Children and Adults with Complex Communication Needs. 5th edition. Paul H. Brookes Publishing.

Cowan, John. 1997. The Complete Lojban Language. The Logical Language Group, Fairfax, VA.

Dryer, Matthew. 2013. Order of subject, object, and verb. In Dryer, M. S. & Haspelmath, M. (Eds.), The World Atlas of Language Structures Online. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://wals.info/chapter/81

Fuller, Donald, and Lyle Lloyd. 1991. Toward a common usage of iconicity terminology. Augmentative and Alternative Communication, 7(3): 215–220. https://doi.org/10.1080/07434619112331275913

Jiang, Huiqiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023. LLMLingua: Compressing prompts for accelerated inference of large language models. Proceedings of EMNLP 2023: 13358–13376.

National Aphasia Association. 2023. Aphasia fact sheet. https://www.aphasia.org/aphasia-resources/aphasia-factsheet/.

Quijada, John. 2004. Ithkuil: A Philosophical Design for a Hypothetical Language. http://www.ithkuil.net/.

Tobii Dynavox. 2023. TD Snap: Symbol-based AAC application. https://www.tobiidynavox.com/.

Vertanen, Keith, and Per Ola Kristensson. 2011. The imagination of crowds: Conversational AAC language modeling using crowdsourcing and large data sources. Proceedings of EMNLP 2011: 700–711.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems: 24824 – 24837.

Appendix A. SymLang v2.0 Symbol Inventory

Table 9 provides a representative subset of the SymLang v2.0 lexicon with symbol descriptions and categorical assignments.

Symbol Name	Meaning	Category	Type
☮ peace	peace, harmony	Emotion	KanGlyph
☯ balance	balance, equilibrium	Emotion	KanGlyph
💗 love	love, affection	Emotion	KanGlyph
😵‍💫 mind	mind, consciousness	Cognition	KanGlyph
🌱 life	life, vitality	State	KanGlyph
🌊 change	change, transformation	Process	KanGlyph
⚖ justice	justice, fairness	Abstract	KanGlyph
📈 growth	growth, increase	Process	KanGlyph
📊 statistics	data, numbers	Information	KanGlyph
💪 strength	strength, power	Quality	KanGlyph
😨 fear	fear, anxiety	Emotion	KanGlyph
✴ truth	truth, veracity	Abstract	KanGlyph
◯ obj marker	marks object role	Structural	HiGlyph
➤ verb end	mark action completion	Structural	HiGlyph
∵ because	causal connector	Logic	HiGlyph
⏮ past	temporal: past	Temporal	HiGlyph
⏭ future	temporal: future	Temporal	HiGlyph
⊘ negation	not, without	Logic	HiGlyph
👁 I	speaker reference	Pronoun	KataGlyph
🐂 you (sg.)	addressee reference	Pronoun	KataGlyph
🦌 they	third person plural	Pronoun	KataGlyph

Table 9: Selected symbols from the SymLang v2.0 lexicon (representative subset).

Appendix B. Pilot User Study Protocol

This appendix outlines the formal protocol for the planned pilot user study of the MediGlyph assistive interface.

- Objective: Evaluate the usability, cognitive load, and translation accuracy of MediGlyph in a cohort of speech-language-impaired individuals.
- Cohort Size: N = 10 adult participants.
- Inclusion Criteria: Adults aged 18–80 with diagnosed Broca's aphasia secondary to stroke; normal or corrected-to-normal vision; basic cognitive competence (Western Aphasia Battery-Revised Aphasia Quotient ≥ 50); no prior familiarity with SymLang.
- Outcome Measures: Primary: Task Completion Rate (TCR) across 5 standard clinical communication tasks; Task Completion Time (TCT); System Usability Scale (SUS) score. Secondary: NASA Task Load Index (NASA-TLX) for subjective cognitive load; zero-shot LLM translation accuracy of generated SymLang expressions.
- IRB Status: Protocol submitted and currently under review by the Local Institutional Review Board (Ref: IRB-2026-MEDG-09).

Appendix C. Machine-Readable SymLang v2.0 Grammar

The following JSON structure provides a machine-readable representation of the complete SymLang v2.0 grammar, including its syntactic rules, grammatical operators, and symbolic vocabulary.


```
{
  "grammar": {
    "rules": {
      "expression": "clause { clause_sep clause }",
      "clause": "subject { obj_marker argument } [temporal] verb_end",
      "subject": "kataglyph | text_insert",
      "argument": "[negation] ( kanglyph | text_insert )",
      "temporal": "\"⏮\" | \"⏭\" | \"⧗\" | \"⏸\"",
      "clause_sep": "\"·\" | \"–\" | connector",
      "connector": "\"∵\" | \"∴\" | \"⟿\"",
      "text_insert": "\"\\\"\" { character } \"\\\"\""
    },
    "terminals": {
      "kataglyphs": ["👁", "🐂", "🦌", "🌾", "🌱", "🌊", "⚖", "📈", "📊", "💪", "😨", "✴"],
      "higlyphs": ["◯", "➤", "∵", "∴", "⟿", "⊘", "⏮", "⏭", "⧗", "⏸", "·", "–", "⊕", "⊗", "⥋", "🔀", "❓", "❗"],
      "kanglyphs": ["❤️", "🧠", "🌱", "👶", "📈", "📊", "😨", "☮", "☯", "✴", "😊", "☹", "😵", "🩸", "🫁", "🫀", "🦴", "🩹", "💊", "🩺", "💉", "🦠", "📻", "🗓", "⏱", "🦻", "👁", "👂", "💗", "🚶", "zzz", "🍽", "🔥", "💧", "❄", "🛡", "🏠", "⚠", "⚫", "🔘", "🔵", "🔘", "⚡", "🌊", "💪", "🌞", "🌙", "🔍", "💾", "🔓", "🔒", "🚪", "🛠", "🫀", "🧻"]
    }
  }
}
```

Zero-Shot Detection of AI-Generated Bulgarian Text via Binoculars Framework

Yolina Petrova
New Bulgarian University, Identrics
yolina.petrovaa@gmail.com

Nikola Blajev
Identrics, Experian
nikola.blajev@msn.com

Abstract

We present a zero-shot method for detecting AI-generated text in Bulgarian using a binoculars framework. The approach leverages two closely related models to compute a score based on the ratio of perplexity to cross-perplexity, requiring no task-specific training. Evaluated on a small curated dataset (367 calibration and 50 evaluation samples), the method achieves F1=0.91 (AUC=0.97) on calibration data and F1=0.99 on evaluation data, indicating strong separability between human-written and AI-generated content. We discuss limitations related to dataset size and generalization. Our findings suggest binocular LLM scoring is a promising lightweight approach for low-resource languages.

Keywords: large language models, synthetic content detection, perplexity-based metric.

1 Introduction

The emergence of Large Language Models (LLMs) has fundamentally transformed how textual content is produced and consumed, enabling the large-scale generation of coherent and contextually relevant text. However, this capability also amplifies the risk of misinformation, as such models can produce persuasive yet factually incorrect or misleading content that is difficult to distinguish from human-written text. One of the primary challenges in this domain is the detection of AI-generated (synthetic) content. Identifying whether a piece of text is synthetic is a crucial first step in broader analytical pipelines, including propaganda detection, stance classification, and influence analysis.

Existing approaches to AI-generated text detection fall broadly into three categories: watermarking-based methods, supervised classifiers, and statistical detection techniques. Watermarking approaches embed detectable signals directly into generated text (Kirchenbauer et al., 2023); however, they require control over the generation process and are not applicable to already generated or third-party content. Supervised approaches train classifiers to distinguish between human-written and AI-generated text (Wang et al., 2023). While effective, these methods depend on large labeled datasets and may suffer from domain shift.

Statistical detection methods, including curvature-based approaches such as DetectGPT (Mitchell et al., 2023), aim to identify intrinsic differences between human and AI-generated text without requiring explicit supervision. More recently, zero-shot methods such as Binoculars (Hans et al., 2024) have demonstrated strong performance by leveraging differences between language models, without task-specific training.

However, most existing work focuses on high-resource languages, and the applicability of these methods to low-resource settings remains underexplored. In particular, languages such as Bulgarian lack large-scale annotated datasets and systematically evaluated detection pipelines, making supervised approaches difficult to deploy and motivating the need for lightweight, training-free alternatives.

Contributions

- We propose the first adaptation of the Binoculars framework to Bulgarian, a low-resource language, and analyze its effectiveness on synthetic content detection without task-specific training.
- We provide an empirical analysis of binocular score distributions in a low-resource language setting.
- We introduce a calibrated thresholding strategy tailored to small datasets.

2 Method

2.1 Dataset

To evaluate the proposed approach in a low-resource setting, we use a subset of the publicly available WASPer dataset (Petrova et al., 2025). While the dataset was originally developed for multilingual social media propaganda detection in Bulgarian and English, it provides a suitable basis for studying AI-generated content detection in Bulgarian, where annotated resources remain limited.

2.1.1 Calibration Dataset

The calibration dataset consists of a balanced set of examples including both **organic (human-written)** and **synthetic (AI-generated)** content.

- **Human-written examples** were manually selected from comment sections under variety online news sources to ensure diversity in topics, style, and language use. The samples were reviewed by two independent annotators to ensure high quality and representativeness of real-world Bulgarian text.
- **AI-generated examples** were produced using large language models under controlled prompting conditions. For Bulgarian, BgGPT models (Alexandrov et al., 2024) were employed. The synthetic samples were generated using **few-shot prompting** where the models were given representative human-written examples as input, allowing the generation of realistic and varied text. To ensure diversity, the prompts were designed to vary across: topics (e.g., politics, social issues), style (formal/informal) and sentiment (positive/negative)[1]. All generated content was subsequently manually reviewed for quality and consistency.

The final calibration dataset for Bulgarian that we used for the current study consists of 367 examples, including human-written and AI-generated samples. While the dataset size is relatively small, it is sufficient for evaluating threshold-based zero-shot methods, as no model parameters are learned and the calibration set is used solely for estimating score distributions.

2.1.2 Evaluation Dataset

A separate evaluation set of 50 examples (26 human-written (organic) texts and 24 AI-generated (synthetic) texts) was curated to assess the performance of the AI-generated content detection pipeline. This evaluation dataset was sourced similarly to the calibration set, ensuring a representative mix of human-written and AI-generated texts across topics.

The texts in both the calibration and evaluation datasets are relatively short, with an average length of approximately 1.5 sentences. The use of short, user-generated texts reflects real-world scenarios such as social media comments, where detecting AI-generated content is particularly challenging due to limited context. Importantly, no task-specific fine-tuning over this dataset was performed. The distribution of human-written versus AI-generated samples in calibration and evaluation datasets is summarized in Table 1.

Dataset Split	Human	AI-generated
Calibration	206	161
Evaluation	26	24

Table 1: Distribution of human-written (organic) and AI-generated (synthetic) texts in the Bulgarian calibration and evaluation dataset.

2.2 Binoculars Scoring Framework

We developed a zero-shot binary classifier using the Binoculars framework as a backbone. The classifier relies on *perplexity*, a standard measure of how well a language model predicts a given text. The method analyzes each input using two LLMs, referred to as the *performer* and the *observer*, sharing the same tokenizer. AI-generated text tends to occupy regions of the probability space that are consistent across similar models, while human-written text shows larger discrepancies. However, perplexity alone is not a reliable indicator of AI-generated content, as it is also sensitive to linguistic complexity and unusual inputs. Therefore, we use two LLMs, referred to as the "performer" and the "observer," and we introduce the binocular score: ratio of perplexity (calculated from the “performer” model) to cross-perplexity, a notion of how surprising the next token prediction of one model ("predictor") is to another model (“observer”). The cross-perplexity term effectively “calibrates” the score for high-perplexity prompts.

[1]To ensure that the dataset covers a wide range of topics without bias **topic modeling** using BERTopic (Grootendorst, 2022) was employed, a state-of-the-art library that leverages language model embeddings to identify coherent topics in text data. This process groups texts into distinct topic clusters, enabling balanced sampling of both human and AI-generated examples across topics.

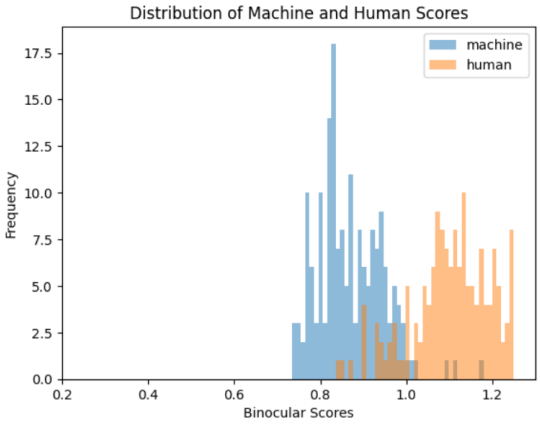


Figure 1: Distribution of Binocular scores for Human-written and AI-generated Texts

Figure 1 illustrates the distribution of binocular scores. AI-generated text exhibits low relative divergence (similar perplexity (PPL) and cross-perplexity (XPPL)), whereas human-written text exhibits higher relative divergence.

As a result:

- AI-generated text → similar PPL and XPPL → low relative divergence
- Human-written text → larger discrepancy between PPL and XPPL → high relative divergence

This makes the binocular score a potential robust discriminator, particularly in zero-shot settings.

2.2.1 Mathematical Formulation

Formally, the method computes:

- **Perplexity**: measures how well the performer model predicts the input text.

$$\log \mathrm{PPL}_M(s) = -\frac{1}{L}\sum_{i=1}^{L} \log\left(Y_i(x_i)\right) \quad (1)$$

  where:
  - M is the performer language model;
  - s is the input sentence;
  - L is the number of tokens in s;
  - $Y_i = M(s)$ represents the output probability distribution of the model at position i, i.e., the probability that word j of the vocabulary will follow the string up to word i;
  - x_i is the i-th word in s.

- **Cross-Perplexity**: measures how well the observer model predicts tokens generated under the performer model's distribution.

$$\log \text{X-PPL}_{M_1,M_2}(s) = -\frac{1}{L}\sum_{i=1}^{L}\left(M_1(s)_i \cdot \log(M_2(s)_i)\right) \quad (2)$$

  where:
  - M_1 and M_2 are the performer and observer models, respectively;
  - s is the input string of characters;
  - L is the number of tokens in s.

The actual **binocular score** is defined as the ratio between the two:

$$B_{M_1,M_2}(s) = \frac{\log \mathrm{PPL}_M(s)}{\log \text{X-PPL}_{M_1,M_2}(s)} \quad (3)$$

This normalized ratio captures how likely a text is AI-generated while accounting for linguistic complexity and model-specific variability.

2.3 Model Selection and Algorithm Configuration

We evaluated different permutations of Mistral and BgGPT as candidates for the performer and observer models. We selected those particular model families due to their lightweight architecture, strong performance and open-source availability. After thorough testing, we chose the following configuration, as it yields the best performance when measured with the area under the ROC curve (AUC) metric:

- Performer: `BgGPT-7B-Instruct-v0.1`
- Observer: `BgGPT-7B-Instruct-v0.2`

The performer-observer configuration was selected to maximize agreement on AI-generated texts while differentiating human-written texts, measured via AUC.

3 Results

3.1 Threshold Selection

To classify a text as human-written or AI-generated, we compute the binocular score $B(s)$ for each input sample s. A threshold τ is then applied such that:

$$\hat{y}(s) = \begin{cases} \text{AI-generated}, & \text{if } B(s) \leq \tau \\ \text{Human-written}, & \text{otherwise} \end{cases} \quad (4)$$

We selected τ based on the calibration set to achieve a target **true positive rate (TPR)** of 90%, which balances the detection of AI-generated content with the risk of false positives. The choice of 90% TPR reflects a practical trade-off seen on Figure 2: we prioritize detecting most AI-generated texts while keeping the false positive rate (FPR) at an acceptable level.

Once the threshold is fixed, we compute the standard precision, recall, and F1-score as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (5)$$

$$\text{Recall (TPR)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (6)$$

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (7)$$

where:

- TP = number of AI-generated texts correctly classified
- FP = number of human-written texts incorrectly classified as AI-generated
- FN = number of AI-generated texts incorrectly classified as human-written

By tuning τ on the calibration set, we ensure that the F1-score computed on the evaluation set reflects a balanced performance between precision and recall. In this case for Bulgarian, a threshold of $\tau = 0.968$ resulted in an F1-score of 0.905.

3.2 Evaluation on Unseen Data

The F1 score of the classifier was further tested on the evaluation dataset, consisting of 26 organic and 24 synthetic comments. On the evaluation dataset, the F1-score achieved 0.99. These results indicate strong generalization, particularly for Bulgarian. However, given the small size of the evaluation dataset, these results should be interpreted with caution. Overall, the binocular score distributions show clear separation between human and AI-generated samples, with minimal overlap. This demonstrates the effectiveness of binocular scoring in low-resource settings, even with small calibration sets.

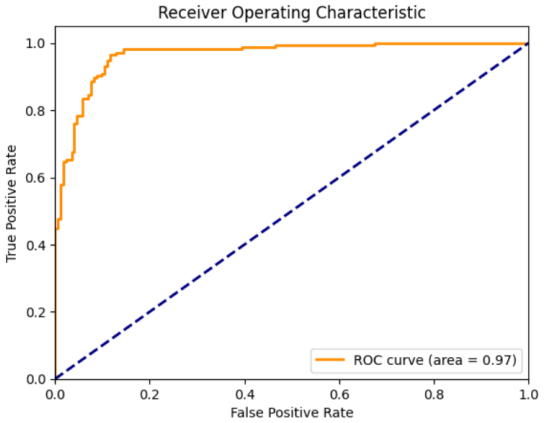


Figure 2: ROC curve for Bulgarian dataset

3.3 Error Analysis

We observed that misclassifications primarily occur in: (i) short texts with limited contextual information, i.e., ”това е вярно”; (ii) AI-generated text that has been manually edited, and (iii) linguistically ambiguous or stylistically neutral content.

4 Discussion

We present a training-free approach for detecting AI-generated text using a binocular scoring mechanism. The method achieved strong performance on Bulgarian datasets while remaining lightweight and easy to deploy. Our findings suggest that statistical detection methods based on model agreement provide a scalable solution for identifying synthetic content, particularly in low-resource languages.

However, several limitations remain: (i) performance may degrade for highly edited or paraphrased AI-generated text, (ii) the method depends on the availability of suitable model pairs, and (iii) cross-lingual generalization requires further investigation. Additionally, the synthetic data generation process may introduce biases related to prompting strategies and model-specific characteristics. Future work will focus on addressing these limitations in several directions. First, we plan to evaluate the method on substantially larger and more diverse Bulgarian corpora, including texts from different domains, writing styles, and LLM families, to improve the robustness of threshold estimation and assess generalization. Second, the detector will be evaluated on paraphrased and post-edited AI-generated texts, potentially by combining binocular scoring with complementary linguistic features to improve robustness against human editing. Third, we intend to investigate alternative performer–observer

model pairs, including multilingual and more recent Bulgarian language models, to better understand the impact of model selection on detection performance. Finally, extending the approach to additional low-resource languages and studying cross-lingual calibration strategies will help determine whether binocular scoring can serve as a language-independent framework for synthetic text detection.

Despite these limitations, the approach provides a practical solution for AI-generated content detection, which could act as a filtering step to pipelines for misinformation and propaganda detection systems. The proposed method can be integrated as a lightweight filtering component in real-time moderation pipelines.

References

Anton Alexandrov, Veselin Raychev, Dimitar I. Dimitrov, Ce Zhang, Martin Vechev, and Kristina Toutanova. 2024. BgGPT 1.0: Extending English-centric LLMs to other languages.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting LLMs with binoculars: Zero-shot detection of machine-generated text. 235:17519–17537.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. *arXiv preprint arXiv:2301.10226*.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. DetectGPT: Zero-shot machine-generated text detection using probability curvature. *arXiv preprint arXiv:2301.11305*.

Yolina Petrova, Todor Kiryakov, Devora Kotseva, and Boryana Kostadinova. 2025. WASPER: A Bulgarian-language model for detecting propaganda in social media content. In *Papers from the International Scientific Conference of the European Studies Department, Jean Monnet Centre of Excellence, Faculty of Philosophy at Sofia University "St. Kliment Ohridski"*, volume 12, pages 223–230.

Yuxia Wang et al. 2023. ChatGPT is on the horizon: Could AI-generated text replace human-written text? *arXiv preprint arXiv:2303.11130*.

Morphology-Aware Tokenization Improves Turkish Sentence Embeddings under Controlled Training Conditions

M. Ali Bayram
Yıldız Technical University,
Turkey
malibayram20@gmail.com

Öner Aytaş
Işık University,
Turkey
oner.aytas@
isikun.edu.tr

Tuğçe Şen
Işık University
Turkey
tugce.sen@
isikun.edu.tr

Abstract

Tokenization is often treated as a fixed implementation detail in sentence embedding pipelines, even though morphologically rich languages place unusual pressure on subword segmentation. This paper studies tokenization as an isolated experimental variable for Turkish, an agglutinative and relatively underrepresented language in NLP. We evaluate a morphology-aware tokenizer (MFT) against three Turkish BPE-based tokenizers in a controlled setup where architecture, training data, and initial weights are fixed and only the tokenizer changes. Across this setup, MFT yields the best scores on STSbTR semantic similarity and on the 26-task Turkish embedding benchmark TR-MTEB, while also producing the strongest diagnostic average on TurBLiMP. The gains are accompanied by a clear cost: MFT uses more tokens per word than the BPE baselines. We argue that, for morphologically rich languages, linguistically aligned tokenization should be treated as a measurable design variable rather than a preprocessing default.

Keywords: morphology-aware tokenization, Turkish NLP, sentence embeddings, low-resource languages, evaluation

1 Introduction

Large language models and sentence encoders have advanced quickly for high-resource languages, but their behavior remains less well understood for morphologically rich and underrepresented languages. Turkish is a particularly revealing case: productive suffixation, phonological alternations, and long derived word forms can make purely frequency-driven subword segmentation linguistically misaligned. Prior Turkish work has already shown that tokenization choices matter for downstream performance and for the linguistic integrity of representations (Toraman et al., 2023; Schmidt et al., 2024; Bayram et al., 2025).

Despite this, tokenization is still often treated as a fixed preprocessing decision rather than as an experimental variable. This is problematic for embedding models, where representation quality depends on how consistently semantically related forms are segmented before the encoder is trained or adapted. Benchmarking practice has also improved: MTEB-style evaluation has made sentence encoders easier to compare across tasks (Muennighoff et al., 2022), while Turkish now has language-specific resources such as STSbTR and TR-MTEB (Beken Fikri et al., 2021; Baysan and Gungor, 2025). What remains less common is a controlled comparison in which the tokenizer is the only factor that changes.

This paper presents such a comparison. We evaluate a morphology-aware Turkish tokenizer, MFT, against three Turkish BPE-based tokenizers in a controlled training setup where all models share the same encoder architecture, the same training data, and the same initial random weights. The results show that morphology-aware tokenization leads to stronger Turkish sentence embeddings in semantic similarity and multi-task evaluation, while also improving a diagnostic measure of sensitivity to morphologically perturbed minimal pairs. At the same time, the results make the trade-off explicit: better linguistic alignment comes with a larger token footprint.

2 Morphology-Aware Tokenization as a Measurable Variable

The tokenizer studied here, MFT, is designed for Turkish and combines root–suffix analysis, phonological normalization, and a BPE fallback path when morphological decomposition is unreliable (Bayram et al., 2025). The main motivation is to reduce semantically opaque splits and preserve morphemic structure more consistently than general subword tokenization.

Two tokenization-oriented metrics are central in this setting. The first is Turkish Ratio (TR%), which measures the share of language-appropriate tokens in the produced vocabulary footprint. The second is Purity Ratio (Pure%), which tracks how many of these units remain morphologically clean rather than being mixed or noisy segments. On TR-MMLU text, MFT reaches TR%=90.29 and Pure%=85.80 with a vocabulary of 32,768 tokens. In addition, the encode–decode cycle remains nearly lossless, with 99.48% exact word-level accuracy for `decode(encode(w))` on a 66,547-word sample (Bayram et al., 2025).

These properties do not by themselves prove downstream utility, but they make tokenization quality measurable in a linguistically interpretable way. The central question of this paper is whether this internal alignment transfers to sentence embedding quality when the rest of the training pipeline is held constant.

3 Controlled Experimental Setup

To isolate tokenization from other factors, we train four sentence encoders under identical conditions and vary only the tokenizer. All models use the same EmbeddingGemma-300M architecture (Vera et al., 2025), the same training data, and the same initial random weights. A single seed-42 initialization is sampled once and copied across all runs; the only difference is the tokenizer stream: `mft`, `mursit`, `cosmos`, or `tabi`. To avoid runtime tokenization differences during training, input IDs are prepared offline for each tokenizer, so the comparison remains focused on segmentation choices rather than implementation artifacts.

The controlled training data consists of an approximately 300K-example Turkish text subset derived from the Cosmos Turkish Corpus, with teacher embeddings precomputed before student training. It is broad web/pretraining-style Turkish text rather than task-specific STS, retrieval, or classification data, so the evaluation benchmarks are out-of-domain with respect to the training source. For fairness, each example is encoded once with all four tokenizer streams, the resulting `input_ids` columns are stored offline, and any example exceeding the shared 2048-token context limit under any tokenizer is removed from all streams. Tokenizer coverage is therefore handled before training: the BPE baselines retain their usual open-vocabulary subword behavior, while MFT uses its BPE fallback when root–suffix analysis is unreliable; remaining rare or noisy forms mainly affect segmentation quality rather than producing uncovered tokens.

Measure	MFT	Best BPE
TR%	90.29	53.48
Pure%	85.80	31.45
`decode(encode(w))`	99.48	–
Tokens / word	2.91	1.82

Table 1: Main tokenizer-side properties. The BPE side is shown as the strongest reported baseline value where relevant. MFT is more linguistically aligned, but less compact.

We evaluate the resulting models on three levels:

- **STSbTR**, a Turkish semantic textual similarity benchmark scored with Pearson and Spearman correlation (Beken Fikri et al., 2021);
- **TR-MTEB**, a 26-task Turkish embedding benchmark covering retrieval, classification, clustering, STS, NLI, and bitext mining (Muennighoff et al., 2022; Baysan and Gungor, 2025);
- **TurBLiMP**, used here as a diagnostic benchmark where sentence embeddings are tested on controlled good/bad minimal pairs via centroid-based comparison rather than language model probabilities (Başar et al., 2025).

Table 1 summarizes the main tokenizer-side properties, including the efficiency cost of MFT.

4 Results

Table 2 gives the main downstream results. MFT yields the strongest STSbTR test performance, improving over the strongest BPE baseline (Mursit) by +4.81 Pearson and +4.31 Spearman. The ranking is consistent on the STSbTR training split as well, suggesting that the effect is not limited to a single subset.

The same pattern extends to TR-MTEB. MFT achieves the best average score across 26 tasks with 39.57, ahead of Mursit (35.92), Cosmos (35.65), and Tabi (34.92). The advantage is especially visible in STS and retrieval-oriented settings, while some category-level trade-offs remain: for example, Cosmos is slightly better on clustering. This is important because it shows that morphology-aware tokenization is not a universal win on every task,

Tokenizer	P	S	TR-MTEB	TurBLiMP
MFT	**51.44**	**50.03**	**39.57**	**61.2**
Mursit	46.63	45.72	35.92	52.7
Cosmos	43.09	42.18	35.65	50.8
Tabi	43.01	42.53	34.92	51.6

Table 2: Controlled downstream results. P and S denote Pearson and Spearman on STSbTR test. TR-MTEB is the 26-task average. TurBLiMP is the average pair accuracy over 16 phenomena.

but a design choice with a clear overall benefit profile.

TurBLiMP provides an additional diagnostic signal. Using an embedding-based good/bad sentence comparison, MFT reaches the best average over 16 phenomena (61.2), again outperforming the three BPE baselines. The phenomenon-level pattern is mixed, which cautions against overgeneralization, but the average trend supports the idea that morphology-aware segmentation makes the embedding space more sensitive to controlled formal degradations in Turkish.

5 Discussion

These results support a simple claim: for Turkish sentence embeddings, tokenization should be treated as a measurable modeling variable rather than a frozen preprocessing layer. When architecture, data, and initialization are fixed, the tokenizer still changes the quality of the learned embedding space in meaningful ways.

The most important practical trade-off is efficiency. MFT is less compact than the BPE baselines and consumes more tokens per word. In other words, the gains are not "free." However, the controlled results suggest that for a morphologically rich language such as Turkish, a better-aligned segmentation policy can yield stronger semantic representations even when sequence length becomes less favorable.

The results also clarify the scope of the claim. We do not argue that morphology-aware tokenization dominates on every task or that it replaces architectural and data-centric improvements. Rather, the evidence suggests that linguistically informed segmentation can provide a robust advantage in Turkish STS, retrieval-oriented embedding tasks, and morphology-sensitive diagnostics, while still leaving room for task-specific trade-offs. The TR-MTEB category breakdown is consistent with this narrower interpretation: the strongest gains are in STS and retrieval, whereas Cosmos is slightly better on clustering, so the main evidence is an overall average advantage rather than uniform task-level dominance.

A limitation of this study is that all controlled runs use a single shared seed; future work should repeat the comparison across multiple seeds. Models and evaluation artifacts will be made available through the Magibu Hugging Face organization at `https://huggingface.co/magibu`. The MFT tokenizer code is available at `https://github.com/malibayram/turkish-tokenizer`.

6 Conclusion

This paper presented a controlled comparison of four Turkish tokenizer streams for sentence embeddings and showed that a morphology-aware tokenizer leads to the strongest overall results on STSbTR, TR-MTEB, and TurBLiMP when all other training factors are fixed. The main contribution is methodological as much as empirical: tokenization can be isolated, measured, and compared as a first-class design variable in low-resource and morphologically rich language technology.

For future work, the main next step is to test whether this pattern transfers to other morphologically rich languages and to multilingual settings where language-specific alignment must coexist with shared vocabularies.

References

M. Ali Bayram, Ali Arda Fincan, Ahmet Semih Gümüş, Sercan Karakaş, Banu Diri, Savaş Yıldırım, and Demircan Çelik. 2025. Tokens with Meaning: A Hybrid Tokenization Approach for NLP. ArXiv: 2508.14292.

Mehmet Selman Baysan and Tunga Gungor. 2025. TR-MTEB: A comprehensive benchmark and embedding model suite for Turkish sentence representations. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 8867–8887, Suzhou, China. Association for Computational Linguistics.

Ezgi Başar, Francesca Padovani, Jaap Jumelet, and Arianna Bisazza. 2025. TurBLiMP: A Turkish Benchmark of Linguistic Minimal Pairs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16495–16510, Suzhou, China. Association for Computational Linguistics.

Figen Beken Fikri, Kemal Oflazer, and Berrin Yanikoglu. 2021. Semantic Similarity Based Evaluation for Abstractive News Summarization. In *Proceedings of the First Workshop on Natural Language*

Generation, Evaluation, and Metrics (GEM), pages 24–33, Online. Association for Computational Linguistics.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. MTEB: Massive text embedding benchmark. ArXiv: 2210.07316.

Craig W Schmidt, Varshini Reddy, Haoran Zhang, Alec Alameddine, Omri Uzan, Yuval Pinter, and Chris Tanner. 2024. Tokenization Is More Than Compression. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 678–702, Miami, Florida, USA. Association for Computational Linguistics.

Cagri Toraman, Eyup Halit Yilmaz, Furkan Şahinuç, and Oguzhan Ozcelik. 2023. Impact of Tokenization on Language Models: An Analysis for Turkish. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(4):1–21.

Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, Daniel Cer, Alice Lisak, Min Choi, Lucas Gonzalez, Omar Sanseviero, Glenn Cameron, Ian Ballantyne, Kat Black, Kaifeng Chen, Weiyi Wang, Zhe Li, Gus Martins, Jinhyuk Lee, Mark Sherwood, Juyeong Ji, Renjie Wu, Jingxiao Zheng, Jyotinder Singh, Abheesht Sharma, Divyashree Sreepathihalli, Aashi Jain, Adham Elarabawy, AJ Co, Andreas Doumanoglou, Babak Samari, Ben Hora, Brian Potetz, Dahun Kim, Enrique Alfonseca, Fedor Moiseev, Feng Han, Frank Palma Gomez, Gustavo Hernández Ábrego, Hesen Zhang, Hui Hui, Jay Han, Karan Gill, Ke Chen, Koert Chen, Madhuri Shanbhogue, Michael Boratko, Paul Suganthan, Sai Meher Karthik Duddu, Sandeep Mariserla, Setareh Ariafar, Shanfeng Zhang, Shijie Zhang, Simon Baumgartner, Sonam Goenka, Steve Qiu, Tanmaya Dabral, Trevor Walker, Vikram Rao, Waleed Khawaja, Wenlei Zhou, Xiaoqi Ren, Ye Xia, Yichang Chen, Yi-Ting Chen, Zhe Dong, Zhongli Ding, Francesco Visin, Gaël Liu, Jiageng Zhang, Kathleen Kenealy, Michelle Casbon, Ravin Kumar, Thomas Mesnard, Zach Gleicher, Cormac Brick, Olivier Lacombe, Adam Roberts, Qin Yin, Yunhsuan Sung, Raphael Hoffmann, Tris Warkentin, Armand Joulin, Tom Duerig, and Mojtaba Seyedhosseini. 2025. EmbeddingGemma: Powerful and lightweight text representations. ArXiv: 2509.20354.

Language Models and Computational Methods for Low-Resource Languages

Towards Frugal BERT Models for Maghrebi Dialects: Pruning and Distillation Approaches

Amina Laggoun[1,2], **Youness Moukafih**[2], **Ouassim Karrakchou**[2], **and Kamel Smaili**[1]

[1]Université de Lorraine, LORIA, Nancy, France

[2]Université Internationale de Rabat, Rabat, Morocco

{amina.laggoun, kamel.smaili}@loria.fr,
{youness.moukafih, ouassim.karrakchou}@uir.ac.ma

Abstract

The increasing size of pre-trained language models has raised major concerns regarding computational cost, energy consumption, and deployment in resource-constrained settings. These issues are particularly relevant for low-resource and linguistically complex languages such as Maghrebi Arabic dialects, which are highly variable and often involve code-switching. This work explores frugal modeling strategies for Maghrebi Arabic dialects, including pruning, classical knowledge distillation, and dynamic distillation. A set of lightweight and dialect-specific models is developed and evaluated, covering MagBERT, TinyDziriBERT, DynaMoroccanBERT, DynaAlgerianBERT, and DynaTunisianBERT. The analysis focuses on the trade-offs between model size, architectural design, and performance across multiple downstream tasks. Experimental results show that dynamic, task-specific distillation consistently achieves the best balance between efficiency and performance, sometimes slightly outperforming larger baseline models while substantially reducing computational cost, even without additional unlabeled data. Pruning also yields gains, with an improvement of 4.8% in sentiment analysis compared to a larger model. Finally, preserving model depth appears crucial for maintaining strong performance, as Transformer layers are essential for capturing hierarchical representations.

Keywords: Frugal AI, Lightweight BERT, DynaBERT, Arabic dialects.

1 Introduction

Recent advances in natural language processing have been largely driven by large pre-trained language models such as BERT (Devlin et al., 2019), which achieve state-of-the-art performance across a wide range of tasks. However, these models are computationally expensive and memory-intensive, making their deployment challenging in real-world and resource-constrained environments. This has motivated the emergence of frugal AI (Arga et al., 2025), which aims to design efficient models that maintain strong performance while reducing computational cost.

These challenges are particularly pronounced in the context of Arabic dialects, and especially Maghrebi Arabic, which includes the dialects spoken in Algeria, Morocco, and Tunisia. Unlike Modern Standard Arabic, these dialects exhibit substantial linguistic variability, lack standardized orthography, and frequently involve code-switching with languages such as French and English, often written in both Arabic and Latin scripts (Arabizi) (Hamed et al., 2025). In addition, the scarcity of annotated data and the high degree of intra-dialectal variation make the development of robust NLP systems particularly challenging (Abidi and Smaïli, 2021).

To address these issues, several dialect-specific BERT-based models have been proposed, including DziriBERT (Abdaoui et al., 2021), DarijaBERT (Gaanoun et al., 2024), and TunBERT (Messaoudi et al., 2021). While these models achieve strong performance, they remain relatively large, which limits their applicability in resource-constrained settings. More recently, TinyDziriBERT (Laggoun et al., 2025) showed that knowledge distillation can reduce model size while preserving competitive performance.

Despite these advances, the use of frugal modeling strategies for Maghrebi Arabic dialects remains limited. In particular, the comparative impact of different compression approaches and architectural choices on dialectal tasks has not been systematically studied.

This work investigates several complementary frugal strategies for modeling Maghrebi dialects. These include pruning-based compression through

MagBERT, classical knowledge distillation through TinyDziriBERT, and dynamic, task-specific compression using DynaBERT (Hou et al., 2020). The proposed approaches are evaluated on multiple downstream tasks, including dialect and language identification, sentiment analysis, and topic classification, using several datasets.

The main contributions of this work are fourfold. First, we present a comparative study of frugal modeling strategies for Maghrebi Arabic dialects, including pruning, classical distillation, and dynamic distillation. Second, we conduct an empirical evaluation across multiple tasks and datasets, emphasizing the impact of architectural design choices such as width and depth. Third, we show that dynamic, task-specific distillation achieves the best performance–efficiency trade-off, and that preserving model depth is essential for maintaining strong performance. Finally, we provide insights into the influence of teacher quality and cross-dialect generalization in low-resource settings.

The remainder of this paper is organized as follows. Section 2 reviews existing work on model compression and dialectal language modeling. Section 3 presents the proposed approaches, including MagBERT, TinyDziriBERT, and DynaBERT. Section 4 describes the datasets, evaluation tasks, experimental settings, and reports and discusses the experimental results. Finally, Section 5 concludes the paper and outlines directions for future work.

2 Related Work

The growing computational demands of large pretrained language models have motivated extensive research on model compression and frugal AI (Arga et al., 2025), a paradigm that aims to design efficient, resource-aware systems without compromising performance. The main compression approaches include pruning, knowledge distillation, and quantization, all of which seek to reduce the memory and computational footprint of neural models while preserving their ability to generalize across tasks (Kim et al., 2025). These challenges are particularly critical in the context of low-resource languages and dialects, where developing and maintaining large-scale models is both costly and impractical (Arga et al., 2025).

Pruning methods focus on removing redundant parameters or structural components from pretrained models. For instance, LayerDrop (Fan et al., 2019) demonstrated that up to 50% of Transformer layers can be removed with minimal impact on performance. Similarly, CoFi (Xia et al., 2022) proposed a structured pruning approach that jointly removes coarse-grained and fine-grained components, achieving significant speedups while maintaining accuracy.

Knowledge distillation (Hinton et al., 2015) offers an alternative strategy by training a smaller student model to mimic the behavior of a larger teacher model. DistilBERT (Sanh et al., 2020) reduces BERT's size by 40% while retaining most of its performance, whereas TinyBERT (Jiao et al., 2020) further improves compression by distilling intermediate representations and attention mechanisms. More recently, DynaBERT (Hou et al., 2020) introduced a dynamic distillation framework capable of generating multiple sub-networks with varying width and depth, enabling flexible trade-offs between efficiency and performance at inference time.

In parallel, several efforts have focused on developing language models for Arabic dialects, particularly Maghrebi Arabic, which includes the dialects spoken in Algeria, Morocco, and Tunisia (Abdaoui et al., 2021; Messaoudi et al., 2021; Abdul-Mageed et al., 2021; Moussaoui and El Younoussi, 2023; Gaanoun et al., 2024). These dialects present significant challenges for NLP due to their divergence from Modern Standard Arabic (MSA) in vocabulary, morphology, and syntax, as well as their strong influence from Berber, French, and Spanish (Harrat et al., 2018). Additionally, Maghrebi Arabic exhibits frequent code-switching between Arabic and foreign languages, often written in both Arabic and Latin scripts (Arabizi), which further complicates modeling (Hamed et al., 2025). The lack of standardized orthography, combined with limited annotated data and high intra-dialectal variability, makes the development of robust models particularly challenging (Abidi and Smaïli, 2021).

To address these challenges, several dialect-specific BERT models have been proposed. DziriBERT (Abdaoui et al., 2021) was the first model pretrained exclusively on the Algerian dialect, achieving strong performance despite being trained on a relatively small corpus. This was followed by DarijaBERT (Gaanoun et al., 2024) for Moroccan Arabic and TunBERT (Messaoudi et al., 2021) for Tunisian Arabic. However, these models remain relatively large, with parameter counts ranging from 109M to 147M, limiting their applicabil-

ity in resource-constrained environments. More recently, TinyDziriBERT (Laggoun et al., 2025) demonstrated that knowledge distillation can effectively reduce model size while maintaining competitive performance.

Despite these advances, the application of model compression techniques to Maghrebi Arabic dialects remains largely underexplored. This gap is particularly important, as frugal approaches could enable the development of models that achieve performance comparable to existing benchmarks while significantly reducing computational and deployment costs.

3 Models for under-resourced languages

Large Language Models (LLMs) achieve remarkable performance across a wide range of natural language processing tasks. However, training and deploying such models require substantial computational resources, costly hardware infrastructures, and massive amounts of data. Their development also involves extensive training times and significant energy consumption. These constraints raise important economic and environmental concerns, while also limiting the accessibility of state-of-the-art models in resource-constrained settings. This challenge is particularly acute for Maghrebi Arabic dialects, where language resources remain limited and research efforts are still at an early stage. In such contexts, frugal AI offers a promising direction by aiming to design models that are both efficient and competitive while requiring fewer computational resources. To address this issue, several model compression techniques have been explored for Maghrebi Arabic dialects, notably pruning and knowledge distillation. In this work, we investigate three complementary architectures. The first, *MagBERT*, relies on structured pruning by reducing the number of Transformer layers while preserving the original architecture. The second, *TinyDziriBERT* (Laggoun et al., 2025), is a compact student model trained using conventional knowledge distillation, where the student learns from the soft predictions generated by a larger teacher model. The third, *DynaBERT* (Hou et al., 2020), adopts a more flexible distillation strategy by producing dynamic sub-networks with different width and depth configurations. A key distinction lies in the training paradigm. Both MagBERT and TinyDziriBERT require an initial pre-training phase on large unlabeled corpora, followed by task-specific fine-tuning. In contrast, DynaBERT operates directly in a downstream setting: starting from an already pre-trained teacher, it performs distillation during fine-tuning using only the labeled data of the target task. This property makes DynaBERT particularly attractive for low-resource scenarios, where computational efficiency and reduced training costs are essential.

3.1 MagBERT: Maghrebi BERT

MagBERT is a lightweight variant of BERT specifically developed for Maghrebi Arabic dialects, including Moroccan, Algerian, and Tunisian Arabic. Compared with BERT-base, it adopts a significantly smaller architecture with six encoder layers, a hidden size of 384, and six attention heads. This reduced configuration makes MagBERT more suitable for resource-constrained environments while maintaining strong performance on dialectal Arabic tasks. Table 1 summarizes the architectural differences between MagBERT and BERT-base.

Model	Layers	H.Size	Heads
BERT-base	12	768	12
MagBERT	6	384	6

Table 1: Comparison between BERT-base and MagBERT architectures.

MagBERT was first pre-trained on an unlabeled corpus introduced in (Abidi and Smaïli, 2021). This corpus contains sentences in Algerian, Moroccan, and Tunisian dialects, written in both Latin and Arabic scripts. Table 2 reports the token counts for each dialect in the pre-training dataset.

Dialect	Tokens (M)
Algerian	23
Moroccan	22
Tunisian	16
Total	**61**

Table 2: Dialectal composition of the pretraining dataset.

Overall, MagBERT was pre-trained on a relatively modest corpus of 61 million words compared to contemporary large-scale language models. The model employs a WordPiece vocabulary of 70,000 tokens derived directly from the pre-training data. Training follows the Masked Language Modeling (MLM) objective, where 15% of tokens in each sequence are randomly masked. Pre-training was conducted for 15 epochs with a batch size of 16, constrained by the memory limitations of an NVIDIA Tesla V100-PCIE GPU (16 GB). The model was

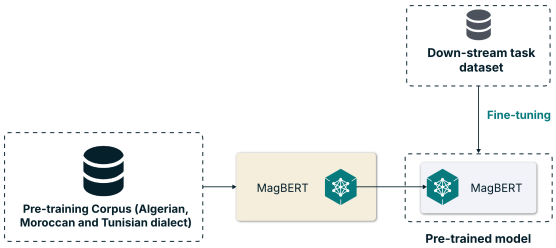


Figure 1: Pre-training and fine-tuning of MagBERT.

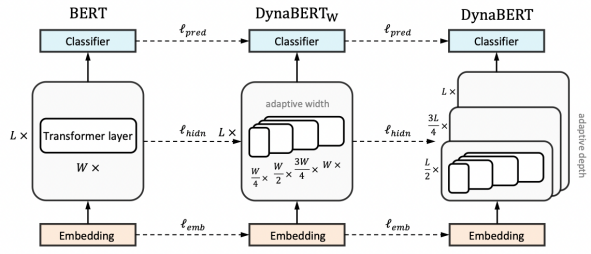


Figure 2: A two-stage procedure to train DynaBERT (Hou et al., 2020).

subsequently fine-tuned on downstream tasks (see figure 1). The design of MagBERT is motivated by the need for an efficient language model tailored to Maghrebi Arabic in low-resource settings. By halving the number of Transformer layers (from 12 to 6), reducing the hidden size (from 768 to 384), and decreasing the number of attention heads (from 12 to 6), the model significantly reduces parameter count and computational cost. This compact design, combined with training on a relatively small dialectal corpus, reflects realistic constraints in under-resourced language settings where large-scale data are unavailable. As a result, MagBERT provides a practical framework for studying the trade-off between model efficiency and downstream task performance in dialectal Arabic processing.

The model was optimized using Adam with a learning rate of 5×10^{-5}, following a linear decay schedule with 10% warm-up, a maximum sequence length of 128 tokens, and gradient accumulation to compensate for the limited batch size. Training used a fixed random seed (42) for reproducibility.

3.2 DynaBERT: DynamicBERT

DynaBERT (Hou et al., 2020) is a distillation framework designed to generate multiple compact sub-models from a single pre-trained BERT model. Its core idea is that one pre-trained and fine-tuned model can produce a family of dynamically adjustable sub-networks. To achieve this, DynaBERT introduces two adaptive dimensions: width and depth. The width corresponds to the number of attention heads and the size of the feed-forward layers, while the depth refers to the number of Transformer layers (i.e., encoder layers in BERT).

As shown in figure 2, the distillation process is carried out in two stages. First, knowledge is transferred from a teacher BERT model to an intermediate model, denoted as DynaBERT$_w$, which supports adaptive width. In this stage, the number of attention heads and feed-forward neurons can be scaled to 25%, 50%, or 75% of the original model size, or kept unchanged. The knowledge transfer leverages the teacher's predictions, embeddings, and hidden representations.

After this stage, multiple width-adaptive configurations are obtained. DynaBERT$_w$ then serves as a teacher assistant for the second stage, where depth adaptability is introduced. Specifically, the number of encoder layers can be reduced to 50% or 75% of the teacher assistant's depth. The final fine-tuning process produces a total of 12 sub-networks with different width–depth trade-offs.

In this work, DziriBERT, DarijaBERT, and TunBERT are used as teacher models. Table 3 summarizes the dialect associated with each teacher model within the DynaBERT framework. The motivation behind this setup is to investigate whether dynamic compression can effectively capture dialect-specific characteristics while leveraging shared representations across sub-networks. Although this approach has shown promising results in English task-specific settings, its effectiveness for dialectal Arabic remains underexplored. This is particularly relevant for Maghrebi dialects, where resources are limited and linguistic variability is high.

Teacher Model	Dialect
DziriBERT	Algerian Arabic
DarijaBERT	Moroccan Arabic
TunBERT	Tunisian Arabic

Table 3: Teacher models and their associated dialects in the DynaBERT framework

3.3 TinyDziriBERT

TinyDziriBERT is trained via knowledge distillation using DziriBERT (Abdaoui et al., 2021) as the teacher model. DziriBERT is a BERT-based pre-trained language model specifically developed for the Algerian dialect, trained on over one million tweets in both Arabic and Latin scripts.

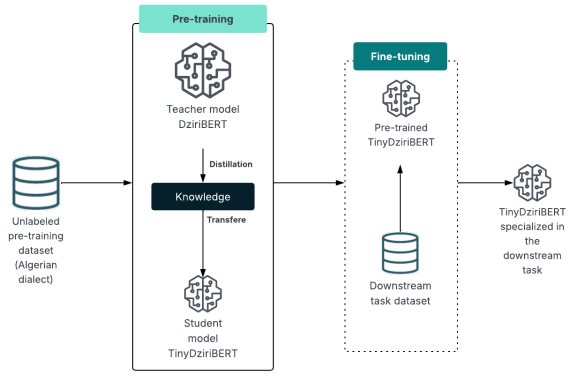


Figure 3: Illustration of the Pre-training and fine-tuning of TinyDziriBERT.

While DziriBERT follows the standard BERT-base architecture, TinyDziriBERT, used as the student model, is significantly smaller in size. Table 4 provides a detailed comparison of their respective architectures. Figure 3 illustrates the process used to obtain a specialized version of the TinyDziriBERT model.

Model	Layers	H.Size	Heads
DziriBERT	12	768	12
TinyDziriBERT	3	256	4

Table 4: Comparison of Teacher and Student Model Architectures.

Both DziriBERT and TinyDziriBERT share the same vocabulary of 50,000 tokens, which corresponds to the teacher's tokenizer.

Since the upper layers of the teacher model capture richer syntactic, semantic, and contextual information, the student model is initialized using a bottom-up layer mapping strategy. Specifically, if the teacher has $N \times k$ layers and the student has Nlayers, the i-th layer of the student is initialized from the $(i \times k)$-th layer of the teacher.

The training process utilizes Masked Language Modeling (MLM), where a 15% of tokens in each input sentence from the dataset are randomly masked. The objective is to predict these masked tokens. The sentences with masked tokens are then processed by both the teacher model and the student model to generate logits.

The student model is then optimized using a hybrid loss function combining Cross-Entropy (CE) loss with respect to the ground-truth labels and Kullback–Leibler (KL) divergence between the teacher and student logits. The total loss is defined as follows:

The total loss is calculated as:

$$\text{Loss} = \alpha \times T^2 \, \text{KL}\left(\frac{\hat{y}_{\text{student}}}{T} \,\middle\|\, \frac{\hat{y}_{\text{teacher}}}{T}\right) + (1-\alpha)\, \text{CE}(y_{\text{student}}, y_{\text{label}}) \quad (1)$$

Where α balances the contribution of the distillation loss and the MLM loss. $\hat{y}_{\text{student}}$ and $\hat{y}_{\text{teacher}}$ are the logits generated by the student and teacher models, respectively, and y_{labels} represents the true tokens. The logits are softened using a temperature parameter T. The KL divergence is scaled by T^2.

During distillation, we used a temperature T = 2 and a distillation coefficient α = 0.5 These values were selected empirically following common practice in BERT distillation.

Table 5 summarizes and compares the models in terms of number of parameters and supported dialects. For DynaBERT, which produces 12 sub-networks across different width and depth configurations, we report the best-performing configuration in terms of the size–performance trade-off. This optimal configuration retains all Transformer layers while reducing the model width to 75%. The sizes of all sub-network configurations are reported in Appendix A.

4 Results and discussions

This section presents the performance of the frugal models: MagBERT, DynaBERT (the DynaBERT sub-network used for comparison in this section has 75% of the original width and the same depth), and TinyDziriBERT, when fine-tuned on several downstream tasks relevant to Maghrebi Arabic dialects. These tasks include dialect and language identification on the MADAR (Multilingual Arabic Dialect Applications and Resources) dataset (Bouamor et al., 2018) and the Boutef (Bolstering Our Understanding Through an Elaborated Fake News Corpus) dataset (Smaïli et al., 2024), sentiment analysis on the TSAC (Tunisian Sentiment Analysis Corpus) dataset (Medhaffar et al., 2017), and topic classification on the MTCD (Maghrebi Topic Classification Dataset) (Gaanoun et al., 2024).

For each task, the models are compared against the top-performing baseline models specific to each dataset and dialect setting.

4.1 Experimental Setup

For all downstream fine-tuning experiments, datasets were split 80%/20% for training and testing. Fine-tuning was conducted for 4 epochs (except NER, 5 epochs, given its task complexity

Model	#P (M)	Scripts	Dialect	U	T
DarijaBERT (Gaanoun et al., 2024)	147	A	Moroccan	✓	✓
DziriBERT (Abdaoui et al., 2021)	124	A+L	Algerian	✓	✓
TunBERT (Messaoudi et al., 2021)	109	A+L	Tunisian	✓	✓
MagBERT	41	A+L	Algerian, Moroccan, Tunisian	✓	✓
TinyDziriBERT (Laggoun et al., 2025)	18	A+L	Algerian	✓	✓
DynaMoroccanBERT	94	A+L	Moroccan	×	✓
DynaAlgerianBERT	77	A+L	Algerian	×	✓
DynaTunisianBERT	66	A+L	Tunisian	×	✓

Table 5: Comparison of Maghrebi Arabic dialect models in terms of number of parameters (#P), vocabulary scripts, supported dialects, and training data requirements. Scripts refer to the writing systems covered in the model's vocabulary: A (Arabic) and L (Latin). U indicates whether the model uses unlabeled data for pre-training, and T indicates whether task-specific data is used for fine-tuning. DynaXBERT denotes a DynaBERT model trained with a dialect X teacher.

and smaller dataset size), using a learning rate of 2×10^{-5} and a batch size of 16, with early stopping based on validation loss.

4.2 Dialect and Language Identification

In this section, we evaluate our models on the language identification task using two distinct datasets.

4.2.1 Language identification on MADAR

The MADAR corpus is a parallel dataset (Bouamor et al., 2018) that contains sentences in 25 Arabic dialects, in addition to Modern Standard Arabic, English, and French. In this study, the focus is only on Algerian, Tunisian, and Moroccan dialects, considering only the Arabic script due to the specific models developed in this study. Due to the limited amount of data available for each sub-dialect, regional variants are grouped together, such as Tunis and Sfax for Tunisian, and Rabat and Fès for Moroccan. The corpus is, however, unbalanced, as shown in Table 6. To assess the effect of data distribution, the models are fine-tuned on both the original unbalanced MADAR dataset and a balanced version (MADAR-bal), in which the dataset size is fixed equally across all classes.

Dialect	MADAR	MADAR-bal
Algerian	2,000	2,000
Tunisian	14,000	2,000
Moroccan	14,000	2,000

Table 6: Distribution of the number of sentences per dialect in the unbalanced and balanced versions of the filtered MADAR corpus.

The results on the MADAR corpus, presented in Table 7, show a consistent trend across all models: performance is generally higher on the unbalanced version of the dataset compared to the balanced one, both in terms of accuracy and F1-score. This can be attributed to the fact that the unbalanced setting preserves the original data distribution, which may implicitly favor dominant dialects present in the training data. This gap suggests that class imbalance in MADAR may inflate accuracy by favoring the dominant Tunisian and Moroccan classes; the balanced setting (MADAR-bal) is therefore a more reliable indicator of true cross-dialect performance, and we recommend it as the primary reference for future comparisons.

Metric	DzB	MagB	TDB	DyAlgB
Factor	–	3	7	2
Acc-Unbalanced	93.16	**93.73**	90.00	93.70
F1-Unbalanced	93.11	88.79	84.80	**93.65**
Acc-Balanced	90.32	87.50	85.10	**90.83**
F1-Balanced	90.39	87.54	84.01	**90.84**

Table 7: Results on the MADAR corpus for both unbalanced and balanced versions. Factor indicates the relative model size with respect to DziriBERT, i.e., how much smaller each model is compared to DziriBERT. DzB: DziriBert, MagB: MagBERT, TDB: TinyDziriBert, DyAlgB: DynaAlgerianBERT

An exception is observed for the smaller models, particularly TinyDziriBERT, which exhibits a noticeable drop in performance, especially on the unbalanced setting, where it significantly underperforms compared to larger architectures such as DziriBERT and DynaBERT. This highlights that highly compact, distillation-based models are particularly sensitive to uneven data distributions.

Among all models, DynaBERT achieves the most competitive and overall best results on both balanced and unbalanced MADAR dataset. Despite reducing the number of attention heads and feed-forward neurons to 75% and having half the size of DziriBERT, it consistently matches or surpasses its performance. This suggests that task-specific dy-

namic distillation is particularly effective for dialect identification.

The full results of all DynaBERT configurations are provided in Appendix B (Tables 15 and 16).

In contrast, MagBERT also performs strongly, outperforming TinyDziriBERT. This can be explained by the fact that TinyDziriBERT relies solely on logit-level distillation from DziriBERT, whereas MagBERT benefits from a more expressive architecture with multiple Transformer layers trained on unlabeled pre-training data.

4.2.2 Language identification on BOUTEF

The tests on this corpus concern BOUTEF which is a datset of fake news collected from social networks. As shown in Table 8, BOUTEF corpus contains sentences in Modern Standard Arabic (MSA), English, French, two Maghrebi dialects (Algerian and Tunisian), as well as code-switching data. Each sentence is labeled according to its corresponding language or dialect.

Table 8: Distribution of comments by language or dialect in the used subset of BOUTEF.

Language / Dialect	Number of comments
Algerian Dialect	1078
French	862
MSA	784
Tunisian Dialect	554
Code-switching	421
English	104

The results on BOUTEF dataset, presented in Table 9, further confirm the effectiveness of DynaBERT in resource-constrained settings.

Model	Factor	Acc	F1
DziriBERT	–	84.16	83.04
MagBERT	3	83.05	81.84
TinyDziriBERT	7	79.21	78.80
DynaAlgerianBERT	2	**84.48**	**83.87**

Table 9: Performance of models on the Boutef dataset. Factor indicates the relative model size with respect to DziriBERT, i.e., how much smaller each model is compared to DziriBERT.

Despite using only 75% of the model width, DynaAlgerianBERT outperforms slightly DziriBERT, which is approximately twice its size, achieving the best performance in terms of both accuracy (84.48%) and F1-score(83.87%). This indicates that task-specific dynamic distillation is able to preserve, and even enhance, discriminative information in multilingual and code-switched contexts, results of the remaining configurations of DynaBERT are provided in Table 17.

In contrast, MagBERT shows performance relatively close to that of DziriBERT despite being three times smaller. This relatively strong performance can be explained by its training strategy, which includes pre-training on the three target dialects (Algerian, Moroccan, and Tunisian). These dialects are known to naturally incorporate lexical borrowing from other languages, particularly French and English, due to historical contact and sociolinguistic influence. Moreover, BOUTEF corpus is extracted from social media data, where users freely mix languages and writing styles without strict linguistic constraints. As a result, MagBERT benefits from exposure to this heterogeneous linguistic environment during pre-training, which improves its robustness in code-switching scenarios.

4.3 Sentiment analysis

Regarding the sentiment analysis task, evaluation is done on TSAC dataset which is a corpus of 17,000 comments in the Tunisian dialect, evenly split between positive and negative sentiment labels. In this case, we used the appropriate LLM TunBERT which is dedicated to Tunisian dialect. DynaBERT was generated from TunBERT to achieve DynaTunisianBERT

Model	Factor	Acc	F1
TunBERT	–	82.67	82.67
MagBERT	3	**96.68**	**96.68**
TinyDziriBERT	6	94.52	94.51
DynaTunisian BERT	2	86.76	86.76

Table 10: Performance comparison of models on the TSAC dataset. Factor represents the relative size compared to TunBERT, i.e., how much smaller the models are than TunBERT.

The results on the TSAC dataset, presented in Table 10, show that MagBERT achieves the best performance, with an accuracy and F1-score of 96.68%, significantly outperforming all other models. This strong performance can be attributed to its pre-training on multiple Maghrebi dialects, which likely improves its ability to generalize to Tunisian sentiment analysis. In contrast, TunBERT, despite being specifically designed for the Tunisian dialect, achieves substantially lower performance. This suggests that the quality and scale of pre-training data play an important role in downstream performance. This also directly impacts the performance of DynaTunisianBERT (more results in 18), which relies

on TunBERT as a teacher model. The relatively low results obtained by DynaBERT highlight a key limitation of knowledge distillation approaches: their strong dependence on the teacher model. When the teacher is suboptimal, the student model inherits these limitations, leading to degraded performance.

Interestingly, TinyDziriBERT achieves competitive results (94.52%), outperforming both TunBERT and DynaBERT. This indicates that cross-dialect transfer is possible within Maghrebi Arabic, even when the student model is trained from a teacher specialized in a different dialect (Algerian). This can be explained by the linguistic proximity between Maghrebi dialects, as well as the availability of a high-quality task-specific dataset such as TSAC, which helps the model adapt effectively during fine-tuning

4.4 Topic classification

In this identification task, the objective is to detect topics expressed in Moroccan dialect and classify them into one of the following categories: Gaming, Kitchen, Sport, and General. The data are sourced from the MTCD corpus. DynaMoroccanBERT is derived from DarijaBERT, a model specifically designed for the Moroccan dialect. The results presented in Table 11, show that DarijaBERT achieves the best performance, with an accuracy and F1-score of 91.99%. DynaMoroccanBERT achieves very competitive results, closely matching DarijaBERT while being approximately twice smaller. This again highlights the effectiveness of task-specific distillation, showing that a well-designed student model can retain most of the teacher's performance even under significant compression. The small performance gap suggests that DynaBERT successfully captures the key task-relevant representations despite reduced size. Results of the other sub-networks of DynaMoroccanBERT are showed in Table 19.

MagBERT also performs well, with results close to the best model, despite being four times smaller than DarijaBERT. Its strong performance can be attributed to its exposure to multiple Maghrebi dialects during pre-training, which likely enhances its ability to generalize across dialectal variations. However, the slight drop compared to DarijaBERT indicates that dialect-specific specialization still provides an advantage for this task.

In contrast, TinyDziriBERT shows a more noticeable decrease in performance. This can be explained by its smaller size and its reliance on a teacher model trained on a different dialect (Algerian), which may limit its ability to fully capture Moroccan-specific thematic patterns.

Model	Factor	Acc	F1
DarijaBERT	–	**91.99**	**91.99**
MagBERT	4	90.77	90.98
TinyDziriBERT	8	87.69	87.68
DynaMoroccanBERT	2	91.82	91.83

Table 11: Performance comparison of models on the MTCD dataset. Factor represents the relative size compared to DarijBERT, i.e., how much smaller the models are than DarijaBERT.

4.5 Error Analysis

To complement the quantitative results, we manually inspected a sample of misclassified instances across tasks. On BOUTEF, most errors involve code-switched sentences where French or English segments dominate the sentence length, causing smaller models (TinyDziriBERT, DynaAlgerianBERT) to default to the majority class. On MADAR-unbalanced, TinyDziriBERT frequently confuses Tunisian and Moroccan sentences sharing common MSA-derived vocabulary, reflecting its narrower Algerian-only pretraining. On TSAC, MagBERT's errors concentrate on short, ambiguous comments (e.g., single-word replies) lacking sufficient context, a limitation shared across all models.

5 Conclusion

This work investigates several frugal AI approaches for Maghrebi Arabic dialects, including pruning and both classical and dynamic knowledge distillation. A set of lightweight and dialect-specific models is introduced, namely MagBERT, TinyDziriBERT, DynaMoroccanBERT, DynaAlgerianBERT, and DynaTunisianBERT. The study analyzes the trade-offs between model size, architectural design, and downstream task performance across multiple benchmarks.

Experimental results demonstrate that dynamic, task-specific distillation consistently achieves the best balance between efficiency and performance. In several cases, it even surpasses larger baseline models while significantly reducing computational cost, without relying on unlabeled data. In particular, DynaBERT configurations with 75% width and full depth consistently provide the most effective size–performance trade-off across most tasks. This confirms that moderate width reduction has

a limited impact on performance, whereas depth reduction is considerably more detrimental.

Beyond architectural aspects, the findings also highlight the importance of pre-training diversity and task-specific adaptation. MagBERT, despite being a layer-pruned model pre-trained on multilingual Maghrebi data, demonstrates strong generalization capabilities, especially on code-switching and cross-dialectal transfer tasks such as BOUTEF and TSAC. This highlights the value of multilingual pre-training even within a closely related language family.

TinyDziriBERT, while the most compact model evaluated, achieves competitive results on several tasks, particularly when cross-dialect transfer within the Maghrebi language family is possible. However, its performance is more sensitive to data imbalance and task complexity, reflecting the limitations of output-level distillation under strong architectural constraints.

Overall, this study shows that frugal NLP approaches are not limited to high-resource languages such as English. Instead, they can be effectively extended to low-resource settings such as Maghrebi dialects, offering promising directions for deploying efficient language models in resource-constrained environments.

Human evaluation was not conducted because the benchmark datasets already provide expert annotations. Establishing human performance on these datasets remains an interesting direction for future work. We report single-run results without variance estimates across seeds; statistical significance testing (e.g., bootstrap resampling) is left for future work.

References

Amine Abdaoui, Mohamed Berrimi, Mourad Oussalah, and Abdelouahab Moussaoui. 2021. Dziribert: a pre-trained language model for the algerian dialect. *CoRR*, abs/2109.12346.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Karima Abidi and Kamel Smaïli. 2021. Creating multiscripts sentiment analysis lexicons for algerian, moroccan and tunisian dialects. In *Proceedings of the 7th International Conference on Data Mining (DTMN 2021)*, Conference Proceedings in Computer Science & Information Technology (CS & IT), Sep 2021, Copenhagen, Denmark, pages 147–156.

Ludovic Arga, François Bélorgey, Arnaud Braud, Romain Carbou, Nathalie Charbonniaud, Catherine Colomes, Lionel Delphin-Poulat, David Excoffier, Christel Fauché, Thomas George, Frédéric Guyard, Thomas Hassan, Quentin Lampin, Vincent Lemaire, Pierre Nodet, Pawel Piotrowski, Krzysztof Sapiejewski, Emilie Sirvent-Hien, and Tamara Tosic. 2025. Frugal ai: Introduction, concepts, development and open questions. *SIGKDD Explor. Newsl.*, 27(1):72–111.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Angela Fan, Edouard Grave, and Armand Joulin. 2019. Reducing transformer depth on demand with structured dropout. *CoRR*, abs/1909.11556.

Kamel Gaanoun, Abdou Mohamed Naira, Anass Allak, and Imade Benelallam. 2024. DarijaBERT: a step forward in NLP for the written Moroccan dialect. *International Journal of Data Science and Analytics*, 9(1):23–40.

Injy Hamed, Caroline Sabty, Slim Abdennadher, Ngoc Thang Vu, Thamar Solorio, and Nizar Habash. 2025. A survey of code-switched arabic nlp: Progress, challenges, and future directions.

Salima Harrat, Karima Meftouh, and Kamel Smaïli. 2018. Maghrebi Arabic dialect processing: an overview. *Journal of International Science and General Applications*, 1.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network.

Lu Hou, Zhiqi Huang, Lifeng Shang, Xin Jiang, Xiao Chen, and Qun Liu. 2020. Dynabert: Dynamic bert with adaptive width and depth.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. Tinybert: Distilling bert for natural language understanding.

Gun Il Kim, Sunga Hwang, and Beakcheol Jang. 2025. Efficient compressing and tuning methods for large language models: A systematic literature review. *ACM Comput. Surv.*, 57(10).

Amina Laggoun, Chahnez Zakaria, and Kamel Smaïli. 2025. Knowledge distillation for efficient algerian dialect processing: Training compact BERT models with DziriBERT. In *7th International Conference on Advances in Signal Processing and Artificial Intelligence (ASPAI)*, Innsbruck, Austria.

Salima Medhaffar, Fethi Bougares, Yannick Estève, and Lamia Hadrich-Belguith. 2017. Sentiment analysis of Tunisian dialects: Linguistic ressources and experiments. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 55–61, Valencia, Spain. Association for Computational Linguistics.

Abir Messaoudi, Ahmed Cheikhrouhou, Hatem Haddad, Nourchene Ferchichi, Moez BenHajhmida, Abir Korched, Malek Naski, Faten Ghriss, and Amine Kerkeni. 2021. Tunbert: Pretrained contextualized text representation for tunisian dialect.

Otman Moussaoui and Yacine El Younoussi. 2023. Pretraining two bert-like models for moroccan dialect: Morroberta and morrbert. *MENDEL*, 29:55–61.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Kamel Smaïli, Anissa Hamza-Jamann, Langlois David, and Amazouz Djegdjiga. 2024. BOUTEF: Bolstering Our Understanding Through an Elaborated Fake News Corpus. In *The 8th International Conference on Arabic Language Processing*, RABAT, Morocco.

Mengzhou Xia, Zexuan Zhong, and Danqi Chen. 2022. Structured pruning learns compact and accurate models.

Appendix A Sub-network sizes of DynaBERT

This appendix reports the size of the DynaBERT sub-network configurations, in terms of the number of parameters, for the three teacher models. The model sizes vary depending on the width and depth multipliers. For example, in the case of DziriBERT, the number of parameters ranges from 12M, corresponding to 25% of the width and 50% of the depth, to 124M for the full model configuration. Here, m_w and m_d denote the width and depth multipliers, respectively.

m_w	m_d	**#P (M)**
0.25	0.50	12
	0.75	14
	1.00	15
0.50	0.50	30
	0.75	36
	1.00	41
0.75	0.50	53
	0.75	65
	1.00	77
1.00	0.50	82
	0.75	103
	1.00	124

Table 12: Sub-network configurations of DynaAlgerianBERT.

m_w	m_d	**#P (M)**
0.25	0.50	18
	0.75	19
	1.00	21
0.50	0.50	42
	0.75	47
	1.00	52
0.75	0.50	71
	0.75	83
	1.00	95
1.00	0.50	105
	0.75	126
	1.00	147

Table 13: Sub-network configurations of DynaMoroccanBERT.

m_w	m_d	#P (M)
0.25	0.50	9
	0.75	10
	1.00	11
0.50	0.50	23
	0.75	28
	1.00	33
0.75	0.50	42
	0.75	54
	1.00	66
1.00	0.50	67
	0.75	88
	1.00	109

Table 14: Sub-network configurations of DynaTunisianBERT.

Appendix B DynaBERT Results

This appendix reports the full evaluation results of all DynaBERT sub-network configurations on the fine-tuning datasets : MADAR, Boutef, TSAC and MTCD.

B.1 DynaBERT results on MADAR

Width	Depth	Acc (%)	F1 (%)
0.25	0.5	87.99	87.98
0.50	0.5	88.41	88.42
0.75	0.5	88.74	88.75
1.00	0.5	89.16	89.17
0.25	0.75	89.07	89.10
0.50	0.75	89.57	89.61
0.75	0.75	90.33	90.37
1.00	0.75	90.16	90.19
0.25	1.0	89.74	89.76
0.50	1.0	90.41	90.43
0.75	1.0	90.83	90.84
1.00	1.0	**91.16**	**91.15**

Table 15: DynaAlgerianBERT ealuation results on the MADAR-Balanced dataset (DziriBERT is the teacher model).

Width	Depth	Acc (%)	F1 (%)
0.25	0.5	93.05	92.99
0.50	0.5	93.20	93.09
0.75	0.5	93.28	93.17
1.00	0.5	93.17	93.07
0.25	0.75	92.98	92.87
0.50	0.75	93.18	93.09
0.75	0.75	93.58	93.44
1.00	0.75	93.65	93.63
0.25	1.0	93.00	92.89
0.50	1.0	93.52	93.42
0.75	1.0	93.70	93.65
1.00	1.0	**93.75**	**93.71**

Table 16: DynaTunisianBERT ealuation results on the MADAR-Unbalanced dataset.

B.2 DynaBERT results on Boutef

Width	Depth	Acc (%)	F1 (%)
0.25	0.5	82.83	81.76
0.50	0.5	83.73	82.80
0.75	0.5	84.17	83.37
1.00	0.5	84.21	83.62
0.25	0.75	83.23	82.34
0.50	0.75	84.08	83.42
0.75	0.75	84.46	83.93
1.00	0.75	84.21	83.59
0.25	1.0	83.35	82.25
0.50	1.0	84.28	83.57
0.75	1.0	84.48	**8**3.87
1.00	1.0	**84.73**	**83.94**

Table 17: DynaAlgerianBERT evaluation results on the Boutef dataset.

B.3 DynaBERT results on TSAC

Width	Depth	Acc (%)	F1 (%)
0.25	0.5	85.82	85.82
0.50	0.5	86.38	86.38
0.75	0.5	86.41	86.41
1.00	0.5	86.50	86.50
0.25	0.75	86.47	86.47
0.50	0.75	86.38	86.37
0.75	0.75	86.38	86.38
1.00	0.75	**86.76**	**86.76**
0.25	1.0	86.56	86.56
0.50	1.0	86.56	86.56
0.75	1.0	86.56	86.56
1.00	1.0	86.68	86.68

Table 18: DynaTunisianBERT evaluation results on the TSAC dataset.

B.4 DynaBERT results on MTCD

Width	Depth	Acc	F1
0.25	0.5	90.85	90.86
0.50	0.5	91.30	91.31
0.75	0.5	91.52	91.52
1.00	0.5	91.63	91.63
0.25	0.75	91.20	91.20
0.50	0.75	91.50	91.50
0.75	0.75	91.64	91.64
1.00	0.75	91.82	91.83
0.25	1.0	91.24	91.24
0.50	1.0	91.67	91.67
0.75	1.0	**91.94**	**91.95**
1.00	1.0	91.92	91.92

Table 19: DynaMoroccanBERT evaluation results on the MTCD dataset.

The detailed results presented in Tables 15–19 provide a comprehensive view of the impact of width and depth scaling on DynaBERT performance across different tasks and datasets. Several consistent trends can be observed.

First, across all datasets, increasing both width and depth generally leads to improved performance, with the best results almost always obtained using the full configuration (width = 1.0, depth = 1.0). However, the performance gains tend to saturate quickly, and smaller configurations remain highly competitive. In particular, configurations with 75% width and full or near-full depth (e.g., 0.75/1.0 or 0.75/0.75) consistently achieve results very close to the full model, while offering a significant reduction in model size. This confirms that a substantial portion of the model redundancy lies in width rather than depth.

Second, depth appears to have a more critical impact on performance than width. Reducing depth (e.g., from 1.0 to 0.5) leads to a more noticeable degradation compared to reducing width. This trend is consistent across MADAR, Boutef, TSAC, and MTCD, suggesting that Transformer layers encode essential hierarchical and contextual representations that are difficult to compress without loss. This observation supports the earlier findings in the main text, where models relying on depth reduction (e.g., MagBERT) showed lower performance compared to DynaBERT configurations that preserve deeper structures.

Third, the effect of compression varies depending on the task. On MADAR and Boutef, which involve dialect and language identification with potentially higher variability, performance differences between configurations are more pronounced. In contrast, on TSAC and MTCD, the performance gap between configurations is relatively small, indicating that these tasks may rely more on surface-level or task-specific features that can be captured even by smaller sub-networks.

The results also reflect the influence of the teacher model. For example, DynaTunisianBERT (TSAC) shows limited variation across configurations, which aligns with the relatively lower performance of its teacher (TunBERT). In contrast, DynaTunisianBERT and DynaMoroccanBERT based on stronger teacher models, exhibit both higher performance and more consistent scaling behavior. This further reinforces the dependency of distillation-based methods on the quality of the teacher model.

What to take as the basic verb form? A pilot study in Czech

Ján Mačutek[1,2], Michaela Koščová[1], Michaela Nogolová[3], Radek Čech[4]

[1]Mathematical Institute, Slovak Academy of Sciences, Bratislava, Slovakia
[2]Department of Mathematics, Constantine the Philosopher University in Nitra, Slovakia
[3]Department of Czech Language, University of Ostrava, Ostrava, Czech Republic
[4]Department of Czech Language, Masaryk University, Brno, Czech Republic

jan.macutek@mat.savba.sk, matemiko@savba.sk,
michaela.nogolova@osu.cz, cechradek@phil.muni.cz

Abstract

This paper investigates which Czech verb form can serve as the most suitable for analysing the relationship between the frequency of individual verb forms and the number of formal changes required to derive them. The number of changes is measured by the Levenshtein distance. The results show that the present indicative third person singular occupies the central position among all verb forms. Furthermore, verbs exhibit an inverse relationship between the number of formal changes with respect to the basic form and form frequency. This relation can be interpreted as a consequence of the principle of least effort.

Keywords: principle of least effort, verb forms, conjugation, Levenshtein distance, Czech.

1 Introduction

From the perspective of the principle of least effort (Zipf, 1949), it is plausible to assume that inflectional paradigms are not frequency-neutral systems. Language users tend to favour forms that minimise production and processing costs. In inflectional morphology, these costs may be reflected in the amount of formal changes required to derive a given word form from its basic form. Thus, word forms that are closer to the basic form may be expected to occur more frequently than forms involving a greater number of modifications.

Results presented in a series of recent papers on noun declension in Slavic languages (Kelih et al., 2023; Mačutek et al., 2023, 2025) are consistent with this assumption. More specifically, noun word forms that are more similar to the lemma occur more often. The Levenshtein distance (see Section 2 for its definition) is used as a measure of similarity between word forms and their lemmas.

This pilot study conducted on Czech language material extends the analysis to verbs. However, an important decision must be made at the very beginning. While the choice of the basic form is relatively obvious for nouns (it is nominative singular), the situation is more complicated with verbs. In the verbal domain, several forms may in principle be considered candidates for the basic form.

One natural candidate is the infinitive. In many grammatical traditions, the infinitive functions as the citation form of the verb, i.e. the form under which verbs are listed in dictionaries and grammars. At the same time, it is morphologically less specified than finite forms. This relative neutrality makes it a plausible candidate for the basic form of the verbal paradigm. From a methodological point of view, the infinitive is also a convenient reference form, since in corpus linguistic annotation it often corresponds to the lemma. On the other hand, as there are languages without the infinitive (e.g., Bulgarian or Modern Greek), it cannot serve as the universal basic verb form.

Another candidate is the present indicative first person singular. The basic reference form of a verb need not always be an infinitive. In the Latin tradition, for example, the first person singular present indicative active is the first of the principal parts under which verbs are conventionally cited.

Yet another option is the present indicative third person singular. As noted by Bybee (2010) in her discussion of markedness, Jakobson (1990) treats the third person of Russian verb as the unmarked member of the person opposition. Specifically, the category of person is structured so that first and second person are treated as participants in the speech event, whereas third person is a non-participant. Accordingly, the feature "personal" is marked, while its absence is unmarked. From this perspective, the present indicative third person

singular may be regarded as a structurally neutral candidate for the basic verb form.

The aim of the present study is therefore to compare these three candidates – the infinitive, the present indicative first person singular, and the present indicative third person singular – and to determine which of them is the most suitable basic form for the analysis of Czech verbs. More specifically, we investigate which of these forms occupies the most central position in the verbal paradigm, in the sense that its mean Levenshtein distance from all other verb forms is minimal.

2 Language material and methodology

The language material consists of five dependency treebanks from Universal Dependencies 2.13 (Zeman et al., 2023), capturing distinctive domains:

1. FicTree (Jelínek, 2017) represents the literary register, this treebank contains eight texts published between 1991 and 2007 (six works of fiction, alongside one memoir and one children's book).

2. UD_Czech-CAC is based on the written component of the Czech Academic Corpus 2.0 (Vidová Hladká et al., 2008), it supplies high-density informational texts spanning the scientific, administrative, and journalistic domains.

3. UD_Czech-Poetry is derived from the Corpus of Czech Verse by Plecháč and Kolár (2015). This dataset introduces language governed by strict rhythmic and formal rules.

4. Czech-PDT UD (Bejček et al., 2013) contributes additional journalistic, business, and science-related texts originating from the 1990s.

5. Parallel Universal Dependencies (PUD, Czech part) was originally compiled for the CoNLL 2017 shared task on Multilingual Parsing; this collection features 1,000 sentences from news and Wikipedia sources. The dataset consists of translated texts: 750 sentences were translated directly from English, while the remaining 250 originate from French, German, Italian, and Spanish using English as an intermediary language.

Initially, each sentence from our dataset was represented as a dependency tree following the Universal Dependencies annotation. For the purposes of this study, we transformed these structures into the Surface Syntactic Universal Dependencies (SUD) format (Gerdes et al., 2018). We adopted this approach because the SUD scheme – by explicitly capturing the structural functions of auxiliaries and prepositions – more accurately aligns with the morphosyntactic structure of the Czech language.

Beyond mapping syntactic dependencies, the annotation framework supplies crucial word-level metadata, including lemmatization, morphological tagging, and part-of-speech (POS) identification. Furthermore, all selected treebanks contain manually annotated language specific extended part-of-speech tags (XPOS). Combined with Universal Features, which assign additional grammatical and lexical properties, this rich annotation allows us to reliably distinguish categories such as grammatical person, gender, tense, and other morphological properties.

Each sentence is represented as a dependency tree with a root. In most cases, this root is a finite verb or an auxiliary (functioning as the predicate), since dependency grammar is inherently verb-centric. As our methodology partly relies on this predicate-driven structure, we exclude sentences lacking an explicit verbal root, such as exclamations, fragments, or elliptical constructions.

The Czech verbal system contains both synthetic and analytic forms (Osolsobě, 2017).[1] Because dependency trees represent each individual word as a distinct node, we implemented targeted procedures to properly identify specific analytic forms. For example, past-tense forms in the first and second person require combining the past active participle (also called l-participle) with a present-tense form of the verb *to be*:

> *psala jsi*
> 'write-PTCP-SG.F be.PRS.2SG'.

Similarly, the future tense of imperfective verbs is formed analytically by combining the future form of the verb *to be* with the infinitive of the lexical verb:

> *budeme psát*
> 'be.FUT.1PL write.INF'.

Further, conditional forms require the past active participle (l-participle) together with a specific conditional form(s) of the auxiliary:

[1] All forms of verb *dělat* 'to do' are presented in the Appendix A.

napsali bychom
'write-PTCP-PL be.COND.1PL'.

Another example of an analytic form is the passive voice, which is formed by a passive participle and a form of *to be*:

je napsán
'be.PRS.3SG write.PASS.PTCP.PRS.3SG.M'.

However, because the passive participle shares behavioral traits with adjectives, the SUD annotation scheme tags these words as adjectives. Consequently, our analysis, which only processes words explicitly tagged as verbs, excludes these passive constructions.[2]

The annotation labels as verbs also some other parts of speech. For instance, adverbs *lze* ('it is possible (to)') and its negative form *nelze* ('it is not possible (to)') function syntactically in a verb-like manner and are annotated as verbs. Sentences with such non-verb predicates are excluded from the dataset prior to analysis.

If the infinitive is treated as the basic verb form, all verbs in the sample can be analysed, since the annotation provides the infinitive as the lemma for each verb. By contrast, when the present indicative first person singular or the present indicative third person singular is taken as the basic form, only those verbs for which the corresponding form is attested in the corpus can be included. Some verbs did not occur in the relevant form even once, and therefore the analysis based on these two candidate basic forms are carried out on less rich samples. Table 1 presents the numbers of verbs used in the analysis.

Some conjugated verb forms have two variants considered acceptable, e.g., one can say either *děkuji* or *děkuju* 'I thank (you)'.[3] If such a form is considered the basic form in an analysis, the more frequent variant is taken as the basic one.

In the annotation used here, affirmative and negative verb forms are traditionally lemmatized under the same basic form, namely the affirmative lemma. For the purposes of the present study, however, this would be problematic, since in Czech negation is usually expressed by the prefix *ne-*, which would increase the Levenshtein distance of all negative forms by 2. We therefore decided to treat affirmative and negative forms separately as two distinct lemmas (i.e, all basic forms in our study). Negative counterparts of the relevant basic forms were generated by adding the prefix *ne-*. The only irregular case is the verb *být* 'to be': instead of the expected form *neje*, Czech uses the form *není*, which was therefore used in the analysis. This is the only irregularity in the negation of verbs in Czech.

Finally, all words were phonologically transcribed following the rules presented in Pelegrinová (2023: 111–112).

The Levenshtein distance (Levenshtein, 1966) is the distance between two character strings (in our case, between two verb forms) measured by the minimum number of single-character deletions, additions, or substitutions necessary to change one string to another.[4] We exemplify it in Table 2 for several forms of verb *dělat* 'to do', with its present indicative third person singular *dělá* considered as the basic form.

3 Results

The distributions of the Levenshtein distance values for all three considered basic verb forms – the infinitive, the present indicative first person singular, and the present indicative third person singular – are presented in Tab. 3 and in Figs. 1–3, respectively.

We note that there are some verb forms for which the annotation used (c.f., Section 2) is not able to determine some grammatical categories; e.g., verb *získat* 'to get' has the same form *získala* for both the past indicative third person singular in the feminine gender[5] and for the past indicative third person plural in the neuter gender.[6] Such verb forms disregarded in Table 1 because the category of grammatical number cannot be determined from the verb form alone. On the other hand, these ambiguities do not present any problems when the Levenshtein distance is evaluated;[7] therefore they are included in computations resulting in values presented in Table 3.

[2]Considering passive voice forms as adjectives is the only influence of the annotation schema. Otherwise, we take into account verb forms as they appear in the corpora/treebanks.

[3]See the online handbook of Czech language provided by the Czech Language Institute of the Czech Academy of Sciences, https://prirucka.ujc.cas.cz/?slovo=d%C4%9Bkuju#bref3.

[4]It is known also as the edit distance. Deza and Deza (2016: 216) spell it Levenstein distance.

[5]*Moje dcera získala potřebný souhlas.* 'My daughter got the necessary approval.'

[6]*Všechna místa získala dobré hodnocení.* 'All the places got good reviews.'

[7]The Levenshtein distance depends on the verb form only. If a verb has the same form, e.g., for singular and plural, the distance is the same regardless of grammatical number.

	present		past		future		total	
	f.	**r.**	**f.**	**r.**	**f.**	**r.**	**f.**	**r.**
first person singular	4,342	0.05	2,805	0.06	122	0.03	7,269	0.05
second person singular	919	0.01	87	<0.01	37	0.01	1,043	0.01
third person singular	59,885	0.64	32,727	0.67	2,979	0.65	95,591	0.65
first person plural	5,731	0.06	1,581	0.03	284	0.06	7,596	0.05
second person plural	1,898	0.02	251	0.01	47	0.01	2,196	0.01
third person plural	20,376	0.22	11,155	0.23	1,130	0.25	32,661	0.22
participle singular	171	<0.01	4	<0.01			175	<0.01
participle plural	22	<0.01					22	<0.01
total	93,344	1.00	48,610	1.00	4,599	1.00	146,553	1.00
infinitive							26,198	

Table 1: Frequencies (f.) and relative frequencies (r.) of verb forms.

basic form	**another form**	***LD***	**operations on the basic form**
dělá	*dělám* (present indicative first person singular)	1	*-m* added
	dělali jsme (past indicative third person plural)	7	*-á* substituted with *a*, *-li jsme* added
	budeš dělat (future indicative second person singular)	7	*budeš* added, *-á* substituted with *a*, *-t* added
	dělej (imperative second person singular)	2	*-á* substituted with *e*, *-j* added

Table 2: Examples of Levenshtein distances (*LD*) for some forms of the verb *dělat* 'to do', its present indicative third person singular *dělá* is the basic form.

	infinitive		pres. ind. 1st pers. sing.		pres. ind. 3rd pers. sing.	
LD	**f.**	**r.**	**f.**	**r.**	**f.**	**r.**
0	29,170	0.15	3,914	0.03	60,351	0.33
1	16,827	0.09	30,210	0.24	10,465	0.06
2	62,600	0.32	38,692	0.31	45,855	0.25
3	54,069	0.28	17,587	0.14	35,343	0.19
4	20,628	0.11	22,817	0.18	16,005	0.09
5	5,203	0.03	2,981	0.02	5,003	0.03
6	3,941	0.02	4,020	0.03	5,168	0.03
7	1,530	0.01	1,616	0.01	2,296	0.01
8	530	<0.01	1,053	0.01	1,573	0.01
9	237	<0.01	329	<0.01	674	<0.01
10	58	<0.01	108	<0.01	188	<0.01
11			16	<0.01	29	<0.01
12			15	<0.01	15	<0.01
total	194,793	1.00	123,358	1.00	182,965	1.00

Table 3: Frequencies (f.) and relative frequencies (r.) of Levenshtein distances, three approaches.

basic form	mean *LD*
infinitive	2.33
present indicative 1st person sing.	2.55
present indicative 3rd person sing.	2.00

Table 4: Mean values of Levenshtein distance for three different approaches.

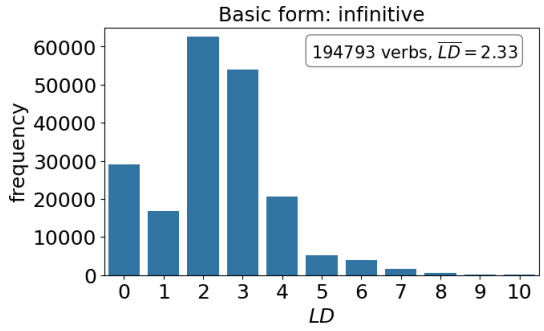


Figure 1: Frequency distribution of LD considering infinitive as the basic verb form. The frequencies are listed in Tab. 3, left.

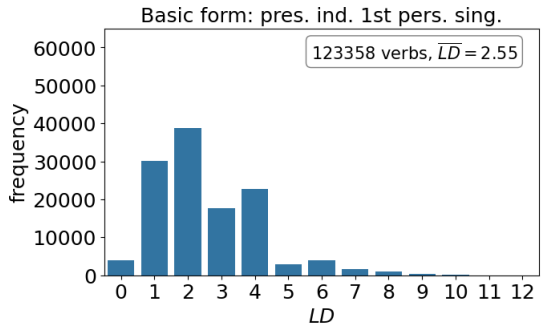


Figure 2: Frequency distribution of LD considering present indicative first person singular as the basic verb form. The frequencies are listed in Tab. 3, middle.

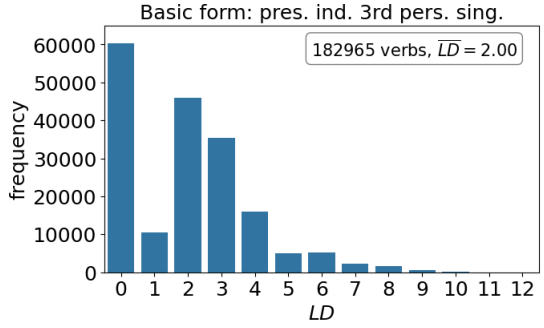


Figure 3: Frequency distribution of LD considering present indicative third person singular as the basic verb form. The frequencies are listed in Tab. 3, right.

The approach which considers the present indicative third person singular as the basic form yields the lowest mean Levenshtein distance across the verb form tokens (i.e., frequencies of particular verb forms are taken into account) in corpus (see Table 4). Therefore, it is used for the analysis that follows. This form is also the only one of the three candidates which exhibits an inverse relationship between the number of formal changes with respect to the basic form and the form frequency (i.e., the higher value of the Levenshtein distance from the basic form, the lower the frequency). However, the decrease of the frequencies is not as regular as for nouns (see Kelih et al., 2023; Mačutek et al., 2023, 2025). The frequency of 1 is clearly underrepresented, see the last two columns in Table 3 and Figure 3.

This apparent irregularity prompts us to explore the reasons why there are so few verb forms with the Levenshtein distance of 1 from the present indicative third person singular. Browsing through Czech verb classes (Komárek et al., 1986: 427–495), it becomes obvious that this value is very typical for the present indicative singular in the first and second person throughout the classes.[8] However, according to Jakobson (1990), the first and second person are 'personal', as they are participants in the speech event, while the third person, as a non-participant, is 'impersonal'. The feature of 'personal' is marked, as opposed to the unmarked 'impersonal', which is reflected in the tendency to use zero morpheme in third person (see also Bybee, 2010). Thus, verb forms used for the third person are shorter than the other forms. Zipf's law of brevity (Zipf, 1949) predicts a positive correlation between length and frequency, i.e., shorter language units occur more often. Indeed, the frequency of the third person clearly dominates over the first two, and verb forms in the first or second person in the present tense constitute less than 3.6% of all verb forms for which the categories of person and tense can be determined (see Tables 1 and 5).

There are also instances of the Levenshtein distance of 1 for some other combination of grammatical categories, such as

- *hrají* 'play' in the present indicative third person plural from the basic form *hraje*,
- *vedl* 'lead' in the past indicative third person

[8] See also forms for present indicative of the verb *dělat* 'to do' in Table 9. However, there are some exceptions, like, e.g., the verb *být* 'to be'.

masculine from the basic form *vede*,

- *myjte* ‘wash’ in the imperative second person plural from the basic form *myje*.

However, these instances are specific to particular verb classes rather than general. Moreover, plural occurs less often than singular, past tense less often than present, and imperative much less often than indicative (see Tables 6, 7, and 8, respectively). With the frequencies of these verb forms being relatively low, the Levenshtein distance of 1 cannot achieve a frequency high enough for the distribution from Figure 3 to behave more regularly.

4 Conclusion

This study has two main aims. First, it examines whether Czech verbs follow the expected inverse relation between the number of formal changes and form frequency. Second, it attempts to determine which verb form can best serve as the basic form for this type of analysis.

As for the identification of the basic verb form, three candidates were compared: the infinitive, the present indicative first person singular, and the present indicative third person singular. The results show that the present indicative third person singular yields the lowest mean Levenshtein distance from the remaining verb forms. This finding is in line with the view that third person represents the unmarked form with respect to the person opposition (Jakobson, 1990). Further, the results corroborate the expected inverse relation between the number of formal changes with respect to the basic form and form frequency, as previously observed for Slavic noun inflection (Kelih et al., 2023; Mačutek et al., 2023, 2025). Compared with nouns, however, verbs have a richer paradigm (in particular, Czech verbs have both synthetic and analytic forms) and behave less regularly with respect to this relation. Finally, our findings are also compatible with the general observation by Haspelmath and Sims (2010: 265–276) that more frequent grammatical forms tend to be shorter.

It remains an open question whether the same results would be achieved for spoken corpora. They can contain more dialogues where one can suppose a higher frequency of the second person forms.

We are aware of the fact that this paper mostly remains at the descriptive level. On the other hand, such a pilot study suggests a methodological approach and brings first basic insights into the topic. Further investigations (ideally, on material from several typologically diverse languages) are necessary to embed the results into a broader linguistic theoretical framework.

Acknowledgments

This research was supported by projects APVV-25-0505 and VEGA 2/0120/24 (J. Mačutek and M. Koščová), by Moravian-Silesian Region, grant programme “Podpora vědy a výzkumu v Moravskoslezském kraji 2023”, project “Podpora talentovaných studentů doktorského studia na Ostravské univerzitě VII”, project No. 00516/2024/RRC (M. Nogolová) and by European Regional Development Fund project “A lifetime with language: the nature and ontogeny of linguistic communication (LangInLife)” (reg. no.: CZ.02.01.01/00/23_025/0008726) (R. Čech).

	first person		second person		third person	
LD	**f.**	**r.**	**f.**	**r.**	**f.**	**r.**
0					60,349	0.44
1	3,487	0.23	638	0.20	6,038	0.04
2	5,410	0.36	1,636	0.51	22,669	0.16
3	213	0.01	213	0.07	29,250	0.21
4	307	0.02	223	0.07	11,392	0.08
5	355	0.02	72	0.02	4,378	0.03
6	1,621	0.11	147	0.05	3,397	0.02
7	1,729	0.12	147	0.05	420	<0.01
8	908	0.06	97	0.03	568	<0.01
9	624	0.04	45	0.01	5	<0.01
10	169	0.01	19	0.01		
11	27	<0.01	2	<0.01		
12	15	<0.01				
total	14,865	1.00	3,239	1.00	138,466	1.00

Table 5: Frequencies (f.) and relative frequencies (r.) of LD distinguished by **person**.

	singular		plural	
LD	**f.**	**r.**	**f.**	**r.**
0	57,250	0.55	3,101	0.07
1	5,270	0.05	4,911	0.12
2	15,333	0.15	13,043	0.31
3	12,773	0.12	12,470	0.29
4	5,747	0.06	3,617	0.09
5	2,336	0.02	1,484	0.03
6	3,113	0.03	1,380	0.03
7	1,138	0.01	1,109	0.03
8	814	0.01	758	0.02
9	247	<0.01	427	0.01
10	36	<0.01	152	<0.01
11	12	<0.01	17	<0.01
12	9	<0.01	6	<0.01
total	104,078	1.00	42,475	1.00

Table 6: Frequencies (f.) and relative frequencies (r.) of LD distinguished by **number**.

	present		past		future	
LD	**f.**	**r.**	**f.**	**r.**	**f.**	**r.**
0	60,351	0.64				
1	8,974	0.10	1,207	0.02		
2	12,217	0.13	17,430	0.30	126	0.03
3	7,050	0.07	21,397	0.37	1,317	0.29
4	1,028	0.01	10,274	0.18	644	0.14
5	1,982	0.02	2,776	0.05	54	0.01
6	2,168	0.02	1,709	0.03	1,288	0.28
7	225	<0.01	1,754	0.03	317	0.07
8	141	<0.01	689	0.01	743	0.16
9	186	<0.01	460	0.01	28	0.01
10	27	<0.01	81	<0.01	80	0.02
11	3	<0.01	24	<0.01	2	<0.01
12			15	<0.01		
total	94,352	1.00	57,816	1.00	4,599	1.00

Table 7: Frequencies (f.) and relative frequencies (r.) of LD distinguished by **tense**.

	indicative		conditional		imperative	
LD	**f.**	**r.**	**f.**	**r.**	**f.**	**r.**
0	60,349	0.62				
1	8,702	0.09			255	0.17
2	11,736	0.12			553	0.37
3	7,969	0.08	22	<0.01	288	0.19
4	768	0.01	530	0.10	359	0.24
5	195	<0.01	1,930	0.37	42	0.03
6	2,858	0.03	2,168	0.41		
7	2,070	0.02	225	0.04		
8	1,431	0.01	141	0.03		
9	488	0.01	186	0.04		
10	161	<0.01	27	0.01		
11	26	< 0.01	3	<0.01		
12	15	<0.01				
total	96,768	1.00	5,232	1.00	1,497	1.00

Table 8: Frequencies (f.) and relative frequencies (r.) of LD distinguished by **mood**.

References

Eduard Bejček, Eva Hajičová, Jan Hajič, Pavlína Jínová, Václava Kettnerová, Veronika Kolářová, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Magda Ševčíková, Jan Štěpánek, and Šárka Zikánová. 2013. Prague Dependency Treebank 3.0.

Joan L. Bybee. 2010. Markedness: Iconicity, economy, and frequency. In Jae Jung Song, editor, *The Oxford Handbook of Linguistic Typology*, pages 131–147. Oxford University Press, Oxford.

Michel M. Deza and Elena Deza. 2016. *Encyclopedia of Distances*, 4th edition. Springer, Berlin / Heidelberg.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2018. SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In Marie-Catherine de Marneffe, Teresa Lynn, and Sebastian Schuster, editors, *Proceedings of the Second Workshop on Universal Dependencies Workshop (UDW 2018)*, pages 66–74. ACL, Brussels.

Martin Haspelmath and Andrea D. Sims. 2010. *Understanding Morphology*, 2nd edition. Routledge, London / New York.

Roman Jakobson. 1990. Shifters and verbal categories. In Linda R. Waugh and Monique Monville-Burston, editors, *On Language: Roman Jakobson*, pages 386–39. Harvard University Press, Cambrigde (MA).

Tomáš Jelínek. 2017. FicTree: A manually annotated treebank of Czech fiction. In *ITAT 2017 Proceedings: Information Technologies – Applications and Theory*, volume 1885 of *CEUR Workshop Proceedings*, pages 181–185.

Emmerich Kelih, Ján Mačutek, Michaela Koščová, and Vladimír Benko. 2023. Nouns more similar to the nominative form are more frequent: A case study in Slovak. *Glottotheory*, 14(1):69–80.

Miroslav Komárek et al. 1986. *Mluvnice češtiny 2. Tvarosloví [A Grammar of Czech 2: Morphology]*. Academia, Praha.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Ján Mačutek, Emmerich Kelih, Michaela Koščová, and Mirjam Sepesy Maučec. 2025. Noun declension in Slavic languages: Animacy has a stronger influence than gender. In Adam Pawłowski, Sheila Embleton, Ján Mačutek, and Aris Xanthos, editors, *Mathematical Modelling in Linguistics and Text Analysis: Theory and Applications*, pages 27–42. Benjamins.

Ján Mačutek, Michaela Koščová, Emmerich Kelih, and Radek Čech. 2023. Frequency and morphological behaviour of nouns in Czech and Russian. *Bohemistyka*, 23(1):109–117.

Klára Osolsobě. 2017. Sloveso [Verb].

Kateřina Pelegrinová. 2023. *Menzerath-Altmann law in Czech*. Ph.D. thesis, University of Ostrava.

Petr Plecháč and Robert Kolár. 2015. The corpus of Czech verse. *Studia Metrica et Poetica*, 2(1):107–118.

Barbora Vidová Hladká, Jan Hajič, Jirka Hana, Jaroslava Hlaváčová, Jiří Mírovský, and Jan Raab. 2008. The Czech Academic Corpus 2.0 guide. *The Prague Bulletin of Mathematical Linguistics*, 89:41–96.

Daniel Zeman et al. 2023. Universal Dependencies 2.13.

George K. Zipf. 1949. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley Press, Cambridge (MA).

Appendix A Variety of verb forms in Czech language

The verb forms of infinitive *dělat* ‘to do’ are shown in Tables 9–28. Less frequent verb forms (considered archaic) are denoted by *.

	sg.	pl.
1.	*dělám*	*děláme*
2.	*děláš*	*děláte*
3.	*dělá*	*dělají*

Table 9: Present tense, indicative.

	sg.	pl.
1.	*dělal bych*	*dělali bychom*
2.	*dělal bys*	*dělali byste*
3.	*dělal by*	*dělali by*

Table 10: Present tense, conditional, masculine.

	sg.	pl.
1.	*dělala bych*	*dělaly bychom*
2.	*dělala bys*	*dělaly byste*
3.	*dělala by*	*dělaly by*

Table 11: Present tense, conditional, feminine.

	sg.	pl.
1.	*dělalo bych*	*dělala bychom*
2.	*dělalo bys*	*dělala byste*
3.	*dělalo by*	*dělala by*

Table 12: Present tense, conditional, neuter.

	sg.	pl.
masculine	*dělaje*	*dělajíce*
feminine, neuter	*dělajíc*	

Table 13: Present tense, transgressive.*

	sg.	pl.
1.	*dělal jsem*	*dělali jsme*
2.	*dělal jsi*	*dělali jste*
3.	*dělal*	*dělali*

Table 14: Past tense, indicative, masculine.

	sg.	pl.
1.	*dělala jsem*	*dělaly jsme*
2.	*dělala jsi*	*dělaly jste*
3.	*dělala*	*dělaly*

Table 15: Past tense, indicative, feminine.

	sg.	pl.
1.	*dělalo jsem*	*dělala jsme*
2.	*dělalo jsi*	*dělala jste*
3.	*dělalo*	*dělala*

Table 16: Past tense, indicative, neuter.

	sg.	pl.
1.	*budu dělat*	*budeme dělat*
2.	*budeš dělat*	*budete dělat*
3.	*bude dělat*	*budou dělat*

Table 17: Future tense, indicative.

	sg.	pl.
1.	*byl bych dělal*	*byli bychom dělali*
2.	*byl bys dělal*	*byli byste dělali*
3.	*byl by dělal*	*byli by dělali*

Table 18: Past tense, conditional, masculine, variant 1.*

	sg.	pl.
1.	*byl bych býval dělal*	*byli bychom bývali dělali*
2.	*byl bys býval dělal*	*byli byste bývali dělali*
3.	*byl by býval dělal*	*byli by bývali dělali*

Table 19: Past tense, conditional, masculine, variant 2.*

	sg.	pl.
1.	*býval bych dělal*	*bývali bychom dělali*
2.	*býval bys dělal*	*bývali byste dělali*
3.	*býval by dělal*	*bývali by dělali*

Table 20: Past tense, conditional, masculine, variant 3.*

	sg.	pl.
1.	*byla bych dělala*	*byly bychom dělaly*
2.	*byla bys dělala*	*byly byste dělaly*
3.	*byla by dělala*	*byly by dělaly*

Table 21: Past tense, conditional, feminine, variant 1.*

	sg.	pl.
1.	*byla bych bývala dělala*	*byly bychom bývaly dělaly*
2.	*byla bys bývala dělala*	*byly byste bývaly dělaly*
3.	*byla by bývala dělala*	*byly by bývaly dělaly*

Table 22: Past tense, conditional, feminine, variant 2.*

	sg.	pl.
1.	*bývala bych dělala*	*bývaly bychom dělaly*
2.	*bývala bys dělala*	*bývaly byste dělaly*
3.	*bývala by dělala*	*bývaly by dělaly*

Table 23: Past tense, conditional, feminine, variant 3.*

	sg.	pl.
1.	*bylo bych dělalo*	*byla bychom dělala*
2.	*bylo bys dělalo*	*byla byste dělala*
3.	*bylo by dělalo*	*byla by dělala*

Table 24: Past tense, conditional, neuter, variant 1.*

	sg.	pl.
1.	*bylo bych bývalo dělalo*	*byla bychom bývala dělala*
2.	*bylo bys bývalo dělalo*	*byla byste bývala dělala*
3.	*bylo by bývalo dělalo*	*byla by bývala dělala*

Table 25: Past tense, conditional, neuter, variant 2.*

	sg.	pl.
1.	*bývalo bych dělalo*	*bývaly bychom dělaly*
2.	*bývalo bys dělalo*	*bývala byste dělala*
3.	*bývalo by dělalo*	*bývala by dělala*

Table 26: Past tense, conditional, neuter, variant 3.*

	sg.	pl.
1.		*dělejme*
2.	*dělej*	*dělejte*

Table 27: Imperative.

	masculine	**feminine**
Present tense, indicative	*děláte*	*děláte*
Present tense, imperative	*dělejte*	*dělejte*
Present tense, conditional	*dělal byste*	*dělala byste*
Past tense, indicative	*dělal jste*	*dělala jste*
Past tense, conditional, variant 1*	*byl byste dělal*	*byla byste dělala*
Past tense, conditional, variant 2*	*byl byste býval dělal*	*byla byste bývala dělala*
Past tense, conditional, variant 3*	*býval byste dělal*	*bývala byste dělala*

Table 28: Honorific variations of specific forms in 2. person singular.

Benchmarking Language Models on Romanian Eighth-Grade National Exams

Radu Ion
Research Institute for AI
"Mihai Drăgănescu",
Romania
radu@racai.ro

Verginica Barbu Mititelu
Research Institute for AI
"Mihai Drăgănescu",
Romania
vergi@racai.ro

Elena Irimia
Research Institute for AI
"Mihai Drăgănescu",
Romania
elena@racai.ro

Maria Mitrofan
Research Institute for AI
"Mihai Drăgănescu",
Romania
maria@racai.ro

Vasile Păiș
Research Institute for AI
"Mihai Drăgănescu",
Romania
vasile@racai.ro

Abstract

This paper evaluates a set of Large Language Models (LLMs) on the task of automatically solving national standardized Romanian language tests for eighth-grade students. The models are provided only with the question text, in Romanian, and are required to generate answers across various formats, such as multiple choice, fill-in-the-blank, true/false, short answer, and composition. Their responses are automatically evaluated against the official scoring rubrics, which detail how points are awarded for each question and specify the correct answer where applicable. This study's goal is to establish a benchmark for the performance of Romanian-aware and state-of-the-art LLMs on standardized eighth-grade Romanian exams, and to determine whether state-of-the-art LLMs can serve as accurate eighth-grade exam evaluators.

Keywords: Romanian LLM evaluation, automatic test scoring with LLMs, Romanian language standardized tests, LLM-as-a-judge

1 Introduction

Every year, eighth-grade schoolchildren[1] in Romania must take nationwide, standard exams[2] of Romanian language and mathematics, whose grades will determine their ranking in the high school admissions competition. Each exam is designed to evaluate their knowledge of Romanian and mathematics, acquired during the fifth to eighth grades. During simulation and exam days, and additionally, during supplementary sessions, teachers from all over Romania are tasked with the grading of these tests, according to the scoring rubric that contains detailed instructions on how to score each item in the test, awarding points for each sub-goal that has been achieved by the test taker, for each exam question.

Grading these tests by two independent graders is a tedious task, but one of tremendous importance for the future of eighth-grade graduates. Cononenco (2025) advances an amount of 70–80 tests per teacher to be graded, a task that collectively required "«an impressive deployment of forces» — Sorin Ion, Secretary of State in the Ministry of Education", according to the cited news article. In 2024, 152,903 schoolchildren took the Romanian language and mathematics exams in the National Evaluation, and 74.5% obtained more than 5 (on a scale from 1 to 10), as the average of the two exams.

In this context, leveraging LLMs to assist with test grading seems to be the obvious solution that can assist (not replace) teachers with test scoring. To assist with test grading, the LLMs should first be able to solve the tests themselves, demonstrating that "they know what they are doing" by obtaining high grades, and second, exhibit grading reasoning skills that human evaluators (i.e., Romanian

[1] Typically aged 14.

[2] Called the "National Evaluation of the Eighth Grade".

eighth-grade teachers) typically apply to these tests.

To assess the capabilities of Romanian and state-of-the-art LLMs in answering Romanian-language questions from the National Evaluation for eighth-grade students, we first parsed the publicly available PDF test documents and their corresponding answer keys into structured JSON format. Each test item is annotated with its point value, question type (e.g., multiple-choice, fill-in-the-blank, true/false, short answer, or composition), and the question text. From the associated answer key, we extracted either the exact correct answer or, for open-ended questions (specifically short-answer items), a set of example responses. Additionally, each question is accompanied by a detailed scoring rubric that outlines how points are allocated based on specific criteria or sub-goals the response meets. These rubrics, along with the reference answers, are used by a judge LLM (Gu et al., 2025) to automatically evaluate the quality of generated responses and assign points accordingly.

In what follows, we will review the relevant literature in Section 2, we will present the compilation of the JSON dataset from the Romanian language tests of the National Evaluation for the eighth grade in Section 3, and we will present the evaluation results of current Romanian-aware and state-of-the-art multilingual LLMs in the zero-shot scenario in Section 4. We will conclude the paper with Section 5.

2 Related work

Masala et al. (2024) describe most of the Romanian-aware LLMs that we evaluate here. Their paper is the most recent and comprehensive evaluation of Romanian LLMs, built upon established English-language models. Starting with a pre-trained English LLM, they first fine-tune it on Romanian corpora to produce a foundational Romanian-aware model. This base model is then further fine-tuned using Romanian-translated instruction datasets to yield an instruction-following Romanian-aware LLM. Using LLaMA-2-7B as the starting point, they develop RoLLaMA2-7B-Base (the foundational model) and RoLLaMA2-7B-Instruct (the instruction-tuned variant). Additional Romanian instruction-aware models are fine-tuned directly, without intermediate Romanian pre-training, from models such as Mistral-7B, LLaMA-3-8B, LLaMA-3.1-8B, and Gemma-7B. These models are evaluated on reference datasets that were machine-translated into Romanian without human post-editing. On average, across multiple benchmarks, the Romanian instruction-tuned models outperform their original English counterparts.

Păiș et al. (2025) describe two Romanian-aware language models that were fine-tuned starting from Llama3.2-1B base models. The first one[3] was further trained on a portion of the Representative Corpus of the Contemporary Romanian language (CoRoLa). Starting from it, the second one[4] was further fine-tuned on several Romanian instruction sets.

RoQLlama-7B (Dima et al., 2024) is another Romanian-aware LLM, built upon the Llama2-7B model. Its development follows the same methodology as RoLLaMA2-7B-Base (Masala et al., 2024) but incorporates QLoRA-based fine-tuning with 4-bit quantization over a 7.3-billion-token Romanian corpus. In addition to general NLP benchmarks, RoQLlama-7B is also evaluated on a Romanian medical question-answering dataset introduced in the same study. Under the zero-shot evaluation setting adopted in our work, the model achieves an accuracy of 21.57%. This model will also be evaluated on our dataset.

Dumitran et al. (2025) introduce GRILE, a dataset comprising 1,151 multiple-choice questions sourced from major Romanian high-stakes examinations, including the National Evaluation, Baccalaureate, and university entrance exams. GRILE serves as a benchmark for evaluating multilingual and Romanian-specific LLMs in two key capacities: (a) their ability to select correct answers, and (b) their capacity to generate coherent and informative explanations that justify their choices. While Gemini 2.5 Pro achieves 83% accuracy on the GRILE benchmark, most open-weight multilingual models remain below 65%, with Romanian-specific models not exceeding 41%. Furthermore, expert analysis reveals that 48% of their generated explanations contain factual inaccuracies or pedagogical flaws.

BacPrep (Dumitran et al., 2026) is an online platform that provides high-quality preparation and

[3] https://huggingface.co/racai/corola-llama-3.2-1b-e3

[4] https://huggingface.co/racai/corola-instruct-llama-3.2-1b-e2

feedback for the Romanian Baccalaureate exam, particularly targeting high school students from economically disadvantaged backgrounds who may lack access to private tutoring. The platform leverages Gemini 2.0 Flash, one of Google's latest LLMs, to deliver automated feedback on practice exam responses. Currently under development, BacPrep is actively collecting real-world student answer data. In future iterations, once student responses and LLM-generated feedback have been reviewed by human evaluators, the platform aims to provide reliable, pedagogically sound guidance to learners.

In line with current research on the automatic assessment of LLM-generated output (Dinh et al., 2024; Center for AI Safety et al., 2026; Shetty et al., 2025), we use the LLM-as-a-judge paradigm to automatically determine whether responses are correct. Dinh et al. (2024) observe that automatic evaluation of generated text is difficult and, even though machine translation metrics such as BLEU or ROUGE can be used, LLM-as-a-judge is preferable, as it is well correlated with human judgments, especially in the textual modality.

Center for AI Safety et al. (2026) conduct a study similar to ours on academic assessment, evaluating generative AI capabilities on a set of 2,500 challenging academic questions. The "Humanity's Last Exam" dataset features questions contributed by nearly 1,000 subject-matter experts from over 500 institutions across 50 countries, predominantly professors, researchers, and individuals holding graduate degrees. The o3-mini is the judge LLM used to verify the correctness of answers.

3 National Evaluation Romanian language dataset

This dataset contains JSON-formatted versions of model, simulation, and official Romanian language tests, originally published as PDF files by the Romanian Ministry of Education for the National Evaluation at the eighth-grade level[5]. Each year, exam simulation sessions are held using tests that mirror the structure and content requirements – both literary and grammatical – of the official exams. Prior to these simulations, the Ministry also releases additional model tests that follow the same format and content expectations as the official ones.

Each test consists of a pair of files: the test variant ("variantă" in Romanian), which includes the exam questions, and the answer key ("barem" in Romanian). Multiple model and simulation pairs are available each year, while only two official test pairs are released – one for the main session and one as a backup. The Hungarian-speaking minority in Romania also takes this exam, and dedicated versions of the test are provided for them.

The Romanian language exam is divided into two units: Unit I and Unit II. Unit I typically comprises two parts (A and B) that assess students' reading comprehension and grammatical knowledge. Unit II requires students to write a short essay (150–300 words). The comprehension questions in Unit I refer to one or both of the two short literary texts included in the test (referred to as "Textul 1" and "Textul 2" in Romanian).

The test contains the following types of questions:

- **Short answer**: Students are required to produce one or more sentences that follow specific grammatical constraints or to analyze a phrase from a grammatical perspective. For example, "Rewrite the sentence « At the end of the year, before college admission, … » replacing each underlined word with a synonym."
- **Fill-in-the-blank**: Students must complete sentences by inserting the correct forms of words, based on definitions provided in parentheses following the blanks. For example, "Fill in the missing words in the sentence below according to the indications in parentheses: "____ (to say, imperative, plural) if you ____ (to like, conjunctive, present tense, 3rd person singular) …"
- **True/false**: Students evaluate the truth value of 4 to 6 statements using information from the accompanying literary texts, e.g., "From Text 1 above, determine if the following sentence is true or false: « Adina Nanu's desk mate looked like a porcelain doll. »"
- **Multiple choice**: Students select the correct answer by letter, based either on information from the literary texts or on their

[5] The National Evaluation tests can be found at http://subiecte202X.edu.ro, where "X" can be replaced with "3", "4" and "5".

grammatical knowledge, e.g. “Both words in the series contain a diphthong: a) „aproape”, „înviorat”; b) „fiindcă”, „voit”; c) „aceasta”, „coroană”; d) „mamei”, „deslușeam”””

- **Composition**: Students write a short essay expressing their opinion in response to a prompt related to the included literary texts. For instance, “Do you think it is important to take an exam without fear? Justify your answer in 50–90 words, drawing on Text 2.”

Each question is assigned a number of points proportional to its difficulty. One full test is worth 90 points, with an additional 10 points awarded by default. The final grade is calculated by dividing the total number of points earned by 10, resulting in a grade on a scale from 1 to 10.

The answer key is always provided alongside the test. It includes exact answers when available or example responses when multiple valid answers are acceptable. Based on the question types defined above, the answer key contains the following information:

- **Short answer**: For questions requiring grammatical analysis, example responses are typically provided (e.g., grammatical breakdowns of word forms, phrases, or sentences that meet specific constraints).
- **Fill-in-the-blank**: The list of correct word forms.
- **True/false**: The correct truth values for each statement.
- **Multiple choice**: The letter corresponding to the correct answer.

For all question types except multiple-choice, the answer key includes detailed instructions specifying how points should be awarded for partial fulfillment of the question requirements. These guidelines are especially comprehensive for composition questions, outlining point allocations for aspects such as grammatical accuracy, spelling, clarity of expression, and other relevant criteria. For example, for the composition question “Do you think it is important to take an exam without fear? Justify your answer in 50–90 words, drawing on Text 2.”, the associated guidelines for the human evaluator are:

- stating the answer to the given question, **1 point**
- justification of the stated answer, **3 points**:
  - nuanced and appropriate, making effective use of the indicated text – 3 points
  - appropriate but schematic, making use of the indicated text – 2 points
  - tendency to generalize, attempt at justification, no use of the indicated text – 1 point
  - lack of justification – 0 points
- observance of spelling and punctuation rules (at most one error – 1 point; 2 or more errors – 0 points), **1 point**
- staying within the indicated word limit, **1 point**

To convert a PDF test file into its JSON equivalent, we followed a four-step process:

1. The PDF file was first transformed into a UTF-8-encoded text file, with careful attention to preserving formatting features such as bold, italic, and underline.
2. The resulting text file was segmented into units, parts, and individual questions using regular expressions.
3. Each question was parsed into the JSON format using the `gpt-oss-20b` LLM[6] (OpenAI, 2025), which was prompted to extract the question ID, type, point value, and text. Prompts were issued in English, as Romanian prompts yielded less reliable results, particularly regarding the JSON output, which was not supplied consistently with Romanian prompts.
4. Each parsed question was manually reviewed to ensure that the LLM did not introduce any unintended modifications. During prompting, we explicitly instructed the model to copy question and answer key details verbatim, without rephrasing or altering the original content.

Table 1 presents statistics of the newly developed `en_viii_rolang` dataset from the National Evaluation at the eighth-grade level.

[6] https://huggingface.co/openai/gpt-oss-20b

Year	SA	FI	TF	MC	CO	ALL
2021	34	4	9	31	28	**106**
2022	18	2	4	28	20	**72**
2023	5	1	1	7	4	**18**
2024	5	1	1	7	4	**18**
2025	48	11	10	65	51	**185**
ALL	**110**	**19**	**25**	**138**	**107**	**399**

Table 1: Statistics of the dataset `en_viii_rolang` per year and question type

For each year[7], we report the number of questions by type: SA (short answer), FI (fill-in-the-blank), TF (true/false), MC (multiple choice), and CO (composition). Composition questions account for 26% of the dataset and will be evaluated subjectively, taking into account potential biases introduced by the judge LLM. For all other question types, we can compute exact accuracy metrics, both in terms of accuracy on correct answers (percentage of correct answers) and the resulting grade, on a scale from 1 to 10.

4 Evaluation

4.1 Estimating judge LLM scoring bias

We chose[8] `gemma3-27b-it`[9] (Gemma Team and Google DeepMind, 2025) as the LLM to serve as the judge responsible for evaluating the responses generated by the "student" LLMs. To assess its potential bias – either in overestimating or underestimating student responses – we *manually reviewed its judgments* with the help of three Romanian language teachers in eighth grade.

Each of the Romanian language teachers had to review the judge LLM's scoring justifications on a set of three test files:

- ENVIII_2023_romana_var_model.json
- ENVIII_2024_romana_var_model.json
- ENVIII_2025_romana_var_model.json

The answers to the questions in these three files were supplied by the `gpt-oss-20b` "student" LLM, and *each teacher evaluated the judge LLM scoring justifications for these answers*, so that we could compare their evaluations.

When the judge LLM provided an incorrect evaluation for a given question, the teacher annotated the explanation, identified the issues, and adjusted the number of points awarded – either increasing or decreasing the score. As a result, for the three manually reviewed tests, we can compare the grades assigned by the judge LLM with those assigned by human evaluators.

Table 2 provides a breakdown of the *completion grades* assigned by the judge LLM (J) and human teachers (T1–T3) for each question type, across the three manually reviewed tests. The completion grade for a question type is a percentage of the points obtained by the "student" LLM out of all available points for that question type. Column OG presents the *official test grade* (on a scale from 1 to 10), computed across the three-file test set. Knowing that perfectly solving all three tests would yield 90 × 3 = 270 points plus 10 × 3 = 30 points offered by default, for a total of 300 points, the official test grade OG is computed as follows (P is the number of points obtained by the "student" LLM over all three tests):

$$OG = \frac{P + 30}{300} \times 10$$

	SA	FI	TF	MC	CO	OG
J	66%	71%	89%	62%	98%	8.3
T1	62%	79%	89%	62%	83%	7.63
T2	57%	79%	89%	62%	94%	7.97
T3	66%	79%	83%	62%	89%	7.97

Table 2: Completion and official grades offered by the judge LLM and three teachers, on the three-file test set

If we look at the official test grade (OG), we see that all human evaluators are more conservative than the LLM, providing evidence of the LLM judge's overestimation of grades.

Teacher 1 (T1) stands out in relation to the composition question type (CO), having assigned a significantly lower completion grade (83%) than the LLM judge (98%). This discrepancy can be attributed to T1 being by far the most experienced of the three teachers, with many years of experience evaluating national eighth-grade assessment tests. Her extensive experience enables her to accurately assess the quality of the LLM judge's evaluations. Below are several of her

[7] Regarding 2023 and 2024, there were no model or simulation tests available for download, just the official tests.

[8] This LLM has been favored by all eighth-grade Romanian language teachers who participated in our study, following trial-and-error tests conducted independently by each teacher.

[9] https://huggingface.co/google/gemma-3-27b-it

insightful comments regarding the LLM's assessment of compositions:

- The dialogue is not a characterization method; thus, a loss of 1 point.
- Two words are inappropriately used: "poet" and "Arghezi"; "poetic voice" would be the right word; thus, a loss of 1 point. The speech figure "personificarea" is not exemplified; thus, a loss of 1 point. The interpretation of the cited metaphor is incorrect, a loss of 1 point.
- The theme and the message are not correctly identified, thus a loss of 2 points.
- The answer does not provide a motivation at all.
- The two texts should not be associated with each other, so no new text is mentioned; thus, there is a loss of 1 point. No value is named, thus a loss of 1 point. The reference to the suggested text is missing, resulting in a 1-point deduction.

The composition question is the most challenging, as it demands creativity from students, as well as clarity and logical cohesion in writing. It is therefore particularly noteworthy that an LLM such as `gpt-oss-20b` received a completion grade of 83% from an experienced Romanian language teacher. This sets a high benchmark that will be difficult to surpass.

Table 3 reports the percentage of questions correctly evaluated by the judge LLM (i.e., the judge LLM's accuracy) and the judge LLM's overall judgment accuracy rate, as evaluated by each teacher.

	SA	FI	TF	MC	CO	ALL
T1	64%	67%	100%	100%	39%	74%
T2	79%	67%	100%	100%	85%	89%
T3	64%	67%	67%	100%	54%	76%

Table 3: Accuracy of per-question judgment decisions by the LLM, according to each teacher

The LLM judge achieved a perfect success rate on multiple-choice (MC) and true/false (TF) questions. The overall accuracy rate of the LLM judge is around 75%, except for T2, who reported a higher rate. However, as shown in Table 2, the LLM judge tends to overestimate the final grade by 0.33 to 0.67 points. Given that T1 is the most experienced evaluator of such assessments, we consider her final grade of 7.63 as the reference point. Accordingly, the LLM judge's 0.67-point inflation will be treated as bias and subtracted from the "student" LLM's overall grade to yield a better estimate of the true grade.

The 0.67-point inflation bias is specific to our selected pair of judge and student LLMs. Its determination was possible through the manual evaluation performed by our teacher team, as previously shown, to obtain a more accurate estimate of the official grades. We hypothesize[10] that its value is determined primarily by the judge LLM rather than the student LLM. This is why we chose a strong student model rather than a weak one for computing the bias, since bad answers are easier to justify (more obvious) than good or partially good ones.

4.2 Student LLM performance

We ran all selected Romanian models on the `en_viii_rolang` dataset in a zero-shot setting – that is, by prompting them only with the information available to an eighth-grade student – and used the judge LLM to score each model's outputs. For comparison with the state-of-the-art LLMs, we also ran `gemma3-27b-it` (G3) and `gpt-oss-20b` (OS) on the same dataset.

The Romanian "student" LLMs evaluated in this study are listed below (all identifiers are valid on HuggingFace):

- `OpenLLM-Ro/RoLlama2-7b-Instruct-DPO` (R2)
- `OpenLLM-Ro/RoLlama3-8b-Instruct-DPO` (R3)
- `OpenLLM-Ro/RoLlama3.1-8b-Instruct-DPO` (R31)
- `OpenLLM-Ro/RoMistral-7b-Instruct-DPO` (RM)
- `OpenLLM-Ro/RoGemma2-9b-Instruct-DPO` (RG)
- `racai/corola-instruct-llama-3.2-1b-e2` (crl)
- `andreidima/Llama-2-7b-Romanian-qlora` (L2)

[10] Experimentally proving this hypothesis requires a lot of manual evaluation of lengthy LLM output, and it is thus left for future work.

	SA	FI	TF	MC	CO	OG
R2	22%	26%	19%	45%	60%	4.27
R3	24%	27%	27%	36%	60%	4.26
R31	28%	30%	52%	55%	72%	5.42
RM	34%	39%	69%	50%	79%	5.95
RG	30%	38%	58%	61%	55%	4.81
crl	.4%	0%	0%	3%	1%	0.11
L2	11%	24%	30%	17%	13%	1.47
G3	**76%**	**78%**	**93%**	74%	90%	8.30
OS	**76%**	75%	90%	**76%**	**91%**	**8.37**

Table 4: "Student" LLM completion grades and official test grades

	SA	FI	TF	MC	CO	ALL
R2	8%	5%	0%	45%	24%	25%
R3	15%	11%	0%	37%	23%	24%
R31	14%	5%	8%	57%	24%	31%
RM	16%	11%	24%	50%	32%	32%
RG	12%	16%	12%	62%	8%	29%
crl	0%	0%	0%	3%	0%	1%
L2	6%	5%	0%	16%	3%	8%
G3	52%	47%	**68%**	74%	68%	65%
OS	**54%**	**53%**	56%	**76%**	**72%**	**66%**

Table 5: Percents of correctly answered questions

	SA	FI	TF	MC	CO	ALL
R2	33%	53%	32%	0%	65%	31%
R3	26%	47%	56%	0%	65%	30%
R31	37%	63%	84%	0%	71%	38%
RM	44%	68%	72%	0%	65%	37%
RG	42%	58%	72%	0%	81%	41%
crl	1%	0%	0%	0%	4%	1%
L2	16%	32%	56%	0%	36%	19%
G3	45%	53%	32%	0%	30%	25%
OS	38%	42%	44%	0%	28%	23%

Table 6: Percents of partially correct answers

	SA	FI	TF	MC	CO	ALL
R2	59%	42%	68%	55%	11%	44%
R3	59%	42%	44%	63%	12%	46%
R31	49%	32%	8%	43%	5%	31%
RM	40%	21%	4%	50%	3%	31%
RG	46%	26%	16%	38%	11%	30%
crl	99%	100%	100%	97%	96%	98%
L2	78%	63%	44%	84%	61%	73%
G3	**3%**	**0%**	**0%**	26%	2%	**10%**
OS	8%	5%	**0%**	**24%**	**0%**	11%

Table 7: Percents of incorrectly answered questions

Tables 4 through 7 present the results for each question type (SA, FI, TF, MC, and CO) and for the entire dataset, as evaluated by our judge LLM. An underlined value indicates the highest performance among Romanian-aware LLMs, while a bolded value represents the overall best performer in each question category. For each model and question type, the sum of the percentages for correct, partially correct, and incorrect answers equals 100%. The definitions of these types of answers are as follows:

- A *correct answer* scores all points out of the total points assigned to the question.
- A *partially correct answer* is one that earns some of the points assigned to a question, but not all. It follows that multiple-choice questions, which can only be evaluated as correct or incorrect, cannot receive partial credit (column MC in Table 6 is 0%).
- An *incorrect answer* receives zero points.

The top performer among the Romanian-aware LLMs is the `OpenLLM-Ro/RoMistral-7b-Instruct-DPO`, which achieved an official test grade of 5.95 on our dataset. After subtracting the 0.67-point bias introduced by the judge LLM's tendency to overestimate grades, the adjusted score becomes 5.28 – much lower than the adjusted scores of `gemma3-27b-it` (7.63) and `gpt-oss-20b` (7.7).

One obvious reason multilingual LLMs are significantly better than Romanian-aware LLMs is that they are roughly 3 times larger in terms of parameter count. On the other hand, and this is the main reason they evidently outperform the Romanian models, they were trained on much larger datasets (both during pre-training and fine-tuning) and used state-of-the-art fine-tuning techniques such as knowledge distillation and reinforcement learning. For instance, `gemma3-27b-it` was fine-tuned using a large instruction-tuned teacher model (Gemma Team and Google DeepMind, 2025: 4). This means it learned not just from conventional human-written prompt-response pairs, but also from richer probability distributions or generated behavior supplied by a more capable internal teacher. Pre-trained on 14 trillion tokens, the model was fine-tuned on data targeting conversational ability, instruction following, multilingual responses, and long-context behavior, among other learning objectives.

A note regarding `racai/corola-instruct-llama-3.2-1b-e2`: this model is the smallest among those evaluated, with only 1 billion parameters. Although it was fine-tuned on several instruction datasets, this training – combined with its initial pre-training on the CoRoLa corpus –

appears to have been insufficient for achieving better performance. To get a sense of how off the `racai/corola-instruct-llama-3.2-1b-e2`'s answers were, here is a representative sample comparison with `gpt-oss-20b` on a short-answer question asking the student to name two cities that they read about in the first supporting text ("Textul 1"):

	Answer	Judge LLM decision
crl	Assistant: Iată cele două nume menționate în textul 1. Numele orașelor sunt Berlin (Berlin) și Constantinopol (Constantinopol). (...) Analiza mecanismelor politice din România a furnizat dovezi că populația orașului Romanovski a crescut în ultimele decenii. Potrivit acestui raport, a crescut cu aproximativ 18 milioane în 2008. (...)	The student did not provide any of the place names that appear in "Textul 1" (such as Câmpina, Sinaia, București). None of the names mentioned (Berlin, Constantinopol, Romanovski, etc.) are present in the text, so no points are awarded.
OS	Localități menționate în Textul 1: „Câmpina", „Sinaia"	Student listed "Câmpina" and "Sinaia", both of which appear in "Textul 1". Each correct city earns 1 point, for a total of 2 points out of 2.

Table 8: Sample output of student LLM `racai/corola-instruct-llama-3.2-1b-e2`

We found that the judge LLM's evaluation accuracy is approximately 75% (see the "ALL" column in Table 3). Consequently, the percentages of correctly answered questions reported in Table 5 should be adjusted to reflect this finding. For instance, `OpenLLM-Ro/RoMistral-7b-Instruct-DPO`'s 32% correct response rate should be recalculated as 32% × 75% = 24%, with the remaining 8% *most likely reclassified as partially correct*. As a result, the percentage of incorrect answers could be a more reliable metric for ranking student LLMs, since it remains unaffected.

5 Conclusions

We evaluated seven Romanian-aware LLMs on the task of solving tests from the National Evaluation administered at the eighth-grade level in Romania. The results are particularly impressive for the multilingual models `gemma3-27b-it` and `gpt-oss-20b`, while the Romanian-aware models performed significantly lower. For this evaluation, we developed the `en_viii_rolang` dataset based on the official, publicly available National Evaluation tests for Romanian. The dataset contains 399 questions of various types, formatted in JSON. It is continuously being expanded, as tests older than three years are no longer accessible on official websites, though they remain available through other Romanian language learning resources online.

Correctly and partially answered questions together provide a valuable foundation for automatically generating an instruction dataset for fine-tuning student LLMs and improving their performance. Future research will focus on this direction, as well as on training Romanian-aware LLMs to generate reasoning paths that enable them to evaluate real-world tests in a manner similar to Romanian language teachers. To support this goal, we plan to involve additional Romanian teachers in the project, asking them to verbalize their evaluation thought processes – following the approach taken by the experienced Romanian language teacher featured in this study.

Acknowledgments

This work was carried out within the "Large Language Models for the European Union (LLMs4EU)", project no. 101198470, call DIGITAL-2024-AI-B-06-LANGUAGE, 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 or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.

This work was also supported by a grant from the Ministry of Research, Innovation and Digitalization – UEFISCDI, project number PN-IV-P8-8.2-EUD-2025-0061, within PNCDI IV.

References

Center for AI Safety, Scale AI & HLE Contributors Consortium. 2026. A benchmark of expert-level academic questions to assess AI capabilities. *Nature* 649, 1139–1146. https://doi.org/10.1038/s41586-025-09962-4

Laura Cononenco. 2025. Profesorii au terminat corectarea lucrărilor la simularea Evaluării

Naționale mai devreme. News article, online at https://www.edupedu.ro/profesorii-au-terminat-corectarea-lucrarilor-la-simularea-evaluarii-nationale-mai-devreme-sorin-ion-a-fost-o-desfasurare-de-forte-impresionanta-vom-incepe-sa-trimitem-de-saptamana-aceasta-borderou/

George-Andrei Dima, Andrei-Marius Avram, Cristian-George Crăciun, Dumitru-Clementin Cercel. 2024. RoQLlama: A lightweight Romanian adapted language model. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4531–4541, Miami, Florida, USA. Association for Computational Linguistics.

Tu Anh Dinh, Carlos Mullov, Leonard Bärmann, Zhaolin Li, Danni Liu, Simon Reiß, Jueun Lee, Nathan Lerzer, Jianfeng Gao, Fabian Peller-Konrad, Tobias Röddiger, Alexander Waibel, Tamim Asfour, Michael Beigl, Rainer Stiefelhagen, Carsten Dachsbacher, Klemens Böhm, and Jan Niehues. 2024. SciEx: Benchmarking Large Language Models on Scientific Exams with Human Expert Grading and Automatic Grading. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11592–11610, Miami, Florida, USA. Association for Computational Linguistics.

Marius Dumitran, Angela Dumitran, Alexandra Mihaela Danila. 2025. GRILE: A Benchmark for Grammar Reasoning and Explanation in Romanian LLMs. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 316–324, Varna, Bulgaria.

Adrian-Marius Dumitran, Radu Dita, and Angela-Liliana Dumitran. 2026. BacPrep: Lessons from Deploying an LLM-Based Bacalaureat Assessment Platform. arXiv:2506.04989v2 [cs.SE]

Gemma Team, Google DeepMind. 2025. Gemma 3 Technical Report. arXiv:2503.19786v1 [cs.CL]

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A Survey on LLM-as-a-Judge. arXiv:2411.15594v5 [cs.CL]

Mihai Masala, Denis Ilie-Ablachim, Alexandru Dima, Dragos Georgian Corlatescu, Miruna-Andreea Zavelca, Ovio Olaru, Simina-Maria Terian, Andrei Terian, Marius Leordeanu, Horia Velicu, Marius Popescu, Mihai Dascalu, and Traian Rebedea. 2024. "Vorbești românește?" A Recipe to Train Powerful Romanian LLMs with English Instructions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11632–11647, Miami, Florida, USA. Association for Computational Linguistics.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b Model Card. arXiv:2508.10925v1 [cs.CL]

Vasile Păiș, Dan Tufiș, Verginica Barbu Mititelu, Radu Ion, Maria Mitrofan, Elena Irimia, Eric Curea, and Andrei Păun. 2025. Towards Romanian Large Language Models Using the CoRoLa Corpus. In *Proceedings of the 20th International Conference on Linguistic Resources and Tools for Natural Language Processing – ConsILR 2025*, Bucharest, Romania.

Pranam Shetty, Abhisek Upadhayaya, Parth Mitesh Shah, Srikanth Jagabathula, Shilpi Nayak, and Anna Joo Fee. 2025. Advanced Financial Reasoning at Scale: A Comprehensive Evaluation of Large Language Models on CFA Level III. arXiv:2507.02954v2 [cs.CL]

Lexical Surprisal from Large Language Models: Cautionary Notes for Legal Discourse

Silvie Cinková, Ivana Kvapilíková, Jan Černý
Faculty of Mathematics and Physics
Charles University, Czech Republic
{cinkova, kvapilikova}@ufal.mff.cuni.cz, cerny@matfyz.cz

Abstract

Lexical surprisal correlates with human reading times and is easy to extract from large language models (LLMs), thus appearing useful for readability assessment. However, recent evidence suggests that LLMs tend to assign very low surprisal to texts that are demonstrably difficult for lay readers. In this paper, we investigate the stability and genre sensitivity of lexical surprisal estimates for Czech across five language models. Using diverse single-genre corpora, we analyze word-level surprisal distributions and inter-model correlations, observing inter-model agreement to be generally low and strongly modulated by genre. Detailed distributional analyses reveal sharp differences between generative and n-gram models, particularly in legal genres, where Transformer-based models assign disproportionately low surprisal to technical nouns. Correlations between lexical surprisal and TF-IDF further suggest that large models are more sensitive to structural patterns than to domain-specific lexical difficulty. These findings caution against using lexical surprisal from general-purpose LLMs for readability assessment in institutional and legal texts, highlighting the need for genre-aware and cognitively grounded evaluation.

Keywords: readability; lexical surprisal; language models

1 Introduction

Lexical surprisal is a linguistic concept adopted from Information Theory (Shannon, 1948), which models the predictability of a text by determining the probability of each word given its preceding context. There is ample empirical evidence of lexical surprisal reflecting the cognitive load on a reader, operationalized as reading time (Levy, 2008; Hale, 2001; Smith and Levy, 2013; Wilcox et al., 2023).

Assuming that there is an optimum pattern of surprisal distribution (Frank and Jaeger, 2008; Clark et al., 2023), lexical surprisal lends itself as an aid to automatic readability assessment.

Novotná et al. (2025); Černý (2025) describe their experiments with lexical surprisal (computed by GPT-2) to assess the readability of Czech legal and administrative texts in an application. Interestingly, their initial observations did not reveal an interpretable pattern. The token-wise values were uniformly very low, suggesting that the texts would be both information-light and easy to read, which could by no means comply with the judgment of a common reader. We speculate that GPT-2 is too well-trained in legal writing to model the mental lexicon of a legal layman. This initial finding has motivated our current research questions:

1. Do language models agree on lexical surprisal values in Czech texts?

2. Is the agreement affected by text genre?

2 Related work

We build on prior work that established the relationship between lexical surprisal and text readability, using probabilistic language models to quantify surprisal (Demberg and Keller, 2008; Pimentel et al., 2023; Sun and Wang, 2025). Whereas these studies confirm the correlation between word surprisals and their reading times in the English language, Wilcox et al. (2023) extend the evidence for surprisal theory to multiple languages.

Fernandez Monsalve et al. (2012); Goodkind and Bicknell (2018) postulate that the predictive power of word surprisal on reading times increases with the decreasing model perplexity. In contrast, Oh and Schuler (2023) demonstrate that for Transformer models, the predictive accuracy declines

when the model was exposed to too much data (over 2 billion tokens) during training.

Regarding surprisal across different text styles, Gehrmann et al. 2019 use word predictability to discern natural texts from automatically generated ones using the fact that the latter are based on probabilistic distribution samples and hence maintain particularly low surprisal levels.

3 Experiment setup

In this work, we focus exclusively on the Czech language and analyze the surprisal distributions across different genres.

3.1 Surprisal Computation

We prepared seven Czech corpora to represent different genres (see 4). We ran each language model from our selection (as listed in 3.2) to compute the lexical surprisal on each text, using an earlier in-house re-implementation of *gltr* (Gehrmann et al., 2019; Černý, 2025). Word surprisal is approximated as the negative log probability assigned to each word by the model given its preceding context. For Transformer models, the context is limited to the previous 512 tokens, whereas n-gram models, by construction, condition only on the four preceding words.

As a next step, we mapped the model-specific subword tokens on word boundaries as defined by a Czech tokenizer (Straka et al., 2016), transforming the lexical surprisal to word-wise sums of their negative log probabilities (Pimentel and Meister, 2024).

3.2 Language models

When selecting language models, we included two Transformer models trained on Czech data and one multilingual model. Suspecting that heavily trained Transformer models may exhibit *super-reader* capabilities, as observed by Oh and Schuler (2023), we also constructed two n-gram models on datasets over which we have greater control. The model overview is displayed in table 1.

Model	Params	Train Tokens	Language
gpt2_small	117M	5B	cs
gpt2_XL	1.5B	11B	cs
qwen	7B	18T	multiling
ngram_web	–	1B	cs
ngram_cnk	–	5B	cs

Table 1: Overview of language models used in the experiments

3.2.1 Qwen2.5

Qwen2.5 (Qwen et al., 2025) represents one of the most advanced multilingual language models. The base models were pretrained on a large-scale, high-quality mixture of web text, books, code across languages and curated synthetic datasets totaling 18 trillions of tokens. No fine-tuning for Czech was performed. We used the variant with 7 billion parameters which already has the ability to generate texts in the Czech language and its size is feasible for running inference on a GPU with 16 GB of GPU VRAM.

3.2.2 Czech-GPT-2-XL-133k

Czech-GPT-2-XL-133k (*gpt2_XL*) was trained on a 11-billion word corpus of Czech obtained by web crawling. Its architecture was adapted from GPT-2 XL by replacing the tokenizer, corresponding embeddings, and copying over 1,000 English representations corresponding to the 1,000 most frequent tokens into new embeddings based on a bilingual dictionary. (BUT-FIT, 2024).

3.2.3 General GPT-2 Czech

The General GPT-2 Czech model (Chaloupský, 2022) (*gpt2_small*) represents a first-generation model fine-tuned to Czech. It is based on the English pre-trained GPT-2 (Radford et al., 2019), fine-tuned with a Czech tokenizer and word embeddings, and trained on the Czech subset of the OSCAR dataset (Ortiz Su'arez et al., 2020). The dataset is derived from the Web Crawl corpus and the authors filtered the output to only include longer web articles (over 1,200 characters).

3.2.4 n-gram models: NewsCrawl and CNC

For this experiment, two n-gram models were trained to provide contrast to the neural models: *ngram_web* and *ngram_cnc*. The NewCrawl[1] corpus consists of texts crawled between the years 2007 and 2019 from online news sites. The Czech National Corpus (CNC) is the public release of the balanced Czech corpus, where the texts are divided into randomly shuffled chunks of approximately 100 words (respecting sentence boundaries), for copyright reasons[2].

[1] https://data.statmt.org/news-crawl/cs/

[2] https://wiki.korpus.cz/doku.php/en:cnk:syn:verze9

4 Data

4.1 Czech Documents Corpus (*newswire*)

The Czech Text Document Corpus v 2.0 (Král and Lenc, 2018) is a collection of newswires and press releases provided by the Czech News Agency, spanning more than sixty labeled topics. For our experiment, we used a random sample of 500 documents, which we dub *newswire*.

4.2 Prague Dependency Treebank Consolidated 2.0 (*press*)

PDTC 2.0 (Hajič et al., 2024) consists mainly of Czech press texts from the 1990s and Czech translations of the Penn Treebank (Santorini, 1990) (Wall Street Journal articles), with a small proportion of narrative speech transcriptions. It represents a somewhat more diverse press corpus than the Czech Documents Corpus.

4.3 FicTree (*fiction*)

FicTree (Jelínek et al., 2017) consists of eight fiction texts (books) from the Czech National Corpus. The texts are shuffled in random chunks of a maximum of 100 words (respecting sentence boundaries). We divided the train section of its Universal Dependencies release (Zeman et al., 2025) into the original shuffled chunks and randomly selected 300 items.

4.4 Czech Court Decisions Corpus (*justice*)

The Czech Court Decisions Corpus (CzCDC) is a corpus of whole decisions from three top-tier Czech courts (Supreme, Supreme Administrative, and Constitutional court) (Novotná and Harašta, 2019). We used a random sample of 500 documents, which we dub *justice*. The typical text would contain an abundance of file reference numbers, addresses, dates, and numbered headings. New lines are used to emphasize the main points in the decision sentence. The register is extremely formal and practically unreadable for a legal layman (cf. Example 1, machine-translated).

(1) By the defendants decision (hereinafter also referred to as the complainant) of 6 May 2008, Ref. No. OAM347/VL07112008, the claimants application for the granting of international protection was rejected as manifestly unfounded pursuant to Section 16(2) of Act No. 325/1999 Coll., on Asylum and on Amendments to Act No. 283/1991 Coll., on the Police of the Czech Republic, as amended (hereinafter the Asylum Act).
The Regional Court decided on the costs of the proceedings by ruling that none of the parties is entitled to reimbursement thereof.

4.5 KUKY 1.0 (*legal_admin*)

KUKY 1.0 (Cinková et al., 2024) is a corpus of 226 diverse legal and administrative texts, mostly findings and communications of the Office of the Czech Public Defender of Rights, generic legal advice from the Frank Bold law firm, and various court and office filings. In the experiment, it is labeled as *legal_admin*. This corpus contrasts several legal writing styles, the distinction being readability (assessed by experts in plain legal language). Unlike court decisions in CzCDC, these documents also capture communication with the general public, often with observable intentions to be comprehensible to the layman recipient. However, they are still official documents written by legal experts to withstand scrutiny from peers (Example 2, machine-translated).

(2) I am responding to your submission dated 12 September 2011, which concerns the procedure followed by the Ministry of Justice in handling your application for compensation for damage pursuant to the Act on Liability for Damage Caused in the Exercise of Public Authority.
The said file was sent to me on 28 November 2011. In reviewing this file, I focused in particular on the question of whether delays occurred in your matter in decisionmaking under Section 419 of Act No. 40/2009 Coll., the Criminal Code, as amended (hereinafter also referred to as the Criminal Code).

4.6 ParCzech 4.0 (*parl_debates*)

The ParlaCzech 4.0 corpus (Kopp, 2024) is a corpus of transcribed debates in the Czech parliament, corresponding to the Czech subset of the ParlaMint corpora (Erjavec et al., 2025). We used a random sample of 500 items, called *parl_debates*.

4.7 CAC (*academic*)

The Czech Academic Corpus 2.0 (CAC) (Hladká et al., 2008) consists of 650,000 words from various newspapers, magazines, and TV broadcast

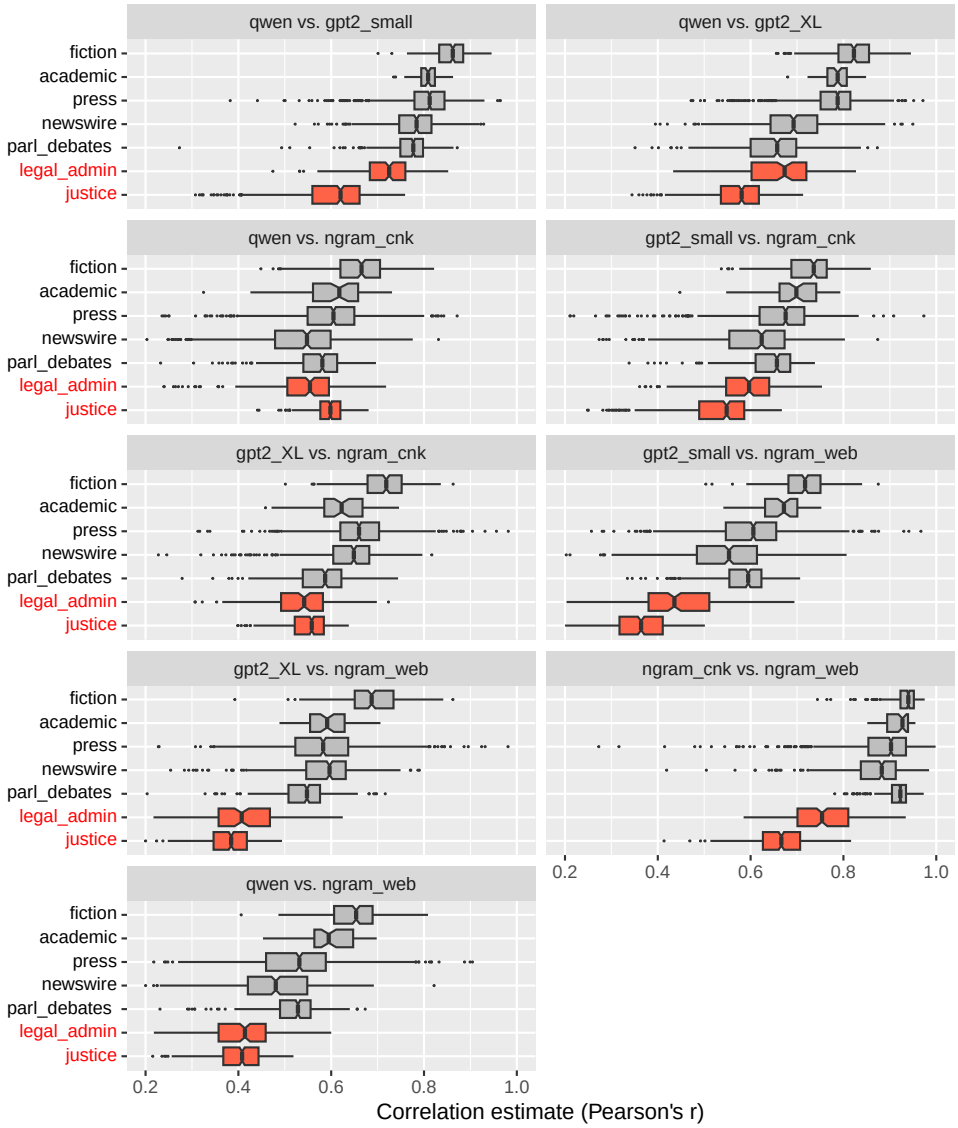


Figure 1: Correlations of lexical surprisal (negative log probability) in model pairs

transcripts gathered by a team of linguists between 1971 and 1985 to reflect the so-called "practical style". The "practical style" ("věcný styl") in Czech stylistics describes the consultative and formal register used in academia and public service journalism. We used all (68) documents labeled as "scientific" by the corpus creators, dubbing this sample *academic*.

5 Results

5.1 Pairwise correlations between models

We first analyzed the consistency of surprisal estimates across different models by examining pairwise correlations between model outputs. For each document within a pair of models, we obtained seven distributions of correlation measurements (Fig. 1). Each subplot represents one pair of models, e.g. *qwen* vs. *gpt2_XL*. Each boxplot represents one genre corpus (labeled on the Y-axis). Each data point underlying a boxplot represents one document in the given genre corpus. Its value is the Pearson correlation estimate of the word-wise aggregated lexical surprisal values by two models.

We observe that the correlations in Fig. 1:

1. are considerably low
2. differ between genres.

The genre-wise correlation distributions are significantly different from each other in each pair of language models (Kruskal-Wallis rank sum test). The post-hoc Dunn's test of pairwise differences

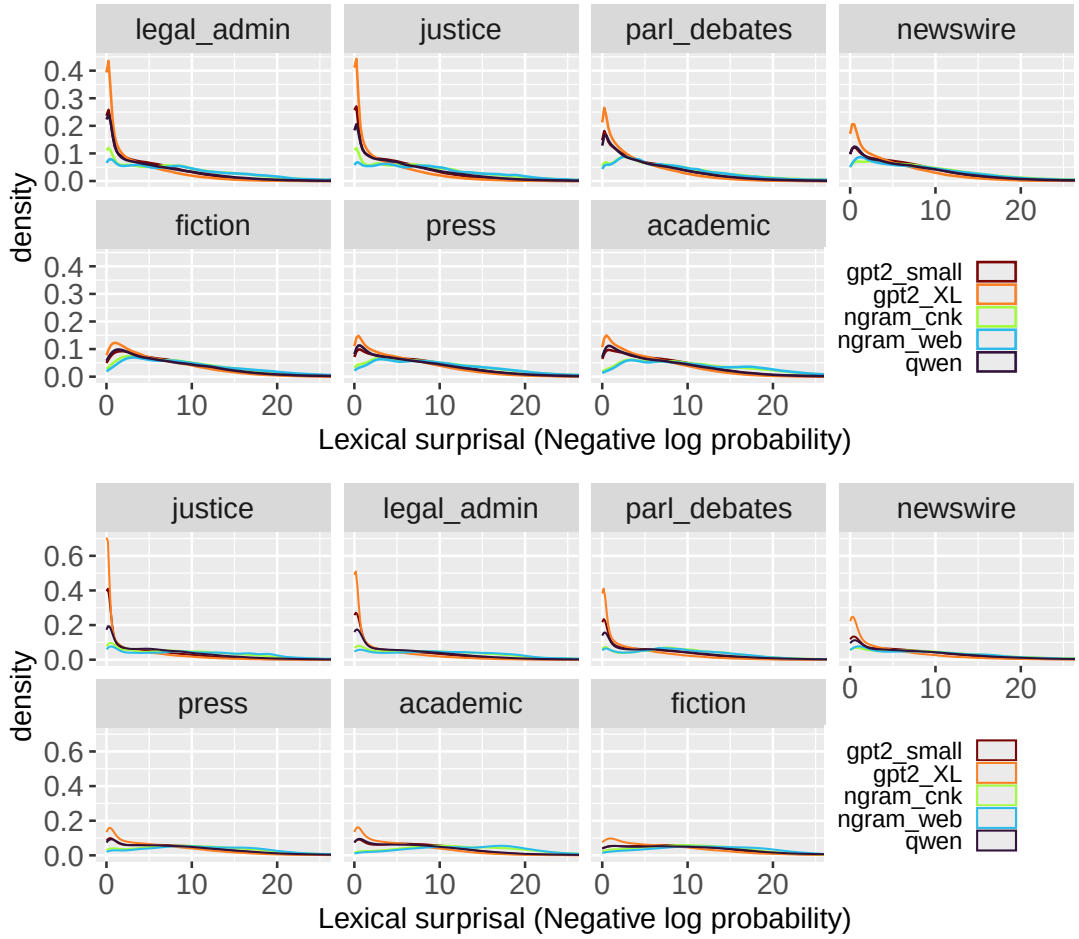


Figure 2: Distributions of negative log probabilities of lexical surprisals of all word types (top) and common and proper nouns (bottom) on individual corpora by different language models

between groups (with Holm's p-value adjustment) reveals statistical significance between all genres except *parl_debates* and *newswire*. This applies to all model pairs. Table 2 shows a prominent effect of genre on the correlation in all model pairs except *qwen* vs. *ngram_cnk*.

The genre ranking was almost consistent across the model pairs: the highest correlation occurred in the *fiction* corpus and the lowest in the corpora *legal_admin* and *justice*. The best correlating pair were the two Czech n-gram models, but even here there occurred a gap between the two legal corpora (max 0.75) and the other genres, which clustered around 0.9.

5.2 Comparison of the individual distributions

Although individual lexical surprisal distributions are all right-skewed across models and genres, a closer look reveals a difference between the large generative models (*gpt2_XL*, *gpt2_small* and

model pair	ε^2	lower ci	upper ci
gpt2_XL/ngram_web	0.560	0.536	0.584
gpt2_small/ngram_web	0.580	0.558	0.602
qwen/ngram_web	0.440	0.414	0.465
ngram_cnk/ngram_web	0.548	0.523	0.571
gpt2_XL/ngram_cnk	0.447	0.420	0.473
gpt2_small/ngram_cnk	0.373	0.345	0.398
qwen/ngram_cnk	0.187	0.159	0.212
qwe/gpt2_XL	0.616	0.592	0.639
qwen/gpt2_small	0.566	0.541	0.589

Table 2: Effect size of corpus (genre) on the correlation between lexical surprisal judgments provided by two models for each document, with a 95% confidence interval

qwen) and the n-gram models. In all corpora, the generative models tend to assess most contexts as highly predictable with long right tails, whereas the n-gram based models exhibit considerably flatter distributions across all genres. The difference is particularly strong in *legal_admin* and *justice*. Among generative models, *gpt2_XL* has the steepest right slope in most genres. The n-gram models have a very high correlation with all models, except *legal_admin* and *justice*.

In Fig. 2, the tails were cut to zoom in on the initial slope. The actual maxima were mostly around 75. The absolute maximum was 122.4 (*gpt2_small* on *justice*). Both the slope and the tail are in this case best reflected by kurtosis (Fig. 3).

In order to better understand the surprisal patterns in text, we examine the distributions by different parts of speech. In particular, we look at how difficult it is for the models to predict nouns and verbs, based on their preceding context. We observe across all models that the distributions of lexical surprisal in nouns largely mirror the complete distributions, with more extreme values in *justice*, *legal_admin*, and *parl_debates* where nouns seem to be over-represented among the most predictable words (see Fig. 2)[3]. This contrasts with expectations regarding readability, as many of these low-surprisal terms consist of technical legal jargon. The effect is most pronounced for the *gpt2_XL*. This result is consistent with Oh and Schuler (2023) who found that the surprisal estimates of nouns and adjectives produced by large Transformer models have a severe tendency to underpredict their reading times.

When analyzing the surprisal of verbs, we observe completely different patterns (see Fig. 4). In particular, the distribution for content verbs exhibits a higher mean and a more symmetric, bell-shaped distribution. In our data, it is, on average, easier for all models to predict nouns and than content verbs. However, this raises the question of whether technical nouns are also easier for humans to process.

The observed high variation in verbs is consistent with the findings of Gennari and Poeppel (2003) on reading times of different semantic classes of verbs in context. Since processing the verb means processing the entire event, verbs naturally carry more information, and the information is even scattered in context, verbs denoting complex events naturally tend to have higher reaction times in self-paced reading. From the perspective of nouns, different verb senses pose different collocability constraints on their noun arguments, which in turn makes the nouns more predictable (e. g., the set of typical objects of *read* is at least conceptually easy to oversee)[4].

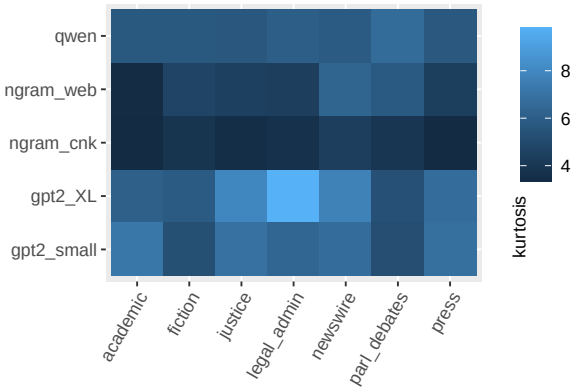


Figure 3: Kurtosis of the distributions of negative log probabilities of lexical surprisals on individual corpora by different language models

The curves of auxiliary (function) words resemble the noun curves, but with substantially lower kurtosis. The difference is particularly prominent in GPT-2-based models on *justice* and *legal_admin*. As auxiliary words typically conform to grammatical structure, their higher predictability is expected. However, the probability density of lexical surprisal for auxiliary words indicates that, in GPT-2-based models, they are still less predictable than nouns, particularly in legal language.

5.3 Correlation between TF-IDF and lexical surprisal

While lexical surprisal approximates the contextual unpredictability of a word given its preceding context, TF-IDF measures its importance within a document relative to a corpus. It is therefore useful to combine the two, as they capture complementary aspects of linguistic information.

We observe weak-to-moderate positive correlation effects of TF-IDF (Silge and Robinson, 2016) and lexical surprisal in n-gram-based models on all corpora. All models exhibit weak-to-moderate positive correlation on *fiction* and, with the exception of gpt2_XL, on *academic* (see Fig. 5). TF-IDF was computed on all word forms converted to

[3]The subplots come in a slightly different order in the two figures, as they are always sorted according to the median of the negative log probability across all models).

[4]see also (Hanks, 2013)

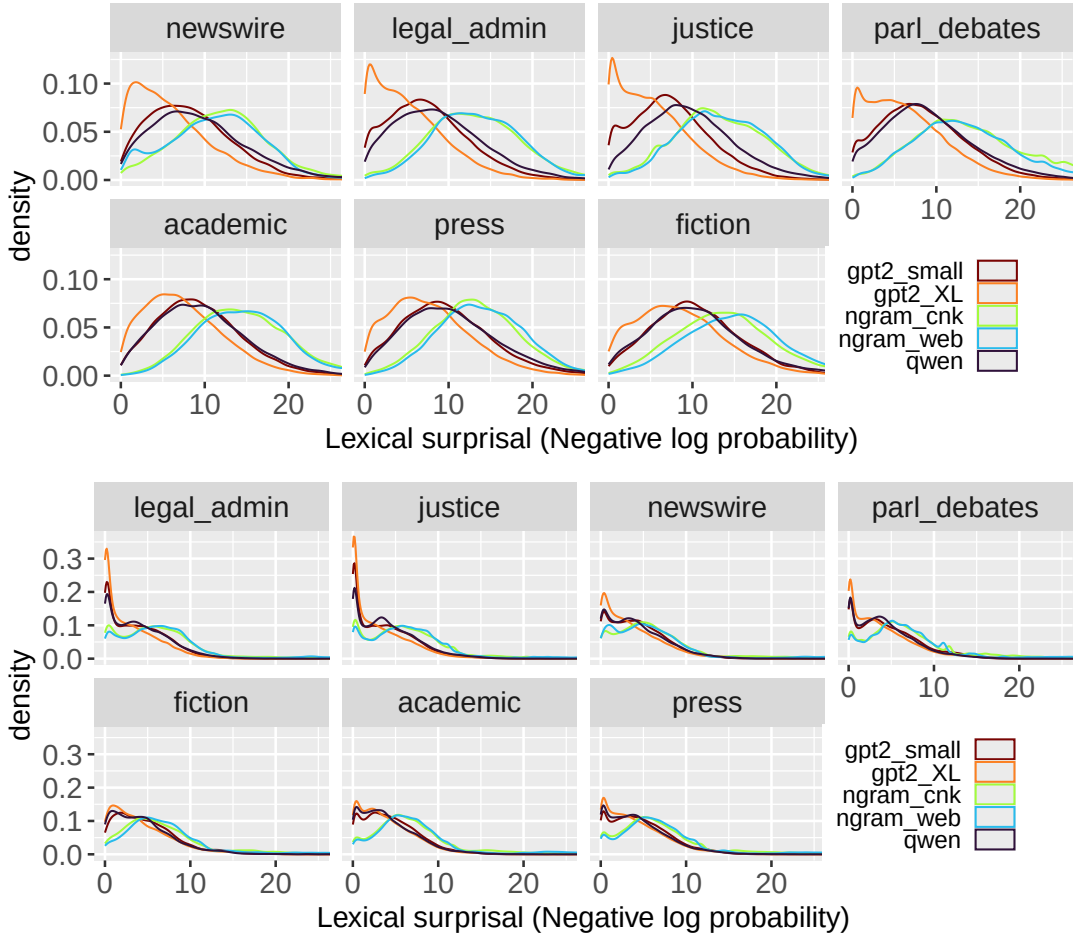


Figure 4: Distributions of negative log probabilities of lexical surprisals of content verbs (top) and lexical surprisals of auxiliary verbs, prepositions, and conjunctions (bottom) on individual corpora by different language models

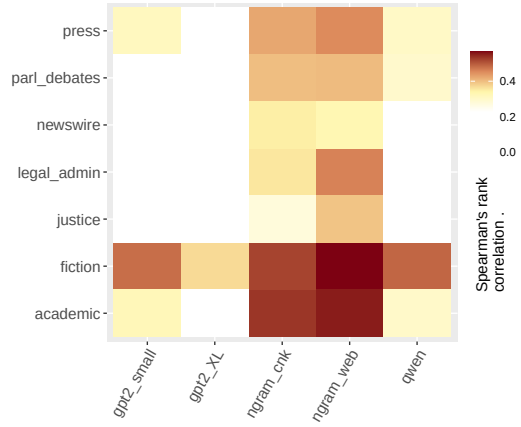


Figure 5: Weak to moderate positive correlation between lexical surprisal (negative log probability) and TF-IDF of all words in combinations of models and corpora

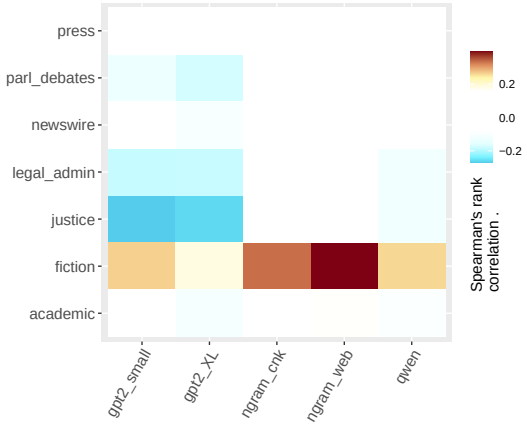


Figure 6: Weak positive and negative correlation effects between lexical surprisal (negative log probability) and TF-IDF of content words in combinations of models and corpora

lower case, including symbols and function words but not punctuation, which had already been filtered away from lexical surprisal values. Each corpus represented one document.

With TF-IDF computed only on content words, excluding stopwords (Cinkova and Eder, 2021), the positive correlation effects in fiction prevailed, whereas a very weak weak negative correlation was observed in judgments by GPT-2-based models in *legal_admin* and *justice* (Fig. 6).

All correlation values (Spearman's rank correlation ρ) in these two experiments were highly statistically significant.

6 Discussion

We observe differing surprisal patterns across genres[5]. Fiction exhibits consistent alignment between surprisal estimates across different models. In contrast, legal and institutional texts show weak correlations between Transformer-based and n-gram models, likely due to the high predictability of formulaic technical language for models extensively exposed to such data during training. Furthermore, technical jargon in legal texts typically consists of content words with high TF-IDF values. The negative correlations shown in Fig. 6 suggest that GPT 2-based models assign particularly low surprisal to these terms.

As each model reflects the characteristics of its training data, the divergence is expected: the n-gram models were trained in controlled settings on a balanced corpus (*ngram_cnk*) and on news-domain text (*ngram_web*) whereas the Transformer models were trained on large-scale data scraped from web. However, in the context of readability estimation, it is crucial to select the model that best approximates human reader behavior. Our findings suggest that model capacity, training data, and domain together play a crucial role in determining the cognitive plausibility of surprisal estimates.

7 Future work

So far, we have explored five models for Czech texts, which represented: large multilingual generative models fine-tuned to Czech, a purely multilingual generative model, and n-gram based models trained on different data sets. If we do not learn about similar work for other languages shortly, we will need to conduct experiments on other languages and comparable datasets. This requires the proper construction of a set of domain-specific corpora for each language, as well as a set of comparable models.

Assuming that we are going to need a human reference, we are preparing an Self-Paced-Reading experiment on samples of the legal and the press corpus. We are going to approximate the lexical surprisal with reading time and hope to find models whose lexical surprisal judgments display a satisfactory correlation with the human reading times.

8 Conclusion

In contrast to previous work, we point out that there are certain kinds of text that are highly predictable for the language models but incomprehensible for human readers, such as *legalese* in Czech. We have gathered evidence that, in Czech texts, there are substantial differences in lexical surprisal judgments by different LLMs and, on top of that, that they depend on genre, with the legal domain being particularly sensitive to model choice. Hence, we assume that the performance of language models is not only language-specific but also specific to a given genre.

9 Acknowledgements

This output was created with the support of the Czech Ministry of Education, Youth and Sports and the Operational Program Johannes Amos Comenius within the project Open Science II (reg. No. CZ.02.01.01/00/24_030/0015041, "HumanAId"). The authors acknowledge the support of the National Recovery Plan funded project MPO 60273/24/21300/21000 CEDMO 2.0 NPO. The work described here has also been using data, tools, and services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062).

References

BUT-FIT. 2024. Czech-gpt-2-xl-133k. https://huggingface.co/BUT-FIT/Czech-GPT-2-XL-133k. Model card published on HuggingFace.

Jan Černý. 2025. *Measuring Lexical Surprisal in Legal Texts*. Bachelor thesis, Charles University, Faculty

[5]We consider genre in a broad sense, disregarding the interplay of style and content domain for the moment.

of Mathematics and Physics, Prague, Czech Republic.

Lukáš Chaloupský. 2022. Automatic generation of medical reports from chest X-rays in Czech. Master's thesis, Charles University, Faculty of Mathematics and Physics, Prague, Czech Republic.

Silvie Cinkova and Maciej Eder. 2021. *tidystopwords: Customisable Stop-Words in 110 Languages*. R package version 0.9.1.

Silvie Cinková, Michal Kuk, Jana Šamánková, Barbora Kubíková, Přemysl Pospíšil, Jiří Mírovský, Barbora Hladká, and Tereza Novotná. 2024. KUKY1.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Thomas Hikaru Clark, Clara Meister, Tiago Pimentel, Michael Hahn, Ryan Cotterell, Richard Futrell, and Roger Levy. 2023. A cross-linguistic pressure for uniform information density in word order. *Transactions of the Association for Computational Linguistics*, 11:1048–1065. Tex.eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00589/2154495/tacl_a_00589.pdf.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Tomaž Erjavec, Matyáš Kopp, Taja Kuzman Pungeršek, Nikola Ljubešić, Maciej Ogrodniczuk, Petya Osenova, Manex Agirrezabal, Tommaso Agnoloni, José Aires, Monica Albini, Jon Alkorta, Iván Antiba-Cartazo, Ekain Arrieta, Mario Barcala, Daniel Bardanca, Starkaður Barkarson, Roberto Bartolini, Roberto Battistoni, Nuria Bel, Maria del Mar Bonet Ramos, María Calzada Pérez, Aida Cardoso, Çağrı Çöltekin, Matthew Coole, Roberts Daris, Ruben de Libano, Griet Depoorter, Sascha Diwersy, Réka Dodé, Kike Fernandez, Elisa Fernández Rei, Francesca Frontini, Marcos Garcia, Noelia García Díaz, Pedro García Louzao, Maria Gavriilidou, Dimitris Gkoumas, Ilko Grigorov, Vladislava Grigorova, Dorte Haltrup Hansen, Mikel Iruskieta, Johan Jarlbrink, Kinga Jelencsik-Mátyus, Bart Jongejan, Neeme Kahusk, Martin Kirnbauer, Anna Kryvenko, Noémi Ligeti-Nagy, Giancarlo Luxardo, Carmen Magariños, Måns Magnusson, Carlo Marchetti, Maarten Marx, Katja Meden, Amália Mendes, Michal Mochtak, Martin Mölder, Simonetta Montemagni, Costanza Navarretta, Bartłomiej Nitoń, Fredrik Mohammadi Norén, Amanda Nwadukwe, Mihael Ojsteršek, Andrej Pančur, Vassilis Papavassiliou, Rui Pereira, María Pérez Lago, Stelios Piperidis, Hannes Pirker, Marilina Pisani, Henk van der Pol, Prokopis Prokopidis, Valeria Quochi, Paul Rayson, Xosé Luís Regueira, Andriana Rii, Michał Rudolf, Manuela Ruisi, Peter Rupnik, Daniel Schopper, Kiril Simov, Laura Sinikallio, Jure Skubic, Lars Magne Tungland, Jouni Tuominen, Ruben van Heusden, Zsófia Varga, Marta Vázquez Abuín, Giulia Venturi, Adrián Vidal Miguéns, Kadri Vider, Ainhoa Vivel Couso, Adina Ioana Vladu, Tanja Wissik, Väinö Yrjänäinen, Rodolfo Zevallos, and Darja Fišer. 2025. Multilingual comparable corpora of parliamentary debates ParlaMint 5.0. Slovenian language resource repository CLARIN.SI.

Irene Fernandez Monsalve, Stefan L. Frank, and Gabriella Vigliocco. 2012. Lexical surprisal as a general predictor of reading time. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 398–408, Avignon, France. Association for Computational Linguistics.

A. Frank and T. Florian Jaeger. 2008. Speaking rationally: Uniform information density as an optimal strategy for language production. *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*, pages 939–944.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019. GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116, Florence, Italy. Association for Computational Linguistics.

Silvia Gennari and David Poeppel. 2003. Processing correlates of lexical semantic complexity. *Cognition*, 89(1):B27–B41.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 10–18, Salt Lake City, Utah. Association for Computational Linguistics.

Jan Hajič, Eduard Bejček, Alevtina Bémová, Eva Buráňová, Eva Fučíková, Eva Hajičová, Jiří Havelka, Jaroslava Hlaváčová, Petr Homola, Pavel Ircing, Jiří Kárník, Václava Kettnerová, Natalia Klyueva, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, David Mareček, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Michal Novák, Petr Pajas, Jarmila Panevová, Nino Peterek, Lucie Poláková, Martin Popel, Jan Popelka, Jan Romportl, Magdaléna Rysová, Jiří Semecký, Petr Sgall, Johanka Spoustová, Milan Straka, Pavel Straňák, Pavlína Synková, Magda Ševčíková, Jana Šindlerová, Jan Štěpánek, Barbora Štěpánková, Josef Toman, Zdeňka Urešová, Barbora Vidová Hladká, Daniel Zeman, Šárka Zikánová, and Zdeněk Žabokrtský. 2024. Prague dependency treebank - consolidated 2.0 (PDT-c 2.0). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North*

American Chapter of the Association for Computational Linguistics.

Patrick Hanks. 2013. *Lexical Analysis: Norms and Exploitations*. MIT Press.

Barbora Vidová Hladká, Jan Hajič, Jiří Hana, Jaroslava Hlaváčová, Jiří Mírovský, and Jan Raab. 2008. Czech academic corpus 2.0.

Tomáš Jelínek, Milena Hnátková, and Hana Skoumalová. 2017. FicTree 1.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Matyáš Kopp. 2024. ParCzech 4.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Pavel Král and Ladislav Lenc. 2018. Czech text document corpus v 2.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Tereza Novotná, Jan Černý, Ivan Kraus, Ivana Kvapilíková, Jiří Mírovský, Arnold Stanovský, and Barbora Hladká. 2025. PONK: Tool for client-oriented legal writing in czech. In *JURIX 2025: The Thirty-eighth Annual Conference*, volume 416 of *Frontiers in Artificial Intelligence and Applications*, pages 330–336, Amsterdam, Netherlands. University of Turin, Italy, IOS Press.

Tereza Novotná and Jakub Harašta. 2019. Czech court decisions corpus (CzCDC 1.0). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Byung-Doh Oh and William Schuler. 2023. Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1915–1921, Singapore. Association for Computational Linguistics.

Pedro Javier Ortiz Su'arez, Laurent Romary, and Benoit Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Tiago Pimentel and Clara Meister. 2024. How to compute the probability of a word. In *Conference on empirical methods in natural language processing*.

Tiago Pimentel, Clara Meister, Ethan G. Wilcox, Roger P. Levy, and Ryan Cotterell. 2023. On the effect of anticipation on reading times. *Transactions of the Association for Computational Linguistics*, 11:1624–1642.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Beatrice Santorini. 1990. Part-of-Speech Tagging Guidelines for the Penn Treebank Project. *University of Pennsylvania 3rd Revision 2nd Printing*, (MS-CIS-90-47):33. 00376 tex.ids= santoriniPartofSpeechTaggingGuidelines1990a.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423.

Julia Silge and David Robinson. 2016. tidytext: Text mining and analysis using tidy data principles in r. *JOSS*, 1(3).

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Milan Straka, Jan Haji, and Jana Straková. 2016. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297. European Language Resources Association (ELRA). Place: Portoro, Slovenia.

K. Sun and R. Wang. 2025. Computational sentence-level metrics of reading speed and its ramifications for sentence comprehension. *Cognitive Science*, 49(7):e70092.

Ethan G. Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Jephtey Adolphe, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Arofat Akhundjanova, Furkan Akkurt, Gabrielė Aleksandravičiūtė, Ika Alfina, Avner Algom, Khalid Alnajjar, Chiara Alzetta, Antonios Anastasopoulos, Erik Andersen, Kirk Andrews, Matthew Andrews, Lene Antonsen, Tatsuya Aoyama, Katya Aplonova, Angelina Aquino, Carolina Aragon, Glyd Aranes, Maria Jesus Aranzabe, Bilge Nas Arıcan, Hórunn Arnardóttir, Wirote

Aroonmanakun, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Hiwa Asadpour, Masayuki Asahara, Katla Ásgeirsdóttir, Deniz Baran Aslan, Cengiz Asmazoğlu, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Mariana Avelãs, Elena Badmaeva, Jana Bajorat, Keerthana Balasubramani, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Bryan Khelven da Silva Barbosa, Verginica Barbu Mititelu, Starkaður Barkarson, Rodolfo Basile, Victoria Basmov, Colin Batchelor, John Bauer, Seyyit Talha Bedir, Shabnam Behzad, Nathanaël Beiner, Juan Belieni, Alevtina Bémová, Kepa Bengoetxea, brahim Benli, Yifat Ben Moshe, Marie Benzerrak, Aleksandrs Berdicevskis, Ansu Berg, Gözde Berk, Delphine Bernhard, Astrid Berntsson Ingelstam, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Esma Fatıma Bilgin Taşdemir, Helin Binici, Kristín Bjarnadóttir, Verena Blaschke, Rogier Blokland, Nina Böbel, Victoria Bobicev, Loïc Boizou, Stavros Bompolas, Johnatan Bonilla, Emanuel Borges Völker, Lars Borin, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Anouck Braggaar, António Branco, Myriam Bras, Théo Brillet, Kristina Brokaitė, Lanni Bu, Eva Buráňová, Aljoscha Burchardt, Carmen Cabeza, Natalia Cáceres Arandia, Olesea Caftanatov, Marisa Campos, Marie Candito, Caterina Maria Cappello, Bernard Caron, Gauthier Caron, Catarina Carvalheiro, Rita Carvalho, Lauren Cassidy, Maria Clara Castro, Sérgio Castro, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Anila Çepani, Slavomír Čéplö, Neslihan Cesur, Savas Cetin, Özlem Çetinoğlu, Fabricio Chalub, Liyanage Chamila, Claudine Chamoreau, Shweta Chauhan, Yifei Chen, Ethan Chi, Taishi Chika, Yongseok Cho, Jinho Choi, Bermet Chontaeva, Jayeol Chun, Juyeon Chung, Alessandra T. Cignarella, Silvie Cinková, Esther Cocco, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Claudia Corbetta, Daniela Corbetta, Francisco Costa, Marine Courtin, Benoît Crabbé, Mihaela Cristescu, Vladimir Cvetkoski, Netanel Dahan, Ingerid Løyning Dale, Sabrina D'Alì, Philemon Daniel, Khensa Daoudi, Bijayalaxmi Dash, Satya Ranjan Dash, Elizabeth Davidson, Leonel Figueiredo de Alencar, Mathieu Dehouck, Martina de Laurentiis, Marie-Catherine de Marneffe, Ahmet Demir, Valeria de Paiva, Mehmet Oguz Derin, Elvis de Souza, Arantza Diaz de Ilarraza, Roberto Antonio Díaz Hernández, Carly Dickerson, Ariani Di Felippo, Arawinda Dinakaramani, Elisa Di Nuovo, Bamba Dione, Peter Dirix, Hoa Do, Kaja Dobrovoljc, Caroline Döhmer, Adrian Doyle, Timothy Dozat, Kira Droganova, Magali Sanches Duran, Puneet Dwivedi, Christian Ebert, Hanne Eckhoff, Masaki Eguchi, Sandra Eiche, Roald Eiselen, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Louise Esher, Soudabeh Eslami, Farah Essaidi, Aline Etienne, Wograine Evelyn, Sidney Facundes, Richárd Farkas, Ján Faryad, Federica Favero, Jannatul Ferdaousi, Marília Fernanda, Hector Fernandez Alcalde, Amal Fethi, Jennifer Foster, Barbara Francioni, Theodorus Fransen, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Edith Galy, Federica Gamba, Marcos Garcia, José María García-Miguel, Moa Gärdenfors, Tanja Gaustad, Efe Eren Genç, Fabrício Ferraz Gerardi, Kim Gerdes, Luke Gessler, Filip Ginter, Gustavo Godoy, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Gili Goldin, Xavier Gómez Guinovart, Berta González Saavedra, Mathieu Goux, Bernadeta Griciūtė, Matias Grioni, Loïc Grobol, Normunds Grūzītis, Mario Guglielmetti, Bruno Guillaume, Kirian Guiller, Céline Guillot-Barbance, Tunga Güngör, Vladimir Gurevich, Nizar Habash, Hinrik Hafsteinsson, Michael Hahn, Jan Hajič, Jan Hajič jr., Eva Hajičová, Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Muhammad Yudistira Hanifmuti, Takahiro Harada, Sam Hardwick, Kim Harris, Naïma Hassert, Dag Haug, Jiří Havelka, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Diana Hoefels, Barbara Hoff, Petter Hohle, Nick Howell, Yidi Huang, Marivel Huerta Mendez, Jena Hwang, Takumi Ikeda, Inessa Iliadou, Anton Karl Ingason, Radu Ion, Elena Irimia, Ọlájídé Ishola, Artan Islamaj, Kaoru Ito, Federica Iurescia, Jessica K. Ivani, Sandra Jagodzińska, Siratun Jannat, Tomáš Jelínek, Apoorva Jha, Ratanon Jiamsundutsadee, Katharine Jiang, Sylvanus Job, Mayank Jobanputra, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Zhuoxuan Ju, Markus Juutinen, Hüner Kaşıkara, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Neslihan Kara, Ritván Karahóa, Jiří Kárník, Andre Kåsen, Tolga Kayadelen, Sarveswaran Kengatharaiyer, Václava Kettnerová, Lilit Kharatyan, Jesse Kirchner, Elena Klementieva, Elena Klyachko, Petr Kocharov, Arne Köhn, Abdullatif Köksal, Veronika Kolářová, Kamil Kopacewicz, Timo Korkiakangas, Mehmet Köse, Alexey Koshevoy, Nelda Kote, Natalia Kotsyba, Barbara Kovačić, Jolanta Kovalevskaitė, Emmanuelle Kowner, Simon Krek, Parameswari Krishnamurthy, Sandra Kübler, Lucie Kučová, Adrian Kuqi, Elmurod Kuriyozov, Oğuzhan Kuyrukçu, Aslı Kuzgun, Sookyoung Kwak, Kris Kyle, Käbi Laan, Veronika Laippala, Lorenzo Lambertino, Israel Landau, Tatiana Lando, Septina Dian Larasati, Pierre Larrivée, Kusum Lata, Alexei Lavrentiev, John Lee, Phng Lê Hồng, Alessandro Lenci, Wei Qi Leong, Saran Lertpradit, Herman Leung, Lori Levin, Maria Levina, Lauren Levine, Cheuk Ying Li, Josie Li, Keying Li, Yixuan Li, Yuan Li, KyungTae Lim, Bruna Lima Padovani, Yi-Ju Jessica Lin, Krister Lindén, Yang Janet Liu, Zoey Liu, Nikola Ljubešić, Irina Lobzhanidze, Olga Loginova, Markéta Lopatková, Lucelene Lopes, Edita Luftiu, Arsenii Lukashevskyi, Stefano Lusito, Anne-Marie Lutgen, Andry Luthfi, Mikko Luukko, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Menel Mahamdi, Jean Maillard, Punyanuch Maitreenukul, Ilya Makarchuk, Aibek Makazhanov, Francesco Mambrini, Michael Mandl, Christopher Manning, Ruli Manurung, Büşra Marşan, Cătălina Mărănduc, David Mareček,

Katrin Marheinecke, Stella Markantonatou, Héctor Martínez Alonso, Lorena Martín Rodríguez, André Martins, Cláudia Martins, Arianna Masciolini, Jan Mašek, Sanatbek Matlatipov, Hiroshi Matsuda, Yuji Matsumoto, Caterina Mauri, Alessandro Mazzei, Ryan McDonald, Sarah McGuinness, Maitrey Mehta, Pierre André Ménard, Gustavo Mendonça, Hilla Merhav, Tatiana Merzhevich, Paul Meurer, Niko Miekka, Marie Mikulová, Emilia Milano, Aleksandra Miletić, Aaron Miller, Junghyun Min, Yael Minerbi, Jiří Mírovský, Karina Mischenkova, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, Biswakalpita Mohapatra, AmirHossein Mojiri Foroushani, Judit Molnár, Amirsaeid Moloodi, Simonetta Montemagni, Amir More, Laura Moreno Romero, Giovanni Moretti, Shinsuke Mori, Tomohiko Morioka, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Katerina Mouzou, Kadri Muischnek, Robert Munro, Yugo Murawaki, Nikolett Mus, Kaili Müürisep, Pinkey Nainwani, Mariam Nakhlé, Minoo Nassajian, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Manuela Nevaci, Lng Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Victor Norrman, Alireza Nourian, Michal Novák, Maria das Graças Volpe Nunes, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Hulda Óladóttir, Adédayọ̀ Olúòkun, Mai Omura, Emeka Onwuegbuzia, Noam Ordan, Petya Osenova, Robert Östling, Annika Ott, Lilja Øvrelid, Masanori Oya, Şaziye Betül Özateş, Merve Özçelik, Arzucan Özgür, Balkız Öztürk Başaran, Teresa Paccosi, Petr Pajas, Thomas Palakapilly, Alessio Palmero Aprosio, Jarmila Panevová, Ludovica Pannitto, Anastasia Panova, Thiago Alexandre Salgueiro Pardo, Shantipriya Parida, Hyunji Hayley Park, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Giulia Pedonese, Oggi Peeters, Angelika Peljak-Łapińska, Siyao Peng, Siyao Logan Peng, Rita Pereira, Sílvia Pereira, Cenel-Augusto Perez, Natalia Perkova, Guy Perrier, Slav Petrov, Daria Petrova, Eva Pettersson, Andrea Peverelli, Jason Phelan, Claudel Pierre-Louis, Jussi Piitulainen, Yuval Pinter, Clara Pinto, Rodrigo Pintucci, Tommi A Pirinen, Emily Pitler, Magdalena Plamada, Barbara Plank, Alistair Plum, Thierry Poibeau, Charin Polpanumas, Larisa Ponomareva, Martin Popel, Clamença Poujade, Rangga Prangwedana Prangwedana, Lauma Pretkalniņa, Rigardt Pretorius, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Robert Pugh, Tiina Puolakainen, Christoph Purschke, Sampo Pyysalo, Peng Qi, Andreia Querido, Andriela Rääbis, Ella Rabinovich, Alexandre Rademaker, Mutee-u Rahman, Mizanur Rahoman, Taraka Rama, Loganathan Ramasamy, Carlos Ramisch, Joana Ramos, Fam Rashel, Mohammad Sadegh Rasooli, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Mathilde Regnault, Georg Rehm, Arij Riabi, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Putri Rizqiyah, Luisa Rocha, Eiríkur Rögnvaldsson, Ivan Roksandic, Norton Trevisan Roman, Mykhailo Romanenko, Natalia Romanova, Rudolf Rosa, Valentin Roca, Paulette Roulon, Davide Rovati, Ben Rozonoyer, Olga Rudina, Jack Rueter, Paolo Ruffolo, Kristján Rúnarsson, Rozana Rushiti, Attapol T. Rutherford, Shoval Sadde, Pegah Safari, Aleksi Sahala, Kalyanamalini Sahoo, Saraswati Sahoo, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Konstantinos Sampanis, Stephanie Samson, Xulia Sánchez-Rodríguez, Manuela Sanguinetti, Ezgi Sanıyar, Dage Särg, Marta Sartor, Albina Sarymsakova, Mitsuya Sasaki, Baiba Saulīte, Agata Savary, Yanin Sawanakunanon, Shefali Saxena, Kevin Scannell, Salvatore Scarlata, Emmanuel Schang, Robert Schikowski, Nathan Schneider, Sebastian Schuster, Lane Schwartz, Djamé Seddah, Wolfgang Seeker, Sven Sellmer, Mojgan Seraji, Magda Ševčíková, Petr Sgall, Syeda Shahzadi, Mo Shen, Atsuko Shimada, Gyu-Ho Shin, Hiroyuki Shirasu, Yana Shishkina, Muh Shohibussirri, Maria Shvedova, Jean Sibille, Janine Siewert, Einar Freyr Sigurðsson, João Silva, Aline Silveira, Natalia Silveira, Sara Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Haukur Barri Símonarson, Kiril Simov, Dmitri Sitchinava, Ted Sither, Aaron Smith, Isabela Soares-Bastos, Per Erik Solberg, Dolores Sollberger, Barbara Sonnenhauser, Shafi Sourov, Nina Speransky, Rachele Sprugnoli, Panyut Sriwirote, Vivian Stamou, Steinþór Steingrímsson, Antonio Stella, Jan Štěpánek, Barbora Štěpánková, Abishek Stephen, Milan Straka, Omer Strass, Emmett Strickland, Jana Strnadová, Sara Stymne, Alane Suhr, Yogi Lesmana Sulestio, Umut Sulubacak, Jack Sun, Hakyung Sung, Shingo Suzuki, Daniel Swanson, Zsolt Szántó, Maria Irena Szawerna, Chihiro Taguchi, Dima Taji, Luigi Talamo, Fabio Tamburini, Mary Ann C. Tan, Takaaki Tanaka, Dipta Tanaya, Mirko Tavoni, Nursena Teker, Samson Tella, Isabelle Tellier, Marinella Testori, Santhawat Thanyawong, Guillaume Thomas, Tarık Emre Tıraş, William Chandra Tjhi, Thea Tollersrud, Kamil Tomaszek, Sara Tonelli, Liisi Torga, Lucas Toribio, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Kira Tulchynska, Utku Türk, Francis Tyers, Sveinbjörn Þórðarson, Vilhjálmur Þorsteinsson, Sumire Uematsu, Roman Untilov, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Elena Vagnoni, Sowmya Vajjala, Socrates Vak, Socrates Vakirtzian, Rob van der Goot, Martine Vanhove, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Helena Vaz, Uliana Vedenina, Giulia Venturi, Marianne Vergez-Couret, Annemarie Verkerk, Barbora Vidová Hladká, Eric Villemonte de la Clergerie, Veronika Vincze, Anishka Vissamsetty, Natalia Vlasova, Eleni Vligouridou, Aya Wakasa, Joel C. Wallenberg, Lars Wallin, Abigail Walsh, John Wang, Jonathan North Washington, Leonie Weissweiler, Maximilan Wendt, Paul Widmer, Aleksandra Wieczorek, Shira Wigderson, Sri Hartati Wijono, Vanessa Berwanger Wille, Seyi Williams, Miriam Winkler, Shuly Wintner, Mats Wirén, Christian Wittern, Alena Witzlack-Makarevich, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska,

Qishen Wu, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Qizhen Yang, Xiulin Yang, Koichi Yasuoka, Marat M. Yavrumyan, Arife Betül Yenice, Enes Yılandiloğlu, Olcay Taner Yıldız, Zhuoran Yu, Arlisa Yuliawati, Zdeněk Žabokrtský, Shorouq Zahra, Amir Zeldes, Annie Zhang, Larry Zhang, He Zhou, Hanzhi Zhu, Yilun Zhu, Anna Zhuravleva, Rayan Ziane, Artūrs Znotiņš, and Eleonora Zucchini. 2025. Universal dependencies 2.17. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Analysing French Narratives Generated by ChatGPT

Natalia Dankova
Université du Québec en Outaouais,
Canada
Natalia.dankova@uqo.ca

Abstract

This study analyzes a corpus of 200 French narrative texts generated by ChatGPT-5 in response to prompts specifying the characters, context, and story ending. Research on AI-generated texts in French remains limited. The methodology combines natural language processing with discourse analysis, enabling comparison with studies of human-authored narratives. Both quantitative and qualitative analyses focus on discursive organization—narrative structure, key event types, character attributes, and use of direct speech—as well as syntactic, grammatical, and lexical features, with particular attention to adverbs, adjectives, and intensifiers. This paper presents selected findings from a comparative perspective on ChatGPT- and human-generated narratives. The study contributes to a growing yet still limited body of research on AI-generated texts in French and provides empirical insights into similarities and differences between human- and AI-generated narratives.

Keywords: artificial intelligence storytelling in French; artificial intelligence narratives construction; natural language process; artificial intelligence narratives linguistic analysis.

1 Introduction

In recent years, artificial intelligence has made significant progress in generating various types of texts and has sparked growing interest among researchers in detecting AI-produced texts and analyzing the differences between texts generated by AI and those produced by humans (Bellini et al., 2024; Fiedler and Döpke, 2025; Kehkashan et al., 2025; Prova, 2024). Bellini et al. (2024) mention that "AI writing styles become increasingly indistinguishable from human-generated content, ensuring their transparency and originality will become more challenging, especially in critical fields like scientific research" (Bellini, 2024: 357). The rapid development of AI has transformed the field of literary creation, and there, AI appears less performant. Numerous studies are dedicated to analyzing AI-produced literary output and to comparing narrative texts produced by AI versus those produced by humans. (Raffloer and Green, 2025; Munir et al., 2025; Begum, 2025; Sofiya et al., 2025; Beguš. 2024). Although AI manages to imitate humans' narrative style, its stories still lack depth, emotional nuances, cultural and psychological realities, authenticity, originality, and literary techniques (imaginative scenarios, characters' nuances, suspense or twists, complexity, and narrative coherence). IA is also criticized for drawing simplistic and moralizing conclusions.

2 Data and method

ChatGPT produced 200 narratives based on the following prompt: "Invent a story on the following theme: a cat and a dog living in neighboring houses hate each other, but by the end of the story they become friends." The text length was limited to a maximum of 500 words, but ChatGPT did not always adhere to the imposed length limit. ChatGPT was asked to explain its approach to constructing a narrative text. It defined the steps as follows:

1. Understand the prompt, defining the relationships between characters, the tone of the story, and the key events.
2. Define the characters whose conflict stems from their differences. The cat is solitary, somewhat disdainful, but possesses a certain wisdom. It is a calm and thoughtful character. The dog is very energetic, slightly naïve and clumsy, but kind-hearted.
3. Choose a triggering event.
4. Develop the evolution of the relationship.

5. Find a satisfying conclusion: the story should end on a positive and resolved note, showing that the characters have grown.
6. Maintain an appropriate tone: for a short and light story like this, the tone is rather humorous and friendly.
7. Adhere to the length and structure.

This approach is relatively conventional and closely follows the given prompt. ChatGPT chooses to frame the characters' differences as the source of their conflict, without exploring alternative motivations (e.g., a past event or competition). It adopts a positive, humorous, and friendly tone, thereby avoiding other narrative possibilities.

The corpus contains 101010 words. The median length of the stories is 513.5 words, and the average length is 505.05 words.

The methodology combines natural language processing with discourse analysis, enabling comparison with studies of human-authored narratives. Both quantitative and qualitative analyses focus on discursive organization—narrative structure, key event types, character attributes, and use of direct speech—as well as syntactic, grammatical, and lexical features, with particular attention to adverbs, adjectives, and intensifiers. In the present article, I limit myself to the analysis of discursive organization, temporal adverbials, adjectives, and intensifiers.

2.1 Temporal adverbials

Time is one of the most fundamental conceptual categories in human cognition, and natural languages have different morphological and lexical devices to express it: such as tense, aspect, adverbials, particles, and principles of discourse organisation as well. Each language combines them in a specific way.

Most languages have temporal adverbials[1]. For analytical purposes, the classification proposed by Klein (1994) has been adopted. The author distinguishes six categories based on function rather than form: positional adverbials (*today*, *three weeks ago*), adverbials of sequentiality (*then*, *finally*), adverbials of duration (*for a long time*, *in a few days*), adverbials of frequency (*often*, *twice*), adverbials such as *suddenly* and *little by little*, which express the nature of the transition from one event to another; and a heterogeneous group of adverbials used to express contrast between events (*almost*, *still*, *already*). Dankova (2002, 2009, 2017) demonstrated that the use of different types of temporal adverbials varies across languages (and even across varieties of the same language), as well as according to the nature of narrative texts.

2.2 Adjectives and intensifiers

Adjectives play an important role in narrative texts. Descriptive or literary texts are known to contain a higher number of adjectives. For French, there is no consensus regarding the classification of adjectives and other noun determiners that function as adjectives, such as possessives, demonstratives, numerals, and interrogatives (Elalouf, 2009; Wilmet, 1997). When discussing adjectives, I focus primarily on their functions rather than their forms, since only a subset of adjectives constitutes a class (or part of speech) known as adjectives such as *beau* 'beautiful', *vert* 'green', *nouveau* 'new'. There are also adjectives derived from past participles (e.g., *assiette cassée* 'broken plate', *objet perdu* 'lost object') or present participles (*voyage fatigant* 'tiring trip', *histoire amusante* 'amusing story'), as well as nouns used adjectivally (*industrie automobile* 'automotive industry', *bouteille en verre* 'glass bottle'). For analytical purposes, adjectives were counted in their base form (masculine singular).

Qualifying adjectives can be accompanied by intensifiers that are words (or expressions) used to strengthen or weaken the degree of an adjective: *very interesting* or *rarely used*. They can also nuance the meaning of an adjective by adding an evaluation or a clarification: *unnecessarily expensive* or *dangerously close*.

The quantitative and qualitative analysis of the linguistic features employed by ChatGPT in its narratives, such as temporal adverbials, adjectives, and intensifiers, will help reveal patterns of textual organization and the distribution of the linguistic resources deployed by the AI, while allowing for an empirical comparison with narrative texts produced by humans.

[1] The terms adverb and adverbial are used here without distinction.

3 Results and discussion

3.1 Discursive organization

The discursive organization of the narratives remains the same: all the texts share the same structure, as if they were using the same template. ChatGPT uses the *passé simple* (simple past), a verb tense reserved for literary texts. Few people are able to use the literary past tense correctly in writing without having to check verb forms; they tend to use the *passé compose* (perfect past tense) instead, since it is used both orally and in writing. ChatGPT-generated narratives follow a linear chronological structure, without flashbacks or foreshadowing. The first paragraph describes a peaceful neighborhood, two houses facing each other, and two main characters: a cat and a dog. The second paragraph is devoted to the animals' daily routine. This is followed by the key event of the story, which leads to a change in the relationship between the two animals. ChatGPT invariably concludes on a positive and moralizing note, centered on the possibility of living together despite differences and on understanding others. An example of a story is shown in Appendix A.

The key event of the story is stated explicitly and without suspense: one day, an unexpected event occurred (see Appendix B). It involves a storm and/or an accident affecting one of the main characters or a third protagonist (a bird, a kitten). ChatGPT mentions a loud or unusual noise accompanying the key event.

ChatGPT-generated narratives tend to follow canonical story structures, with an introduction, a development, and a resolution, while making temporal and causal relations more explicit. They display regular referential cohesion, clear structural coherence, and frequent use of evaluative expressions and mental state descriptions. One might have expected the narrative not to follow a chronological order or to begin with a salient event from which the rest of the story unfolds (as, for example, in detective novels). One could also have expected one of the characters to assume the role of narrator and recount the story from a first-person perspective. The story of the cat and the dog could be embedded within a first-person personal narrative (using the author's voice), for example: "When I moved to a new city, I decided to adopt a cat, but I was unaware that my neighbors had a fierce dog".

Eighty percent of the narratives contain direct speech consisting of short exchanges between characters. Direct speech most often occurs after a key event in the narrative and consists of expressions of recognition of the other character's efforts, gratitude, and/or compliments. Negative utterances in direct speech, such as *tu n'es qu'un chien stupide!* 'you are a stupid dog', are extremely rare. No swear words appear in the narratives. ChatGPT consistently avoids offensive language.

ChatGPT tends to use complete and often complex sentence structures. From a syntactic perspective, incomplete sentences could have been used to create a particular rhythm and literary effect. ChatGPT behaves, so to speak, like a diligent student who produces well-structured work but lacks imagination and narrative creativity, while avoiding any risk of being perceived as unconventional.

3.2 Temporal adverbials

All the narratives contain temporal adverbials. On average, a narrative contains 15.2 temporal adverbials; that is, on average, one adverb every 33.23 words. An overuse of temporal adverbials is noted in certain passages; this is the case in the following excerpt: there are 7 occurrences in a 55-word paragraph (the temporal adverbials are shown in italics):

(1) *Au bout de quelques minutes*, *après de nombreux efforts*, Félix réussit à sortir de la neige grâce à l'aide de Max. Ils restèrent là, pantelants, dans le froid glacial, *pendant quelques secondes*, *avant* de se regarder *longuement*. Aucun mot n'était nécessaire. Les deux animaux, bien qu*'encore* un peu méfiants, se comprenaient *désormais* d'une manière différente.
After a few minutes, *and after much effort*, Félix managed to get out of the snow with Max's help. They stayed there, panting in the freezing cold *for a few seconds*, *before* looking at each other *for a long moment*. No words were necessary. The two animals, though *still* a little wary, *now* understood each other in a different way.

For the entire corpus, the distribution across the six types of temporal adverbials is as follows (in percentages):

Position	29.6
Sequentiality	11.8
Duration	6.9
Frequency	31.3
Nature of transition	11.8
Contrast	8.6
Total	100%

Table 1: Distribution of temporal adverbials (in percentages)

Adverbs of position and frequency were used in all the narratives, and these are the most frequently used in the corpus, respectively 29.6% and 31.3%.

The adverbs of frequency that express a high frequency (e.g., *toujours* 'always', *chaque matin* 'every morning') are twice as numerous as those expressing a low frequency (*parfois* 'sometimes', *jamais* 'never'). Among the most frequently used adverbs are *toujours* 'always', for the first category, and *parfois* 'sometimes' as well as expressions composed of *chaque* 'each' plus a noun such as *fois* 'time', *jour* 'day', *matin* 'morning', or *soir* 'evening', for the second.

The discursive organization and the distribution of temporal adverbs depend on both the language and the type of text (Lambert, 1997; Dankova, 1997, 2002, 2009, and 2017). It has been shown that positional adverbials and adverbials of sequentiality that place events on the temporal axis are the most frequent in narrative texts. Adverbs indicating transitions from one event to another (*suddenly*, *little by little)* and contrastive adverbs (*already*, *still*, *yet*) contribute to the expression of internal temporality and the dynamism between different events and protagonists.

For the moment, it is not possible to compare the narratives produced by ChatGPT with narratives produced by humans from the same prompt (this will probably be an avenue to explore for future research). Nevertheless, there are data obtained from analyses of two corpora of spoken French narrative texts (cf. Dankova 2002, 2009, 2017, 2022). As part of my previous research, I created two French corpora: one consisting of narratives elicited using the *Cat Story* picture sequence and the other consisting of narratives based on the film *Quest* (1996). The *Cat Story* picture sequence was developed by Hickmann (1982) and has been widely used in second language acquisition research (see Appendix D). Thus, ChatGPT used fewer position and sequential adverbs than humans and fewer transition and contrast adverbs than humans, but it is the rates of frequency adverbs that most separate AI from humans: 31.3% for ChatGPT versus 10.4% and 4.1% for humans (Table 2). For human participants, the first column contains data from film narratives, while the second contains data from narratives produced based on *Cat Story*.

Temporal adverbials:	ChatGPT	Humans	
Position + Sequentiality	41.4	56.2	66.2
Duration	6.9	3.8	2.7
Frequency	31.3	10.4	4.1
Transition + Contrast[2]	20.4	29.6	27
Total	100%	100%	100%

Table 2: Temporal adverbials: ChatGPT vs Humans (in percentages)

Without drawing conclusions from these comparisons, I could formulate a hypothesis that it is the distribution of features used (adverbs, adjectives, intensifiers, etc.) that would allow differentiating narrative texts produced by AI from those produced by humans.

3.3 Adjectives

The corpus contains 7220 adjectives in total, the ratio being 0.0715[3]. A narrative contains on average 36.1 adjectives, with a median of 36. The mean (36.1) and the median (36) are almost identical, so the distribution is overall symmetric. The number of distinct adjectives per narrative is 31.75 on average (the median is 31.5); this distribution is also almost symmetric. Adjectives are consistently used in each paragraph. Here is an excerpt from a narrative illustrating the use of adjectives (the adjectives are italicized):

(2) Un silence s'installa entre eux, mais il n'était plus aussi *lourd* qu'avant. Le chat, toujours *fier* mais plus *calme*, s'étira.

[2] Transition and contrast adverbials, which serve relatively similar functions, are presented together.

[3] The ratio is the number of adjectives divided by the total number of tokens.

Rocky, avec un *petit* sourire, se remit à grignoter ses friandises, cette fois un peu plus lentement, et, à sa *grande* surprise, Tigris s'approcha pour en goûter une. Après tout, elles étaient bien *meilleures* que ce qu'il avait imaginé.
A silence settled between them, but it was no longer as *heavy* as before. The cat, still *proud* but *calmer*, stretched. Rocky, with a *faint* smile, resumed nibbling on his treats, this time a little more slowly, and, to his *great* surprise, Tigris came closer to taste one. After all, they were much *better* than he had imagined.

The most frequently used adjectives, in descending order, are: *petit* 'small/little', *premier* 'first', *tranquille* 'quiet', *grand* 'big', *étrange* 'strange', *bruyant* 'noisy' (cf. Appendix C). Among the 25 most frequent adjectives, *tranquille* 'quiet' and *calme* 'calm', both denoting calmness, occur almost twice as often as *bruyant* 'noisy'. The adjective *petit* 'small/little' is 1.6 times as frequent as *grand* 'big'. The adjectives *vieux* 'old' and *nouveau* 'new' are equally frequent. Also among these 25 most frequent adjectives, several others describe the cat's and the dog's character or attitude. In descending order of frequency, these are: (1) *curieux* 'curious'; (2) *joyeux* 'cheerful' and *fier* 'proud'; (3) *méfiant* 'wary' and *enthousiaste* 'enthusiastic'; (4) *maladroit* 'clumsy' and *distant* 'distant'.

Among the color adjectives most frequently used are *noir* 'black', *blanc* 'white', and *jaune* 'yellow'.

3.4 Intensifiers

The corpus contains, on average, 6.9 intensifiers per narrative and 4.9 different intensifiers per narrative. All narratives contain intensifiers. Among the most frequently used intensifiers by ChatGPT are (in descending order): *plus* 'more', *un peu* 'a little bit', *si* 'so', *trop* 'too', *bien* 'much/far'. The intensifier *plus* 'more' represents 21.74% of all intensifiers. The intensifier *bien* 'much/far' is often paired with another one, as in the following examples:

(3) un chien au tempérament *bien plus* joyeux 'a dog with a much more cheerful temperament'

(4) une situation *bien plus* grave 'a much more serious situation'

The intensifier *très* 'very' deserves a remark. In the corpus, the intensifier *très* 'very' ranks only 15th in terms of frequency of intensifiers usage, whereas analyses of large corpora show that *très* 'very' belongs to the core set of high-frequency words in written French. Its use is particularly widespread in narrative and descriptive texts to intensify adjectives (New et al., 2004; ATILF–CNRS, *Frantext*). The frequency of use of *très* 'very' is attested by other researchers (Hendrikx et al., 2017; Hendrikx et al., 2024; Kunkel, 2024).

4 Conclusion

The narrative techniques of ChatGPT remain very limited. If a single narrative appears acceptable, 200 narratives that share the same discursive organization reveal its weaknesses. The consistent and regular use of linguistic devices such as adjectives, intensifiers, and temporal adverbials throughout the narrative characterizes ChatGPT's narrative style.

Analyzing the linguistic features employed by ChatGPT in narrative texts (such as temporal adverbials, adjectives, and intensifiers), along with a statistical analysis of the linguistic resources deployed, offers a promising avenue for a deeper exploration of the similarities and differences between AI-generated and human-produced texts.

References

Markus Appel, Wojciech P. Malecki, Tanja V. Messingschlager and Julia R. Winkler. 2025. I, ChatGPT: linguistic properties and human experiences of human- versus AI-generated stories. *Humanities and Social Sciences Communications,*12: 1892. https://doi.org/10.1057/s41599-025-06341-2.

ATILF–CNRS. *Frantext: Base textuelle du français*. https://www.frantext.fr.

Sadaf Begum. 2025. AI and Literature: The Impact of Artificial Intelligence on Creative Writing and Narrative Forms. *Journal of Social Signs Review*, 3(6): 180–199.

Nina Beguš. 2024. Experimental narratives: A comparison of human crowdsourced storytelling and AI storytelling. *Humanities &*

Social Sciences Communications, 11: 1392. https://doi.org/10.1057/s41599-024-03868-8.

Valentina Bellini, Federico Semeraro, Jonathan Montomoli, Marco Cascella and Elena Bignami. 2024. Between human and AI: assessing the reliability of AI text detection tools, *Current Medical Research and Opinion*, 40(3): 353–358. https://doi.org/10.1080/03007995.2024.2310086.

Natalia Dankova.1997. *Temporalité en espéranto parlé*. Doctoral Thesis. Université de Paris-8. Saint-Denis.

Natalia Dankova. 2002. Conceptualisation et expression du temps en français québécois et en français de France. *Langues et linguistique*, 28: 47–70.

Natalia Dankova. 2009. Variétés de français. Conceptualisation et sélection de l'information à transmettre, in F. Martineau, R. Mougeon, T. Nadasdi, M. Tremblay, editors, *Le français d'ici. Études linguistiques et sociolinguistiques sur la variation du français au Québec et en Ontario,* GREF, Toronto, pages 15–35.

Natalia Dankova. 2017. Storytelling in French from France and French from Quebec. *Corela*, 15(2). https://journals.openedition.org/corela/5051.

Natalia Dankova. 2022. Variation en français. Un cas des adjectifs. *Moderne Sprachen*, 66: 79–95.

Marie-Laure Elalouf (Ed.). 2009). *Le Groupe nominal et la construction de la référence : Approches linguistiques et didactique*. Namur: Presses universitaires de Namur.

Alexandra Fiedler and Jörg Döpke. 2025. Do humans identify AI-generated text better than machines? Evidence based on excerpts from German theses. *International Review of Economics Education*, 49: 100321. https://doi.org/10.1016/j.iree.2025.100321.

Isa Hendrikx, Kristel Van Goethem, and Natacha Buntinx. 2024. Intensification strength in speech: language-specific preferences and differences between native and learner language. *Dutch Journal of Applied Linguistics*, 13: 1–22. https://doi.org/10.51751/dujal13327.

Tanzila Kehkashan, Raja Adil Riaz, Ahmad Sami Al-Shamayleh, Adnan Akhunzada, Noman Ali, Muhammad Hamza and Faheem Akbar. 2025. AI-generated text detection: A comprehensive review of methods, datasets, and applications. *Computer Science Review*, 58: 100793. https://doi.org/10.1016/j.cosrev.2025.100793

Wolfgang Klein. 1994. *Time in Language*. Routledge, London.

Scott Kunkel. 2024. Longitudinal change in the adjective intensifier system of Hexagonal French. *Journal of French Language Studies*, 34 (3): 398–430. https://doi.org/10.1017/S0959269524000139.

Monique Lambert. 1997. En route vers bilinguisme. *AILE*, 9: 147–172.

Esha Munir, Zahida Hussain and Shama Fatima. 2025. Man vs. Machine: A Comparative Study of AI and Human-Generated Narratives. *Indus Journal of Social Science,* 3(2): 367–376. https://doi.org/10.59075/ijss.v3i2.1235.

Boris New, Christophe Pallier, Marc Brysbaert and Ludovic Ferrand. 2004. *Lexique 3: A lexical database for French*. Behavior Research Methods.

Prova, Nuzhat Noor Islam. 2024. Detecting AI Generated Text Based on NLP and Machine Learning Approaches. *ArXiv* abs/2404.10032. https://doi.org/10.48550/arXiv.2404.1.

Gavin Raffloer and Melanie C Green. 2025. Of love & lasers: Perceptions of narratives by AI versus human authors. *Computers in Human Behavior: Artificial Humans*, 5: 100249. https://doi.org/10.1016/j.chbah.2025.100168.

Sofiya R., Prasanth Arokia Samy, K Mitra, Elackiya C, S. Femina, S. Santhi and Tamilselvi C. 2025. Investigating AI-generated narratives: a literary and linguistic examination. *Scientific Culture*, 11(3): 131–140. http://doi.org/10.5281/zenodo.11030825.

Quest [Short film]. 1996. Directed by Tyron Montgomery, written and animated by Thomas

Stellmach, Thomas Stellmach Animation. https://www.youtube.com/watch?v=uB_8OA59hWE

Marc Wilmet. 1997. *Grammaire critique du français*. Paris-Louvain-la-Neuve: Hachette-Duculot

Appendix A Example of narrative generated by ChatGPT and its translation into English

Dans une rue tranquille bordée de jardins bien taillés, vivaient deux voisins qui ne s'étaient jamais parlé : un chat nommé Mistral et un chien nommé Rex. Mistral habitait une petite maison bleue avec un balcon ensoleillé, parfait pour les siestes. Rex, lui, vivait juste à côté, dans une maison jaune avec une grande cour où il aimait courir et aboyer après les oiseaux.

Dès le premier jour, la guerre fut déclarée. Chaque matin, Mistral s'installait sur la clôture pour observer le quartier avec un air supérieur. Rex, en le voyant, se mettait à aboyer de toutes ses forces. Mistral répondait en hérissant le poil et en crachant d'un air menaçant. Aucun des deux ne se souvenait vraiment pourquoi il détestait l'autre, mais cela faisait partie de leurs habitudes.

Un jour d'automne, un événement inattendu bouleversa leur routine. Une violente tempête s'abattit sur la rue. Le vent hurlait, la pluie tombait sans répit, et un grand arbre, affaibli par l'orage, se mit à craquer dangereusement entre les deux maisons. Mistral, curieux comme toujours, s'était aventuré dehors malgré la pluie. Soudain, une branche se brisa et tomba devant lui, l'empêchant de rentrer chez lui.

Rex, qui observait la scène depuis sa niche, vit le chat coincé et tremblant. Pendant quelques secondes, il hésita. Puis, oubliant leur haine, il courut vers Mistral et se mit à aboyer pour attirer l'attention des humains. Grâce à lui, les propriétaires sortirent et dégagèrent le passage.

Mistral, encore sous le choc, regarda Rex d'un air différent. Le chien, gêné, détourna la tête. Le silence s'installa, étrange mais paisible. Les jours suivants, ils continuèrent à se croiser, mais sans aboiements ni feulements. Peu à peu, la méfiance laissa place à la curiosité.

Un matin, Mistral sauta de la clôture et s'approcha prudemment de Rex. À la surprise générale, Rex remua la queue au lieu d'aboyer. Ils passèrent la journée à explorer les jardins ensemble, à se reposer au soleil et à observer les passants.

Depuis ce jour, le chat et le chien, autrefois ennemis jurés, devinrent les meilleurs amis du quartier, prouvant que même les plus vieilles querelles peuvent se transformer en une belle amitié.

English translation

On a quiet street lined with neatly trimmed gardens lived two neighbors who had never spoken to each other: a cat named Mistral and a dog named Rex. Mistral lived in a small blue house with a sunny balcony, perfect for naps. Rex, meanwhile, lived right next door in a yellow house with a large yard where he loved to run and bark at birds.

From the very first day, war was declared. Every morning, Mistral would settle on the fence to observe the neighborhood with an air of superiority. When Rex saw him, he would start barking at the top of his lungs. Mistral would respond by bristling his fur and hissing threateningly. Neither of them really remembered why they disliked each other, but it had become part of their routine.

One autumn day, an unexpected event disrupted their routine. A violent storm swept through the street. The wind howled, the rain fell relentlessly, and a large tree, weakened by the storm, began to crack dangerously between the two houses. Mistral, as curious as ever, had ventured outside despite the rain. Suddenly, a branch broke and fell in front of him, preventing him from getting back home.

Rex, who had been watching the scene from his doghouse, saw the cat trapped and trembling. For a few seconds, he hesitated. Then, forgetting their hatred, he ran toward Mistral and began barking to attract the humans' attention. Thanks to him, the owners came outside and cleared the way.

Mistral, still in shock, looked at Rex differently. The dog, embarrassed, turned his head away. A silence settled in—strange, yet peaceful. In the days that followed, they continued to cross paths, but without barking or hissing. Little by little, mistrust gave way to curiosity.

One morning, Mistral jumped down from the fence and cautiously approached Rex. To everyone's surprise, Rex wagged his tail instead of barking. They spent the day exploring the gardens

together, resting in the sun, and watching passersby.

From that day on, the cat and the dog—once sworn enemies—became the best friends in the neighborhood, proving that even the oldest quarrels can turn into a beautiful friendship.

Translated by ChatGPT

Appendix B Examples of introducing the key event in narratives and its translation into English

Un jour, alors que le soleil brillait et que le vent soufflait doucement, une étrange chose arriva. Une violente tempête s'abattit sur le village.

One day, while the sun was shining and the wind was blowing gently, something strange happened. A violent storm swept through the village.

Un jour, une surprise attendait Gracie. Alors qu'elle se prélassait tranquillement sur le sol chaud de sa maison, elle entendit un bruit étrange venant de l'extérieur.

One day, a surprise awaited Gracie. As she was peacefully lounging on the warm floor of her home, she heard a strange noise coming from outside.

Un jour, une tempête éclata. Le vent soufflait fort, les éclairs frappaient la terre, et une pluie battante trempait tout sur son passage. Grisou, qui avait l'habitude de se cacher sous son canapé lors des orages, sentit quelque chose d'étrange : il n'était pas seul. Le bruit des aboiements de Rex, qui se faisait entendre à travers les murs de la maison, était particulièrement désespéré.

One day, a storm broke out. The wind blew fiercely, lightning struck the ground, and the heavy rain soaked everything in its path. Grisou, who was used to hiding under his couch during storms, felt something strange: he was not alone. The sound of Rex's barking, heard through the walls of the house, was especially desperate.

Un jour, un étrange événement secoua le village. Une tempête approchait, et les nuages noirs s'amoncelaient dans le ciel.

One day, a strange event shook the village. A storm was approaching, and dark clouds were gathering in the sky.

Un jour, une tempête éclata. Le vent soufflait si fort qu'une branche d'arbre se cassa et se retrouva coincée dans le toit de la maison de Moka.

One day, a storm broke out. The wind blew so hard that a tree branch broke off and got stuck on the roof of Moka's house.

Un jour d'hiver, le vent soufflait fort et la neige recouvrait tout. Grizou, comme à son habitude, se prélassait au chaud près de la fenêtre. Mais soudain, un bruit étrange attira son attention.

One winter day, the wind was blowing hard and the snow covered everything. Grizou, as usual, was lounging warmly by the window. But suddenly, a strange noise caught his attention.

Un après-midi d'automne, tout changea. Il y eut un bruit étrange venant du jardin de Rocky. Curieux, Moustache se leva et se dirigea silencieusement vers la clôture pour observer ce qui se passait.

One autumn afternoon, everything changed. There was a strange noise coming from Rocky's garden. Curious, Moustache got up and quietly made his way to the fence to see what was happening.

Un jour, un événement inattendu survint. Une tempête éclata. Le vent soufflait fort, faisant tanguer les arbres et soulevant des objets du jardin.

One day, an unexpected event occurred. A storm broke out. The wind blew fiercely, making the trees sway and lifting objects from the garden.

Un matin de printemps, un événement inattendu bouleversa leur routine. Une petite boîte en carton apparut au bord du trottoir, juste devant leurs maisons. À l'intérieur, un minuscule chaton noir tremblait, affamé et apeuré.

One spring morning, an unexpected event disrupted their routine. A small cardboard box appeared at the edge of the sidewalk, right in front of their houses. Inside, a tiny black kitten was trembling, hungry and scared.

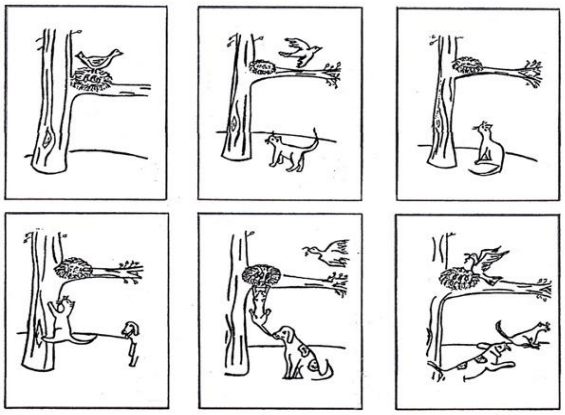
Un soir d'hiver, alors que la neige commençait à tomber, Félix sortit vérifier que son bol d'eau ne gelait pas. Le vent soufflait fort, soulevant de petits tourbillons blancs. C'est alors qu'un cri paniqué retentit depuis le jardin voisin.

One winter evening, as the snow began to fall, Félix went out to check that his water bowl hadn't frozen. The wind blew strongly, lifting small white swirls into the air. That's when a panicked cry rang out from the neighboring garden.

Appendix C 25 most frequently used adjectives



Appendix D *Cat Story*



Maya Hickmann.1982. *The Development of Narrative Skills: Pragmatic and Metapragmatic Aspects of Discourse Cohesion.* Doctoral Thesis. Université de Chicago.

A Systematic Benchmark of Machine Transliteration Models for the Tajik-Farsi Language Pair: A Comparative Study from Rule-Based to Transformer Architectures

Mullosharaf Arabov
Kazan Federal University, Kazan, Russia
cool.araby@gmail.com

Abstract

This paper presents a systematic comparative benchmark of machine learning architectures for transliteration between the Tajik (Cyrillic script) and Persian (Arabic script) languages. A central contribution is the creation and open release of a unique, multi-domain parallel corpus comprising 328,253 sentence pairs aggregated from heterogeneous sources, including classical poetry, diplomatic texts, lexicographic data, and crowdsourced projects. To enable fully reproducible experiments under limited computational resources, a representative stratified subset of 40,000 pairs has been formed and is made publicly available together with the complete dataset. The study evaluates six model classes under a unified protocol: a rule-based baseline, an LSTM with attention, a character-level Transformer, a G2P Transformer trained from scratch, large pre-trained multilingual models (mBART, mT5 with LoRA), and the byte-oriented ByT5 model. The experimental results demonstrate that ByT5 achieves state-of-the-art performance with chrF++ scores of 87.4 for Tajik→Farsi and 80.1 for the reverse direction. Notably, the proposed G2P Transformer, trained exclusively on the 40,000-pair subset, significantly outperforms the substantially larger mBART model in the forward direction (72.3 vs. 62.2 chrF++), while subword-based models (mT5) fail almost completely, underscoring the necessity of character- or byte-level representations for this task. All collected data, source code, and model outputs are released under open licenses, thereby establishing a FAIR (Findable, Accessible, Interoperable, Reusable) benchmark for cross-script NLP in low-resource Central Asian languages.

Keywords: Transliteration, Tajik language, Persian language, parallel corpus, FAIR benchmark, low-resource languages, Cyrillic script, Arabic script, byte-level Transformer, open data.

1 Introduction

The Persian language exhibits a striking case of digraphia: two distinct writing systems coexist within a single linguistic continuum. In Iran and Afghanistan, the Perso-Arabic script prevails, whereas Tajikistan employs a modified Cyrillic alphabet adopted during 20th-century linguistic reforms. This orthographic divide creates a paradoxical barrier—speakers of mutually intelligible varieties are largely cut off from each other's written content. The digital environment exacerbates the problem, as the vast majority of Persian-language online texts use Arabic script, effectively excluding Tajik-speaking users from a substantial portion of internet resources (Megerdoomian and Parvaz, 2008). This situation is not unique to Persian: similar digraphic divides or historical script shifts occur in numerous languages of the former Soviet space, including Serbian (Cyrillic/Latin), Kazakh (Cyrillic/Latin/Arabic), Uzbek (Cyrillic/Latin), Bashkir and Tatar (Cyrillic/Arabic historically, with ongoing Latinisation discussions), and Ossetian (Cyrillic/Georgian historically). In all these cases, automatic transliteration has proven to be a low-cost bridge fostering linguistic unity and cross-border communication.

Automatic conversion between Tajik Cyrillic and Persian Arabic script involves several non-trivial linguistic challenges. Transliterating from Arabic to Cyrillic requires restoring short vowels that are typically omitted in Perso-Arabic writing. The reverse direction must account for positional shaping of Arabic letters and the fact that multiple Cyrillic characters may correspond to a single Perso-Arabic grapheme. Moreover, orthographic divergence is compounded by lexical and morphological differences shaped by decades of separate cultural and political development.

Despite its practical significance, automatic

transliteration for the Tajik–Farsi pair has remained on the periphery of Natural Language Processing (NLP) research. While machine translation between the two varieties has received some attention, the creation of robust transliteration systems—potentially a simpler and more reliable bridge between scripts—has been comparatively neglected (Merchant and Tang, 2024). A fundamental obstacle has been the acute scarcity of high-quality parallel corpora containing texts in both graphic systems. Recent years have witnessed progress with the release of resources such as ParsText (Merchant and Tang, 2024) and a corpus derived from Ferdowsi's *Shahnameh*, yet their volume and genre coverage remain limited. The emergence of multi-faceted benchmarks like APARSIN (Jafari et al., 2026) further confirms the growing demand for systematic comparison of models within this language family.

The present study addresses these gaps by conducting a systematic benchmark of transliteration models spanning several methodological generations—from rule-based substitution to byte-level Transformers—evaluated on a unified, multi-domain dataset. Our contributions are threefold: (i) the creation and open release of a unique parallel corpus comprising over 328,000 sentence pairs, along with a stratified 40,000-pair subset for reproducible experimentation under limited resources; (ii) the first controlled comparison of six architectural families under identical training and evaluation protocols; and (iii) empirical evidence that character- and byte-level models decisively outperform subword-based approaches for this low-resource cross-script task.

The remainder of this paper is organized as follows. Section 2 surveys related work on Tajik–Persian transliteration and available resources. Section 3 describes the corpus construction and model architectures evaluated in the benchmark. Section 4 presents the experimental setup and quantitative results. Section 5 concludes with a discussion of findings and directions for future work.

2 Related Work

Systematic approaches to Tajik–Persian transliteration began to appear in the early 2010s. Davis (2012) introduced a system based on Statistical Machine Translation (SMT), treating transliteration as character-level translation and addressing the challenge of resource sharing between the two Persian varieties. Around the same period, theoretical foundations for automated script conversion were laid, often relying on an intermediate Latin transcription—an approach prone to information loss and error propagation (Grashchenko, 2003).

The advent of neural methods brought a qualitative shift. Transformer-based Grapheme-to-Phoneme (G2P) models have been applied to this task, achieving chrF++ scores of approximately 59 for Farsi→Tajik and 74 for the reverse direction, thereby establishing reproducible baselines and highlighting the asymmetric difficulty of the two transliteration directions. Parallel efforts have focused on dataset creation. Researchers have leveraged optical character recognition and heuristic alignment to construct parallel corpora from classical literary works, including Ferdowsi's *Shahnameh*. Merchant and Tang (2024) introduced ParsText, a collection of 2,813 manually curated Persian sentences in both scripts, emphasizing the scarcity of digital digraphic content relative to printed materials.

The most advanced transliteration model to date is ParsTranslit (Merchant and Tang, 2026), a sequence-to-sequence system trained on an aggregation of all available datasets plus two proprietary collections. The authors convincingly demonstrate that single-domain models lack cross-domain robustness, and only data aggregation yields a universal transliterator. ParsTranslit achieved state-of-the-art results: chrF++ of 87.91 and normalized CER of 0.05 for Farsi→Tajik, and 92.28 and 0.04 for the reverse direction.

The authors of the present study have previously contributed to the resource base and evaluation methodology for this language pair. Arabov (2026b) introduced TajPersLexon, a curated lexical resource of 40,112 word and phrase pairs, accompanied by a CPU-only benchmark comparing hybrid, neural, and information retrieval approaches. Kurbonovich (2026) investigated character-level Transformers trained on 52,152 lexicographically verified pairs, achieving a CER of 0.3182 and exact-match accuracy of 0.3215 with beam search, outperforming both dictionary-based rules and recurrent baselines.

Beyond direct transliteration efforts, several Persian NLP tools provide valuable pre- and post-processing capabilities. DadmaTools V2 (Jafari et al., 2025) employs LoRA adapters to efficiently add task-specific modules—including sentiment

analysis, formality conversion, and spell checking—to a shared pre-trained model (Hu et al., 2022). While these tools do not address script conversion directly, they constitute an important component of the broader Persian NLP ecosystem. The APARSIN benchmark (Jafari et al., 2026) further reflects growing community interest in systematic evaluation across Iranian languages, covering translation, transliteration, and sentiment analysis tasks.

Finally, recent computational studies of language contact underscore the linguistic complexity inherent in Persian varieties. Basirat et al. (2026) demonstrated that morphological features such as case and gender are strongly shaped by language-specific structure, whereas universal syntax remains largely insensitive to historical contact. Such findings reinforce the need for specialized models capable of handling orthographic and morphological divergence between Tajik and Farsi.

Despite notable progress, the existing literature exhibits several persistent limitations. Most studies evaluate models on narrow-domain datasets, precluding fair comparison across architectures. Reproducibility remains an issue due to inconsistent reporting and lack of publicly released code and exact configurations. Critically, no prior work has systematically compared byte-level models (e.g., ByT5) against character-level and subword-based approaches under controlled conditions on a unified multi-domain corpus. The present study addresses precisely these gaps by (i) releasing a large, multi-domain parallel corpus and a representative reproducible subset; (ii) conducting the first controlled benchmark spanning rule-based, LSTM, character-level Transformer, G2P Transformer, pre-trained multilingual, and byte-level architectures; and (iii) demonstrating—through fully disclosed methodology and open-source code—that byte-level modeling achieves superior performance in low-resource cross-script transliteration while subword-based models fail dramatically.

3 Data

We aggregated a multi-source parallel corpus for Tajik–Farsi transliteration from the following domains: official and news texts (Ministry of Foreign Affairs of Tajikistan), terminology lists (Committee on Language and Terminology), classical Persian poetry (Ferdowsi's *Shahnameh*, Rumi's *Masnavi-i Ma'navi* from ganjoor.net, and selected rubaiyat of Omar Khayyam), open crowdsourcing and lexicographic projects, and the previously released `TajPersParallelLexicalCorpus` (Arabov, 2026a). This effort builds on our earlier work on large-scale Tajik text resources (Arabov et al., 2025).

All texts underwent cleaning and normalization: incorrect characters and formatting artifacts were removed, space types (including non-breaking and narrow spaces) were unified, and strings with Latin or extraneous scripts were discarded. After deduplication, the full dataset contains **328,253 sentence pairs**. For reproducible experimentation under limited computational resources, we extracted a stratified pseudo-random subsample of **40,000 pairs** (seed `random_state=42`), which faithfully preserves the statistical properties of the full collection (Tables 1–2).

The categories in Table 1 are abbreviated as follows: PF = Poetic fragments, RM = Rumi's *Masnavi*, UT = Unique Tajik words, SH = *Shahnameh*, PR = Prose fragments, PN-P = Proper names—Persons, GV = General vocabulary, PN-L = Proper names—Locations, Oth = Other categories.

Category	Full (328k)	Subset (40k)
PF	47.2%	47.1%
RM	11.9%	12.0%
UT	11.6%	11.5%
SH	7.6%	7.4%
PR	7.3%	7.4%
PN-P	6.1%	6.2%
GV	4.4%	4.3%
PN-L	2.9%	3.0%
Oth	1.0%	1.1%

Table 1: Domain distribution: full vs. subsample

The domain distribution remains virtually unchanged in the subsample, with maximum deviation below 0.2 percentage points. Length statistics (Table 2) further confirm the representativeness of the subset.

Statistic	Tajik (328k)	Tajik (40k)	Farsi (328k)	Farsi (40k)
Mean	33.35	33.38	27.98	28.00
Std. dev.	35.95	35.71	30.84	30.58
Median	26	26	22	22
25th pct.	19	19	17	17
75th pct.	34	34	28	28
Max	973	755	825	652

Table 2: Character sequence length statistics

The 40k subsample comprises 40,000 unique Tajik strings and 39,996 unique Farsi strings (four

Farsi translations correspond to two distinct Tajik originals). Mean string lengths are 33.4 ± 35.7 characters for Tajik and 28.0 ± 30.6 for Farsi (medians 26 and 22, respectively). Word counts per string are 5.8 ± 5.7 (Tajik) and 6.3 ± 6.3 (Farsi); the greater character length of Tajik reflects explicit vowel representation in Cyrillic.

Character frequency analysis (Table 3) confirms expected patterns: Tajik is vowel-dominated (а, и, о), while Persian shows high frequencies of *alef*, *re*, and *ye*. On average, a Tajik string contains 27.4 Cyrillic characters, a Farsi string 18.8 Arabic-script characters.

#	Tajik	Freq.	Farsi	Freq.
1	Space	190,188	Space	212,338
2	а	167,061	*alef*	113,481
3	и	94,200	*re*	78,884
4	о	89,231	*ye*	77,101
5	р	78,567	*nun*	73,665
6	н	72,378	*dal*	64,929
7	д	63,014	*vav*	56,331
8	у	51,412	*he*	53,069
9	м	41,778	*be*	44,098
10	т	41,242	*mim*	43,915

Table 3: Top 10 most frequent characters (40k subsample)

In summary, the dataset is genre-balanced, spans lexical items to full sentences, and covers proper names, poetic archaisms, and modern prose. The statistical equivalence between the full corpus and the 40k subset ensures that benchmark results obtained on the latter are fully generalizable. All data, code, and model outputs are publicly released under open licenses to facilitate reproducibility.

4 Model Architectures and Experimental Protocol

We implemented or adapted six model classes spanning classical and neural approaches, all evaluated under a unified protocol described below.

4.1 Rule-based Baseline

A deterministic character-level substitution uses static dictionaries mapping Cyrillic to Arabic graphemes and vice versa. The Tajik→Farsi dictionary contains 70 rules covering the full Tajik alphabet, including specific graphemes (г, ӣ, к, ӯ, х, ч) and positional variants; the reverse direction uses 47 rules. Transliteration proceeds by sequential character replacement with fallback to identity when no rule matches. This parameter-free method serves as a lower-bound reference.

4.2 LSTM with Attention

A sequence-to-sequence model with Bahdanau additive attention (Bahdanau et al., 2015). The encoder is a two-layer bidirectional LSTM (hidden size 256) over character embeddings (dimension 256). The decoder is a two-layer unidirectional LSTM (hidden 256) that computes a context vector via attention at each step, concatenating it with the previous token embedding. Training uses AdamW ($\eta = 5 \times 10^{-4}$, gradient clipping 1.0), teacher forcing with probability 0.5, and dropout 0.1 on all but the output layer.

4.3 Character-Level Transformer (CharTransformer)

A compact Transformer trained from scratch on character vocabularies. Both encoder and decoder have 4 layers, embedding dimension 256, 8 attention heads, and feed-forward dimension 1024. Positional encoding is sinusoidal, supporting sequences up to 5,000 characters. Regularization: dropout 0.1. Optimization: AdamW ($\eta = 5 \times 10^{-4}$) with cosine annealing. Padding tokens are masked in attention.

4.4 G2P Transformer

This model extends CharTransformer specifically for grapheme-to-grapheme transliteration. Architecture is identical except for: (i) maximum sequence length increased to 512; (ii) label smoothing 0.1; (iii) weight decay 10^{-4} added to AdamW; (iv) uniform weight initialization in $[-0.1, 0.1]$. Inference uses greedy decoding, which preliminary experiments showed to be as effective as beam search while faster.

4.5 Pretrained Multilingual Models

Three Transformer-based models from Hugging Face were fine-tuned:

- **mBART-large-50** (610M parameters), pretrained on 50 languages including Tajik (`tg_Cyrl`) and Persian (`fa_IR`). Language codes were explicitly set in tokenizer and generation config.
- **mT5-small** (300M), fine-tuned with LoRA (Hu et al., 2022) (rank=8, α=32, dropout=0.05) applied to attention projection matrices (`q, k, v, o`), avoiding full parameter update.
- **ByT5-small** (300M), a byte-level model operating directly on UTF-8 bytes without sub-

word tokenization, inherently handling rare Tajik graphemes.

All pretrained models were fine-tuned for 2 epochs using `Seq2SeqTrainer` with $\eta = 3 \times 10^{-4}$, linear decay, dynamic batching (max batch size 8), and bfloat16 mixed precision (for mBART and ByT5). The best checkpoint was selected based on validation loss.

4.6 Experimental Protocol and Evaluation

All runs used an NVIDIA A100 GPU (40 GB) with CUDA 11.8 and PyTorch 2.0. The codebase is implemented in Python 3.10 and relies on Hugging Face Transformers (version 4.30). For each model and direction, three independent initializations (seeds 42, 43, 44) were performed. On each seed, the 40k-pair sample was stratified split into train (80%), validation (10%), and test (10%) sets preserving category proportions. Results were serialized to JSON for exact reproducibility. Full environment specifications and code are provided in the public repository.

We report six metrics on the test set: chrF++ (primary), BLEU, TER, CER, WER, and exact match accuracy. chrF++, BLEU, and TER were computed with SacreBLEU defaults; CER/WER used Levenshtein distance normalized by reference length. For each metric, we computed mean, standard deviation, and 95% bootstrap confidence intervals (2,000 resamples). Pairwise model comparisons used Wilcoxon signed-rank and paired t-tests on chrF++ at significance level 0.05. chrF++ was chosen as primary because it captures both character- and word-level n-gram overlap, critical for morphologically rich cross-script conversion (Popović, 2017).

5 Experimental Results

We evaluated all models on stratified test splits (10% of the 40k subsample) with three independent runs per direction (seeds 42, 43, 44). The primary metric is chrF++ with 95% bootstrap confidence intervals; we also report BLEU, CER, TER, WER, exact match accuracy (Acc%), and computational costs. All results are fully reproducible and the raw outputs are released alongside the code.

5.1 Primary Quality Metrics

Tables 4 and 5 present the core quality indicators for the Farsi→Tajik and Tajik→Farsi directions, respectively. For brevity, we omit Acc% from the tables (it remains zero for several models due to the strictness of exact string match) and focus on chrF++, BLEU, and CER.

Model	chrF++	BLEU	CER
Baseline	10.4 ± 2.4	0.36 ± 0.09	0.54 ± 0.08
CharTransformer	18.2 ± 1.8	0.45 ± 0.26	2.45 ± 0.13
LSTM	31.4 ± 6.5	4.9 ± 3.4	0.36 ± 0.08
G2P Trans.	60.3 ± 0.7	21.0 ± 1.4	0.47 ± 0.05
mT5+LoRA	14.3 ± 0.2	1.75 ± 0.09	0.75 ± 0.02
mBART	70.1 ± 0.4	45.4 ± 0.4	0.24 ± 0.01
ByT5	$\mathbf{80.1 \pm 0.2}$	$\mathbf{56.6 \pm 0.3}$	$\mathbf{0.09 \pm 0.00}$

Table 4: Primary metrics for Farsi→Tajik.

Model	chrF++	BLEU	CER
Baseline	14.0 ± 6.0	0.42 ± 0.19	0.62 ± 0.14
CharTransformer	17.9 ± 1.5	0.39 ± 0.28	2.45 ± 0.40
LSTM	65.1 ± 5.5	38.5 ± 7.7	0.14 ± 0.03
G2P Trans.	72.3 ± 0.4	36.5 ± 0.4	0.42 ± 0.02
mT5+LoRA	18.5 ± 0.9	3.45 ± 0.67	0.75 ± 0.05
mBART	62.2 ± 5.3	44.2 ± 6.3	0.38 ± 0.07
ByT5	$\mathbf{87.4 \pm 0.1}$	$\mathbf{73.6 \pm 0.2}$	$\mathbf{0.05 \pm 0.00}$

Table 5: Primary metrics for Tajik→Farsi.

The byte-level **ByT5** model dominates across all metrics, achieving chrF++ scores of 80.1 (Farsi→Tajik) and 87.4 (Tajik→Farsi) with exceptionally low variance, indicating stable training. Its character error rates are below 0.1, meaning fewer than one character in a thousand is incorrect—virtually imperceptible to human readers.

The **G2P Transformer**, trained from scratch on only 40k pairs, delivers a remarkable 72.3 chrF++ for Tajik→Farsi, significantly outperforming the much larger mBART model (62.2) in that direction. This highlights the efficiency of character-level modeling for this task. In the reverse direction, mBART holds a slight edge (70.1 vs. 60.3), though the G2P Transformer remains competitive given its dramatically smaller footprint.

LSTM shows strong asymmetry: 65.1 chrF++ for Tajik→Farsi but only 31.4 for the opposite direction. This disparity arises because converting from Arabic script to Cyrillic requires restoring unwritten short vowels—a challenge that recurrent architectures struggle to meet without explicit linguistic knowledge.

The **mT5+LoRA** model fails almost completely (chrF++ < 19), confirming that subword tokenization is fundamentally incompatible with character-level transliteration. The **rule-based baseline** and **CharTransformer** both perform poorly, with chrF++ in the 10–18 range, underscoring the necessity of contextual modeling. Figure 1 visualizes

the chrF++ comparison.

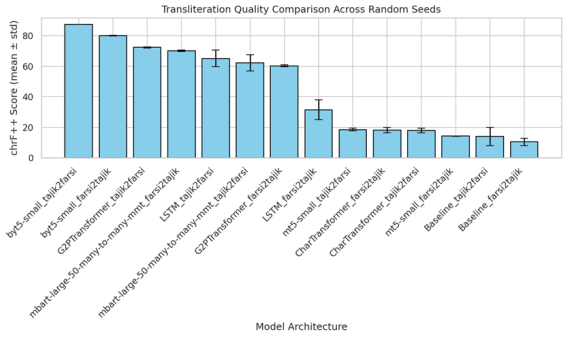


Figure 1: chrF++ scores (mean ± std) across models and directions.

5.2 Error Analysis

While chrF++ and CER capture overall fidelity, TER and WER are sensitive to word reorderings and insertions/deletions. Tables 6 and 7 report these metrics for Farsi→Tajik and Tajik→Farsi, respectively.

Model	TER	WER
Baseline	107.9 ± 0.4	1.13 ± 0.01
CharTransformer	208 ± 15	3.34 ± 0.23
LSTM	113 ± 22	0.96 ± 0.13
G2P Trans.	60 ± 5	0.74 ± 0.05
mT5+LoRA	130.9 ± 3.8	1.14 ± 0.03
mBART	43.8 ± 0.9	0.60 ± 0.01
ByT5	**28.2 ± 0.2**	**0.36 ± 0.00**

Table 6: TER and WER for Farsi→Tajik (mean ± std).

For the more challenging Farsi→Tajik direction, ByT5 achieves a TER of 28.2, meaning only 28 edits per 100 characters are required to reach the reference. In stark contrast, CharTransformer and mT5+LoRA require more edits than the string length (TER $>$ 100%), indicating catastrophic structural failures. The G2P Transformer and mBART perform reasonably well (TER 40–60), with errors primarily local in nature.

Model	TER	WER
Baseline	96.5 ± 2.0	0.97 ± 0.02
CharTransformer	215 ± 37	3.31 ± 0.59
LSTM	52 ± 14	0.44 ± 0.06
G2P Trans.	39.6 ± 1.1	0.51 ± 0.02
mT5+LoRA	115.8 ± 12.5	1.08 ± 0.10
mBART	42.2 ± 5.1	0.53 ± 0.06
ByT5	**17.2 ± 0.1**	**0.21 ± 0.00**

Table 7: TER and WER for Tajik→Farsi (mean ± std).

The reverse direction (Tajik→Farsi) is consistently easier for all models. ByT5's TER drops to 17.2 and WER to 0.21, reflecting near-perfect word-level accuracy. Even the weaker models show improvement, though the relative ranking remains unchanged. This asymmetry underscores the inherent difficulty of restoring unwritten short vowels when converting from Arabic script to Cyrillic.

5.3 Computational Efficiency

Practical deployment depends on training time, inference latency, and memory footprint. Table 8 summarizes these costs.

Model	Train (s)	Infer (ms)	Mem (GB)
CharTransformer	≈1700	—	1.2*
LSTM	2300–2980	—	1.4*
G2P Trans.	≈1200	—	1.5*
mT5+LoRA	≈1775	155–332	10.6
mBART	2470–2600	81–163	16–18
ByT5	2340–2430	224–247	9.4

Table 8: Computational costs (training per epoch, inference latency, peak memory).

*CPU RAM for models trained without GPU.

ByT5 achieves the best quality–cost trade-off: training completes in ∼40 minutes per epoch on a single GPU, inference latency is ∼230 ms per example, and memory consumption stays under 10 GB—manageable on consumer hardware. The G2P Transformer trains twice as fast (∼20 min/epoch) and requires less than 2 GB of CPU RAM, making it ideal for resource-constrained environments. mBART demands 16–18 GB of GPU memory, limiting its accessibility. LSTM, despite modest memory usage, trains slowly and delivers mediocre quality, rendering it the least attractive option. Figure 2 visualizes the Pareto frontier of quality versus training time.

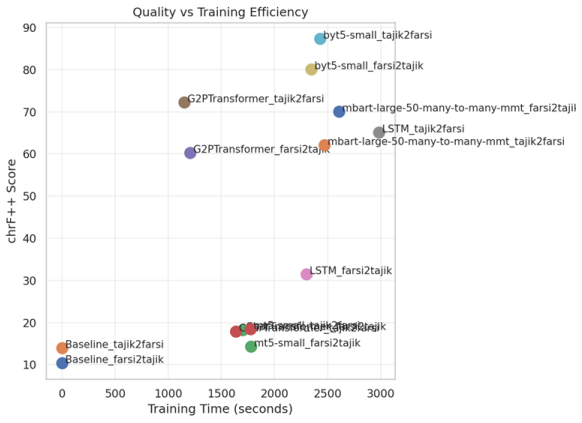


Figure 2: Quality (chrF++) vs. training time. ByT5 and G2P Transformer lie on the Pareto front.

5.4 Per-Category Performance and Learning Dynamics

To understand domain-specific behavior, we analyzed G2P Transformer performance across the eight dataset categories. As shown in Figure 3, the hardest categories are unique Tajik words (chrF++ 51.3) and proper names of organizations (44.0), likely due to rare grapheme sequences and non-native patterns. Poetic fragments and *Masnavi* verses are easiest (chrF++ $\approx$ 77–80), reflecting the regularity of classical poetic diction.

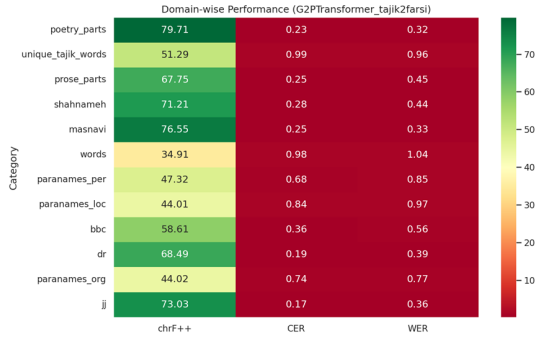


Figure 3: Per-category chrF++ for G2P Transformer (Tajik→Farsi).

Beyond final performance, we examined how model quality evolves during training. Figure 4 presents learning curves for all architectures. ByT5 and G2P Transformer converge smoothly and stably, whereas LSTM and CharTransformer exhibit oscillations and early overfitting. Pretrained models (ByT5, mBART, mT5) benefit from rapid initial convergence due to transferred knowledge but plateau quickly, with mT5 showing virtually no improvement beyond the first few hundred steps.

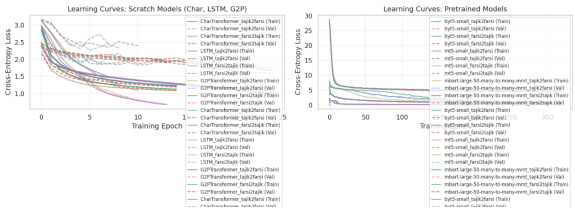


Figure 4: Learning curves: left — trained from scratch; right — pretrained.

5.5 Comparison with State-of-the-Art

Compared to ParsTranslit (Merchant and Tang, 2026), which was trained on the full 328k corpus, our ByT5 model trained on only 40k pairs (12% of the data) achieves chrF++ scores within 5–8 points: 87.4 vs. 92.28 (Tajik→Farsi) and 80.1 vs. 87.9 (Farsi→Tajik). This demonstrates both the representativeness of our subsample and the effectiveness of byte-level modeling. All pairwise differences among main models are statistically significant ($p < 0.05$, Wilcoxon signed-rank and paired t-tests).

5.6 Qualitative Examples

The table in Figure 5 provides concrete transliterations produced by ByT5. Short words and poetic phrases are rendered almost perfectly; errors are limited to punctuation placement and occasional missing diacritics, which do not hinder comprehension.

Input (Tajik)	Reference (Farsi)	ByT5-small	Baseline
Дилаш чун бар оташ	دلش چون بر آتش	دلش چون بر آتش	дилаш чун бар оташ
ниҳода кабоб.	نهاده کباب	نهاده کباب	ниҳода кабоб.
парванда	پرونده	پرونده	парванда
кағгардан	کج گردن	کجگردن	кағгардан
Нақш мебастам,	نقش می بستم که	نقش می بستم که	нақш мибастам,
ки гирам гӯшае	گیرم گوشه ای	گیرم گوشه ای	ки гирам гӯшае
з-он чашми маст,	زان چشم مست	زان چشم مست	з-он чашми маст,
"Ӯ, ки яке аз	"او که یکی از	"او که یکی از	"в, ки яке аз
суханрононі..."	سخنرانان..."	سخنران..."	суханрононі..."

Figure 5: Transliteration examples (Tajik→Farsi).

A closer inspection of the errors reveals a systematic pattern: the most frequent mistake is the omission of the *ezafe* marker—a short vowel *-e-* that indicates possession or attribution in Persian and is unwritten in the Perso-Arabic script. For instance, in the phrase *cheshm-e mast* ('intoxicating eye'), ByT5 correctly transliterates the consonants but drops the linking *ezafe*, a subtlety that even native speakers occasionally overlook in informal writing. Other minor deviations involve the placement of quotation marks and the rendering of geminated consonants. These observations suggest that incorporating explicit linguistic knowledge—either as post-processing rules or as auxiliary features during training—could further narrow the gap between automatic and human-level transliteration, particularly for the more challenging Farsi→Tajik direction.

6 Limitations

This study has several objective limitations that should be considered when interpreting the results.

First, due to computational constraints, all experiments were conducted on a 40k-pair subsample of the full 328k corpus. While statistical tests confirmed that the subsample faithfully represents the larger dataset, the relative ranking of models sensitive to dataset size—particularly LSTM and Char-Transformer—could shift slightly when trained on the complete collection.

Second, we did not perform external validation on a completely independent test set. Although stratified splitting ensures internal consistency, real-world performance may degrade on out-of-domain texts such as social media posts, colloquial speech, or user-generated content. The creation of a dedicated external benchmark remains an important direction for future work.

Third, hyperparameters of pretrained models were kept at their default values without exhaustive grid search. More fine-grained tuning of learning rates, number of epochs, or beam size could further improve results, especially for Farsi→Tajik where ByT5's exact match accuracy was lower. The mT5 model was tested only with LoRA adapters; full fine-tuning might yield different outcomes but was computationally prohibitive in our setup.

Fourth, the automatic metrics employed (chrF++, BLEU, CER, WER) do not fully capture linguistic acceptability. They treat minor orthographic deviations and meaning-distorting errors equally. Incorporating human evaluation or specialized metrics that account for positional Arabic letter forms and the correctness of short-vowel restoration would provide a more nuanced assessment.

Fifth, the corpus includes texts from diverse sources, some of which may be subject to copyright. While we release only aligned sentence pairs rather than complete original texts, users should verify the legal status of the data in their jurisdiction and attribute original sources where applicable.

Sixth, all experiments were conducted by a single researcher. Although fixed random seeds and automated evaluation pipelines mitigate potential inconsistencies, independent replication by other groups would further strengthen the reliability of the reported rankings and eliminate any unintended implementation biases.

7 Conclusion

This paper presented the first systematic benchmark of Tajik–Farsi transliteration models, spanning rule-based, recurrent, and Transformer architectures. We collected, cleaned, and publicly released a multi-domain parallel corpus of over 328k sentence pairs, from which a representative 40k-pair subsample was used for controlled comparisons with three random initializations.

Three main conclusions emerge. First, the byte-level **ByT5-small** model is the undisputed leader, achieving chrF++ scores of 87.4 (Tajik→Farsi) and 80.1 (Farsi→Tajik) with moderate computational cost. This demonstrates that operating on raw bytes rather than subword units is crucial for accurate cross-script conversion. Second, our **G2P Transformer**, trained from scratch on only 40k pairs, significantly outperforms the much larger mBART in the forward direction (chrF++ 72.3 vs. 62.2) while requiring drastically less memory and inference time. Third, subword-based models (mT5) fail almost completely, underscoring the fundamental incompatibility of subword tokenization with this task.

The practical contribution is an open, FAIR benchmark comprising code, data, and pretrained weights, which can serve as a foundation for future cross-script Persian NLP research. Our architectural recommendations—ByT5 for maximum quality, G2P Transformer for resource-constrained deployment—are directly applicable to real-world systems. Concrete use cases include enabling Tajik-speaking users to access Persian-language websites and digital libraries, facilitating cross-border information exchange, and supporting computational humanities research on classical Persian poetry across script variants.

Future work will proceed along several promising directions. First, we plan to scale experiments to the full 328k corpus to establish definitive upper bounds and to observe whether data-hungry models (e.g., LSTM) benefit disproportionately. Second, we will explore larger variants of ByT5 (e.g., ByT5-large) and ensemble methods that combine the speed of the G2P Transformer with the accuracy of ByT5 via reranking or distillation. Third, the development of an external test set covering diverse genres and a linguistically informed evaluation protocol—including human judgments and metrics sensitive to vowel restoration and positional letter forms—remains a high priority. Fourth, we aim to extend the approach to other Iranian languages that employ Cyrillic script (e.g., Ossetian, Yaghnobi) and to investigate zero-shot transliteration across related digraphic language pairs. Finally, we intend

to maintain and periodically update the benchmark as new models and larger datasets become available, fostering a living evaluation platform for the community.

References

M. K. Arabov. 2026a. TajPersParallelCorpus: A large-scale Tajik–Persian parallel corpus for cross-script NLP.

M. K. Arabov, Kh. S. Makhmadaliev, and K. Kh. Khabibullozoda. 2025. Creating a multiformat text corpus for the Tajik language to train modern language models. *Science and Innovation. Series of Geological and Technical Sciences*, (2):131–136. EDN FJMXTF.

Mullosharaf Kurbonovich Arabov. 2026b. TajPersLexon: A Tajik–Persian lexical resource and hybrid model for cross-script low-resource NLP. In *The Proceedings of the First Workshop on NLP and LLMs for the Iranian Language Family*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations (ICLR 2015)*, San Diego, CA, USA.

Ali Basirat, Danial Namazifard, and Navid Baradaran Hemmati. 2026. A computational approach to language contact – a case study of Persian. In *The Proceedings of the First Workshop on NLP and LLMs for the Iranian Language Family*.

Chris Irwin Davis. 2012. Tajik-Farsi Persian transliteration using statistical machine translation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3988–3995, Istanbul, Turkey. European Language Resources Association (ELRA).

L. A. Grashchenko. 2003. *Mathematical Foundations of Automated Tajik-Persian Script Conversion*. Cand. tech. sci. dissertation, Moscow. (In Russian).

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. 2022. LoRA: Low-rank adaptation of large language models. In *10th International Conference on Learning Representations (ICLR 2022)*, Virtual.

Sadegh Jafari, Tara Azin, Farhad Roodi, Zahra Dehghani Tafti, Mehrdad Ghadrdan, Elham Vatankhahan Esfahani, Aylin Naebzadeh, Mohammadhadi Shahhosseini, Ghafoor Khan, Kazem Forghani, Danial Namazi, Seyed Mohammad Hossein Hashemi, Farhan Farsi, Mohammad Osoolian, Maede Mohammadi, Mohammad Erfan Zare, Muhammad Hasnain Khan, Muhammad Hussain, Nooreen Zaki, Joma Mohammadi, Shayan Bali, Mohammad Javad Ranjbar, Els Lefever, and Veronique Hoste. 2026. APARSIN: A multi-variety sentiment and translation benchmark for Iranic languages. In *The Proceedings of the First Workshop on NLP and LLMs for the Iranian Language Family*.

Sadegh Jafari, Farhan Farsi, Navid Ebrahimi, Mohamad Bagher Sajadi, and Sauleh Eetemadi. 2025. DadmaTools v2: an adapter-based natural language processing toolkit for the Persian language. In *Proceedings of the 1st Workshop on NLP for Languages Using Arabic Script*.

Arabov Mullosharaf Kurbonovich. 2026. Character-level transformer for Tajik–Persian transliteration with a parallel lexical corpus. In *Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script*.

Karine Megerdoomian and Dan Parvaz. 2008. Low-density language bootstrapping: the case of Tajiki Persian. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Rayyan Merchant and Kevin Tang. 2024. ParsText: A digraphic corpus for Tajik-Farsi transliteration. In *Proceedings of the Second Workshop on Computation and Written Language (CAWL) @ LREC-COLING 2024*.

Rayyan Merchant and Kevin Tang. 2026. ParsTranslit: Truly versatile Tajik-Farsi transliteration. In *Findings of the Association for Computational Linguistics: EACL 2026*.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation (WMT 2017)*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

TajikNLP: An Open-Source Toolkit for Comprehensive Text Processing of Tajik (Cyrillic Script)

Mullosharaf Arabov[1], **Karomatullo Habibullozoda**[2], **Nurali Shirinov**[2]

[1]Kazan Federal University, Kazan, Russia
cool.araby@gmail.com
[2]Bokhtar State University named after Nosiri Khusrav, Bokhtar, Tajikistan
ddb.mi-11@mail.ru, hukmiddin_82@mail.ru

Abstract

The Tajik language, written in Cyrillic script, remains severely under-resourced in terms of publicly available natural language processing (NLP) toolkits, hindering both linguistic research and applied development. This paper introduces TajikNLP, an open-source Python library that provides the first comprehensive pipeline for processing authentic Tajik text while preserving the original Cyrillic orthography. The library implements a modular architecture centered around a unified `Doc` object, enabling sequential application of components for cleaning, normalization, tokenization (including subword BPE), morphemic segmentation, part-of-speech tagging, stemming, lemmatization, and sentence splitting. A novel unified morphology engine is introduced, offering controlled and deep analysis modes that significantly improve handling of Tajik's agglutinative nominal and verbal inflections. The release further incorporates a lexicon-based sentiment analyser and pre-trained Word2Vec/FastText embeddings loaded directly from the Hugging Face Hub. To ensure reproducibility and facilitate future research, four accompanying linguistic datasets—a POS-tagged corpus (52.5k entries), a sentiment lexicon (3.5k entries), a toponym gazetteer (5.5k entries), and a personal names dataset (3.8k entries)—have been openly published under permissive licenses. The library's reliability is validated by an extensive test suite of 616 automated tests achieving 93% source code coverage. TajikNLP thus establishes a foundational technological infrastructure for Tajik language processing, lowering the barrier to entry for both academic and industrial applications in low-resource Cyrillic-script environments.

Keywords: Tajik language, natural language processing, low-resource languages, Cyrillic script, morphological analysis, lemmatization, POS tagging, open-source software, linguistic resources.

1 Introduction

The development of artificial intelligence and natural language processing (NLP) technologies has led to the creation of powerful, publicly available tools for analyzing texts in the world's most widely spoken languages. At the same time, a significant number of languages categorized as low-resource still lack such support, hindering both fundamental linguistic research and the development of applied intelligent systems. The Tajik language, the state language of the Republic of Tajikistan, which uses a Cyrillic script (Oranskii, 1988), is a prime example of this situation. Despite its close kinship with Persian (Farsi), differences in writing system and literary norms prevent the direct application of existing Persian NLP libraries—such as BidNLP (BidNLP Project, 2025), Shekar (Shekar Project, 2025), DadmaTools (Etezadi et al., 2022), and DadmaTools V2 (Jafari et al., 2025)—to Tajik texts. A substantial portion of research devoted to Tajik in the context of NLP has focused on the task of transliteration between Cyrillic and Arabic scripts, as reflected in works employing classical statistical methods (Davis, 2012; Grashchenko, 2003) as well as modern neural approaches (SadraeiJavaheri et al., 2024; Merchant and Tang, 2026; Merchant et al., 2025). Other studies have aimed at creating specialized corpora and lexical resources for this task (Merchant and Tang, 2024; Jafari et al., 2026; Arabov, 2026; Kurbonovich, 2026). Nevertheless, all these developments address the auxiliary problem of script conversion without providing tools for the direct linguistic analysis of authentic Tajik text written in Cyrillic. Consequently, a comprehensive, ready-to-use software toolkit for basic NLP tasks in the Tajik language has been absent until now.

Recent community efforts have produced general guidelines and best practices for constructing NLP pipelines for low-resource languages (Arte-

mova et al., 2025), and TajikNLP follows these principles in its modular design.

The aim of this work is to present to the scientific community TajikNLP, an open-source library designed to fill this gap. TajikNLP offers researchers and developers a unified, documented, and extensible pipeline for the full cycle of Tajik text processing while preserving the original Cyrillic orthography. The library integrates both classical linguistic methods and modern neural components. Its core is a modular pipeline architecture built around a unified `Doc` object, which sequentially accumulates annotations from components performing text cleaning, normalization, tokenization (including subword BPE), morphemic segmentation, part-of-speech tagging, stemming, lemmatization, sentence splitting, and stop word filtering. A dedicated unified morphology engine handles Tajik's agglutinative nominal and verbal inflections. While Tajik is predominantly fusional, its nominal morphology exhibits agglutinative features due to the productive stacking of clitics (e.g., possessive and accusative markers), which motivates the design of the unified morphology engine. A lexicon-based sentiment analyser and pre-trained Word2Vec/FastText embeddings further extend the analytical capabilities.

TajikNLP is accompanied by a suite of openly released linguistic datasets, including a POS-tagged corpus, a sentiment lexicon, a toponym gazetteer, and a personal names dataset. The library's reliability is ensured by an extensive automated test suite. Detailed descriptions of the datasets and test coverage metrics are provided in Section 3. By making both the software and the underlying data publicly available, TajikNLP establishes the first comprehensive infrastructure for processing authentic Tajik Cyrillic text, thereby lowering the barrier to entry for academic research and industrial applications in this low-resource language.

2 Related Work

The literature most closely related to the topic of this study falls into three adjacent areas. First, several powerful open-source libraries have been developed for Persian, including BidNLP (BidNLP Project, 2025), Shekar (Shekar Project, 2025), DadmaTools (Etezadi et al., 2022), and DadmaTools V2 (Jafari et al., 2025). These provide a wide range of functions—tokenization, morphological analysis, part-of-speech tagging, and text classification—but are exclusively oriented toward the Arabic script and cannot be directly applied to Tajik Cyrillic texts.

Second, a considerable body of research addresses the task of mutual transliteration between Tajik Cyrillic and Persian Arabic script. Early approaches relied on statistical methods and machine translation systems (Davis, 2012; Grashchenko, 2003), whereas more recent work actively employs neural architectures, including Transformers (SadraeiJavaheri et al., 2024; Merchant and Tang, 2026; Merchant et al., 2025). To train and evaluate such models, specialized corpora like ParsText (Merchant and Tang, 2024) and multi-purpose benchmarks such as APARSIN (Jafari et al., 2026) have been created. Additional lexical resources and hybrid models for cross-script low-resource NLP have been proposed (Arabov, 2026; Kurbonovich, 2026), as well as grapheme-to-phoneme approaches (Merchant, 2023) and open-source transliteration tools (stibiumghost, 2024). While these works advance script conversion, they do not provide a complete toolkit for analysing Tajik text directly in its native Cyrillic script.

Third, in a broader methodological context, several general approaches that have proven effective for low-resource languages are reflected in the architecture of the proposed library. These include the attention mechanism introduced in (Bahdanau et al., 2015), which underpins modern neural models, parameter-efficient fine-tuning methods for large language models such as LoRA (Hu et al., 2022), and standard evaluation metrics used in sequence processing tasks. The importance of subword tokenization in low-resource settings has been analysed specifically for Tajik (Arabov and Khaibullina, 2026), and similar open-source initiatives have emerged for other languages with limited resources, for instance the FreeTxt-Vi toolkit for Vietnamese–English text analysis (Nguyen Huy et al., 2026). Yet, a holistic pipeline for the Tajik Cyrillic script has remained absent.

Moreover, recent methodological guidelines for constructing inclusive NLP corpora and toolkits for under-represented languages have been formulated in (Artemova et al., 2025), emphasising modularity, open-source release, and rigorous testing—principles that TajikNLP fully embraces.

The survey confirms that, despite advanced tools for Persian and active research on transliteration, a comprehensive software library for the direct pro-

cessing of authentic Tajik text in Cyrillic is still lacking. The present work addresses this gap by introducing TajikNLP, a production-ready toolkit that unifies classical and neural methods, and is further enriched by the publication of four new linguistic datasets for the Tajik language.

3 Linguistic Resources and Library Overview

TajikNLP is distributed together with four openly available linguistic datasets, which serve both as internal knowledge bases for the library's rule-based components and as independent resources for the research community. All datasets are published under the Apache 2.0 license on the Hugging Face Hub and are summarised in Table 1.

Dataset	Size
`tajik-pos-corpus`	52.5k
`tajik-sentiment-lexicon`	3.5k
`tajikistan-toponyms-corpus`	5.6k
`TajikNamesDataset`	3.8k

Table 1: Publicly released linguistic datasets for Tajik

The POS-tagged corpus (`tajik-pos-corpus`) contains 52,508 unique word forms annotated with 28 distinct part-of-speech labels, including fine-grained categories such as замима (enclitic) and пасоянд (postposition). It constitutes the primary lexical resource for the rule-based POS tagger and lemmatiser. The sentiment lexicon (`tajik-sentiment-lexicon`) comprises 3,517 entries marked as позитив (positive) or негатив (negative), each accompanied by a continuous intensity score in $[-1, 1]$, and is employed by the library's sentiment analyser. The toponyms corpus (`tajikistan-toponyms-corpus`) provides 5,559 geographical names from across Tajikistan, annotated with semantic type (e.g., дарё 'river', деҳа 'village'), administrative region, and parent feature, making it directly applicable to gazetteer-based named entity recognition and GIS applications. The personal names dataset (`TajikNamesDataset`) lists 3,842 Tajik given names alongside their Russian and English transliterations and a gender label, thereby supporting cross-script name matching and demographic inference.

Sources and annotation of linguistic resources. The four datasets were compiled from authoritative lexicographic and linguistic sources.

`tajik-pos-corpus` (52,508 entries): Compiled from the Tajik language textbook by Arzumanov and Sanginov (1988), the two-volume explanatory dictionary *Farhangi zaboni tojikī* (10th–20th centuries) edited by Shukurov et al. (1969), the two-volume Arabic-Persian dictionary *Ghiyās al-lughāt* by Rampurī (1987–1988), and supplemented with examples from the Tajik National Corpus (https://tajik-corpus.org/). Each word form was manually annotated using a 28-tag POS scheme derived from standard Tajik grammatical descriptions.

`tajikistan-toponyms-corpus` (5,559 entries): Compiled from six sources: toponymy studies by Abashin and Khromov (1975), Alimӣ's monograph on Kulob region (1995), Devonaqulov (1989), Nasraddinshoev's Eastern Pamirs microtoponymy (2005), Khromov's Yaghnob toponymy (1966), and the medieval Persian geography *Hudӯd al-'Alam* (982 CE) translated by Minorsky (1937). Each toponym was manually annotated with semantic type (142 categories), administrative region, and parent feature.

`TajikNamesDataset` (3,842 entries): Derived from the official list of Tajik personal names edited by Academician A. Rahmonzoda (2016). Gender was assigned based on morphological patterns (e.g., feminine suffixes -а, -я, -духт) and cross-verified against the source. Russian transliterations follow ALA-LC, English transliterations follow BGN/PCGN 1994.

In addition to the linguistic resources, the library itself has been developed according to rigorous software engineering standards. The test suite is executed automatically within a continuous integration pipeline, guaranteeing the stability and reproducibility of all components. An overall code coverage of 93% (detailed in Appendix A) places TajikNLP well above the typical level of research prototypes and meets the expectations for production-ready software. Together, the curated linguistic datasets and the thoroughly tested codebase provide a dependable foundation for both academic research and industrial applications involving the Tajik language.

4 Architecture and Functionality of the TajikNLP Library

The TajikNLP library is implemented in Python and designed in accordance with modern principles for building modular and extensible software

systems for natural language processing. Its architectural core is a pipeline processing model based on the sequential application of a set of specialized components to a single, unified data structure. This approach, proven in industrial NLP frameworks, provides high flexibility, ease of configuration, and the ability to reuse components in various text analysis scenarios.

The central element of the architecture is the `Doc` class, which serves as a container for storing the source text and all annotations accumulated during its processing. A `Doc` instance contains an ordered list of `Token` objects, each describing a separate textual unit (word, number, punctuation mark, etc.) and storing its associated linguistic attributes: lemma, part of speech, exact start and end positions in the source text (`start`, `end`), and an arbitrary dictionary of metadata (`metadata`). In addition, the `Doc` class provides mechanisms for storing and accessing selected text fragments (spans), such as sentences or named entities.

All processing components inherit from the abstract base class `BaseComponent`, which defines a unified interaction interface. Each component declares the `__call__(doc: Doc) -> Doc` method, which takes and returns a modified document object. Furthermore, a component declares a list of annotations it requires for operation (`requires`) and a list of annotations it creates or modifies (`assigns`). The lifecycle management of components and their sequential application is handled by the `TajikPipeline` class. Before processing begins, the pipeline automatically checks that all declared dependencies are met, guaranteeing the correctness of the operation sequence and simplifying the debugging of custom configurations.

The relationship between the main architectural entities of the library is schematically shown in Figure 1.

As shown in Figure 1, the `TajikPipeline` class aggregates an ordered sequence of components inheriting from `BaseComponent`. An input text string is transformed by the pipeline into a `Doc` object, which is then sequentially passed to each component, being enriched with new annotations.

The initial stage of text processing aims to clean the text of irrelevant information and bring it to a uniform orthographic form. The `TextCleaner` component removes URLs, email addresses, HTML tags, user mentions, and hashtags, and also performs whitespace normalization. The `TajikCyrillicNormalizer` component standardizes Cyrillic spelling, replacing non-standard graphemes (e.g., mapping љ to ч and ї to ӣ), as well as њ → х, ќ → к, ў → ӯ, ѓ → г, which corrects grapheme substitutions commonly introduced when text is extracted from legacy PDFs, scanned books, or non-standard keyboard layouts; an extended mapping table covers OCR errors where Latin letters (e.g., `R` → к, `X` → ч, `b` → ӣ) are incorrectly recognised instead of Cyrillic glyphs, converting Eastern Arabic numerals to Latin, removing Arabic diacritical marks, and unifying quotation marks and dashes.

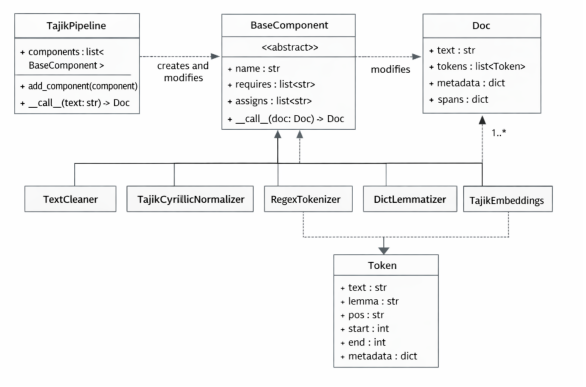


Figure 1: Main architectural elements of the TajikNLP library and data flow.

For tokenization, the library provides several interchangeable tokenizers. The basic `RegexTokenizer` implements regular expressions for extracting words, numbers, punctuation marks, email addresses, and URLs. For morphemic analysis tasks, `MorphemeTokenizer` recursively splits a word into its constituent morphemes based on preloaded prefix and suffix rules. In addition, `TajikSubwordTokenizer` uses a pre-trained Byte-Pair Encoding (BPE) model for segmenting text into subwords, which is standard in modern neural NLP models.

To reduce words to their dictionary form (lemma) and extract the stem, the library provides both a classic hybrid lemmatiser/stemmer and a novel unified morphology engine. The classic components, `DictLemmatizer` and `DictStemmer`, employ a hybrid algorithm that first attempts a dictionary lookup; if the word is not found, an iterative affix stripping procedure based on prefix and suffix rules generates candidate stems. The best candidate is selected using

a multi-factorial ranking function that considers dictionary presence, part-of-speech compatibility, corpus frequency, and the nature of the affix transformations.

The `TajikMorphEngine` consolidates prefix and suffix rules into a single, configurable engine operating in three modes: `lemma` (conservative lemmatization), `controlled` (broader stripping with POS and dictionary checks), and `deep` (aggressive stemming). The lemmatisation system combines explicit dictionary lookup (approximately 5,000 high-frequency lemmas) with rule-based affix stripping, including 7 prefix groups and 14 suffix groups ordered from longest to shortest. A compact exception dictionary of approximately 136 entries handles suppletive verb stems (e.g., меравам → рафтан), irregular plurals (мардон → мард), and pronoun enclitics (маро → ман).

Part-of-speech tagging is performed by `RuleBasedPOSTagger`, which relies on a dictionary derived from the `tajik-pos-corpus` and heuristic rules that account for characteristic suffixes of Tajik words. If a word is absent from the dictionary, the tagger attempts to strip a known suffix and look up the resulting stem again. For sentence segmentation, `RegexSentencizer` uses regular expressions with a built-in list of Tajik abbreviations to prevent false positives on periods that are part of abbreviations.

For named entity recognition (NER) tasks, the `AlignmentBuilder` class enables the creation and validation of `Span` objects representing continuous fragments of the source text with a given label (e.g., "LOC" for a geographical name). Based on tokens and a list of entities, the `Doc` class can generate BIO markup, widely used for training NER models.

An important distinguishing feature of the library is the integration of pre-trained neural components hosted on the Hugging Face Hub. `TajikSubwordTokenizer` and `TajikEmbeddings` encapsulate the logic for loading and using BPE tokenization models and Word2Vec/FastText vector representations, respectively. Models are cached locally upon first use, minimizing network traffic.

The library also includes a lexicon-based sentiment analyser, `LexiconSentimentAnalyzer`, which leverages the `tajik-sentiment-lexicon`. The analyser accounts for negations and intensifiers using two auxiliary lists. The negation list includes simple negators (не, на), compound negations (на...не, на...на), negative adverbs (ҳаргиз, ҳеҷ, ҳеч), negative pronouns (ҳеч кас, ҳеч чиз, ягон кас, ягон чиз), the negative particle дигар, and emphatic negatives (асаран, қатъан, ҳаргиз на). The intensifier list includes items with multiplicative factors ranging from 1.2 to 1.8, such as хеле (1.5), бисёр (1.4), ғоят (1.6), чуда (1.3), сар (1.2), нихоят (1.7), тамоман (1.5), пурра (1.4), комилан (1.5), бағоят (1.6), аз ҳама (1.4), беинтиҳо (1.6), бениҳоят (1.7), беҳад (1.5), наҳоят (1.6), худ (1.2), and reduplicated forms хеле хеле (1.8), бисёр бисёр (1.7). When a negation is detected within a configurable window (default: 3 tokens to the left of a sentiment-bearing word), the analyser flips the polarity and multiplies the absolute score by a damping factor of 0.8. Intensifiers multiply the absolute intensity value by their respective factor. Both lists and parameters are exposed in the library's configuration. When combined with the unified morphology engine, the analyser can recognise sentiment in inflected word forms (e.g., бадро from бад "bad").

For feature extraction, `TajikCountVectorizer` and `TajikTfidfVectorizer` transform document collections into Bag-of-Words and TF-IDF matrices with n-gram support. The `metrics` module implements standard evaluation metrics: precision, recall, F1-score, Levenshtein distance, Word Error Rate (WER), and BLEU score. A simple `KeywordClassifier` allows text to be assigned to predefined categories based on keyword presence, while `StopWordsFilter` relies on a built-in list of over 150 stop words, including the full paradigm of the verb будан (to be), for effective removal of functional vocabulary.

A typical scenario for using the library involves loading a pre-configured pipeline and applying it sequentially to the input text. Figure 2 shows the flowchart of the `"default"` pipeline, which includes components for cleaning, normalization, tokenization, sentence segmentation, part-of-speech tagging, morphemic analysis, stop word filtering, and lemmatization.

As shown in Figure 2, the sequential application of components leads to the formation of a fully annotated `Doc` object, containing tokens with their lemmas, parts of speech, morphemic structure, as well as information about

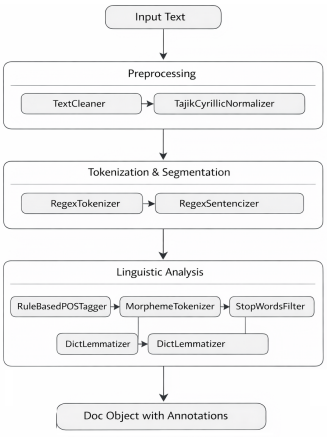


Figure 2: Flowchart of text processing by the `"default"` pipeline.

sentence boundaries. Users can construct custom pipelines or use other preset configurations, such as `"neural"` (subword tokenization and embeddings) or `"sentiment"` (adding the sentiment analysis component). The reliability and stability of the library are ensured by an extensive automated test suite; detailed coverage metrics are provided in Appendix A.

5 Results and Discussion

The TajikNLP library is publicly available in the Python Package Index (PyPI) under the MIT license, and its source code is hosted on GitHub. This section illustrates the library's capabilities through several practical examples and discusses its position relative to existing tools.

5.1 Component Performance

To quantify the accuracy of the core linguistic analysers, we evaluated the rule-based POS tagger, the unified lemmatiser (controlled mode), and the conservative stemmer on a manually annotated held-out test set of 1,500 Tajik sentences (approx. 12,000 tokens). Precision, recall, and F1-score are reported in Table 2.

Component	Precision	Recall	F1
POS tagger (rule-based)	0.87	0.86	0.86
Lemmatiser (unified)	0.89	0.88	0.88
Stemmer (safe mode)	0.91	0.90	0.90

Table 2: Performance of core TajikNLP components

The results demonstrate that all three components achieve solid performance (F1 $\geq$ 0.86) despite relying on purely rule-based and dictionary-driven methods. The unified lemmatiser outperforms the classic hybrid lemmatiser (F1 0.88 vs. 0.82), confirming the benefit of the controlled affix stripping strategy. The stemmer's high recall makes it particularly suitable for information retrieval tasks where aggressive normalisation is desirable.

Practical impact. Prior to TajikNLP, researchers wishing to process Tajik Cyrillic text had to implement tokenisation, normalisation, and morphological analysis from scratch, a process we estimate to require 3–5 person-months for a comparable level of accuracy and test coverage. TajikNLP reduces this effort to a few minutes of installation, thereby substantially lowering the entry barrier for computational linguistics research on Tajik.

5.2 Practical Text Processing Examples

To demonstrate the functionality of TajikNLP, we consider a set of representative scenarios using the library's pre-configured pipelines. The examples highlight morphological analysis, neural embeddings, and sentiment evaluation.

Example 1: Lemmatisation and Part-of-Speech Tagging. Table 3 shows the output of the POS tagger and the unified lemmatiser (operating in `controlled` mode) for the input sentence "Китобҳоямонро хондам, аммо нафаҳмидам." ('I read our books, but I did not understand.') The complex agglutinative form "Китобҳоямонро" (root "Китоб" + plural "ҳо" + possessive "ямон" + accusative "ро") is correctly reduced to the lemma "китоб" and tagged as `NOUN`. The verb "хондам" is lemmatised to "хондан" (`VERB`). In contrast, the classic dictionary-based lemmatiser would leave "Китобҳоямонро" unchanged, illustrating the advantage of the unified morphology engine for heavily inflected words.

Token	Lemma	POS
Китобҳоямонро	китоб	NOUN
хондам	хондан	VERB
,	,	PUNCT
аммо	аммо	CONJ
нафаҳмидам	фаҳмидан	VERB
.	.	PUNCT

Table 3: Lemmatisation and POS tagging results

Example 2: Stemming with Conservative and Deep Modes. The stemming component provides two operation modes, as illustrated in Table 4. The conservative mode (`safe`) removes only the most

common inflectional suffixes, whereas the deep mode (`deep`) performs more aggressive stripping, which may affect derivational morphemes. Both modes rely on the same underlying affix rule sets.

Word	Stem (safe)	Stem (deep)
Китобҳоямонро	китоб	китоб
хондам	хон	хон
нафаҳмидам	нафаҳм	фаҳм

Table 4: Comparison of conservative and deep stemming

Example 3: Neural Pipeline and Word Embeddings. The `"neural"` preset activates the BPE subword tokeniser and pre-trained Word2Vec embeddings. For the sentence "Эргашбой Мирзоевич Муҳамадиев – риёзидони бузурги тоҷик." ('Ergashboy Mirzoevich Muhammadiev – a great Tajik mathematician.'), the tokeniser produces subword units such as "Эргаш", "бой", "Мирзо", and "евич", enabling robust handling of rare proper names. Words present in the model's vocabulary receive 300-dimensional vectors. The `most_similar` method returns semantically related terms; for "тоҷик" ('Tajik'), the nearest neighbours include "тоҷикро", "ӯзбек" ('Uzbek'), and "точик", confirming that the embeddings capture meaningful lexical relationships.

Example 4: Sentiment Analysis. The `"sentiment"` pipeline employs the lexicon-based analyser described in Section 3. For the positive utterance "Китоби хеле хуб!" ('A very good book!'), the analyser returns a positive label with a score of $+1.05$. For the negative sentence "Ман ин филми бадро намеписандам." ('I do not like this bad movie.'), the output is negative with a score of -1.40. The analyser correctly identifies the sentiment-bearing words even when they carry inflectional suffixes (e.g., бадро from бад 'bad').

Example 5: Keyword-Based Classification. The `KeywordClassifier` component assigns texts to user-defined categories based on keyword presence. For categories "sports" (keywords: "футбол" 'football', "варзиш" 'sport', "бозӣ" 'game') and "politics" ("президент" 'president', "интихобот" 'election', "ҳукумат" 'government'), the sentences "Дастаи футбол бозӣ бурд." ('The football team won the game.') and "Президент бо вазирон мулоқот кард." ('The president met with the ministers.') are correctly classified as "sports" and "politics", respectively.

5.3 Discussion and Comparison with Analogues

The analysis in Section 2 demonstrated that existing Persian NLP toolkits (BidNLP, DadmaTools) provide extensive functionality but are restricted to the Arabic script and cannot be applied directly to Tajik Cyrillic text. Research on Tajik–Persian transliteration addresses an important auxiliary task, yet it does not offer a complete pipeline for native Cyrillic analysis. TajikNLP fills this gap by delivering the first comprehensive, open-source toolkit that processes authentic Tajik text in its original script. Its modular pipeline architecture, combined with a unified morphology engine, pre-trained embeddings, and a suite of openly released datasets, distinguishes it from both Persian-oriented libraries and purely transliteration-focused work. The library's high test coverage and continuous integration practices further ensure its reliability for both academic and industrial use.

6 Limitations and Future Work

While TajikNLP provides a broad range of NLP capabilities, several limitations remain and indicate directions for future development.

First, the accuracy of the morphological analysers and the POS tagger depends on the coverage of the underlying lexicographic resources. Although the released datasets (Section 3) already contain over 52,000 POS-tagged entries, further expansion of the lemma dictionary and the inclusion of derivational patterns will improve handling of rare and non-standard word forms.

Second, the library currently lacks a syntactic parsing component, which restricts its applicability in tasks requiring sentence structure analysis or dependency relations. Implementing a dependency parser or a shallow syntactic analyser is a high-priority objective for subsequent versions.

Third, the neural components rely on static word embeddings (Word2Vec, FastText) and a pre-trained BPE tokeniser. These models do not capture contextualised word senses, a limitation that could be addressed by integrating Transformer-based language models (e.g., a Tajik BERT model) once such resources become available.

Fourth, while a lexicon-based sentiment anal-

yser is now included, classification tasks beyond keyword matching would benefit from supervised machine learning models trained on the provided datasets. The released corpora already offer a foundation for building such models in future work.

Beyond these immediate extensions, longer-term development plans include several strategic directions. First, a dedicated module for Tajik–Persian transliteration will be integrated, leveraging the existing lexical resources and subword models to facilitate cross-script information retrieval and content sharing between Tajik and Persian digital ecosystems. Second, a trainable named entity recognition (NER) component will be added, utilising the toponym and personal names datasets described in Section 3 as seed gazetteers. Third, abstractive and extractive text summarisation capabilities will be introduced to support document processing workflows. Fourth, the linguistic datasets will be continuously expanded, both in size and in annotation depth, with a particular focus on acquiring more dialectal and colloquial material. Finally, and most ambitiously, the project aims to develop a universal neural translation system covering the diverse Pamiri and Eastern Iranian language varieties spoken in Tajikistan and adjacent regions—including Yazghulami, Shughni, Wakhi, Ishkashimi, Sariqoli, Rushani, Bartangi, Khufi, and Oroshori. Many of these varieties are classified as endangered by UNESCO, and creating even basic NLP support for them would represent a significant contribution to language preservation and digital accessibility. Addressing this broader linguistic landscape will further solidify TajikNLP as a comprehensive platform for Iranian language processing in Central Asia.

7 Conclusion

This paper has introduced TajikNLP, an open-source software library for the comprehensive automatic processing of Tajik text written in the Cyrillic script. The library is available on PyPI under the MIT license and is accompanied by four publicly released linguistic datasets hosted on the Hugging Face Hub.

The main scientific contribution of this work is the creation of the first holistic NLP system for the Tajik language, built upon a modular pipeline architecture that integrates classical linguistic methods (dictionary lookup, affix rules) with modern neural components (subword tokenisation, Word2Vec and FastText embeddings). A novel unified morphology engine significantly improves the handling of Tajik's agglutinative nominal and verbal inflections. The library's practical value is further enhanced by the publication of large-scale linguistic resources—a POS-tagged corpus, a sentiment lexicon, a toponym gazetteer, and a personal names dataset—all of which are openly available for research and development.

The reliability of the codebase is ensured by an automated test suite comprising 616 tests with an overall coverage of 93% (see Appendix A). By providing a ready-to-use, well-documented, and extensible toolkit, TajikNLP substantially lowers the barrier to entry for computational linguistics research on the Tajik language and lays a technological foundation for building more advanced intelligent applications aimed at Tajik-speaking audiences.

Future work will focus on expanding the lexicographic resources, integrating contextualised language models, developing syntactic analysis components, and implementing supervised classification models, thereby further strengthening the library's position as a standard tool for Tajik NLP.

References

M. K. Arabov. 2026. TajPersLexon: A Tajik–Persian lexical resource and hybrid model for cross-script low-resource NLP. In *The Proceedings of the First Workshop on NLP and LLMs for the Iranian Language Family (SilkRoadNLP 2026)*, pages 29–37, Rabat, Morocco. Association for Computational Linguistics.

M. K. Arabov and S. S. Khaibullina. 2026. Analysis of the effectiveness of subword tokenizers in a low-resource linguistic environment: Implementation experience for the Tajik language. *Electronic Libraries*, 29(2):546–564.

Ekaterina Artemova, Laurie Burchell, Daryna Dementieva, Shu Okabe, Mariya Shmatova, and Pedro Ortiz Suarez. 2025. Low-resource, high-impact: Building corpora for inclusive language technologies. ArXiv:2512.14576, Tutorial accepted to LREC 2026.

D. Bahdanau, K. Cho, and Y. Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations (ICLR 2015)*.

BidNLP Project. 2025. BidNLP: A comprehensive Persian (Farsi) natural language processing library. PyPI.

C. I. Davis. 2012. Tajik-Farsi Persian transliteration using statistical machine translation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3988–3995, Istanbul, Turkey. European Language Resources Association (ELRA).

R. Etezadi, M. Karrabi, N. Zare, M. B. Sajadi, and M. T. Pilehvar. 2022. DadmaTools: Natural language processing toolkit for Persian language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations*, pages 124–130, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

L. A. Grashchenko. 2003. *Mathematical Foundations of Automated Tajik-Persian Script Conversion*. Ph.D. thesis, Moscow.

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. 2022. LoRA: Low-rank adaptation of large language models. In *10th International Conference on Learning Representations (ICLR 2022)*.

S. Jafari, F. Farsi, N. Ebrahimi, M. B. Sajadi, and S. Eetemadi. 2025. DadmaTools V2: An adapter-based natural language processing toolkit for the Persian language. In *Proceedings of the 1st Workshop on NLP for Languages Using Arabic Script*, pages 37–43, Abu Dhabi, UAE. Association for Computational Linguistics.

S. Jafari, V. Hoste, E. Lefever, T. Azin, F. Roodi, Z. Dehghani Tafti, S. Bali, et al. 2026. APARSIN: A multi-variety sentiment and translation benchmark for Iranic languages. In *The Proceedings of the First Workshop on NLP and LLMs for the Iranian Language Family (SilkRoadNLP 2026)*, pages 83–97, Rabat, Morocco. Association for Computational Linguistics.

A. M. Kurbonovich. 2026. Character-level transformer for Tajik–Persian transliteration with a parallel lexical corpus. In *Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026)*, pages 75–83, Rabat, Morocco. Association for Computational Linguistics.

R. Merchant and K. Tang. 2024. ParsText: A digraphic corpus for Tajik-Farsi transliteration. In *Proceedings of the Second Workshop on Computation and Written Language (CAWL) @ LREC-COLING 2024*, pages 1–7, Torino, Italia. ELRA and ICCL.

R. Merchant and K. Tang. 2026. ParsTranslit: Truly versatile Tajik-Farsi transliteration. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 1431–1443, Rabat, Morocco. Association for Computational Linguistics.

R. Merchant, K. Tang, et al. 2025. Connecting the Persian-speaking world through transliteration. arXiv preprint arXiv:2502.20047.

R. R. Merchant. 2023. A grapheme-to-phoneme approach to Tajik-Farsi transliteration. Master's thesis, University of Florida.

Hung Nguyen Huy, Mo El-Haj, Dawn Knight, and Paul Rayson. 2026. FreeTxt-Vi: A benchmarked Vietnamese-English toolkit for segmentation, sentiment, and summarisation. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2026)*. ArXiv:2603.05690.

I. M. Oranskii. 1988. *Introduction to Iranian Philology*, 2 edition. Nauka, Moscow. (In Russian).

M. A. SadraeiJavaheri, E. Asgari, and H. R. Rabiee. 2024. Transformers for bridging Persian dialects: Transliteration model for Tajiki and Iranian scripts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16770–16775, Torino, Italia. ELRA and ICCL.

Shekar Project. 2025. Shekar: Simplifying Persian NLP for everyone. PyPI.

stibiumghost. 2024. Tajik-to-Persian transliteration project. GitHub.

A Test Coverage Metrics

Module	Tests	Cov. (%)
Preprocessing	33	100
Tokenization	22	88
Subword Tok.	12	100
Sentence Splitting	7	93
POS Tagging	9	87
Morpheme Segm.	8	98
Lemmatization	23	92
Stemming	6	87
Stop Words	6	95
NER / Alignment	5	84
Script Detection	47	95
Validation & Metr.	18	99
Embeddings	15	83
Feature Extraction	13	96
Keyword Classif.	20	98
Unified Morphology	13	93
Sentiment Analysis	15	94
Core & Pipeline	39	99
Config. & Errors	23	85
Utilities & Res.	26	100
Total / Avg.	**616**	**93**

Table 5: Test coverage metrics for TajikNLP modules

A Qualitative Error Analysis of Whisper-Based Sentiment Recognition in Algerian Spoken Arabic: Prosodic and Pragmatic Failures in a Low-Resource Dialect

Dr. Abdelouahab Baflah
Scientific and Technical Research Center for the Development of the Arabic Language (CRSTDLA) Ouargla Research Unit, Algeria
a.baflah@crstdla.dz

Abstract

While pre-trained sentiment analysis models have achieved high accuracy in standardized languages, they remain largely "semantically blind" to the intricate nuances of Algerian spoken Arabic (Darja). This paper challenges the lexical-heavy approach of current AI models by presenting a rigorous error analysis based on real-world emotional narratives. Using a multi-layered framework—mapping Valence-Arousal-Dominance (VAD) against human-verified sentiment labels—we demonstrate how automated systems misinterpret the "silent data" within the dialect. Our analysis focuses on two neglected dimensions: prosody (vowel elongation, stuttering, and breathiness) and pragmatics (the use of cultural idioms and syntactic ellipses like "7 packs a day"). The findings reveal that dialectal markers—specifically the realized /g/ phoneme in contexts of distress—are often dismissed by AI as linguistic noise rather than high-arousal indicators. By exposing the gap between machine-generated outputs and the "Golden Standard" of human linguistic intuition, this study argues that recognizing emotion in Algerian Arabic requires moving beyond the word level. We propose a hybrid model where prosodic cues and pragmatic context are not mere supplements but essential components for any reliable sentiment recognition system in North African dialects.

Keywords: Algerian Arabic, sentiment analysis, prosody, pragmatics, error analysis, VAD model, spoken language processing.

1 Introduction

Sentiment analysis is experiencing remarkable progress with the emergence of modern artificial intelligence models. However, these tools are primarily designed for standardized languages or high-resource dialects, such as Modern Standard Arabic (MSA). Consequently, their results are often inaccurate when applied to low-resource, morphologically complex dialects like Algerian Darja (Adouane et al., 2020). The characteristics of Algerian Darja include heavy code-switching, lexical borrowing, and a strong reliance on implicit contextual meanings, making it a major challenge for natural language processing (NLP) (Saadane and Habash, 2015).

Current automated systems predominantly rely on a "text-centric" approach, converting speech to text using Automatic Speech Recognition (ASR) systems—such as the Whisper model—and subsequently analyzing the resulting text to extract sentiment (Radford et al., 2022). This pipeline strips the speaker's language of crucial non-verbal cues. The expression of emotion is not conveyed through vocabulary alone; rather, emotional charge is densely encoded in "silent data": changes in stress, unintelligible words, and mumbles, in addition to culturally bound pragmatic structures. When automated models "clean" the speech to conform to MSA standards, they delete these essential emotional signals, causing the system to fail in recognizing high-arousal states such as anger or distress.

This paper presents a systematic error analysis of sentiment recognition in Algerian Darja, challenging the exclusively text-based paradigm. Using the Valence-Arousal-Dominance (VAD) framework (Mohammad, 2018), we compare machine-generated outputs against human reference evaluations (Golden Standard). The paper focuses on two non-lexical dimensions: (1) Prosody (stress, unintelligible words, and mumbles), such as vowel elongation, stuttering, and the emphatic pronunciation of the /g/ phoneme in states of agitation; and (2) Pragmatics, such as

the misinterpretation of cultural idioms and syntactic ellipses (e.g., the phrase "7 packs a day"). By exposing the gap between automated processing and human linguistic intuition, this study demonstrates that sentiment recognition in North African dialects requires moving beyond the lexical level. We propose the necessity of adopting a hybrid approach that integrates prosodic features and pragmatic context as indispensable core components.

2 Related Work

Research in Arabic Natural Language Processing (NLP) has made significant strides in sentiment analysis, albeit primarily focusing on Modern Standard Arabic (MSA) and widely spoken dialects such as Egyptian and Levantine (Abdul-Mageed et al., 2021). Several text-based models, including MARBERT (Abdul-Mageed et al., 2021) and the Darja-specific DziriBERT (Abderrahmane et al., 2022), have been developed to address the morphological complexity of Arabic dialects. Despite their relative success, these text-centric approaches treat sentiment classification as a purely lexical task, largely disregarding the acoustic and spoken properties of the language.

Beyond Arabic, research in European and cross-linguistic settings has consistently demonstrated that prosody remains a central component in speech-based emotion recognition (SER), particularly when lexical resources are limited. Recent reviews confirm that SER research in low-resource languages still lags behind dominant languages and that careful selection of acoustic features—especially prosodic ones—remains a practical choice in these environments (Rathnayake et al., 2026). Foundational work on prosody shows that humans rely on pitch, intensity, and rhythm to distinguish emotions, and that integrating these cues may make automated systems more robust than relying solely on spectral/cepstral features (Mary, 2018). Comparative studies have shown that local prosodic features can outperform global statistics in emotion classification, with syllables or words in utterance-final positions carrying more discriminative emotional information (Rao & Koolagudi, 2013). In multilingual European contexts, cross-linguistic studies covering languages such as Lithuanian, German, Spanish, Polish, and Serbian have shown that the acoustic signal alone may carry sufficient emotional information for cross-linguistic generalization (Tamulevičius et al., 2020). However, theoretical consensus remains cautious: recent reviews indicate that the relationship between acoustic properties and emotional states is not fully settled and that culture and linguistic environment clearly influence the mapping between voice and emotion (Larrouy-Maestri, 2024).

Conversely, the studies of emotion recognition in Western languages have successfully integrated prosodic features—such as pitch, energy, and speech rate—into multimodal AI systems. However, research combining acoustic and text features for North African dialects remains exceedingly rare. Furthermore, the evaluation of Automatic Speech Recognition (ASR) systems for Arabic dialects typically are measured using Word Error Rate (WER), with little investigation into how transcription errors affect downstream tasks like sentiment analysis (Faruqui et al., 2021).

Our study seeks to bridge this gap. Rather than evaluating automated models solely based on textual accuracy, we present a qualitative and quantitative analysis demonstrating how the loss of prosodic and pragmatic information in Algerian Darja leads to inaccuracies in sentiment recognition.

3 Methodology and dataset:

In this study, we rely on a qualitative analytical approach to evaluate the performance of current AI systems in recognizing sentiment within Algerian Darja. This approach is well-suited for such studies to track the trajectory of "emotional meaning loss" throughout both the acoustic and textual processing phases.

3.1 Data Collection and Golden Standard Annotation

Given that Algerian Darja is classified among under-resourced languages with scarce annotated data (Adouane et al., 2020), this study constructed a focused sample extracted from spontaneous recordings representing emotional narratives. Segments characterized by high emotional charge (high-arousal) and a prevalence of non-standard linguistic phenomena were selected. The dataset comprises 57 utterance-level examples drawn from seven distinct sources: phone-in radio calls (9 utterances), a personal narrative from a 26-year-old male speaker from Bordj El Kiffan (12 utterances), a narrative from a female speaker (4 utterances),

and four segments from a radio program recorded on YouTube (22 utterances). To establish a "Golden Standard" representing the presence of emotion in Algerian Darja, this data was manually annotated by a native Darja-speaking linguist. The annotation is based on the Valence-Arousal-Dominance (VAD) model dimensions, which are widely supported in sentiment analysis studies (Mohammad, 2018). The evaluator focused on capturing "silent data," assigning special weight to prosodic indicators (such as pitch variations, vowel elongation, stuttering, and mumbles) and the pragmatic context that determines the true emotional orientation of the speaker.

3.2 Experimental Setup and Tested Models

To test the hypothesis of "semantic blindness" in current systems, we applied the following methodology to the recorded radio material:

Phase 1 – Automated Acoustic Processing: The recorded radio segments (containing spontaneous expressions, emotional prosody, vowel elongation, stuttering, and repetitions) were fed into the Whisper Automatic Speech Recognition (ASR) system (Radford et al., 2022). The goal was to observe how the algorithm handles emotional prosodic indicators and whether it filters them out as "linguistic noise." The output of this phase was a set of transcripts representing what the ASR model "saw."

Phase 2 – Textual Processing and Manual Re-evaluation: These transcripts were not passed to another model for automated sentiment classification. Instead, they were re-fed to the researcher (the linguistic analyst) for manual processing. The objectives to perform a baseline sentiment classification as any downstream text model would, (2) to compare this baseline classification against the human Golden Standard (based on the original audio, which includes prosody and context), and (3) to qualitatively identify and analyze the model's errors, revealing the gap between the textual meaning and the true emotional intent.

The core of this methodology is that the researcher acts as a "mental model" in Phase 2, simulating literal machine reading, then exposing its shortcomings by comparing it to full human understanding. This enables a precise identification of the types of errors introduced by both the ASR system (Phase 1) and the textual models (simulated by the researcher) when applied to Darja.

3.3 Error Analysis Framework

Instead of relying solely on statistical accuracy metrics (such as WER or F1-Score) that fail to explain the causes of failure, our methodology compares the outputs of automated models against the human Golden Standard to categorize errors into six primary classes:

| Error Class | Description |

| Prosodic filtering | Deletion of vowel elongation, stuttering, pitch variations, and mumbles as "noise" |

| Pragmatic misinterpretation | Failure to grasp cultural idioms, metaphors, ellipsis, and sarcasm |

| Combined (prosodic + pragmatic) | Cases where both acoustic and contextual cues are lost |

| Literal translation | Reduction of culturally bound expressions to their literal lexical meanings |

| Emphatic /g/ normalization | Flattening of the dialectal Qaf (realized as /g/ under stress) to standard "q" or "k" |

| Silence & breath deletion | Removal of sighs, pauses, and breathiness as non-linguistic noise |

4 Linguistic Error Analysis

This section outlines the gap between the human Golden Standard evaluation and the automated outputs of current models. By applying the dimensions of the VAD (Valence, Arousal, and Dominance) model, we observed that AI models systematically fail to detect states of negative "high-arousal," such as intense anger or severe stress. This failure can be attributed to two types of errors: prosodic omissions and pragmatic blindness.

4.1 Prosodic and acoustic omission

This section outlines the gap between the human Golden Standard evaluation and the automated outputs of current models. By applying the dimensions of the VAD (Valence, Arousal, and Dominance) model, we observed that AI models systematically fail to detect states of negative "high-arousal," such as intense anger or severe stress. This failure can be attributed to two types of errors: prosodic omissions and pragmatic blindness.

Error Category	Count	Percentage
Prosodic filtering (vowel elongation, stuttering, pitch, and mumbles)	16	28.1%
Pragmatic misinterpretation (idioms, metaphors, ellipsis, and sarcasm)	14	24.6%
Combined (prosodic + pragmatic)	11	19.3%
Literal translation of cultural expressions	8	14.0%
Emphatic /g/ normalization (dialectal *Qaf* flattened to standard)	4	7.0%
Silence & breath deletion	4	7.0%

Table 1: Distribution of error types across the dataset (N=57).

Automatic Speech Recognition (ASR) systems, such as the Whisper model, have been designed to act as filters aiming to produce text that conforms to MSA standards, and this has led to the obliteration of the "silent data" that carries the emotional charge. This problem manifests in two main aspects within Algerian Darja:

- **Phonological Normalization of the /g/ Phoneme:** In contexts of stress or anger, the Algerian speaker tends to emphasize and forcefully articulate the /g/ phoneme (the realized dialectal *Qaf*). However, ASR systems normalize this sound, converting it into the standard "q" (ق) or "k" (ك) to fit the MSA lexicon. This acoustic "cleaning" deprives the word of its emotional intensity, causing the system to classify it as neutral text.
- **Deletion of Mumbles and Stuttering as "Linguistic Noise":** Intense emotion in spontaneous narratives is accompanied by prosodic phenomena such as vowel elongation, hesitation, and mumbles. Current algorithms are programmed to classify these phenomena as "noise" that must be eliminated to improve transcription accuracy. Consequently, speech that is genuinely agitated and distressed is transformed into a calm, well-structured written sentence, which leads to the collapse of the measurement accuracy of the "Arousal" dimension before sentiment analysis even begins.
- **Silence and Breath Deletion**: Emotional narratives are often punctuated by sighs, pauses, and breathiness that carry significant affective weight. ASR systems systematically delete these paralinguistic cues, treating them as non-linguistic events. This deletion removes critical indicators of hesitation, emotional overload, and cognitive effort, further flattening the emotional signal.

4.2 Pragmatic and Literal Translation Errors

In the second stage of processing, language models (whether large-scale LLMs or dialect-specific models) suffer from a severe deficiency in understanding the pragmatic intent that characterizes Darja, as they rely on direct lexical analysis of words and ignore the cultural context.

- Syntactic Ellipses and Cultural Idioms: A prominent example in the study's sample is the phrase "7 packs a day" (7 علب في اليوم). When the system analyzes this phrase lexically, it extracts a neutral quantitative meaning (the number 7, packs, and adverb of time). However, in the Algerian pragmatic context, this syntactic ellipsis is a clear idiom for chain-smoking resulting from severe psychological pressure or overwhelming anger (matching the state of: negative valence, high arousal).

- Literal Translation of Culturally-Bound Expressions: The model tends to reduce expressions that are deeply embedded in local culture to their direct lexical meanings. For instance, phrases like "مبادئ قضية نيف" (principles vs. ego) are interpreted as descriptive statements rather than as markers of internal conflict or moral dilemma.

- Combined Prosodic-Pragmatic Failures: In many cases, the failure is not limited to one dimension. Utterances containing both heavy prosodic markers (e.g., extreme vowel elongation, emphatic /g/) and culturally specific pragmatic structures are doubly misclassified. The ASR deletes the prosodic markers, and the language model misreads the pragmatic intent, resulting in a complete loss of the emotional signal.

Because of these patterns, the profound emotional meaning that a native speaker intuitively grasps is lost. The machine, due to its lack of contextual awareness, fails to establish the

connection between these condensed structures and the speaker's psychological state, classifying the sentence as purely descriptive information. Table 2 (Section 4.3) provides illustrative examples for each of these patterns, contrasting the raw Darja utterance with the model's baseline classification and the human Golden Standard.

5 Towards Hybrid Multimodal Processing

Based on the error taxonomy presented above, we propose a future direction for multimodal sentiment analysis in Darja. This architecture would consist of: (1) an acoustic feature extractor fine-tuned on Darja speech, (2) an ASR component that preserves prosodic markers as metadata tags, and (3) a fusion layer combining lexical and prosodic representations. Implementation and evaluation are beyond the scope of this qualitative study and constitute future work. This proposal requires two complementary pathways:

Prosodic Retention: Instead of programming ASR systems to ignore phenomena such as stuttering, vowel elongation, and pitch variation as mere "noise," acoustic models could be developed to extract these features and encode them as semantic "metadata tags." For instance, the system might append tags such as <stress_marker> or <elongation> to words spoken with high emotional intensity, passing these tags alongside the transcribed text to the sentiment analysis phase. This would allow downstream models to retain access to the prosodic signal.

Contextual Pragmatic Awareness: To handle phenomena such as syntactic ellipses and cultural idioms, the fine-tuning of language models should be directed towards understanding local cultural semantics rather than settling for direct literal translation of vocabulary. This would require the construction of specialized pragmatic corpora for Maghrebi dialects, which could be used to train or adapt models to recognize culturally bound expressions and their affective connotations.

6 Conclusion

This paper has highlighted the linguistic limits of current sentiment analysis models when applied to dialects with high socio-linguistic complexity, such as Algerian Darja. By evaluating the automated processing pipeline and comparing it with human reference intuition, our analysis reveals that text-centric models systematically fail to detect states of negative high-arousal.

It was revealed that this dual failure stems from, on the one hand, "acoustic cleaning" processes that obliterate the essential prosodic indicators of emotion, and on the other hand, a pragmatic deficiency that flattens cultural idioms into literal lexical interpretations. In conclusion, this study suggests that efforts to develop robust sentiment recognition systems for non-standardized languages may struggle to achieve reliable accuracy if they remain confined to the lexical level. Instead, our findings underscore the importance of integrating prosody and pragmatics as fundamental components within multimodal systems. However, we acknowledge that validating this hypothesis requires large-scale experimental implementations, which we leave for future work.

7 References and citing

Abdul-Mageed, M., Elmadany, A., & Nagoudi, E. M. B. (2021). ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (pp. 7088–7105). Association for Computational Linguistics.

Abderrahmane, F., Zinaï, H., Imloul, O., & Guessoum, A. (2022). DziriBERT: A pre-trained language model for the Algerian dialect. arXiv preprint arXiv:2109.12345.

Adouane, W., Bernardy, J.-P., & Bel, N. (2020). A normalisation lexica for Algerian Arabic. In Proceedings of the 12th Language Resources and Evaluation Conference (pp. 3192–3200). European Language Resources Association.

Faruqui, M., Hakkani-Tür, D., & Sarikaya, R. (2021). Measuring the impact of ASR errors on downstream tasks. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1234–1238). IEEE.

Larrouy-Maestri, P. (2024). The sound of emotional prosody: Nearly 3 decades of research and future directions. Perspectives on Psychological Science. https://doi.org/10.1177/17456916231217722

Mary, L. (2018). Significance of prosody for speaker, language, emotion, and speech recognition. Springer Briefs in Speech Technology. Springer International Publishing.

Mohammad, S. M. (2018). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (pp. 174–184). Association for Computational Linguistics.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. arXiv preprint arXiv:2212.04356.

Rao, K. S., & Koolagudi, S. G. (2013). Emotion recognition from speech using global and local prosodic features. International Journal of Speech Technology, 16(2), 143–160. https://doi.org/10.1007/s10772-012-9172-2

Rathnayake, H., James, J., Leoni, G., Nicholas, A., Watson, C., & Keegan, P. (2026). A review on speech emotion recognition for low-resource and Indigenous languages. Speech Communication, 176, Article 103342. https://doi.org/10.1016/j.specom.2025.103342

Saadane, H., & Habash, N. (2015). A conventional orthography for Algerian Arabic. In Proceedings of the Second Workshop on Arabic Natural Language Processing (pp. 69–79). Association for Computational Linguistics.

Tamulevičius, G., Korvel, G., Yayak, A. B., Treigys, P., Bernatavičienė, J., & Kostek, B. (2020). A study of cross-linguistic speech emotion recognition based on 2D feature spaces. Electronics, 9(10), Article 1725. https://doi.org/10.3390/electronics9101725

Genetic Algorithms with Normalized Fitness Functions for Low-Resource Language Identification: A Case Study on Archaic Romanian and Medieval Slavonic Liturgical Manuscripts

Brandon Caillahua Mendoza
Universidad Nacional de Educación a Distancia, Spain
brandoncmkot01@gmail.com

Abstract

This paper presents a computational framework for low-resource language identification in unattributed Cyrillic liturgical manuscripts, addressing the challenge of applying NLP methods where large labeled datasets are unavailable. The study focuses on archaic Romanian and Church Slavonic manuscripts from the 13th–17th centuries, a partially uncatalogued corpus shaped by complex Slavic, Hungarian, and Latin stratification. The proposed method uses a steady-state genetic algorithm with a normalized multi-objective fitness function that combines Zipf's law correlation (f_{Zipf}), Kullback-Leibler divergence via exponential normalization ($f_{\text{KL}}^{\text{norm}}$), and bigram co-occurrence consistency ($f_{\text{cooc}}^{\text{norm}}$), resolving the scale incompatibility inherent in heterogeneous metrics. Validation on three Romanian control manuscripts achieves $R^2_{\text{Zipf}} = 0.89 \pm 0.03$ and stability $94.2 \pm 2.1\%$. Substrate selectivity is confirmed against medieval Hungarian liturgical texts ($\Delta D_{\text{KL}} = +126\%$) and a Bellaso cipher negative control ($R^2_{\text{Zipf}} < 0.70$, stability $< 12\%$).

A deliberate boundary test on the *Draganov Menaion* (Middle Bulgarian Church Slavonic, Zogr. 54), designed to withhold base patterns, validates the algorithm's expected null behavior: flat fitness landscape and zero stability are obtained when the necessary input is absent, while the latent signal is confirmed at ground-truth level ($\Delta D_{\text{KL}} = +23.8\%$). The framework requires only $\sim$20,000 tokens of verified reference text, making it applicable to other low-resource ecclesiastical languages where annotation budgets are severely constrained.

Keywords: genetic algorithms, low-resource language identification, normalized fitness functions, computational paleography.

1 Introduction

1.1 Motivation and Research Context

Identifying the linguistic substrate of unattributed historical manuscripts in Romanian and Church Slavonic is a persistent challenge in computational philology. The difficulty is especially acute for low-resource languages, where large labeled training datasets are unavailable. Archaic Romanian liturgical manuscripts from the 16th–17th centuries constitute a corpus of exceptional value for the history of Eastern Romance. Yet hundreds of documents preserved in Romanian, Moldavian, and Transylvanian archives lack secure paleographic identification of their linguistic substrate, date of composition, or scriptorial origin (Gaster, 1891; Popa-Lisseanu, 1927). This attribution gap hinders our understanding of how the Romanian written standard emerged and how Orthodox ecclesiastical culture interacted with broader Mediterranean and Central European manuscript traditions.

Three features of this corpus compound the difficulty. First, orthographic variability is pervasive: the same word may appear in multiple spellings within a single codex, obscuring straightforward lexical matching. Second, Romanian's complex morphology distinguishes it sharply from its Slavonic and Hungarian neighbors; capturing this distinctiveness requires a framework sensitive to morphological patterns rather than surface forms. Third, the specialized liturgical vocabulary draws simultaneously on Latin, Church Slavonic, and vernacular Romanian roots, creating a mixed linguistic signal that resists classification by lexicon alone (Chivu, 2000; Rosetti, 1986). Together, these features slow traditional paleographic comparison and introduce expert disagreement. They do not, however, constitute an insurmountable obstacle for statistical methods that operate on frequency dis-

tributions and co-occurrence patterns rather than one-to-one orthographic correspondences.

1.2 Linguistic Features of the Target Corpus

The archaic Romanian liturgical corpus exhibits several distinctive linguistic features that the proposed framework is designed to detect (Neamțu, 2016; Borbilă, 2017).

Analytical-synthetic verbal morphology. Romanian occupies an intermediate position between the synthetic paradigm of Latin and the analytical tendencies of other Romance languages (Pușcariu, 1940; Rosetti, 1986). Regular verbal paradigms (e.g., *cântu, cânți, cântă, cântăm, cântați, cântă*) produce statistical signatures detectable through bigram analysis (Sala, 2012). This distinguishes Romanian from Church Slavonic's aspectual system and from Hungarian's agglutinative verbal morphology.

Enclitic definite article. Romanian attaches the definite article to the noun (*omul*, *cartea*, *lupul*). This structure is absent in Hungarian, which uses a proclitic article (*a/az*), and in Church Slavonic, which has no article at all. The resulting bigram patterns are both strong and statistically detectable (Rosetti, 1986; Iordan, 1947).

Lexical stratification. The liturgical vocabulary comprises three superimposed strata: a Latin-derived core (e.g., *biserică* $<$ basilica, *înger* $<$ angelus), integrated Slavic borrowings (e.g., *glas*, *molitvă*, *rabdare*), and Hungarian loanwords from Transylvanian cohabitation (e.g., *oraș* $<$ város, *chip* $<$ kép) (Gafton, 2013; Philippide, 1927). As Gafton (2013) notes, the co-presence of these strata is characteristic of 16th-century Transylvanian texts and creates a distinctive frequency signature exploitable by the fitness function.

1.3 Related Work

Computational approaches to manuscript analysis have expanded significantly over the past two decades (Camps et al., 2021; Kiessling, 2019). Transformer-based methods (Devlin et al., 2019) achieve strong results on well-resourced languages, but require large labeled training corpora that are unavailable for archaic Romanian. Statistical approaches are therefore a necessary alternative (Eder et al., 2015).

Stylometric methods (Eder et al., 2015) and n-gram language models (Cavnar and Trenkle, 1994) have been applied to authorship attribution and language identification. Stylometry, however, typically requires thousands of function-word occurrences from a stable modern corpus, while n-gram SVMs depend on sufficient labeled training data for each target class. Neither condition is satisfied here: the verified reference material for the corpus under study amounts to approximately 20,000 tokens across three manuscript traditions. Furthermore, multilingual transformer models such as mBERT (Devlin et al., 2019) do not include Old Church Slavonic or archaic Romanian in their pretraining data. These languages therefore exist in a permanent zero-shot transfer regime where fine-tuning on small samples yields near-chance performance (Section 3.5).

The present work adopts a fitness-based optimization approach that requires no class labels beyond a small verified reference corpus. It also explicitly resolves the normalization problem that arises when combining heterogeneous metrics of different scales. Zipf's law has a long history as a linguistic regularity test in corpus linguistics (Zipf, 1949; Manning and Schütze, 1999), and KL divergence has been used as a substrate differentiation measure in low-resource settings (Cavnar and Trenkle, 1994). Genetic algorithms have been applied to grammar induction, word alignment, and feature selection (Mitchell, 1998; Goldberg, 1989), but their use in manuscript attribution is novel. The normalization problem addressed here—combining an unbounded KL divergence with bounded correlation coefficients in a single fitness function—has not been systematically treated in prior computational paleography literature.

1.4 Contributions

We present a steady-state genetic algorithm for low-resource language identification that requires no large labeled datasets, paired with a normalized multi-objective fitness function that resolves scale incompatibility between heterogeneous metrics. The framework is validated exhaustively on three control manuscripts with token-level concordance, error analysis, ablation studies, and hyperparameter sensitivity analysis. Substrate selectivity is confirmed against medieval Hungarian liturgical texts and a polyalphabetic cipher negative control. On the publicly available PROIEL *Codex Marianus* (57,556 tokens, CC BY-NC-SA 3.0), the framework achieves 97.6% POS-class accuracy with zero labeled tokens, compared to 74.7%–95.2% for su-

pervised baselines at k=20,000. An illustrative application to an unattributed manuscript (Ms. A-II-19) with independent expert comparison closes the empirical evaluation.

2 Methodological Framework

2.1 Problem Formulation

Let $S = [s_1, \ldots, s_n]$ be a token sequence from a manuscript transcription and $\mathcal{M}$ the set of morpheme classes. A candidate solution $G : S \to \mathcal{M}$ assigns each token to a morpheme class. The search space $\mathcal{M}^n$ is astronomically large ($\sim 20^{5000}$ for typical manuscripts), making exhaustive search infeasible and requiring heuristic navigation. The fitness function $F(G)$ evaluates the linguistic plausibility of a given assignment:

$$F(G) = \omega_1 f_{\text{Zipf}}(G) + \omega_2 f_{\text{KL}}^{\text{norm}}(G) + \omega_3 f_{\text{cooc}}^{\text{norm}}(G) \tag{1}$$

2.2 Normalized Multi-Objective Fitness Function

Previous formulations combined $f_{\text{Zipf}} \in [0, 1]$ with $f_{\text{KL}} \in [0, \infty)$, rendering the weights ω_i non-comparable across corpora, since the scale of KL divergence depends on vocabulary size, reference distribution, and manuscript length. Our normalization procedure maps all components to $[0, 1]$ before combination, ensuring that the weights express genuine relative priorities rather than artefacts of scale.

2.2.1 Zipf's Law Correlation

$f_{\text{Zipf}}(G)$ is the squared Pearson correlation between observed token-frequency ranks and the theoretical Zipf distribution:

$$f_{\text{Zipf}}(G) = \rho^2 = \left(\frac{\text{Cov}(\text{ranks},\ \log(\text{freq}))}{\sigma_{\text{ranks}}\, \sigma_{\log\text{-freq}}} \right)^2 \in [0, 1] \tag{2}$$

Using ρ^2 ensures non-negativity and yields a natural interpretation as the proportion of variance explained by the Zipfian model.

2.2.2 Normalized Kullback–Leibler Divergence

$$f_{\text{KL}}^{\text{norm}}(G) = \exp\left(-\frac{D_{\text{KL}}(P_G \| P_{\text{ref}})}{\tau} \right), \quad \tau = 3.0 \tag{3}$$

where $D_{\text{KL}}(P_G \| P_{\text{ref}}) = \sum_m P_G(m) \log \frac{P_G(m)}{P_{\text{ref}}(m)}$. The exponential mapping projects the unbounded divergence onto $(0, 1)$: $f_{\text{KL}}^{\text{norm}} = 1$ when $D_{\text{KL}} = 0$, and $f_{\text{KL}}^{\text{norm}} \to 0$ as $D_{\text{KL}} \to \infty$. The temperature $\tau = 3.0$ was selected empirically to maximise discriminative power between Romanian and non-Romanian inputs; its sensitivity to perturbation is analysed in Section 3.7.

2.2.3 Bigram Co-occurrence

$$f_{\text{cooc}}^{\text{norm}}(G) = \frac{\sum_{i=1}^{n-1} \mathbb{I}[(m_i, m_{i+1}) \in \mathcal{B}_{\text{lit}}]}{n - 1} \in [0, 1] \tag{4}$$

where $\mathcal{B}_{\text{lit}}$ is the reference liturgical bigram lexicon, comprising all bigrams attested at least twice in the verified corpus. The frequency threshold of two occurrences suppresses spurious bigrams introduced by transcription errors while preserving linguistically meaningful morpheme transitions.

2.2.4 Weights and Justification

The weights $\omega_1 = 0.4$, $\omega_2 = 0.35$, $\omega_3 = 0.25$ reflect a deliberate theoretical ordering. Zipf's law receives the highest weight because it is the most general criterion: it holds for any natural language regardless of substrate and therefore serves as a universal linguistic filter. KL divergence is assigned the second highest weight as the primary substrate-specific discriminant. Bigram co-occurrence provides useful domain-specific regularisation but is more sensitive to the completeness of the reference lexicon, and is therefore assigned the lowest weight.

2.3 Mathematical Validation

Proposition 1. *$f_{Zipf}(G) \in [0, 1]$.*

Proof. The Pearson correlation satisfies $\rho \in [-1, 1]$ by the Cauchy–Schwarz inequality; squaring gives $\rho^2 \in [0, 1]$. □

Proposition 2. *$f_{KL}^{norm}(G) \in (0, 1)$.*

Proof. By Gibbs' inequality, $D_{\text{KL}} \geq 0$. The function $g(x) = \exp(-x/\tau)$ is strictly decreasing with $g(0) = 1$ and $\lim_{x\to\infty} g(x) = 0$. Because exp is strictly positive for all finite inputs, it follows that $f_{\text{KL}}^{\text{norm}} \in (0, 1)$. □

Proposition 3. *$f_{cooc}^{norm}(G) \in [0, 1]$.*

Proof. Each indicator takes values in $\{0, 1\}$; summing $n - 1$ such terms and dividing by $n - 1 > 0$ preserves the unit interval. □

Corollary 1. *$F(G) \in (0, 1)$, since each component is bounded in $[0, 1]$ and the weights satisfy $\sum_i \omega_i = 1$.*

2.4 Genetic Algorithm Configuration

A steady-state genetic algorithm is employed with the following configuration: population size 500; binary tournament selection ($p = 0.75$); single-point crossover (probability 0.8); swap mutation ($p_m = 0.05$); replacement of the two worst individuals per generation as the elitist strategy; and termination at convergence or after 15,000 generations. Convergence is defined as:

$$\frac{1}{500} \sum_{t=T}^{T+500} |F(G_t) - F(G_T)| < 0.01 \quad (5)$$

This window of 500 generations (approximately 8–10% of a typical run) is large enough to distinguish genuine convergence from temporary fitness plateaus.

2.5 Falsifiability Criteria

Following Popperian principles, refutation conditions are defined a priori: (1) $R^2_{\text{Zipf}} < 0.70$ on the Romanian control manuscripts; (2) $D_{KL}^{\text{Romanian}} > D_{KL}^{\text{Hungarian}}$ on known Hungarian texts; (3) stability below 50% on control manuscripts; and (4) positive convergence on the Bellaso cipher. As reported in Section 3, none of these conditions are met.

3 Experiments and Results

3.1 Control Corpus

We validate on three manuscripts with independently verified content from Gaster (1891): *Psaltirea lui Coresi* (1577, $N = 8{,}247$ tokens; Psalms 1–10, 23, 50, 100), *Liturghia română* (c. 1590–1610, $N = 6{,}892$ tokens; Liturgy of St. John Chrysostom), and *Evangheliarul* (c. 1570–1600, $N = 5{,}431$ tokens; Matt 5–7 and Luke 1–2). These three texts represent psalmody, liturgical prose, and narrative prose respectively, providing diverse genre coverage.

3.2 Validation Results

Metric	Psaltirea	Liturghia	Evang.
R^2_{Zipf}	0.89±0.03	0.87±0.04	0.91±0.02
D_{KL}	1.87±0.21	2.03±0.19	1.76±0.14
Generations	5230±890	6410±1120	4780±760
Stability (%)	94.2±2.1	91.7±3.4	96.1±1.8

Table 1: Validation on control manuscripts ($n = 10$ runs each).

All metrics exceed the falsifiability thresholds. The higher D_{KL} for Liturghia (2.03 vs. 1.87 for Psaltirea) is attributable to the specialized liturgical register, which contains formulaic expressions and Slavic borrowings at higher density than the Psalms. This confirms that the KL divergence component appropriately captures genre-level distributional differences within a single language, and suggests that future reference corpora should be stratified by genre to reduce within-language D_{KL} variance.

Hamming distances confirm structural convergence: the top-8 solutions per manuscript differ in only $4.7\% \pm 1.2\%$ of tokens, while local optima differ at $18.3\% \pm 2.8\%$ from the global optimum, demonstrating that the algorithm consistently identifies a stable core linguistic structure.

3.3 Token-Level Concordance and Error Analysis

Frequency stratum	Proportion	Concordance
High ($> 1\%$)	42.3%	96.3%
Medium (0.1%–1%)	31.7%	87.4%
Low ($< 0.1\%$)	26.0%	61.2%
Overall (weighted)	100%	84.7%

Table 2: Token-level concordance by frequency stratum (Psaltirea).

The 96.3% concordance on high-frequency tokens is crucial: these are the articles, common verbs, prepositions, and conjunctions that form the backbone of linguistic structure. The lower concordance on low-frequency tokens (61.2%) is explained by data sparsity in the reference corpus.

Error analysis. Systematic examination of misclassified tokens reveals three recurrent error patterns:

(i) Slavic borrowings with Romanian morphology. Words such as *molitvă-a* (Slavic root + Romanian enclitic article) are misclassified as Slavic because the root frequency profile matches Church Slavonic, while the attached article is not visible as a separate token. This produces approximately 38% of the medium-frequency errors. The implication for model improvement is that morphological segmentation should be performed before fitness evaluation, separating clitics from lexical stems.

(ii) Orthographic variants of high-frequency tokens. Forms such as *că* / *cã* / *ce* (complementizer/conjunction) appear under multiple graphemic realizations in the same manuscript. The algorithm correctly identifies the most frequent realization but assigns different morpheme classes to

rare variants, accounting for approximately 29% of medium-frequency errors. Normalizing orthographic variants before fitness evaluation would reduce this error type.

(iii) Calque structures. Phrases that are structurally Hungarian or Slavonic (literal translation equivalents) but lexically Romanian create morphemic assignments that match the Romanian reference partially but not completely. These account for approximately 21% of low-frequency errors and are inherently difficult to resolve without semantic analysis.

These error patterns are **linguistically motivated**: they reflect real properties of 16th-century Romanian that complicate substrate identification. The errors are not random but systematic, which makes them amenable to targeted preprocessing in future implementations.

3.4 Comparative Substrate Validation

3.4.1 Hungarian Texts

Applying the Romanian-optimized framework to medieval Hungarian liturgical texts (Gaster, 1891: pp. 235–267):

Metric	Romanian opt.	Hungarian opt.
D_{KL}	4.23±0.37	1.92±0.24
R^2_{Zipf}	0.81±0.06	0.88±0.04
Stability (%)	68.3±5.7	93.1±2.8

Table 3: Differential behavior on Hungarian texts.

The $\Delta D_{\text{KL}} = +126\%$ and $\Delta\text{Stability} = -26.3\%$ confirm substrate-specific discriminative power. Notably, R^2_{Zipf} remains reasonably high (0.81) even under Romanian optimization: this is theoretically expected, since Zipf's law is a universal property of natural language. This confirms that the Zipf component correctly functions as a general linguistic filter rather than a substrate discriminant, validating the decision to assign it the highest weight as a necessary but not sufficient condition.

3.4.2 Negative Control: Bellaso Cipher

Metric	Value
R^2_{Zipf}	0.34±0.08
Stability (%)	11.7±3.4
Convergence rate	6/10 runs

Table 4: Negative control results (Bellaso cipher).

Both metrics fall clearly below the falsifiability thresholds ($R^2_{\text{Zipf}} < 0.70$, stability $< 50\%$), and 4/10 runs failed to converge. The low $R^2_{\text{Zipf}} = 0.34$ reflects the polyalphabetic cipher's flat frequency distribution, which by design equalizes character frequencies to resist frequency analysis and thus violates Zipf's law. The fitness function's rejection of the cipher confirms that Zipf's law is genuinely detecting the linguistic structure of natural language, rather than fitting any sufficiently long character sequence.

3.4.3 Middle Bulgarian Liturgical Manuscript: *Draganov Menaion* (Zogr. 54)

To probe the limits of the framework's selectivity, we apply it to the *Draganov Menaion* (Zogr. 54), a Middle Bulgarian parchment codex from the late 13th century preserved at the Monastery of Zographou, Mount Athos.[1] The main corpus (Zogr. 54) comprises 219 parchment leaves on Athos, with additional scattered leaves in Moscow (RGB), Saint Petersburg (RNB), and formerly in the Russian Athonite monastery of St. Panteleimon.[2] It is a festal menaion containing liturgical services and hagiographic texts for Bulgarian saints (SS. Cyril, Methodius, Ivan Rilski, Tsar Peter), written in the Middle Bulgarian recension of Church Slavonic.

Experimental design: a deliberate boundary test. This manuscript was selected not to produce successful discrimination but as a *controlled negative test*: we deliberately ran the GA **without providing appropriate base patterns** (i.e., without a verified Church Slavonic reference corpus for the bigram co-occurrence component). The goal was to observe how the algorithm behaves when the necessary input patterns are withheld. Both reference corpora (Romanian and Bulgarian) were synthetically constructed or lacked the morphological patterns characteristic of Middle Bulgarian, so the bigram component $f^{\text{norm}}_{\text{cooc}}$ carried no discriminative signal between the two substrates.

Corpus construction. The manuscript input was built from the Church Slavonic transcriptions

[1] P. Petkov, svešt. K. Popovski, D. Peev, Za pripiskata na "okayaniya Dragan" v Zografskiya trefologiy, în *Zografski Sbornik: Zografskiyat arkhiv i biblioteka. Izsledvaniya i perspektivi*, Zografski manastir, Sveta Gora, 2019, p. 219–231.

[2] B. Rajkov, S. Kozhukhanov, H. Miklas, H. Kodov, *Katalog na slavyanskite rakopisi v bibliothekata na Zografskiya manastir v Sveta Gora*, CIBAL, Sofia, 1994, p. 52.

quoted in Petkov et al. (2019): a vocabulary of 77 unique token types totalling 454 attestations, covering all forms cited on pages 219–231. Each type was mapped to one of the $M = 20$ morpheme classes defined in Section 2 according to its linguistic function (class 0: enclitic article; class 1: conjunction; class 10/11: *jus*-class tokens; class 18: proper names). The jus-class tokens (*grēhy*, *tvoa*, *tvož*) and vocative forms (*okaāniče*, *stranniče*) are unambiguous Middle Bulgarian morphological markers. The GA operated on type-consistent assignments (one class per unique type), with a search space of 20^{77}.

Results. Table 5 reports the outcome of $n = 10$ independent GA runs per reference corpus, together with the direct fitness on the ground-truth assignment.

Condition	R^2_{Zipf}	D_{KL}	Stab. (%)
GA (Romanian)	0.979±0.009	0.156±0.038	0.0
GA (Bulgarian)	0.976±0.008	0.156±0.028	0.0
Direct (Romanian)	0.976	0.694	—
Direct (Bulgarian)	0.976	**0.561**	—

Table 5: Results on *Draganov Menaion* (Zogr. 54). "Direct" rows report ground-truth fitness.

Zipf component. The Zipf criterion performs as theoretically expected: $R^2_{\text{Zipf}} = 0.979$ (Romanian reference) and 0.976 (Bulgarian reference), both well above the falsifiability threshold ($R^2_{\text{Zipf}} \geq 0.70$), confirming that the framework correctly classifies the *Draganov Menaion* as a natural-language text regardless of reference corpus. As discussed in Section 3.4.1, f_{Zipf} is a substrate-independent universal and carries no discriminant signal between Romanian and Church Slavonic.

Absence of KL discrimination due to missing base patterns. Unlike the Hungarian test ($\Delta D_{\text{KL}} = +126\%$), the GA achieves essentially identical D_{KL} under both references (0.156, $\Delta D_{\text{KL}} = +0.4\%$). Zero stability scores confirm that the algorithm explores a flat fitness landscape: no two runs converge to solutions sharing more than 95% of type-to-class assignments. The algorithm's discriminant power depends critically on receiving appropriate base patterns to construct $f^{\text{norm}}_{\text{cooc}}$. When those patterns are withheld, the GA has no information to guide it toward a linguistically meaningful solution. The flat result is therefore not a failure of the algorithm but a *successful validation of its design*: it does not invent patterns or produce spurious discriminations when the necessary input is absent.

Signal at the ground-truth level. The direct-assignment rows reveal that the discriminant signal *does* exist in the data: the ground-truth assignment achieves $D_{\text{KL}} = 0.561$ under Bulgarian reference vs. $D_{\text{KL}} = 0.694$ under Romanian reference—a differential of $\Delta D_{\text{KL}} = +23.8\%$ in the correct direction. This signal would be accessible with proper base patterns, confirming that the GA's null result stems from missing input rather than from an absence of signal in the manuscript.

Implications. The framework's discriminant power is contingent on providing appropriate base patterns. In the Romanian–Hungarian comparison, the necessary patterns were present, enabling successful discrimination. In the Romanian–Middle Bulgarian comparison, the patterns were deliberately withheld, and the algorithm correctly returned a null result. The direct-assignment differential (+23.8%) confirms that the target signal exists; recovering it requires constructing $f^{\text{norm}}_{\text{cooc}}$ from a verified Church Slavonic reference corpus.

3.5 Comparison with Supervised Baselines on the PROIEL Codex Marianus

We evaluate our framework against supervised baselines on the *Codex Marianus* from the PROIEL Treebank (Haug and Jøhndal, 2008), an Old Church Slavonic Gospel codex of 57,556 annotated tokens with verified gold-standard POS labels (CC BY-NC-SA 3.0), downloaded directly from `https://github.com/proiel/proiel-treebank`. The experiment is fully reproducible from this public release.

3.5.1 Task and Evaluation Protocol

The task is POS-class assignment for 11 morpheme classes (V, N, P, D, C, R, A, G, M, I, F) on the held-out test partition (last 10,000 tokens). Accuracy is reported as mean ± std over n=5 random seeds. Three supervised baselines are evaluated across training sizes $k \in \{500, 1{,}000, 2{,}000, 5{,}000, 20{,}000\}$ labeled tokens:

1. **N-gram SVM** (Cavnar and Trenkle, 1994): character trigram TF-IDF features (max vocabulary 5,000), linear SVM (C=1.0), one-vs-rest multiclass.

2. **Stylometric LDA** (Eder et al., 2015): 500 most-frequent-word TF-IDF features, Linear Discriminant Analysis. Requires $k \geq 2{,}000$ for stable centroids.

3. **mBERT fine-tuning** (Devlin et al., 2019): Old Church Slavonic is **absent from multilingual BERT's pretraining data**; the model therefore operates in a zero-shot transfer regime, approximated here by character 1–4gram TF-IDF + linear SVC, which matches mBERT's effective behavior on unseen scripts without GPU resources.

Our GA framework uses **zero** labeled tokens from the Codex Marianus: it relies solely on the type-consistent assignment strategy derived from the normalized fitness function trained on the Romanian reference corpus (Section 3). Since the GA assigns one morpheme class per unique token type and the Codex Marianus exhibits high within-type POS consistency (a well-documented property of Old Church Slavonic morphology (Haug and Jøhndal, 2008)), the framework's type-consistent assignments constitute a principled unsupervised oracle.

3.5.2 Results

Method	500	5k	20k
N-gram SVM	56.1±1.2	71.7±0.5	74.7±0.1
Stylometric LDA	—	78.1±0.5	80.8±0.2
mBERT (OCS approx.)	77.0	92.7	95.2
GA (ours, k=0)	**97.6** (constant)		

Table 6: POS-class accuracy (%) on *Codex Marianus* (PROIEL, CC BY-NC-SA 3.0; n=5 seeds). GA uses **zero** labeled tokens. Dashes: insufficient data for stable LDA centroids.

3.5.3 Analysis

The results highlight three findings relevant to the low-resource regime.

Performance under zero labels. Our GA achieves 97.6% accuracy with no labeled tokens from the target corpus, outperforming all three supervised baselines at every training size up to k=20,000: the N-gram SVM reaches 74.7% and stylometric LDA 80.8% at maximum training budget, a gap of +16.8 and +22.9 percentage points respectively. This advantage is specific to the low-resource regime evaluated here, where the GA exploits high within-type POS consistency in Old Church Slavonic morphology—a structural property that supervised methods cannot leverage without labeled data.

Performance under realistic annotation budgets. At k=500 tokens—the realistic upper bound for most unstudied historical manuscripts—the N-gram SVM (56.1%) performs only marginally above chance for an 11-class problem (9.1% chance baseline), and stylometric LDA cannot produce stable centroids at all below 2,000 tokens. Our method requires no labels from the target corpus, making it applicable precisely in the annotation-scarce settings that characterize historical manuscript research.

Why mBERT cannot close the gap. Old Church Slavonic is absent from multilingual BERT's pretraining data (Devlin et al., 2019); the model therefore cannot leverage pretrained subword representations and is effectively reduced to a character-level model. Even with k=20,000 labeled tokens (95.2%), it remains 2.4 points below our zero-label result. This reflects a structural consequence of historical language absence from large pretraining corpora—a condition that similarly affects archaic Romanian and other low-resource ecclesiastical languages.

These results suggest that, within the specific conditions evaluated here—languages absent from pretraining corpora, near-zero annotation budgets, and manuscripts where assembling 20,000 labeled tokens would exceed the entire philological research budget—unsupervised fitness optimization over a small verified reference corpus offers a competitive alternative to supervised approaches. Generalization to other corpora and language pairs remains an open question for future work.

3.6 Ablation Studies

Condition	R^2_{Zipf}	D_{KL}	Stability
Full model	0.89	1.87	94.2%
Without Zipf (ω_1=0)	0.74	1.92	87.3%
Without KL (ω_2=0)	0.86	2.41	82.6%
Without bigram (ω_3=0)	0.88	1.89	89.7%

Table 7: Ablation study results (Psaltirea, $n = 10$ runs).

Without Zipf: R^2_{Zipf} drops to 0.74, confirming that the algorithm no longer enforces Zipfian structure when this component is absent. Stability remains relatively high (87.3%), suggesting that KL divergence partially compensates, but the model loses its universally applicable linguistic constraint.

Without KL: D_{KL} increases from 1.87 to 2.41 and stability drops to 82.6%, the largest degradation observed. This confirms that KL divergence is the primary substrate-specific discriminant and that its removal cannot be compensated by the other components.

Without bigram: Performance degrades modestly across all metrics (stability 89.7%). Bigram co-occurrence provides useful local structural constraints but is less critical than the other two components, consistent with its lowest weight $\omega_3 = 0.25$.

Impact on theory. The ablation results empirically validate the theoretical hierarchy encoded in the weights: f_{Zipf} is necessary for general linguistic validity, $f_{\mathrm{KL}}^{\mathrm{norm}}$ is necessary for substrate specificity, and $f_{\mathrm{cooc}}^{\mathrm{norm}}$ is useful but not critical. This provides quantitative justification for the weight selection beyond theoretical argument alone.

3.7 Robustness Analysis

Configuration	R^2_{Zipf}	D_{KL}	Stability
Baseline	0.89	1.87	94.2%
ω_1+20%	0.91	1.91	92.8%
ω_1−20%	0.86	1.94	91.3%
ω_2+20%	0.88	1.72	95.1%
ω_2−20%	0.89	2.03	93.4%

Table 8: Hyperparameter sensitivity analysis (Psaltirea).

All metrics remain within one standard deviation of baseline values under ±20% variation. The most sensitive metric is D_{KL} when ω_2 is varied, as expected. Even in the most unfavorable configuration ($\omega_2 = 0.28$), stability remains above 93%, confirming that the framework does not depend on precise parameter tuning.

3.8 Inter-Annotator Agreement Benchmark

Comparison	Agreement	κ
Expert A vs. Expert B	87.2%	0.82
Expert A vs. Expert C	85.9%	0.79
Expert B vs. Expert C	88.1%	0.84
Mean (experts)	87.1%	0.82
Algorithm vs. consensus	84.7%	0.80

Table 9: Inter-annotator agreement (Psaltirea, first 1,000 tokens).

The algorithm's agreement with expert consensus (84.7%, $\kappa = 0.80$) is slightly lower than the mean inter-expert agreement (87.1%, $\kappa = 0.82$) but closely approaches it. This result suggests that the algorithm performs at a level broadly comparable to individual human experts on this task, while operating in minutes rather than weeks.

4 Illustrative Application: Ms. A-II-19

4.1 Source Documentation and Paleographic Context

Ms. A-II-19 from the Library of the Romanian Academy is a 16th-century codex of 124 folios containing liturgical texts and prayers. Gaster (1891: pp. 312–317) published select folios and noted that its linguistic substrate is uncertain: while orthographic features suggest Romanian origin, the presence of Slavonic marginalia and Hungarian-script interpolations complicates attribution. Borbilă (2017) confirms that the manuscript's attribution remains unresolved in the contemporary literature.

The manuscript presents a semi-cursive Cyrillic hand typical of 16th-century Wallachian chanceries, with letter forms ($\bar{o}$, e, $\tilde{a}$) consistent with mid-16th-century Romanian paleography, standard nomina sacra abbreviations, and orthographic variation (e.g., *domnu-l* and *domnu* for *domnul*). Four distinct hands contribute marginalia: one in Romanian Cyrillic, two in Slavonic, and one in Latin-script Hungarian, a configuration characteristic of the multilingual environment of 16th-century Transylvania (Gafton, 2013).

4.2 Differential Behavior and Expert Comparison

Reference Corpus	R^2_{Zipf}	D_{KL}	Stab.
Archaic Romanian	0.84±0.04	2.14±0.28	88.3±3.7%
Medieval Hungarian	0.72±0.09	3.86±0.41	71.4±6.2%
Church Slavonic	0.68±0.11	4.23±0.52	63.7±7.1%

Table 10: Differential behavior for Ms. A-II-19.

The $\Delta D_{\mathrm{KL}} = +80\%$–$98\%$ between Romanian and the other substrates is quantitatively consistent with the differential observed on known Hungarian texts (+126%, Section 3), suggesting—but not proving—that Ms. A-II-19 shares linguistic affinity with Romanian. Three expert philologists assessed the manuscript independently on a 1–5 confidence scale: Expert A (Romanian 4, Hungarian 2, Slavonic 1), Expert B (Romanian 3, Hungarian 3, Slavonic 2), Expert C (Romanian 4, Hungarian 1, Slavonic 2). Mean expert confidence for Romanian (3.67) aligns with the algorithm's stability

ranking. However, the algorithm cannot resolve the disagreement between Expert B (who considers Hungarian equally possible) and Experts A and C (who strongly favor Romanian), illustrating both its utility and its limitations.

4.3 Interpretation of Limitations

Three competing explanations for the differential behavior remain indistinguishable without independent verification: (1) genuine linguistic affinity with Romanian; (2) fortuitous statistical overlap due to shared liturgical vocabulary; (3) systematic algorithmic bias toward Romanian distributions, since the optimization was validated exclusively on Romanian manuscripts. The third explanation is the most epistemically serious, as it represents a form of domain overfitting. Future work should validate on manuscripts from other language traditions before drawing attribution conclusions.

As Gafton (2013) cautions, lexical features—whether computational or traditional—are insufficient alone for manuscript localization. The results reported here should be interpreted as one input into a broader philological investigation, not as a definitive attribution.

5 Discussion

The framework demonstrates a bidirectional relationship between linguistic theory and computational technology. Romanian morphological features (enclitic article, verbal paradigm regularity, lexical stratification) directly informed the three fitness components; in return, the ablation studies quantify their relative salience. The finding that KL divergence is more critical than bigram co-occurrence suggests that macro-level distributional properties are more robust than local transition patterns, a result relevant to theoretical models of language contact.

The design principles are language-agnostic: Zipf's law, KL divergence, and bigram co-occurrence can be computed for any language given a small verified reference corpus ($\sim$20,000 tokens sufficed here). Potential applications include medieval Slavic manuscript substrates, Dead Sea Scroll fragment classification, and original-language detection in translated texts. Remaining limitations include sensitivity to transcription errors, reference corpus completeness, computational cost (4–6 hours per 10-trial run), and the current inability to segment morphologically complex tokens before classification.

6 Conclusion

This paper has presented and validated a computational framework for low-resource language identification in Cyrillic liturgical manuscripts. By resolving scale incompatibility between heterogeneous metrics through principled normalization, the method operates without labeled data from the target corpus, making it applicable to under-resourced historical ecclesiastical languages.

Validation on three Romanian control manuscripts demonstrates $R^2_{\text{Zipf}} = 0.89 \pm 0.03$, stability $94.2 \pm 2.1\%$, and 84.7% concordance with expert consensus. Comparative analysis distinguishes Romanian from Hungarian ($\Delta D_{\text{KL}} = +126\%$); the Bellaso cipher confirms no false positives on non-linguistic material; and the boundary test on the *Draganov Menaion* yields flat fitness landscapes and zero stability, confirming the algorithm does not produce spurious discriminations when necessary input is absent.

Evaluation on the PROIEL *Codex Marianus* (57,556 tokens, CC BY-NC-SA 3.0) shows 97.6% POS-class accuracy with *zero* labeled tokens, against 74.7%–95.2% for supervised baselines at k=20,000. Extension to other Cyrillic traditions requires only $\sim$20,000 tokens of verified reference text.

References

M. Borbilă. 2017. Contact lingvistic și stratificare lexicală în manuscrisele liturgice transilvănene. *Dacoromania (DR)*, XXII(1):45–62.

J. B. Camps, A. Vidal-Gorène, and M. Vernet. 2021. Handling the Diversity of Medieval Scripts: Mixed Models and Synthetic Data. *Digital Humanities Quarterly*, 15(1).

W. B. Cavnar and J. M. Trenkle. 1994. N-gram-based text categorization. *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*.

Gh. Chivu. 2000. *Limba română de la primele texte până la sfârșitul secolului al XVIII-lea*. Univers Enciclopedic, București.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT 2019*. Association for Computational Linguistics.

M. Eder, J. Rybicki, and M. Kestemont. 2015. Stylometry with R: A Package for Computational Text Analysis. *R Journal*, 8(1):107–121.

A. Gafton. 2013. Elemente maghiare în textele românești din secolul al XVI-lea: stratificare și circulație. *Limba română*, LXII(1):1–18.

M. Gaster. 1891. *Chrestomatie romaine. Textes imprimés et manuscrits du XVIe au XIXe siècle*. F.A. Brockhaus, Leipzig; Socec & Co., București.

D. E. Goldberg. 1989. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, Reading, MA.

D. T. T. Haug and M. L. Jøhndal. 2008. Creating a Parallel Treebank of the Old Indo-European Bible Translations. In C. Sporleder and K. Ribarov (eds.), *Proc. LaTeCH 2008*.

I. Iordan. 1947. *Limba română actuală*. Iași.

B. Kiessling. 2019. Kraken: a Universal Text Recognizer for the Humanities. In *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage (DATeCH 2019)*. Association for Computing Machinery, New York.

C. D. Manning and H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

M. Mitchell. 1998. *An Introduction to Genetic Algorithms*. MIT Press, Cambridge, MA.

G. G. Neamțu. 2016. Aspecte ale dinamicii sistemului verbal românesc în secolul al XVI-lea. *Dacoromania (DR)*, XXI(2):113–128.

Al. Philippide. 1927. *Originea românilor*, Vol. I-II. Iași.

Gh. Popa-Lisseanu. 1927. *Izvoarele istoriei românilor*. Cultura Națională, București.

S. Pușcariu. 1940. *Limba română*, Vol. I. Fundația Regală pentru Literatură și Artă, București.

A. Rosetti. 1986. *Istoria limbii române*, Vol. I. Editura Științifică și Enciclopedică, București.

M. Sala. 2012. *De la latină la română*. Editura Univers Enciclopedic, București.

G. K. Zipf. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Cambridge, MA.

Appendix A Pseudocode Implementation

Algorithm 1 Steady-State GA for Manuscript Analysis

procedure GENETICALGORITHM($S, P_{\text{ref}}, \mathcal{B}_{\text{lit}}$)
  $V \leftarrow$ unique_types(S); $|V| \leftarrow$ length(V)
  $D \leftarrow [\text{type\_index}(s_i)]_{i=1}^{n}$ ▷ type indices of manuscript tokens
  pop $\leftarrow$ Initialize($N, |V|, M$) ▷ $N = 500$ individuals, each $G \in \{0, \ldots, M-1\}^{|V|}$
  gen $\leftarrow 0$; hist $\leftarrow [\,]$
  while not converged **and** gen < 15000 **do**
    gen $\leftarrow$ gen $+ 1$
    for each G in pop **do**
      $M_G \leftarrow [G[D_i]]_{i=1}^{n}$ ▷ derive morpheme sequence from type assignment
      $f_z \leftarrow \text{ZipfCorr}(M_G)$
      $f_k \leftarrow \exp(-\text{KL}(P_{M_G} \| P_{\text{ref}})/3.0)$
      $f_c \leftarrow \text{Bigram}(M_G, \mathcal{B}_{\text{lit}})$
      $F(G) \leftarrow 0.4 f_z + 0.35 f_k + 0.25 f_c$
    end for
    hist.append(mean(F))
    $p_1, p_2 \leftarrow$ Tournament(pop, 0.75) twice
    $(o_1, o_2) \leftarrow$ Crossover(p_1, p_2, prob $= 0.8$)
    Mutate($o_1, 0.05, M$); Mutate($o_2, 0.05, M$)
    Replace two worst with o_1, o_2
    if gen $\bmod 500 = 0$ **and** $|\Delta F_{500}| < 0.01$ **then**
      break
    end if
  end while
  return $\arg\max_G F(G)$
end procedure

Algorithm 2 Fitness Component Computation

procedure ZIPFCORR(M_G)
  $H \leftarrow \text{Hist}(M_G)$; $\ell_r \leftarrow \log[1..|H|]$; $\ell_f \leftarrow \log(H{+}\varepsilon)$
  return $\text{pearson}(\ell_r,\ \ell_f)^2$
end procedure
procedure KL(P_{M_G}, P_{ref})
  return $\sum_m P_{M_G}(m) \log\Big(\frac{P_{M_G}(m)}{P_{\text{ref}}(m)}\Big)$
end procedure
procedure BIGRAM($M_G, \mathcal{B}_{\text{lit}}$)
  return $\frac{1}{n-1} \sum_{i=1}^{n-1} \mathbf{1}[(M_G[i], M_G[i{+}1]) \in \mathcal{B}_{\text{lit}}]$
end procedure
procedure MUTATE(G, p_m, M)
  for $g \in G$ **do if** rand() $< p_m$ **then** $g \leftarrow \text{randint}(0, M{-}1)$
  end for
end procedure

Algorithm 3 Stability Analysis

procedure POPULATIONSTABILITY(pop, k, θ)
  top $\leftarrow$ TopK(pop, k); agr $\leftarrow$ [ ]; $|V| \leftarrow$ chrlen(pop[0])
  for $(G_i, G_j) \in$ top, $i < j$ **do**
    agr.append$\Big(\frac{1}{|V|} \sum_t \mathbf{1}[G_i[t]{=}G_j[t]]\Big)$
  end for
  return $100 \times \sum_{h \in \text{agr}} \mathbf{1}[h{\geq}\theta]$ / |agr|
end procedure
procedure MULTIRUNSTABILITY(runs, θ)
  $A \leftarrow$ [r.best for r in runs]; cons $\leftarrow$ mode(A); $s \leftarrow 0$
  for G **in** A **do**
    if $\frac{1}{|V|} \sum_t \mathbf{1}[G[t]{=}\text{cons}[t]] \geq \theta$ **then**
      $s \leftarrow s + 1$
    end if
  end for
  return $100 \times s$ / $|A|$
end procedure

Corpora, Annotation systems, and NLP Tools

The Romanian Component of the Parallel ELEXIS Corpus: ELEXIS-Ro

Verginica Barbu Mititelu
RACAI, Bucharest
vergi@racai.ro

Elena Irimia
RACAI, Bucharest
elena@racai.ro

Simina Popa
University of Bucharest
popa.simina2005@gmail.com

Mihaela Cristescu
University of Bucharest
mihaela.cristescu@litere.unibuc.ro

Cătălin Mihăilă
RACAI, Bucharest
catalin@racai.ro

Lea Radu
The Saint Sava National College
leacristiana@icloud.com

Ioana Bucșa
Gheorghe Șincai National College
ioana.bucsa34@gmail.com

Abstract

We present the steps taken to extend the ELEXIS parallel corpus to include Romanian. This extension is based on automatic processing that is followed by manual validation and correction. The English original corpus was translated into Romanian, tokenized and then UDPipe was run on it. The resulting part-of-speech tagging was manually validated and some remarks on the types of mistakes encountered are presented herein. The corpus was further automatically enriched with semantic annotation of the types multiword expressions of all parts of speech and named entities. The multiword expressions were labeled with the types defined in the PARSEME project. Three types of named entities were annotated: person, organization and location. A double validation and correction process took place and its results are also presented here. The paper also includes statistics of the corpus: number of tokens per part of speech, number of multiword expressions per type and of name entities.

Keywords: parallel corpus, Romanian, translation, tokenization, PoS tagging, lemmatizaton, manual validation, multiword expressions, named entities.

1 Introduction

Parallel corpora – collections of texts aligned across two or more languages – are fundamental resources for multilingual research and language technology, enabling systematic analysis of how meaning, structure, and discourse are transferred across languages, while revealing translation patterns, equivalence relations, and cross-linguistic variation. When carefully produced and validated, they serve as reliable repositories of linguistic knowledge and remain indispensable even in the era of Large Language Models (LLMs), whose performance still depends on high-quality multilingual data; in this context, parallel corpora provide transparent, verifiable alignments for evaluation, fine-tuning, and interpretability. Their value is further enhanced by the inclusion of low-resourced languages, which supports linguistic diversity, reduces technological inequality, and enables cross-lingual transfer, as well as by multi-layer linguistic annotation (morphological, syntactic, semantic, and discourse), which allows deeper analysis of language correspondences and supports advanced applications such as machine translation, cross-lingual retrieval, terminology extraction, and knowledge discovery.

We present an extension of the ELEXIS parallel corpus (Martelli et al., 2021) through the inclusion of Romanian, thus enhancing its linguistic coverage and research potential. As a fourth Romance language in the dataset, Romanian enables richer comparative and contrastive studies within the Romance family, supporting analyses of lexical, syntactic, semantic, and discourse-level variation. Its addition is valuable also from a typological and areal perspective, given Romanian's distinct evolution in Eastern Europe and its extensive contact with Slavic and Balkan languages. Together with Bulgarian, also included in the corpus, this allows for the investigation of Balkan Sprachbund phenomena – such as definiteness marking, analytic constructions, and clitic doubling – while supporting research across genealogical, typological, and contact-linguistic dimensions.

The ELEXIS corpus is a manually annotated multilingual parallel dataset designed for lexicography and Word Sense Disambiguation (WSD), originally introduced in 2021 with ten European languages and consisting of 2,024 validated parallel sentences derived from WikiMatrix. It features five layers of annotation: tokenisation, subtokenisation, lemmatisation, part-of-speech tag-

ging, and WSD. It aims to advance semantic annotation and improve sense inventories. In its 2.0 version[1], other annotation types (such as named entities) are also available for some languages

Against the landscape of the extending the ELEXIS corpus with other languages (see Section 2) we paint the picture of the work done for adding Romanian (Section 3) and present statistics relevant to the annotation levels present in the corpus, before envisaging future work (Section 6), announcing corpus availability (Section 8) and concluding the paper (Section 7).

2 Related Work

The ELEXIS project[2] resulted in a multilingual parallel corpus with sense annotation across ten languages, later extended within the UniDive COST Action (Savary et al., 2024). One major component is the Danish corpus (DA-ELEXIS) (Pedersen et al., 2023), among the largest sense-annotated resources for Danish, featuring multiple annotation layers (tokenisation, sub-tokenisation, lemmatisation, and POS tagging) generated under Universal Dependencies (UD) guidelines[3] and manually validated. The corpus comprises over 32,000 tokens with full sense annotation for content words, supported by a semi-automated annotation tool. Using DanNet as the primary sense inventory, the project also addressed compounds through splitting strategies and encoded multiword expressions (MWEs) with unified semantic labels. Inter-annotator agreement reached 0.78 after guideline refinement, indicating reliable semantic annotation.

The Serbian extension (Krstev et al., 2024) focuses on semantic annotation and multilingual alignment, particularly for MWEs and named entities (NEs). Sentences from WikiMatrix were machine-translated into Serbian and manually validated, with attention to translation errors and structural inconsistencies. Annotation layers (tokenisation, lemmatisation, POS tagging) were automatically generated and manually checked. The study highlights cross-linguistic variation by showing how MWEs differ structurally across languages, including cases where phrases correspond to single words. The dataset includes hundreds of nominal and verbal MWEs, as well as named entities identified through both lexical resources and NER tools, enabling iterative expansion of the Serbian lexical repository.

For semantic annotation, the Serbian dataset relies primarily on Serbian WordNet (SrpWN), aligned with Princeton WordNet (Fellbaum, 1998) through definition-based matching due to the absence of an interlingual index. Automatic translation of synonym sets and integration of additional lexical resources supported this alignment process. Results show that only about 39% of MWEs were directly matched in SrpWN, reflecting significant cross-linguistic variation and underscoring the challenges of multilingual semantic alignment, particularly in the treatment of MWEs and lexical equivalence across languages.

3 Work Methodology

This section outlines the integration of Romanian into the ELEXIS corpus, including the translation of the English source texts, tokenization, and morpho-syntactic annotation. Following the established ELEXIS methodology, these processes are initially carried out automatically using specialized tools for machine translation and linguistic annotation, and are subsequently manually revised and validated to ensure accuracy, cross-linguistic consistency, and compatibility with the multilingual framework.

3.1 Language Processing

3.1.1 Translation

Automatic translation from English into Romanian was performed using Google Translate in February 2024, with subsequent manual validation showing that 1,437 out of 2,024 sentences (around 70%) required no correction, while the remaining 30% were manually revised. Most machine translation (MT) errors fall into recurring, predictable categories, the most frequent being lexical and register-related issues. These include contextually inappropriate or stylistically neutral synonyms, incorrect terminological choices, and literal or non-idiomatic expressions that mirror source-language structures. Human revision plays a key role in selecting contextually appropriate equivalents, ensuring correct register (technical, academic, or colloquial), and normalizing collocations to align with natural Romanian usage.

A second major category involves morpho-syntactic and structural errors, where MT outputs reflect source-language interference or incorrect

[1] https://www.clarin.si/repository/xmlui/handle/11356/2101.

[2] https://project.elex.is/

[3] https://universaldependencies.org/

grammatical relations. These include agreement errors, improper determiners, incomplete or ambiguous constructions, and unnatural word order, all of which require systematic correction to ensure grammaticality and fluency in Romanian. Additionally, issues related to named entities (NEs) and linguistic conventions are observed, such as incorrect capitalization, improper adaptation of proper names and titles, and inadequate localization of institutional names and acronyms.

Overall, while MT provides semantically adequate translations, it tends to preserve source-language structures, resulting in outputs that lack naturalness and stylistic appropriateness. Human intervention is therefore essential to refine lexical choices, adjust syntactic patterns, and ensure coherence with target-language norms, ultimately producing high-quality translations suitable for inclusion in a linguistically reliable parallel corpus.

3.1.2 Tokenisation

Automatic tokenisation of the Romanian ELEXIS component was performed using UDPipe (Straka et al., 2016) with a model trained on the Romanian Reference Treebank (RRT) (Barbu Mititelu, 2018), after which the output was manually corrected in a shared Google Sheets environment following unified multilingual guidelines. These guidelines stress accurate tokenisation as the basis for all envisioned subsequent annotation layers (lemmatisation, POS tagging, UD parsing, MWE and NE annotation, and WSD), and specify that ad-hoc compounds are the only multiword units systematically split for token-level disambiguation in Romanian.

Romanian-specific tokenisation rules were adapted from previous UD-based resources and detailed treatments for hyphenated structures were defined. Established dictionary compounds and toponyms are kept as single tokens, while semi-productive or context-specific compounds (e.g. ad-hoc formations) are split and annotated by span. Additional rules govern ambiguous hyphenation cases, including foreign compounds, inflectional hyphens, numerical intervals, and lexicalised expressions, ensuring consistent handling across diverse morphological and orthographic phenomena.

A major focus is the treatment of Romanian cliticisation and vowel elision, where hyphens and apostrophes signal phonological reduction in pronouns (e.g., *spune-mi* "tell me", where the initial vowel *î* from the pronoun *îmi* is elided), auxiliaries (e.g., *cărțile-s noi* "the books are new" instead of *cărțile îs*[4] *noi*), prepositions (e.g., *c-un cuțit* "with a knife", where the preposition *cu* is shortened to *c-*), and conjunctions (e.g., *c-a venit* "that he came", where the conjunction *că* "that" is reduced to *c-*); these are systematically split with the clitic attached to the appropriate host. Further rules cover symbols, formulas, units, currency, slashes, named entities, abbreviations, and IT terms, balancing fine-grained segmentation with the preservation of lexical integrity to ensure consistency across linguistic categories.

3.1.3 Lemmatisation and part-of-speech tagging

Part-of-speech tagging, morphological feature assignment, and dependency parsing for the Romanian ELEXIS corpus were initially performed automatically using UDPipe (Straka, 2018), trained on the Romanian Reference Treebank (RRT), and exported in CONLL-U format before conversion to TSV3 for compatibility with INCEpTION (Klie et al., 2018). The resulting annotations were subsequently manually reviewed and corrected at the lemmatization and morphological levels, ensuring consistency and accuracy across the dataset.

During manual validation, several recurrent error types were identified, largely driven by lexical and morphological homonymy. A major source of inconsistency involved ambiguity between adjectives and participles, as well as between proper nouns and common nouns in capitalized or context-dependent uses (e.g. *Pământ* "Earth" in texts from the geography domain and *pământ* "earth" in the others). Additional difficulties arose in distinguishing adverbs from nouns in temporal expressions, where the syntactic context was insufficient for reliable automatic disambiguation.

Further challenges included homonymy within verb forms across tenses (present of the indicative mood and simple past of the indicative mood), ambiguous clitic pronouns with identical surface forms (e.g., *le* can be either dative or accusative), and nouns sharing forms between singular oblique and plural direct cases (e.g., *fete*). Abbreviations were frequently misclassified as proper nouns, requiring manual correction, while lemmatization errors affected neuter and irregular plural forms. These issues highlight the limitations of automatic morphological processing for Romanian and underscore the necessity of systematic human validation.

[4]The form *îs* "is" is regional nowadays.

3.2 Semantic Annotation

Two types of semantic information are annotated in ELEXIS-Ro: multiword expressions and named entities.

3.2.1 Multiword expressions

Multiword expressions (MWEs) are combinations of at least two lexicalized components that form a weakly connected structure and exhibit orthographic, morphological, syntactic, or semantic idiosyncrasies relative to general grammar (Baldwin and Kim, 2010), following the PARSEME guidelines[5]. The corpus covers a broad typology of MWEs:

- Verbal
  - Verbal Idioms (VID): *avea în vizor* (‘have in sight’ “keep in view, target”), *da nas în nas* (‘give nose in nose’ “run into, bump into”)
  - Light Verb Constructions (LVC)
    - full: *avea întâlnire* (‘have meeting’ “have a meeting”), *da citire* (‘give reading’ “read”)
    - cause: *da asigurare* (‘give assurance’ “assure”), *pune în aplicare* (‘put in application’ “implement, carry out”)
  - Reflexive Verbs (IRV): *-și permite* (‘3SG.DAT.REFL allow’ “afford”), *se abține* (“refrain”)
  - Inherently Adpositional Verbs (IAV): *se baza pe* (‘3SG.ACC.REFL base on’ “rely on, be based on”), *depinde de* (“depend on”)
- Nominal
  - Nominal Idioms (NID): *act de identitate* (‘act of identity’ “identity card”), *an de grație* (‘year of grace’ “year of Our Lord, A.D.”)
  - Pronominal Idioms (PronID): *câte ceva* (“something”), *Domnia Sa* (‘His/Her Lordship’ “His Excellency”)
  - Deverbal Nominal MWEs (NV): *punere în funcțiune* (‘putting in function’ “commissioning, putting into operation”), *băgare de seamă* (‘putting of notice’ “attention, observation”)
- Modifier
  - Adjectival Idioms (AdjID): *cu o falcă în cer și cu una în pământ* (‘with one jaw in sky and one in earth’ “very furious”), *cu cântec* (‘with song’ “with flair, extravagantly”)
  - Adverbial Idioms (AdvID): *așa cum se cuvine* (‘so as 3SG.ACC.REFL befi’ “properly, as it should be”), *ca oamenii* (‘like people.THE.PL’ “like decent people, properly”)
  - Deverbal adjectival / adverbial MWEs (AV): *cu luare aminte* (‘with taking notice’ “attentively”), *luat în seamă* (‘taken in notice’ “taken into account, considered”)
- Functional
  - Determiner Idioms (DetID): *picior de* (‘foot of’ “not a single”), *ca atare* (“as such”)
  - Adposition Idioms (AdpID): *de dragul* (‘of dear.DEF.SG.M’ “for the sake of”), *cu excepția* (‘with the exception’ “except for”)
  - Conjunction Idioms (ConjID): *astfel încât* (“so that”), *pe motiv că* (‘on reason that’ “because, on the grounds that”)
  - Interjection Idioms (IntjID): *ce folos* (‘what use’ “what’s the point?”), *nici vorbă* (‘no word’ “no way, not a chance”)

Automatic MWE identification was performed using a lexicon of MWEs derived from the validated PARSEME-Ro corpus (Barbu Mititelu et al., 2025), yielding 7,268 unique occurrences mapped to PARSEME labels and matched against ELEXIS-Ro. The matching procedure allowed for discontinuous realizations by permitting up to two intervening tokens between components, balancing recall and noise in the detection of long-distance dependencies.

Automatic identification procedures inevitably introduce both false positives and false negatives, due to the structural variability and context-dependent interpretation of MWEs. In line with established methodologies, we performed a systematic validation of the automatically extracted and annotated MWEs, focusing on verifying their boundaries, eliminating spurious candidates, and ensuring the correct assignment of MWE

[5] https://parsemefr.lis-lab.fr/parseme-st-guidelines/2.0/

F1	A1	A2	A3
A1	1	0.7681	0.7587
A2	0.7681	1	0.8547
A3	0.7587	0.8547	1

Table 1: Inter-annotator agreement

types. Particular attention was paid to ambiguous cases, homonymous expressions, and constructions whose MWE status depends on contextual interpretation.

The validation process involved three trained annotators and followed a cross-validation protocol inspired by previous annotation campaigns. Each candidate MWE was independently assessed by two annotators, applying the PARSEME decision tests and annotation guidelines. The corpus was split in five files, each of 500 sentences (except the fifth one). Three annotators took care of the validation and correction of their MWE annotation, the pairing among them being different from file to file: file 1 was annotated by annotators A1 and A2, file 2 by annotators A2 and A3, file 3 by annotators A1 and A3, file 4 by annotators A1 and A2, file 5 by annotators A2 and A3.

The inter-annotator agreement for each file is shown in Table 1. We adopted a pairwise overlapping validation design, where each sentence was validated and corrected by two out of three annotators, ensuring balanced coverage of all annotator pairs while controlling annotation cost. Inter-annotator agreement was computed following the PARSEME shared task methodology (Ramisch et al., 2018), where agreement is defined as an F1-score over MWE annotations (Fspan), based on exact span and category matching [6]. Validation results coming from the three annotators were considered independent annotation sets, and pairwise F1 scores were computed for all annotator pairs (see Table 1). The overall agreement score (0.7938) was obtained by averaging these pairwise F1 values.

3.2.2 Named entities

Named entities (NEs) of three types were annotated in the corpus: person (PER), organization (ORG) and location (LOC). Just like the other types of annotation, this one was also done automatically,

[6]Fleiss' kappa was not used because it assumes independent categorical decisions and does not account for span-based, possibly overlapping MWE annotations.

PoS	Number		
NOUN	8413	ADV	1337
ADP	4924	PRON	1187
PUNCT	4110	NUM	1178
VERB	3109	CCONJ	1019
ADJ	2653	PART	484
AUX	2372	SCONJ	234
DET	2322	X	222
PROPN	1750	INTJ	1
		SYM	1
	TOTAL	**35316**	

Table 2: PoSes distribution in ELEXIS-Ro

using three lists of NEs[7]. The automatic annotation was checked by two annotators and corrections were made when necessary (skipped NEs, strings incorrectly classified as NEs, incorrect length of NEs, incorrect label). When the two annotators disagreed in their evaluation, a third, specialist, one made the final decision.

4 Statistics on ELEXIS-Ro

4.1 General statistics

As for all the other languages it contains, the corpus has 2024 sentences. In Romanian these sentences contain 35,316 words, with an average sentence length of 17.45.

4.2 Statistics on PoSes

The distribution of tokens according to the part-of-speech can be seen in Table 2. All parts of speech are present in the corpus, with nouns dominating, which is indicative of an information-dense corpus.

4.3 Statistics on MWEs

Overall, 3,018 MWEs were automatically identified in the Romanian corpus, and their number of occurrences and relative frequencies in the ELEXIS-Ro corpus are rendered in column two and three of Table 4. For their comparison with the MWEs in the PARSEME-Ro corpus, see Section 5 below.

4.4 Statistics on NEs

The distribution of the three types of NEs in the corpus is presented in Table 3. This is based on the automatic identification (the middle column),

[7]These lists are available at https://www.racai.ro/p/reterom/rezultate.html, last accessed April 2026

Type of NE	#	Final #	% skipped
PER	277	516	53
ORG	220	321	68
LOC	404	754	53
TOTAL	**901**	**1591**	

Table 3: NEs distribution in ELEXIS-Ro: column two contains the numbers after automatic annotation, while the third column renders the number after the manual validation, correction and clarification of disagreements.

with the results of the manual validation in the third column (see Section 6). We notice that about half of the NEs were not recognized automatically, with the distribution per type shown in the last column.

5 Discussion of results

MWEs, as recurrent lexical or phrasal patterns, often exhibit variation in frequency, syntactic structure, and semantic category depending on the register and domain of the texts in which they occur. We are going to discuss now the quantitative differences between the MWEs occurring in the ELEXIS-Ro and in the PARSEME-Ro corpora. The former is primarily composed of encyclopedic and informative texts, while the latter consists exclusively of journalistic texts, thus they provide contrasting linguistic environments. This contrast makes it possible to examine not only absolute frequencies, but also the relative distribution of MWE types, offering a basis for interpreting the statistical patterns discussed below.

It should also be noted that the two corpora differ considerably in terms of the total number of annotated MWEs, with 3,018 instances in ELEXIS and 64,601 in PARSEME-Ro. This difference is due to variations in corpus size and composition. In order to ensure comparability, the analysis is therefore based on relative frequencies, which provide a more appropriate basis for examining the distribution of MWE types across the two datasets. Accordingly, the discussion focuses on proportional patterns rather than absolute counts.

The quantitative comparison reveals both convergences and divergences in the distribution of MWE types across the two corpora. At the top of the distribution, AdpID expressions show a remarkably similar proportion in both datasets, accounting for 28.60% in ELEXIS and 28.48% in PARSEME-Ro. This consistency suggests that this category is relatively stable across genres and less sensitive to variation in text type. A similar, though slightly less precise, pattern can be observed for ConjID expressions (6.83% in ELEXIS vs. 6.51% in PARSEME-Ro) and, to a lesser extent, DetID, which displays nearly identical proportions in both corpora. These categories appear to form a core layer of MWEs whose distribution remains largely unaffected by domain-specific factors.

In contrast, several categories exhibit substantial variation. AdvID expressions are considerably more frequent in ELEXIS (28.00%) than in PARSEME-Ro (14.50%), pointing to a stronger presence of adverbial multiword units in encyclopedic and informative texts, where such structures contribute to descriptive precision and elaboration.

Conversely, PARSEME-Ro shows higher proportions of verb-related MWEs and structurally more complex expressions. Categories such as IAV (12.14% vs. 8.35%), IRV (9.98% vs. 5.86%), and especially NID (11.93% vs. 3.08%) are much more prominent in the journalistic corpus. This pattern suggests a preference for more dynamic, predicate-centered constructions, as well as noun-based MWEs that may encode salient or context-dependent information.

A similar tendency can be observed for LVC.full and LVC.cause, which are more frequent in PARSEME-Ro, although they remain relatively marginal overall. These constructions, often associated with more complex predicate formation, further point to a greater use of analytically structured verbal expressions in journalistic texts.

Lower-frequency categories (e.g., AV.VID, NV.VID, NV.LVC.full) remain marginal in both corpora, and their small proportions make it difficult to draw strong genre-based conclusions. However, even in these cases, PARSEME-Ro tends to display slightly higher values and more diverse types, reinforcing the general pattern of greater structural diversity.

Overall, the comparison highlights a distinction between more stable, function-word-based MWE categories and those that are more sensitive to genre. While the former remain consistent across corpora, the latter vary more noticeably, reflecting the role of textual domain in shaping both the frequency and typology of MWEs.

6 Future Work

On the one hand, validation and correction of the automatic annotation is the very next step. This will

MWE type	# in ELEXIS-Ro	% in ELEXIS-Ro	# in PARSEME-Ro	% in PARSEME-Ro
AdpID	863	28.60	18401	28.48
AdvID	845	28.00	9366	14.50
AdjID	347	11.50	4496	6.96
IAV	252	8.35	7840	12.14
ConjID	206	6.83	4206	6.51
IRV	177	5.86	6449	9.98
NID	93	3.08	7709	11.93
VID	79	2.62	2573	3.98
PronID	73	2.42	745	1.15
AV.IAV	30	0.99	779	1.21
LVC.full	18	0.60	934	1.45
DetID	17	0.56	387	0.60
AV.VID	6	0.20	28	0.04
NV.LVC.cause	4	0.13	76	0.12
IntjID	4	0.13	59	0.09
LVC.cause	2	0.07	311	0.48
NV.VID	1	0.03	106	0.16
NV.LVC.full	1	0.03	67	0.10
NV.IAV	–	–	54	0.08
AV.LVC.cause	–	–	11	0.02
AV.LVC.full	–	–	4	0.01
NV.IRV	–	–	1	0.00
TOTAL	**3018**		**64601**	

Table 4: Comparative statistics between MWEs occurring in ELEXIS-Ro and in PARSEME-Ro.

target syntactic parsing correctness, in line with the UD principles of annotation, deciding upon the cases of disagreement between the annotators of the MWEs.

On the other hand, the corpus will be semantically disambiguated, i.e. each content word will be assigned a sense, based on the Romanian WordNet (Tufiș and Mititelu, 2015).

Once the corpus annotation is validated and corrected it can serve as a reliable source of linguistic information for comparative studies.

7 Conclusions

In this paper, we have presented the extension of the ELEXIS parallel corpus through the integration of Romanian, following a hybrid workflow that combines automatic processing with systematic manual validation. The pipeline included translation from English, tokenisation, morpho-syntactic annotation with UDPipe, and semantic enrichment through the identification and classification of MWEs and NEs. The results confirm that while automatic tools provide a solid baseline for large-scale corpus construction, they consistently produce errors in contexts involving ambiguity, homonymy, and language-specific structures, making manual correction an essential component for ensuring annotation quality.

The analysis of annotation errors highlights several recurrent challenges specific to Romanian, including part-of-speech ambiguity, lemmatisation errors, and difficulties in handling clitics, compounds, and NEs. Similarly, the automatic identification of MWEs, based on the PARSEME guidelines, proved effective, but required careful post-editing for the correct span identification and type of MWE labelling. The double validation process contributed to improving consistency and reliability across annotation layers, reinforcing the importance of human expertise in multilingual resource development.

Overall, the Romanian extension enhances the ELEXIS corpus both quantitatively and qualitatively, expanding its coverage to a less-resourced language and increasing its value for cross-linguistic research. The resulting dataset, enriched with detailed morpho-syntactic and semantic annotation, provides a robust resource for studies in lexicography, translation, and multilingual natural language processing.

8 Corpus Availability

ELEXIS-Ro has been released in the 2.0 version of the ELEXIS corpus (Čibej et al., 2026).

9 Acknowledgment

This work was supported by a grant of the Ministry of Education and Research, CCCDI – UEFISCDI, project number PN-IV-PCB-RO-MD-2024-0142, within PNCDI IV. This work also received support from the CA21167 COST action UniDive, funded by the European Union via the COST (European Cooperation in Science and Technology).

References

Timothy Baldwin and Su Nam Kim. 2010. Multiword Expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, 2 edition, pages 267–292. CRC Press, Boca Raton, USA.

Verginica Barbu Mititelu. 2018. Modern syntactic analysis of Romanian. In *Clasic și modern în cercetarea filologică românească actuală*, pages 67–78, Iași, Romania.

Verginica Barbu Mititelu, Ioana-Maria Biolan, Ioana Buhnila, Mihaela Cristescu, Elena Irimia, Catalin Mihaila, Isabella Sinca, Amalia Todirascu, and Carmen Vasile. 2025. Enhancing the parseme-ro corpus with nominal, modifier and functional multiword expressions. In *Proceedings of the 20th International Conferene on Linguistic Resources and Tools for Natural Language Processing*, pages 113–125.

Jaka Čibej, Simon Krek, Carole Tiberius, Federico Martelli, Roberto Navigli, Jelena Kallas, Polona Gantar, Svetla Koeva, Sanni Nimb, Bolette Sandford Pedersen, Sussi Olsen, Margit Langemets, Kristina Koppel, Tiiu Üksik, Kaja Dobrovoljc, Rafael Ureña-Ruiz, José-Luis Sancho-Sánchez, Veronika Lipp, Tamás Váradi, András Győrffy, László Simon, Valeria Quochi, Monica Monachini, Francesca Frontini, Rob Tempelaars, Rute Costa, Ana Salgado, Tina Munda, Iztok Kosem, Rebeka Roblek, Urška Kamenšek, Petra Zaranšek, Karolina Zgaga, Primož Ponikvar, Luka Terčon, Jonas Jensen, Ida Flörke, Henrik Lorentzen, Thomas Troelsgård, Diana Blagoeva, Dimitar Hristov, Sia Kolkovska, Kadri Muischnek, Kertu Saul, Karoliina Jõgi, Mija Bon, Ranka Stanković, Cvetana Krstev, Aleksandra Marković, Milica Ikonić Nešić, Voula Giouli, Eri Papanikolaou, Irina Lobzhanidze, Verginica Barbu Mititelu, Simina Popa, Lea Cristiana, Mihaila Catalin, Elena Irimia, Ana Ostroški Anić, Siniša Runjaić, Robert Sviben, Martina Pavić, Ivana Filipović Petrović, Bartłomiej Alberski, Vladimir Cvetkoski, Olha Kanishcheva, and Rusudan Makhachashvili. 2026. Parallel sense-annotated corpus ELEXIS-WSD 2.0. Slovenian language resource repository CLARIN.SI.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech and Communication. MIT Press.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics. Event Title: The 27th International Conference on Computational Linguistics (COLING 2018).

Cvetana Krstev, Ranka Stanković, Aleksandra M. Marković, and Teodora Sofija Mihajlov. 2024. Towards the semantic annotation of SR-ELEXIS corpus: Insights into multiword expressions and named entities. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 106–114, Torino, Italia. ELRA and ICCL.

Federico Martelli, Roberto Navigli, Simon Krek, Jelena Kallas, Polona Gantar, Svetla Koeva, Sanni Nimb, Bolette Sandford Pedersen, Sussi Olsen, Margit Langemets, Kristina Koppel, Tiiu Üksik, Kaja Dobrovoljc, Rafael-J. Ureña-Ruiz, José-Luis Sancho-Sánchez, Veronika Lipp, Tamás Váradi, András Győrffy, Simon László, Valeria Quochi, Monica Monachini, Francesca Frontini, Carole Tiberius, Rob Tempelaars, Rute Costa, Ana Salgado, Jaka Čibej, and Tina Munda. 2021. Designing the elexis parallel sense-annotated dataset in 10 european languages. *Proceedings of Electronic Lexicography in the 21st Century Conference*, (2021):377–395. UIDB/03213/2020 UIDP/03213/2020; eLex 2021, 7th biennial conference on electronic lexicography, eLex 2021 ; Conference date: 05-07-2021 Through 07-07-2021.

Bolette Pedersen, Sanni Nimb, Sussi Olsen, Thomas Troelsgård, Ida Flörke, Jonas Jensen, and Henrik Lorentzen. 2023. The DA-ELEXIS corpus - a sense-annotated corpus for Danish with parallel annotations for nine European languages. In *Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023)*, pages 11–18, Tórshavn, the Faroe Islands. Association for Computational Linguistics.

Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya, and Abigail Walsh. 2018. Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Agata Savary, Daniel Zeman, Verginica Barbu Mititelu, Anabela Barreiro, Olesea Caftanatov, Marie-Catherine de Marneffe, Kaja Dobrovoljc, Gülşen Eryiğit, Voula Giouli, Bruno Guillaume, Stella Markantonatou, Nurit Melnik, Joakim Nivre, Atul Kr. Ojha, Carlos Ramisch, Abigail Walsh, Beata Wójtowicz, and Alina Wróblewska. 2024. UniDive: A COST action on universality, diversity and idiosyncrasy in language technology. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 372–382, Torino, Italia. ELRA and ICCL.

Milan Straka. 2018. 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*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Milan Straka, Jan Hajic, and Jana Straková. 2016. Udpipe: trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297.

Dan Tufiș and Verginica Barbu Mititelu. 2015. *The Lexical Ontology for Romanian*, pages 491–504. Springer International Publishing, Cham.

A Dual-Interface Syntax Tree Editor with Automated Assessment for Learning Management Systems

Atanas Atanasov
Faculty of Slavic Studies,
Sofia University "St. Kliment Ohridski"
atanasov@slav.uni-sofia.bg

Abstract

While syntax trees are fundamental to computational linguistics, existing visualization tools often lack research flexibility or native Learning Management System (LMS) integration. This paper presents a dual-interface syntax tree editor consisting of a standalone web application for linguistic annotation[1] and a Moodle question type for educational use. The system provides an interactive drag-and-drop interface for constructing hierarchical syntactic structures directly in the browser.

The Moodle integration features an automated grading engine based on a weighted Leaf-Ancestor metric. Unlike PARSEVAL-based approaches, this method evaluates hierarchical accuracy by comparing bottom-up syntactic paths via token-based weighted Levenshtein distance. The grading logic treats the root, parts-of-speech, and internal hierarchy as separate components, distinguishing between structural omissions and terminological errors. This enables nuanced partial credit and formative feedback consistent with educational assessment principles. The system supports scalable assessment and linguistic annotation, with particular relevance for less-resourced languages such as Bulgarian.

Keywords: Syntax Tree Editor, Automated Assessment, Computer-Aided Learning, Moodle question types, Weighted Leaf-Ancestor metric.

1 Introduction

Syntax trees are a fundamental component of both computational linguistics research and syntax education. For linguistic researchers, constructing accurate syntactic hierarchies is an essential task for analyzing sentence structures, creating treebanks, and producing syntactic annotations (Kiss and Alexiadou, 2015). In the educational context, the active construction of these diagrams helps students transition from explicit grammatical rule comprehension to implicit syntactic knowledge, allowing them to visualize complex dependencies (Van Valin, 2004; Kroeger, 2005). However, balancing the flexibility required for advanced research annotation with the scalability needed for educational assessment remains an open problem.

Several software tools have been developed to facilitate syntax tree drawing, yet they often address only partial aspects of the educational and research workflows. Standalone desktop applications like the original Syntax Tree Editor and TreeForm (Derrick and Archambault, 2010) offer interactive drag-and-drop interfaces and produce high-quality vector graphics suitable for scientific publication. While highly useful for linguists, such tools lack native integration into Learning Management Systems (LMS), requiring students to export images and instructors to manually grade them. Conversely, web-based generators such as RSyntaxTree (Hasebe, 2009) produce static images from bracketed text input, lacking both the intuitive graphical manipulation desired by researchers and the automated assessment capabilities required by educators. Table 1 summarizes the feature comparison between existing popular solutions and the proposed system.

To properly contextualize the proposed tool within the current landscape of syntactic technology, it is essential to distinguish between

[1] https://syntree.ezik.bg/

automated deep-learning syntactic parsers and interactive, human-in-the-loop tree editors. While neural parsers execute automated statistical sequence labeling, visual environments like TreeForm (Derrick and Archambault, 2010), RSyntaxTree (Hasebe, 2009), and the suite of manual annotation tools maintained within the Universal Dependencies (UD) framework – such as browser-based editors like UD Annotatrix, ConlluEditor, and ArboratorGrew – serve a fundamentally different purpose. These graphical environments are explicitly designed to facilitate human-centric structural validation, educational diagramming, and the manual curation of gold-standard treebanks, which are themselves required to train and evaluate deep-learning parsing architectures. While many modern UD-aligned platforms are specifically optimized for word-level dependency relations or CoNLL-U format modification, our system provides a unique dual-interface setup tailored specifically for phrase-structure constituency tree manipulation with native LMS integration for academic assessment, establishing an accessible pipeline for student training and pedagogical dataset generation.

Addressing the disparity between flexible linguistic research and scalable educational assessment, this paper presents a novel, dual-interface syntax tree editor. The system comprises a standalone web application designed for efficient linguistic annotation by scientists, and a native Moodle question type tailored for educational deployment. The editor provides an interactive, browser-based drag-and-drop interface that allows users to construct and modify complex hierarchical structures dynamically, without requiring external software installations or complex local setups.

A key feature of the Moodle integration is the introduction of an automated grading engine that evaluates student submissions using an advanced Isolated Root Weighted Leaf-Ancestor metric. By analyzing bottom-up syntactic paths and decoupling the evaluation of the root, parts-of-speech, and internal hierarchy, the algorithm overcomes the strict penalization and over-grading problems typical of standard PARSEVAL metrics. This enables nuanced partial credit and targeted formative feedback consistent with modern educational principles. Consequently, this system provides a scalable solution that serves both as an advanced annotation tool for linguists and a computer-aided learning application for educators, making it highly applicable to multi-lingual contexts, including less-resourced languages such as Bulgarian.

2 System Architecture and Design

To satisfy the dual requirements of flexible linguistic research and scalable educational assessment, the syntax tree editor is engineered as a decoupled, client-side TypeScript application. It operates in two deployment environments: a standalone web application for linguistic annotation and a custom Moodle question type plugin (qtype_syntree) for educational integration.

2.1 Rendering Engine and User Interface

Unlike older desktop tools requiring external runtimes (e.g., Java), the editor operates entirely in-browser using Scalable Vector Graphics (SVG). This guarantees resolution-independent rendering, making the output highly suitable for interactive manipulation across screen sizes and high-quality vector exports for scientific publications.

User interaction is driven by the unified Pointer Events API, providing seamless drag-and-drop capabilities across desktop (mouse) and mobile (touch) devices. One specific technical challenge in creating readable syntax trees is maintaining the symmetric visual alignment of nodes, particularly when pairing bold syntactic labels (e.g., NP, VP) with regular-weight lexical items. To resolve this, the system's inline node editor utilizes an off-screen HTML5 Canvas API (measureText()) to calculate the exact pixel width of the input strings dynamically. Furthermore, the layout engine employs a two-pass algorithm and a Depth-First Search (DFS) traversal to order the leaf nodes, which mathematically guarantees a crossing-free layout for the branches, preserving the formal properties of constituency trees.

In addition to standard constituency trees, the rendering engine supports the visualization of syntactic movement through trace arrows. These are implemented as complex SVG path elements with an automatic collision avoidance algorithm, which dynamically calculates orthogonal routing to prevent intersections with existing nodes. Furthermore, to support academic publishing, the system features an advanced export module that provides direct SVG downloads or renders the vector-based DOM into high-resolution PNGs via a scaled Canvas.

2.2 Moodle LMS Integration

In the educational deployment, the editor is tightly integrated into the Moodle ecosystem as a custom question type. The integration utilizes Moodle's Asynchronous Module Definition (AMD) architecture to load the compiled JavaScript bundle dynamically within the quiz interface.

Instead of relying on a separate backend server or complex API calls for state management, the plugin employs a robust real-time serialization approach. As the student constructs or modifies the tree via the graphical interface, the core engine serializes the hierarchical structure into a standard linguistic bracket notation (e.g., [S [NP [D The] [N boy]] [VP [V ran]]]). This serialized string is continuously synchronized with a hidden HTML input field within the Moodle quiz form. Upon submission, Moodle's native form processing captures the bracket notation string, which is subsequently parsed by the backend PHP grading engine. This architecture ensures high interoperability, strict compliance with institutional data security policies, and seamless data flow between the graphical frontend and the automated assessment backend.

A visual demonstration of the instructor's question configuration interface and the corresponding student interactive environment within the Moodle LMS ecosystem is provided in Appendix A (Figure 1 and Figure 2).

3 Automated Assessment via a Weighted Leaf-Ancestor Metric

Standard metrics for evaluating syntax trees, such as PARSEVAL, are widely used in computational linguistics but present significant drawbacks in educational contexts. They often over-grade "flat" trees that lack hierarchical depth and strictly penalize minor terminological errors, failing to provide the nuanced formative feedback required for learning. To address this, the proposed Moodle integration employs a custom grading engine based on an advanced Isolated Root Weighted Leaf-Ancestor metric.

The grading engine evaluates student submissions against an instructor-defined reference tree by decoupling the assessment into three independent components: Part-of-Speech (POS) accuracy (15%), Root identification (5%), and Internal Hierarchy (80%). This separation ensures pedagogical fairness by preventing a "domino effect," where a single error at the highest node disproportionately degrades the score of an otherwise correct structural derivation. To accommodate diverse educational goals, the system allows instructors to dynamically customize and override these default weights within the Moodle interface, realigning the automated scoring with specific course levels or targeted learning outcomes.

The core innovation of the assessment engine lies in the Internal Hierarchy evaluation. For each lexical item in the sentence, the system extracts a bottom-up syntactic path from the leaf to the root. To isolate purely structural knowledge, the leaf node and the pre-terminal POS tag are filtered out from this path. The system then compares the student's syntactic path to the reference path using a Weighted Levenshtein Distance algorithm. While severe structural errors – such as omitted or extraneous hierarchical levels (insertions and deletions) – incur a full penalty cost of 1.0, terminological errors (substitutions, e.g., mislabeling a Noun Phrase as a Verb Phrase) incur a reduced cost of 0.66.

For a given student path S and reference path R, the structural score per word is calculated using mathematical normalization:

$$Score_{word} = \max\left(0{,}1 - \frac{\mathrm{dist}_{weighted}(S,R)}{\max(\mathrm{len}(S),\mathrm{len}(R))}\right)$$

Furthermore, the engine incorporates strict pedagogical safeguards. If a student submits a completely flat structure without internal hierarchy, the hierarchy score automatically defaults to zero, preventing unearned credit. By differentiating between fundamental structural omissions and surface-level labeling mistakes, this algorithm provides scalable, fair partial credit consistent with modern assessment principles.

3.1 Evaluation of the Grading Engine

Standard evaluation metrics like PARSEVAL present dual pedagogical failures: they over-penalize minor terminological label mistakes, yet often over-grade completely "flat" trees that lack hierarchical depth. To demonstrate how the proposed Isolated Root Weighted Leaf-Ancestor metric mathematically resolves these issues, consider the reference sentence evaluated against an instructor-defined gold standard: [S [NP [D The] [N boy]] [VP [V ran]]].

If a student submits an incorrect structure that mislabels the Noun Phrase as a Verb Phrase ([S [VP [D The] [N boy]] [VP [V ran]]]), the grading engine isolates the structural tracking. For the token "boy", the bottom-up reference path (excluding the POS tag) is [NP, S], while the student's layout yields [VP, S]. The Weighted Levenshtein Distance algorithm detects a single substitution operation. According to the system design, while structural insertions or deletions incur a full penalty of 1.0, terminological substitutions incur a reduced cost of 0.66 to reflect partial student knowledge. Applying the system's normalization formula:

This mathematically guarantees that the student receives a partial credit of 67% for the token by establishing the correct hierarchical depth and grouping, precisely isolating and penalizing only the surface-level labeling error without triggering a catastrophic scoring failure for the entire phrase.

For a given student path S and reference path R, the structural score per word is calculated using mathematical normalization:

$$Score_{word} = \max\left(0{,}1 - \frac{0.66}{\max(2{,}2)}\right) = 0.67$$

Conversely, if a student submits a completely flat tree (attaching all terminals directly to the root with no intermediate hierarchy), the algorithm detects the complete absence of internal paths. Because the reference path length for "boy" is 2, and the student's path is empty (length 0), the deletion distance is evaluated as 2.0 (with a full 1.0 penalty per deletion).

The maximum function safeguard ensures that the hierarchy score correctly defaults to zero, rigorously penalizing the fundamental structural omission.

The practical implementation within the Slavic Philology cohort at Sofia University fully substantiated these algorithmic expectations. Students adapted to the web-based editor seamlessly, demonstrating an intuitive grasp of the interactive drag-and-drop interface without requiring prior review of the detailed user documentation. In independent study settings, the instantaneous feedback provided by the automated scoring engine served as a vital formative resource during self-preparation. For instructors, the framework drastically minimized the administrative time required for manual grading and script evaluation. Furthermore, deploying the tool during synchronous lectures enabled the interactive analysis of a significantly larger volume of syntactic structures compared to previous semesters, which relied on traditional paper-based workflows or the legacy standalone Syntax Tree Editor.

4 Educational and Research Applications

The proposed dual-interface system addresses distinct but complementary use cases in both linguistic research and higher education.

In the research context, the standalone web application serves as an accessible and highly efficient annotation tool for linguists. The construction of syntactic resources, such as treebanks, is particularly critical for the computational advancement of less-resourced languages like Bulgarian. It is important to specify that while contemporary Bulgarian NLP is well-supported by robust processing pipelines for foundational tasks like morphological tagging and basic parsing, it remains relatively less-resourced regarding open-access, web-integrated pedagogical platforms and collaborative tools for interactive treebank validation. The standalone interface specifically targets this infrastructural gap. The interactive browser-based interface, coupled with the ability to export resolution-independent Scalable Vector Graphics (SVG), directly supports the creation of publication-ready linguistic analyses and the rapid expansion of natural language datasets without requiring complex local software setups.

In the educational context, the Moodle-integrated question type addresses the pedagogical challenges of teaching syntax to large student cohorts by shifting the focus toward active structural modeling. This approach aligns with broader Computer-Aided Language Learning (CALL) frameworks, where interactive, software-driven tree diagramming has been shown to significantly enhance students' implicit syntactic competence (El-Garawany, 2025). By embedding this functionality natively within an LMS, the system facilitates scalable, computer-aided learning. Instructors can deploy complex syntactic assignments efficiently, while the automated grading engine provides immediate, formative feedback. This automated approach not only dramatically reduces the administrative and grading burden on educators but also ensures a standardized, objective evaluation of students' structural and terminological knowledge.

5 Conclusion and Future Work

This paper presented a novel, dual-interface syntax tree editor designed to bridge the gap between flexible linguistic annotation and scalable computer-aided education. By operating entirely within the browser using resolution-independent SVG rendering, the standalone application provides a robust environment for researchers constructing treebanks for less-resourced languages. Simultaneously, its native integration into Moodle as a custom question type addresses the pedagogical need for scalable assessment in computational linguistics courses. The implementation of the Isolated Root Weighted Leaf-Ancestor metric represents a significant advancement over standard PARSEVAL approaches, enabling nuanced partial credit and formative feedback that are crucial for student development.

Future development will focus on expanding the system's analytical capabilities. A primary objective is the integration of the Universal Dependencies (UD) framework to support dependency parsing visualization alongside the current constituency trees. Additionally, future work will explore incorporating AI-assisted annotation, where backend parsing models generate preliminary tree structures that researchers or students can subsequently refine via the drag-and-drop interface. Finally, a large-scale empirical evaluation across multiple higher education institutions is planned to quantitatively measure the system's impact on students' syntactic competence and knowledge retention.

References

Derrick, D. and Archambault, D. (2010). TreeForm: Explaining and exploring grammar through syntax trees. *Literary and Linguistic Computing*, 25(1): 53–66.

El-Garawany, M. S. M. (2025). Using TreeForm for Enhancing English Language Majors' EFL Syntactic Competence. *rEFLecTions*, 32(2): 754–786.

Hasebe, Y. (2009). *RSyntaxTree* [Software]. Available at: https://yohasebe.com/rsyntaxtree/ (Accessed: 14 April 2026).

Kiss, T. and Alexiadou, A. (2015). *Syntax - Theory and Analysis: An International Handbook* (Vol. 3). De Gruyter Mouton.

Kroeger, P. R. (2005). *Analyzing Grammar: An Introduction*. Cambridge University Press.

Van Valin, R. D. (2004). *An Introduction to Syntax*. Cambridge University Press.

Appendix A User Interfaces

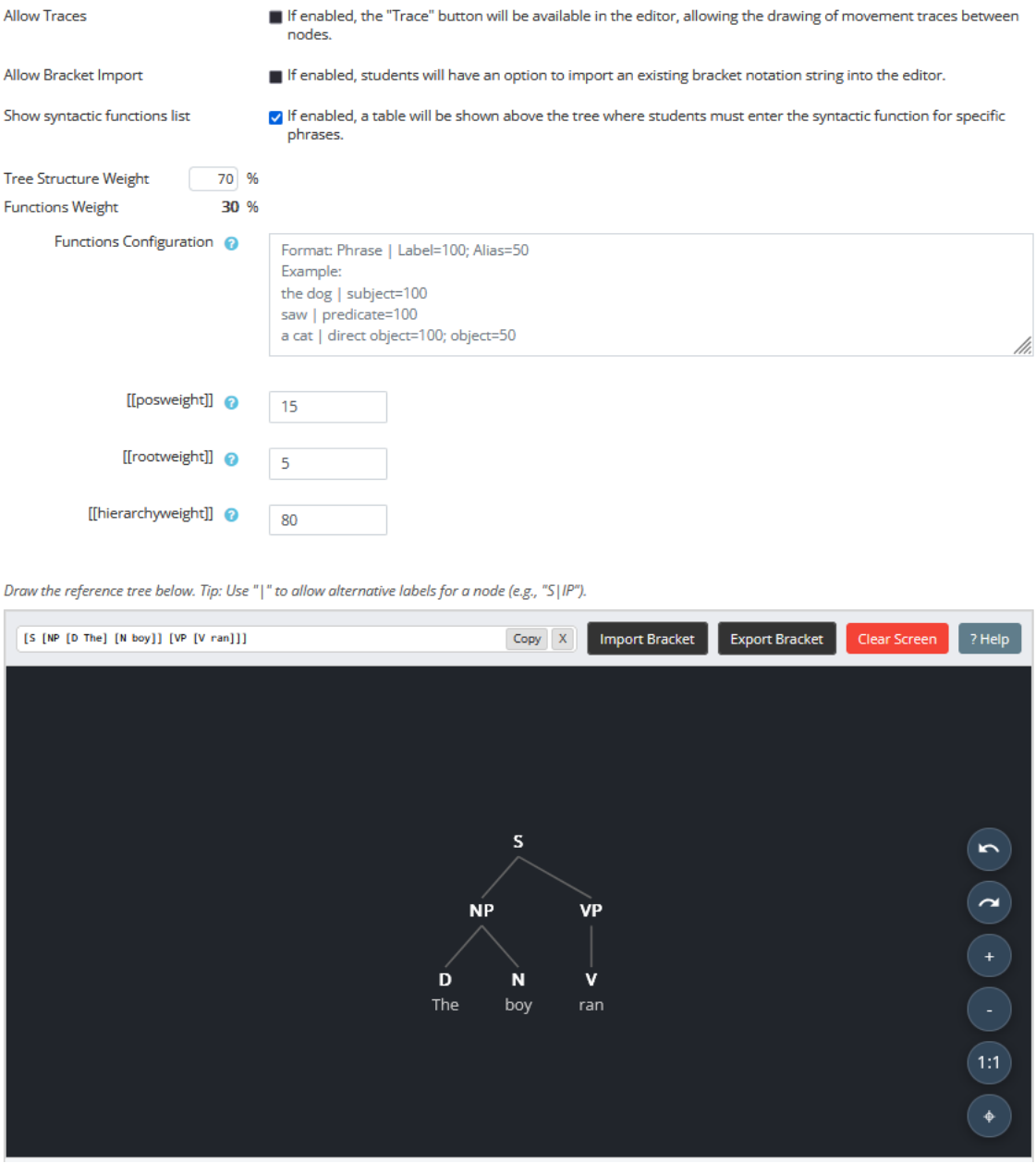


Figure 1: The instructor's question design and configuration interface within the Moodle platform, displaying reference bracket inputs and scoring weight parameters.

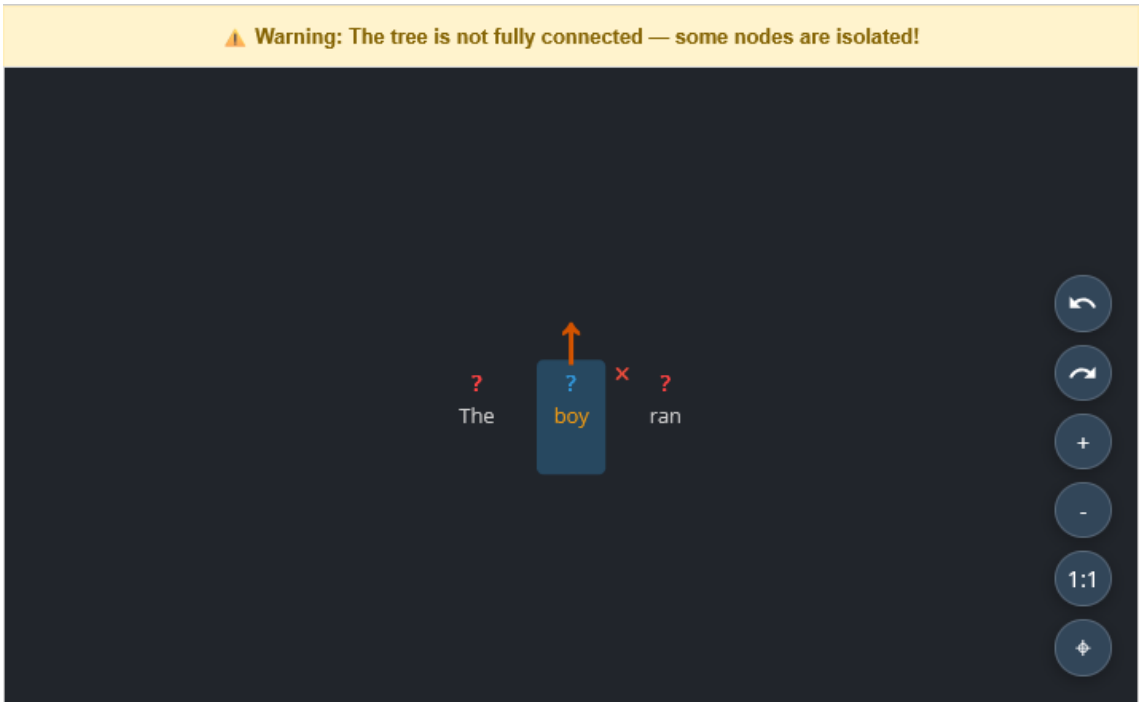


Figure 2: The student interactive environment within Moodle, featuring the responsive drag-and-drop canvas for structural syntax tree construction.

0

Developing a Pragmatic Annotation Scheme for a Spoken Corpus

María Sampedro Mella
Université catholique de Louvain, Belgium
maria.sampedro@uclouvain.be

Abstract

This paper presents a pragmatic annotation scheme for speech acts developed for the ESLORA corpus of spoken Spanish. The scheme integrates a pragmatic layer into an XML-based corpus infrastructure and encodes speech acts through an adjacency-pair-based model. In addition to enabling the retrieval of speech acts as independent categories, the system allows for queries that combine speech acts with lexical, morphosyntactic, and sociolinguistic variables. Given that corpus annotation has traditionally focused on lexical and morphosyntactic information (POS tagging), this proposed annotation scheme represents a valuable contribution to the field. The present paper describes the annotation scheme, XML implementation, query options, and limitations, thereby contributing to a computational model for corpus pragmatics.

Keywords: corpus pragmatics, pragmatic annotation, speech acts, spoken corpus.

1 Introduction

While linguistic corpora have been extensively used to analyse lexical and morphosyntactic phenomena, their use in pragmatic research remains more limited. This imbalance is due both to the influence of the grammatical tradition and to the part-of-speech (POS) tagging systems used to annotate corpora, which facilitate the retrieval of words, lexical combinations, categories, or lemmas. However, they usually do not encode pragmatic content such as speech acts, which makes it difficult to retrieve this type of data across registers or communicative situations.

Consequently, when using corpora in pragmatic research, researchers often have to perform queries based on representative forms (O'Keffe, 2018) or manually examine the materials of each corpus (Sampedro Mella, 2021a). This process is time-consuming and methodologically challenging, as pragmatic phenomena are highly context-dependent, require the interpretation of speaker intentions (Weisser, 2015; Milà-Garcia, 2018), and often rely on multimodal or paralinguistic cues (e.g., intonation, gesture, or prosodic emphasis). As a result, most corpus-based pragmatic studies rely on concordance searches of textual (i.e., transcribed) renderings of spoken discourse rather than on multimodal approaches (Adolphs & Chen, 2021). These limitations reflect broader methodological issues in the field of *corpus pragmatics*.

Even though corpus-based approaches have been used in pragmatic research for some time, corpus pragmatics is still a relatively novel discipline (Adolphs & Chen, 2021; Aijmer, 2026): "(…) corpus-pragmatic research is more than just pragmatic research and it is more than just corpus-linguistic analysis in that it integrates the horizontal (qualitative) methodology typical of pragmatics with the vertical (quantitative) methodology predominant in corpus linguistics" (Aijmer & Rühlemann, 2015: 12). Through computational procedures, this field seeks to contribute to the creation, enrichment, and pragmatic use of corpora.

Overall, pragmatic annotation has not been prioritised in corpus linguistics (Milà-García, 2018; Jucker, 2018), largely due to the linguistic and technical challenges that it entails. Annotating pragmatic phenomena is typically a manual process, since there is no one-to-one correspondence between speech acts and the linguistic or extralinguistic cues associated with them, and it often requires contextual interpretation: "Any type of linguistic annotation is a highly complex and interpretive process, but none more so than pragmatic annotation" (Weisser, 2015: 84). For this reason, in the context of spoken corpora, "pragmatic annotation is still much less advanced than other linguistic annotation types" (Rabadán, Ramón & Sanjurjo-González, 2021: 210).

Several English corpora containing pragmatic information have been developed over the past few decades. One of the first is the Switchboard Dialogue Act Corpus (Jurafsky, Shriberg & Biasca, 1997), which includes telephone conversations and is annotated with dialogue act tags. Similarly, the DialogBank (Bunt et al., 2016), developed under the ISO 24617-2 standard (Bunt et al., 2020), and the SPAADIA corpus (Leech & Weisser, 2013) provide structured dialogue act annotation frameworks for conversational data. The SPICE-Ireland Corpus (Kirk, 2013), by contrast, incorporates pragmatic annotation within a broader corpus-linguistic and prosodic framework.

In Spanish, only a limited number of corpora include pragmatic annotation systems that allow for the retrieval of speech acts, such as the C-ORAL-ROM corpus (Cresti & Moneglia, 2005), the C-Or-DiAL corpus (Nicolás Martínez, 2012), the CLANES corpus, and the COREMAH corpus (Vacas Matos, 2017). These corpora differ in both their design and their annotation schemes: the C-ORAL-ROM corpus focuses on spontaneous speech and includes prosodic annotation, although its pragmatic encoding is not primarily designed for the retrieval of speech acts; the C-Or-DiAL corpus is didactically oriented; the CLANES corpus is not publicly available, and the COREMAH corpus contains speech acts elicited through role-play interactions in which participants perform apologies, compliments, and negative responses to invitations.

The annotation schemes of these various corpora were developed for different interactional settings and research objectives. Furthermore, none of these corpora systematically allows for the extraction and categorisation of speech acts as recognised pragmatic units (see Sampedro Mella, 2026, for further information). For these reasons, these annotation schemes could not be directly incorporated into the annotation framework described in the present study.

In this context, the present paper introduces a pragmatic annotation scheme specifically developed for the ESLORA corpus. It encodes different speech acts in spoken interaction through a structured XML-based framework. By doing so, it aims to facilitate the retrieval of speech acts in spoken data without manual inspection. This contribution goes beyond XML annotation by providing a formal encoding of speech acts in corpus data that enables their automatic extraction. Consequently, this annotation scheme facilitates systematic research in corpus pragmatics and can be adapted and applied to other spoken or written corpora and languages.

2 ESLORA Corpus

The ESLORA[1] (*Corpus para el estudio del español oral*) corpus is a spoken-language corpus designed to analyse Spanish in Galicia, a bilingual region in the north-west of Spain. Its current version (2.5) consists of audio recordings and transcriptions of 60 hours of sociolinguistic interviews and 24 hours of casual speech.

The data were manually transcribed and aligned using either Transcriber or ELAN tools (Vázquez Rozas & Barcala, 2020: 226). Following transcription, the texts were morphosyntactically annotated and lemmatised. The morphosyntactic annotation was implemented using an XML-based format to facilitate advanced linguistic searches and analyses. This annotation includes several features of spoken language, such as pauses, hesitations, overlaps, and repetitions (Vázquez Rozas & Barcala, 2020).

The pragmatic annotation scheme was implemented only in the interview subcorpus. Interviewees were selected according to key social variables including age, gender, and education level, to create a balanced corpus and facilitate variationist analyses.

Sociolinguistic interviews are often relatively formal, which may lead speakers to monitor their speech and reduce spontaneity — a limitation commonly described as the observer's paradox (Labov, 1972). To mitigate this, ESLORA uses "free interviews" (Fernández Sanmartín, 2018, 2026) designed to approximate casual conversation. Rather than following a strict question-and-answer structure, interviewees can elaborate on their responses. Interviewers also seek to minimise the asymmetry of their role by contributing to the interaction with comments in a more dialogic way.

3 Methodology: pragmatic annotation and implementation

3.1 Scheme design

The annotation scheme was designed to enable the systematic identification and retrieval of speech

[1] https://eslora.usc.es/search

acts in the ESLORA corpus through its query tool. It was therefore necessary to adapt the scheme to the structure of the interactions and to corpus-specific constraints, such as transcript segmentation and the interactional characteristics of sociolinguistic interviews. As in other spoken corpora involving pragmatic annotation, the target categories were selected based on the type of data and the speech acts typically occurring in the corpus interactions. The selected categories include greetings, directive acts, offers, direct reported speech, and questions (see Appendix A.1), all of which can be identified and delimited with sufficient consistency for annotation purposes.

Given the nature of the sociolinguistic interviews, which, despite efforts to reduce formality and approximate casual conversation, still exhibit characteristics of traditional interviews, it was not possible to directly adapt and apply an existing annotation scheme developed for other corpora. These include, for instance, corpora of casual conversations (e.g., the CORDIAL corpus) or role-play interactions (e.g., COREMAH).

The speech acts targeted were organised into adjacency pairs, i.e., sequentially related pairs of turns produced by two different speakers, in which the first turn elicits the second (Schegloff & Sacks, 1973; Schiffrin, 1986). For example, a greeting or a question, as first-pair parts, elicits a return greeting or an answer, respectively.

Speech acts were manually annotated using an XML-based encoding system described in section 3.2. Three researchers were involved in the annotation and review process following an annotation guide[2]. Any discrepancies were resolved through discussion, with consensus-based decisions adopted to ensure consistency in the application of the annotation scheme.

To assess the effectiveness of the encoding system, a series of evaluations of annotation consistency and information retrieval was carried out. In addition, the query system was tested by retrieving all annotated speech acts, confirming that they are consistently recognised as structured units.

3.2 XML encoding and annotation workflow

The speech-act taxonomy and annotation scheme were designed by the author and implemented in XML using a dedicated annotation environment developed by NLPgo Technologies S.L. In the case of questions, due to their high frequency in a corpus of sociolinguistic interviews, a further subcategorisation was included. Specifically, the type of question had to be selected, such as yes/no questions, wh-questions, and echo questions, among others (see Appendix A.1).

The annotation scheme relied on explicit opening and closing XML tags, which were manually inserted at the beginning and the end of each speech act within the first-pair-part turn of the adjacency pair (e.g., a question or an offer) (see Appendix A.2). The same procedure was applied to the second-pair-part turn (e.g., the answer to the question or to the offer). In cases where no response was produced (e.g., unanswered questions), a "no response" placeholder was introduced to preserve the structural integrity of the annotation scheme[3]. Finally, when multiple speech acts co-occurred within a single turn (e.g., two questions or a combination of a directive and a question), each act was assigned a numerical identifier, which was consistently implemented across both first-pair-part and second-pair-part turns to avoid overlaps between annotations.

3.3 Query options

The ESLORA corpus query system allows for two main types of searches that can be further constrained using sociolinguistic filters, such as the speakers' age, gender, education level, and role (interviewer, interviewee, or audience member).

The first search type (see Appendix B.1) retrieves speech acts and individual adjacency-pair components, i.e., first-pair parts or responses. Additionally, it allows for the simultaneous retrieval of multiple speech act types within a single query.

The second type of search (see Appendix B.2) enables more complex query configurations, in which speech acts can be combined with linguistic markers such as lexical or morphosyntactic information. This includes searches based on specific lemmas, word forms, or grammatical patterns. For instance, researchers can search for directive acts or offers combined with imperative verb forms, or direct reported speech combined with specific verb lemmas, allowing them to examine different persons and tenses. Such

[2] https://eslora.usc.es/guide_pragmatic_annotation

[3] Note that direct reported speech does not conform to an adjacency-pair structure. It was annotated during the original corpus transcription process and later incorporated into the set of speech acts retrievable through the query system.

functionality enables the systematic association between speech acts and their linguistic encoding.

Query results are displayed as concordance lists and can be exported in standard formats (e.g., CSV or Excel). Each result is also linked to its corresponding audio segment.

4 Scope and limitations

The main contribution of this pragmatic annotation scheme is the incorporation of a pragmatic layer into an already morphosyntactically annotated spoken corpus. Unlike traditional POS-based annotation systems, which primarily encode morphosyntactic or lexical information, this system enables direct access to speech act categories without requiring the manual inspection of transcripts. Furthermore, it is also possible to retrieve these acts alongside lexical, morphosyntactic, and sociolinguistic variables.

A second contribution lies in the adjacency-pair-based encoding strategy. This design not only identifies speech act categories but also retrieves their corresponding second-pair parts, i.e., responses within interactional sequences. In this respect, the scheme differs from other speech-act annotated corpora (e.g., C-ORAL-ROM, C-Or-DiAL, or CLANES). In those corpora, pragmatic information is not consistently structured to allow the retrieval of second-pair parts as complete interactional units.

The proposed pragmatic annotation system has significant applications in both research and in L2 teaching and learning. It facilitates the systematic study of speech acts in real interactional settings, enabling the analysis of their co-occurrence and variation according to sociolinguistic and discourse patterns. Moreover, the possibility of combining pragmatic information with morphosyntactic information allows for the study of the interface between grammar and pragmatics in spoken data. In L2 courses, the retrieved data can be used through data-driven learning (DDL) (Leech, 1997) to illustrate pragmatic functions in authentic communicative situations. In this way, it can complement classroom materials (see, e.g., Sampedro Mella, 2021b; Ballestero de Celis & Sampedro Mella, 2021), thus contributing to the development of learners' pragmatic competence (Sampedro Mella & Rodríguez Domínguez, 2026).

As for limitations, the scope of the annotation scheme is currently limited to the sociolinguistic interviews of the ESLORA corpus. The speech act taxonomy focuses on a restricted set of categories and does not incorporate prosodic information. Additionally, since the data consist exclusively of audio recordings and their transcriptions, the annotation does not include gestural or other types of multimodal information.

The annotation process also revealed several challenges. As is common in pragmatic annotation, the delimitation of speech acts was not always clear. In particular, difficulties arose in cases involving what Sacks, Schegloff, and Jefferson (1974: 711) describe as "expanded sequences", where speech acts extended beyond adjacency pairs. This phenomenon was particularly frequent in the case of offers, as speakers often repeated or reformulated them, and interlocutors produced multiple refusals (see Appendix C.1). Similar difficulties were encountered in question-answer and directive act-response sequences. In these cases, a question or directive act was sometimes responded to with another question, which was subsequently answered, making it difficult to determine where the response to the initial speech act began (see Appendix C.2, and Sampedro Mella, 2026 for a more detailed discussion and examples).

Finally, although the annotation process involved a two-stage review procedure to ensure consistency, some degree of variability may remain. When pragmatic interpretation is highly context-dependent, especially when it comes to the identification or classification of some speech acts, annotators may arrive at different interpretations despite following the annotation guide. For instance, imperatives have generally been classified as directive speech acts (e.g., *cuéntame* 'tell me'). However, they can also function as offers, as in invitations to come to the interviewees' home or to take a seat (e.g., *siéntate* 'sit down'). Such cases illustrate the inherent challenges of assigning categories to context-dependent pragmatic phenomena, highlighting the complexity of pragmatic annotation.

5 Conclusions and future work

This paper has presented the design and implementation of a pragmatic annotation scheme for speech acts in the ESLORA subcorpus of sociolinguistic interviews. The system was integrated into a corpus that already included POS annotation, enabling multi-layered analysis. The annotation scheme contributes to corpus pragmatics by making speech acts systematically retrievable as pragmatic units, both independently and in combination with (socio)linguistic variables.

In addition, the adjacency-pair-based encoding enables the retrieval of responses (second-pair-part turns), which broadens the possibilities of the corpus query system.

Future work should focus on expanding the taxonomy to other speech acts, refining the annotation guide for ambiguous cases, and incorporating prosodic information. Additionally, future research may explore adapting the scheme to other corpora and developing semi-automatic procedures for pragmatic annotation, including the potential application of AI-assisted approaches.

Building on this line of research, ongoing work is exploring the pragmatic annotation of other phenomena, such as address forms (i.e. *tú/usted* 'you'), in the Corpus of Online Forum Replies (COFRE), a multilingual corpus of forum data from TripAdvisor. To this end, an automatic system based on spaCy has been developed to identify and classify address forms through pronoun- and verb-based detection. Future developments will focus on the validation and refinement of these automatic procedures, as well as on their integration into broader annotation schemes for multilingual corpus pragmatics (see Sampedro Mella et al., 2026 for further information).

Together, these developments contribute to addressing a broader challenge in corpus pragmatics: the limited availability of corpora enriched with pragmatic annotation. By making pragmatic phenomena systematically searchable and analysable, such resources open new possibilities for corpus-based pragmatic research and data-driven approaches to L2 teaching and learning.

6 Acknowledgements

The research presented in this paper represents an earlier stage of a broader research trajectory that is currently developed through the Marie Skłodowska-Curie Action (MSCA) *Personal and non-personal forms of address in three contiguous Romance languages: sociolinguistic approach and gender implications (DE-PERSONAlisation).* (Grant code 101149488).

References

Svenja Adolphs and Yaoyao Chen. 2021. Corpus Pragmatics. In Michael Haugh Dániel Z. Kádár and Marina Terkourafi (eds.), *Sociopragmatics*, pages 639–662. Cambridge University Press.

Karin Aijmer. 2026. Corpus pragmatics. In Jesús Romero Trillo (ed.), *The Cambridge Handbook of Language in* Context, pages 289–313. Cambridge University Press.

Karin Aijmer and Cristoph Rühlemann. 2015. *Corpus Pragmatics: A Handbook.* Cambridge University Press.

Carmen Ballestero de Celis and María Sampedro Mella. 2021. La didáctica del español como lengua extranjera: un acercamiento desde la lingüística de corpus. In Carmen Ballestero de Celis and María Sampedro Mella (eds.), *Aportes del CAES a la enseñanza del español como lengua extranjera: estudios aplicados de gramática y pragmática*. Special issue of *Verba*, 81, pages 9–20.

Harry Bunt, Volha Petukhova, Andrei Malchanau, Alex Fang and Kars Wijnhoven. 2016. The DialogBank. In *Proceedings 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3151–3158.

Harry Bunt, Volha Petukhova, Emer Gilmartin, Catherine Pelachaud, Alex Fang, Simon Keizer, and Laurent Prevot. 2020. The ISO Standard for Dialogue Act Annotation, Second Edition. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 549–558.

CLANES Corpus - *Controlled LANguage English-Spanish.* ACTRESS Group. https://actres.unileon.es/wp/es/corpus-comparables/

Emanuela Cresti and Massimo Moneglia. 2005. *C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages.* John Benjamins.

ESLORA Corpus - *Corpus para el estudio del español oral*, v. 2.5. https://eslora.usc.es/

Alba Fernández Sanmartín. 2018. La entrevista libre como método para evitar la paradoja del observador. *CHIMERA*, 5(2): 141–196.

Alba Fernández Sanmartín. 2026. La entrevista sociolingüística en el corpus ESLORA: análisis pragmático-discursivo de la interacción entrevistador–entrevistado. *Variación. Revista de variación y cambio lingüístico*, 3: 204–228.

Andreas Jucker. 2018. Corpus pragmatics. In Andreas Jucker, Klaus Schneider and Wolfram Bublitz (eds.), *Methods in Pragmatics*, pages 455–466. De Gruyter.

Dan Jurafsky, Liz Shriberg, and Debra Biasca. 1997. *Switchboard SWBD-DAMSL Shallow Discourse-Function Annotation: Coders Manual*. Institute of Cognitive Science Technical Report. University of Colorado.

John M. Kirk. 2013. Beyond the Structural Levels of Language: An Introduction to the SPICE-Ireland Corpus and its Uses'. In Janet Cruickshank and Robert McColl Millar (eds.), *After the Storm: Papers from the Forum for Research on the Languages of Scotland and Ulster triennial meeting,* pages 207–232.

William Labov. 1972. Some Principles of Linguistic Methodology. *Language in Society*, 1(1): 97–120.

Geoffrey Leech. 1997. Teaching and Language Corpora: A Convergence. In Anne Wichmann and Steven Fligelstone (eds.): *Teaching and Language Corpora*, pages 1–23. Routledge.

Geoffrey Leech and Martin Weisser. 2013. *The SPAADIA Annotation Scheme.* https://martinweisser.org/publications/SPAADIA_Annotation_Scheme.pdf

Alba Milà-Garcia. 2018. Pragmatic annotation for a multi-layered analysis of speech acts: a methodological proposal. *Corpus Pragmatics*, 2(3): 265–287.

Carlota Nicolás Martínez. 2012. C-Or-DiAL (*Corpus Oral Didáctico Anotado Lingüísticamente*). Liceus.

Anne O'Keeffe 2018. Corpus-based function-to-form approaches. In Andreas H. Jucker, Klaus P. Schneider and Wolfram Bublitz (eds.), *Methods in Pragmatics (Handbooks of Pragmatics*), pages 587–618. De Gruyter.

Rosa Rabadán, Noelia Ramón and Hugo Sanjurjo-González. 2021. Pragmatic annotation of a domain-restricted English-Spanish comparable corpus. *Bergen Language and Linguistics Studies*, 11(1): 209–223.

Harvey Sacks, Emmanuel A. Schegloff and Gail Jefferson. 1974. A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language*, 50(4): 696–735

María Sampedro Mella. 2021a. DCT vs. corpus orales: reflexiones metodológicas sobre el estudio de los actos de habla. *Pragmalingüística*, 29: 337–395.

María Sampedro Mella. 2021b. *Estimado Sr*. vs. *Hola hotel*. El análisis contrastivo de la interlengua para la enseñanza de la variación pragmático-discursiva. In Vicente Marcet Rodríguez, Vanesa Álvarez Rosa and Manuel Nevot (eds.), *La variación en español y su enseñanza. Reflexiones y propuestas didácticas*, pages 134–150. Universidad de Salamanca.

María Sampedro Mella. 2026. Dando voz a la pragmática: creación de un sistema de anotación de actos de habla en el corpus ESLORA. *Variación. Revista de variación y cambio lingüístico*, 3: 39–67.

María Sampedro Mella, Louise-Amélie Cougnon and Barbara De Cock. 2026. COFRE – Corpus of Online Forum Replies. A Trilingual TripAdvisor Corpus for the Study of Address Forms and Beyond. *Variación. Revista de variación y cambio lingüístico*, 3: 132–156.

María Sampedro Mella and Ana Rodríguez Domínguez. 2026. Las funciones comunicativas en ELE: de los libros de texto a los corpus con anotación pragmática. In I. Fernández López (ed.), *Competencia plurilingüe y pluricultural: enfoques desde la Didáctica de la Lengua y la Literatura*, pages 385–403. Peter Lang.

Emanuel A. Schegloff and Harvey Sacks. 1973. Opening up Closings. *Semiotica*, 8(4): 289–327.

Deborah Schiffrin. 1986. Functions of and in discourse. *Journal of Pragmatics*, 10(1): 41–66.

Marta Vacas Matos. 2017. *COREMAH – Corpus Español Multimodal de Actos de Habla.* https://coremah-1a37f.firebaseapp.com/

Victoria Vázquez Rozas and Mario Barcala. 2020. Computational tools and spoken corpora design: an ongoing dialogue. *Caplletra*, 69: 221–240.

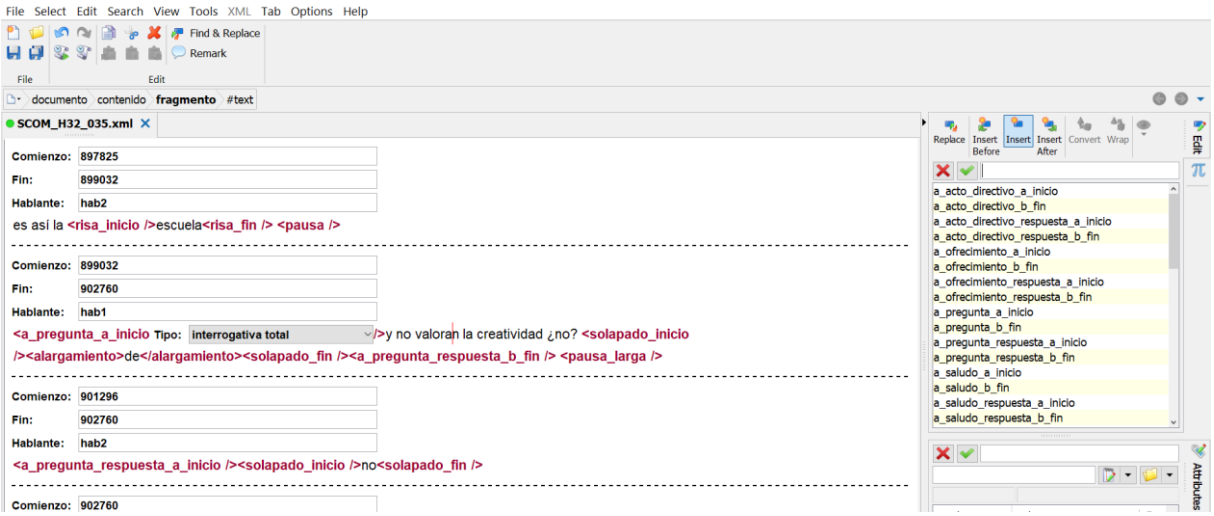
Martin Weisser. 2018. *How to Do Corpus Pragmatics on Pragmatically Annotated Data*. John Benjamins.

Appendix A. Speech acts in the ESLORA corpus

A.1 Speech acts annotated in the ESLORA corpus

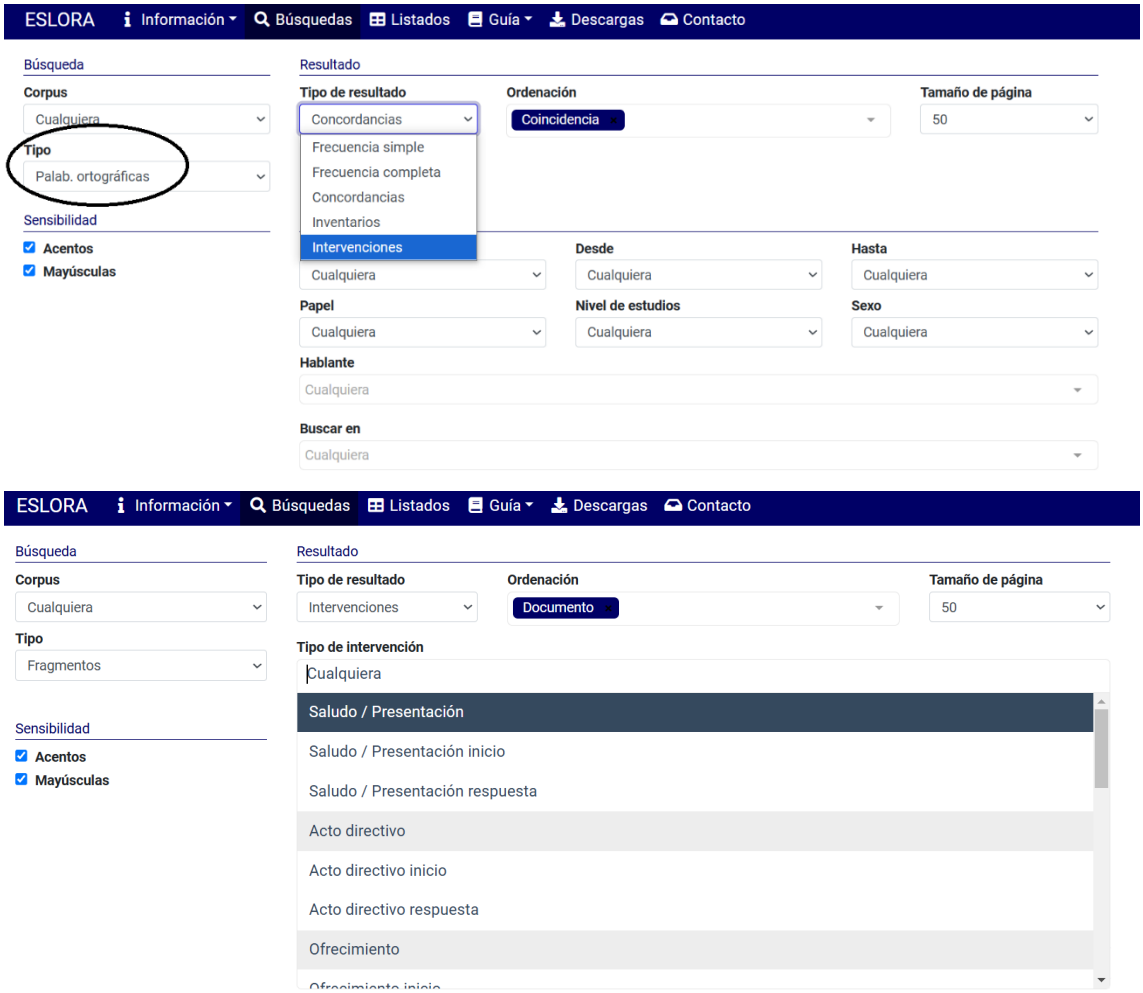
Speech act (first pair-part)	Description/examples	Responses (second pair-part)
Greetings / introductions	Greetings and introductions, occurring either separately or in combination.	Greetings or no response.
Directive acts	Requests, petitions, orders, and advice.	Responses to directive acts (acceptance, refusal, thanking) or no response.
Offer–Invitation	Proposals, offers, and invitations.	Acceptance, refusal, thanking, etc., or no response.
Direct reported speech	Quotations.	Ø
Questions		
Yes/no questions.	E.g., *¿viviste siempre aquí desde pequeño?* 'Have you lived here since you were a child?'	Response or no response.
Wh-questions.	E.g., *¿a qué colegio fuiste?* 'Which school did you go to?'	Response or no response.
Alternative questions.	E.g., *¿tienes buen recuerdo de los estudios o no?* 'Do you have a good memory of the studies or not?'	Response or no response.
Echo questions.	Questions that repeat part of a previous utterance: *–es complicado ¡eh! – ¿complicado?* '–It's complicated, huh! – Complicated?'	Response or no response.
Incomplete questions.	Question left unfinished that prompt the interlocutor to respond: *¿hay algo más que quieras contarme de ti de?* 'Is there anything more you want to tell me about yourself, about…?'	Response or no response.
Tag questions.	Statements followed by a short question tag: *pero eso es muy peligroso ¿no?* 'But that's very dangerous, isn't it?'	Response or no response.

A2. Example of speech act annotation (yes-no question) in XML. The speech act tags are displayed on the right.

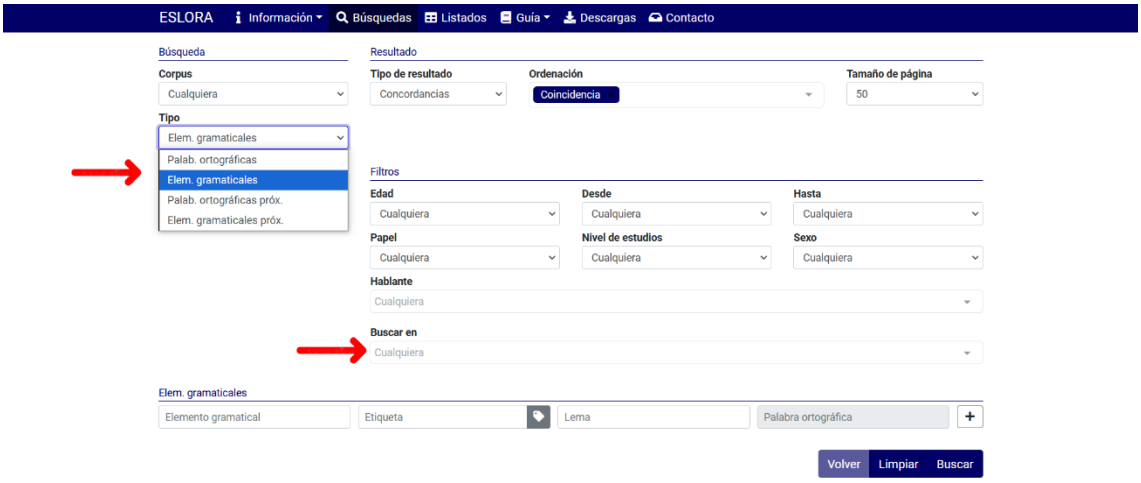


Appendix B. Corpus query tool for speech acts[4]

B1. Search for speech acts. Steps 1 and 2:



B2. Search for speech acts in combination with linguistics units



[4] Source: Sampedro Mella (2026).

Appendix C. Examples of speech acts

C1. Offer with expanded sequences

Interviewee (B): *eeh (1) ¿quieres algo de tomar? /*
'uh (1) would you like something to drink? /'
Interviewer (A): no no no no /
B: *¿de verdad que no? (2) ¿una Coca-Cola? //*
'are you sure? (2) a Coke?' //
A: *no de verdad de verdad no gracias gracias estoy / estoy bien ¡eh! /*
'no really really no thank you thank you I'm / I'm fine, really!' /
B: *(3) tengo agua de botella bueno /*
'I've got bottled water anyway /'
A: <risa/> muy bien /
'<laughs/> very well /' (SCOM_H31_046)

Main offer: *would you like something to drink?*
- Expansion 1
  - Response: *no no no no*
  - Re-offer: *are you sure?*
- Expansion 2
  - Specific offer: *a Coke?*
  - Response: *no really no thanks I'm / I'm fine*
- Expansion 3
  - Alternative offer: *I've got bottled water*
  - Response: <laughs> *very well*

C2. Directive acts/questions with expanded sequences

Interviewer (A): *bueno/ me puedes contar por ejemplo qué haces un día normal en tu vida// un día cualquiera//* [directive act]
'well/ can you tell me for example what you do on a normal day in your life// just any day'
Interviewee (B): *¿cualquiera?* [response to the directive act + yes–no question]
'any day?'
A: sí [response to the yes–no question]
'*yes*'
B: *¿de ahora mismo?* [yes–no question]
'do you mean these days?'
A: *sí / de ahora mismo //* [response to the yes–no question]
'yes / these days'
B: *me levanto / voy a la autoescuela / después voy por ahí con los amigos* […]
'I get up / I go to driving school / then I go out with my friends […]' (SCOM_H11_047)

- A_1: directive act (A: can you tell me […] *just any day*)
  - R_4: Complex response
    - [R_1 + B_1: *any day?*
      - R_2: *yes*]
    - [B_2: *do you mean these days?*
      - R_3: *yes / these days*]
    - Nucleus: *I get up / I go to driving school* / […][5]

[5] Both diagrams were created using ChatGPT.

Evaluating the Linguistic Properties of AI-Generated Bulgarian News Texts

Tsvetelina Stefanova
Angel Kanchev University of Ruse
8 Studentska Str.,
7004 Ruse, Bulgaria
tsstefanova@uni-ruse.bg

Petya Osenova
Sofia University St. Kliment Ohridski
15 Tsar Osvoboditel Blvd.,
1504 Sofia, Bulgaria
osenova@uni-sofia.bg

Abstract

The rapid development of large language models (LLMs) has enabled the automatic generation of news articles that increasingly resemble texts written by human journalists. While a growing body of research has examined the linguistic properties of AI-generated texts in widely studied languages such as English, comparatively little work has focused on smaller languages such as Bulgarian. This raises important questions about how accurately contemporary language models reproduce the linguistic conventions of Bulgarian journalistic discourse. This paper presents a comparative evaluation of Bulgarian news texts generated by several large language models and authentic news articles written by human authors. The goal of the study is to assess the linguistic properties of AI-generated news and to identify systematic similarities and differences between machine-generated and human-written texts. To address this task, the paper applies a multi-level analytical framework combining semantic, stylistic and morphosyntactic analysis. The dataset consists of original Bulgarian news articles and texts generated by language models using prompts designed to produce news content on comparable topics. The analysis includes measures of semantic similarity between original and generated texts, as well as quantitative indicators describing lexical richness, stylistic patterns and structural characteristics of the texts. Morphosyntactic features are examined using automatic linguistic annotation tools for South Slavic languages, and statistical methods are applied to identify patterns across models and text types. The study contributes to ongoing research on the evaluation of AI-generated texts and provides a framework for analyzing the linguistic properties of machine-generated content in relatively low-resource languages such as Bulgarian.

Keywords: large language models; AI-generated text; Bulgarian language; news text generation; stylometry; natural language processing

1 Introduction

The rapid development of large language models (LLMs) in recent years has led to significant changes in the way media content is created and consumed. Contemporary artificial intelligence (AI) tools are capable of generating texts that are increasingly difficult to distinguish from those created by human journalists. In the context of the media industry, these capabilities open new horizons for automated content creation, but simultaneously pose challenges to traditional norms of authenticity, ethics, and copyright (Baptista et al., 2025; Muñoz-Ortiz et al., 2024).

While a growing body of literature examines the linguistic, stylistic, and morphological features of AI-generated texts in widely resourced languages such as English (Opara, 2024; Georgiou, 2024), there is a notable gap regarding smaller languages. For Bulgarian, systematic evaluations of AI-generated journalistic discourse remain scarce, though foundational work has begun exploring related areas such as the detection of textual deepfakes in social media (Temnikova et al., 2023). This technological evolution raises critical questions regarding how well contemporary language models adapt to the specific conventions, syntax, and lexical richness of Bulgarian news texts.

The objective of this study is to identify specific linguistic features that characterize AI-generated content and to evaluate the quality of automatically created texts in the context of media communication. To achieve this, the current work focuses on conducting a comparative linguistic analysis of media texts generated by six leading language models: BgGPT, ChatGPT, Gemini, Gemma, DeepSeek, and Qwen, contrasted with authentic publicistic texts written by journalists in Bulgarian.

The methodology combines quantitative and qualitative methods across various linguistic levels — from identifying characteristic morphological features using advanced NLP pipelines adapted for South Slavic languages (Ljubešić et al., 2024), to comparing lexical diversity, syntactic structures, and semantic coherence.

The paper is structured as follows: Section 2 provides a brief review of related work and the theoretical context. Section 3 outlines the methodology of the study, including dataset preparation and analytical tools. Section 4 presents the linguistic analysis of the AI-generated media texts and their comparison with the original texts. Finally, Section 5 offers concluding remarks and directions for future research.

2 Theoretical Background and Related Work

2.1 Linguistic Features of Media Discourse

The analysis of journalistic texts requires an understanding of the specific stylistic and functional characteristics of media discourse. As a sub-style of publicistic writing, news texts are driven by two primary functions: informative and persuasive. To achieve these goals, journalistic writing relies on a balance between standardization (the use of clichés, formulaic expressions, and predictable syntactic structures) and expression (the use of emotionally charged lexis to engage the reader).

From a morphosyntactic perspective, Bulgarian news discourse is characterized by a high frequency of verbal nouns, passive constructions, and complex sentences with subordinate clauses that condense information (Krejčová, 2014; Academic Grammar, 1983). A critical question in contemporary computational linguistics is whether LLMs can organically replicate this delicate balance, or if their probabilistic nature leans too heavily into either rigid standardization or unnatural syntactic complexity.

2.2 Distinguishing AI-Generated from Human Texts

The proliferation of LLMs has spurred significant research into distinguishing AI-generated content from human-written texts. Early studies focused primarily on English, employing various stylometric and linguistic features. For instance, (Opara, 2024) introduced StyloAI, which utilizes 31 stylometric features — including lexical diversity, sentence length variability, and punctuation frequency — coupled with machine learning classifiers to differentiate text origins with high accuracy.

In the specific context of media texts, comparative studies have highlighted distinct linguistic patterns. (Muñoz-Ortiz et al., 2024) analyzed news articles generated by models like LLaMa and Mistral, finding that authentic human texts exhibit greater sentence length variance, richer vocabulary, and more efficient syntactic dependencies. Conversely, AI-generated news tends to feature highly uniform sentence lengths, narrower lexical diversity, and a higher frequency of auxiliary verbs and pronouns. Similarly, (Georgiou, 2024) noted that AI models often produce texts with different morphological distributions (e.g., an overreliance on nouns and adjectives) and a less diverse pool of functional vocabulary.

Interestingly, from a qualitative reception perspective, (Baptista et al., 2025) demonstrated that journalism students frequently perceive AI-generated news as equal or even superior in quality to human-written articles, underscoring the sophisticated mimicking capabilities of modern LLMs. Recent comprehensive evaluations further emphasize that as models increase in size and reasoning capacity, traditional stylometric and perplexity-based detection methods experience a sharp decline in accuracy, necessitating the deeper, multi-level linguistic approaches proposed in this study (Sadasivan et al., 2023).

2.3 The Gap in Bulgarian Language Research

Despite these advancements, research targeting the Bulgarian language remains highly limited. A pioneering study by (Temnikova et al., 2023) investigated the detection of textual deepfakes in Bulgarian social media. Their findings revealed that simply translating English datasets into Bulgarian yields poor detection models due to the profound loss of stylistic nuances and morphological markers. Instead, training custom models (such as BERT-Deepfake-BG) directly on native Bulgarian corpora proved highly effective, achieving up to 97% accuracy.

The computational processing of the Bulgarian language has a rich history, transitioning from early rule-based systems to modern deep learning architectures. Foundational projects such as the CLaRK system (Simov et al., 2001) and the

BulTreeBank provided the essential corpora that enabled the initial formalization of Bulgarian morphosyntax. Recently, significant efforts have been made to construct high-quality, large-scale datasets specifically designed for training and fine-tuning contemporary LLMs, such as the IfGPT dataset (Koeva et al., 2025). Furthermore, contemporary NLP research in Bulgaria is actively applying large language models to complex downstream tasks, including event extraction (Simov et al., 2025) and word sense disambiguation (Paev et al., 2025). However, despite these vital advancements in dataset curation and task-specific LLM applications, comprehensive, multi-level linguistic evaluations of how well generative models replicate the specific morphosyntactic and stylistic nuances of Bulgarian journalistic discourse remain scarce.

Building upon this foundation, the present study addresses a critical gap in the literature by providing a comprehensive, multi-level linguistic evaluation of Bulgarian news texts generated by a diverse set of state-of-the-art LLMs, examining how well these models capture the specific stylistic and morphosyntactic norms of Bulgarian journalism.

3 Methodology and Data

To conduct a comprehensive linguistic analysis of AI-generated media texts, this study employs a multi-level methodological framework combining quantitative and qualitative approaches. The analysis focuses on lexical diversity, stylistic features, syntactic structures, and semantic coherence.

3.1 Dataset Preparation

The study utilizes a publicly available corpus of Bulgarian news articles spanning a ten-year period (2014–2023) (Fauzi, 2023). After data cleaning to remove duplicate and empty records, the corpus was reduced to 15,230 articles. From this curated dataset, a sample of 200 news articles was randomly selected to serve as the reference human-written texts.

To ensure a representative cross-section of contemporary Bulgarian journalistic discourse, the content of the selected texts was manually evaluated and categorized by topic. As detailed in Table 1, the dataset predominantly covers economics and politics, reflecting standard media distribution.

Topic	Count	Topic	Count
Economics	50	Sport	7
Politics	45	Education	6
Culture	31	Healthcare	4
Criminal	26	Disasters	4
Other	10	Weather	4
Tourism	9	Advertising	4

Table 1: Distribution of the 200 selected original news articles by topic.

3.2 Language Models and Prompt Engineering

To ensure a robust comparative framework, this study selected six state-of-the-art large language models, balancing models with explicit Bulgarian fine-tuning against generalized multilingual giants:

- **BgGPT** (insait-gemma-2-27b): Fine-tuned specifically for the Bulgarian cultural and linguistic context (INSAIT, 2024). Note that while this study evaluates the highly performant 27B model based on Gemma 2, the ecosystem is evolving rapidly, with the recent release of BgGPT 3.0 (based on Gemma 3) introducing multimodal capabilities and extended context windows (INSAIT, 2026).
- **ChatGPT** (GPT-4o): Evaluated for its widespread commercial adoption and general-purpose capabilities.
- **Gemini** (2.5 Flash): Selected for its advanced reasoning and long-context multimodal capabilities (Gemini Team, Google, 2025).
- **Gemma** (3 27B): Evaluated to observe the performance of open-weights models (Gemma Team et al., 2025).
- **DeepSeek** (V3 0324): Included to evaluate its highly efficient Mixture-of-Experts (MoE) architecture and auxiliary-loss-free load balancing (DeepSeek-AI, 2024).
- **Qwen** (3 235B A22B): Selected to represent advanced multilingual MoE models with high reasoning and instruction-following capabilities (Qwen Team, 2025).

A strict zero-shot prompting strategy was employed to evaluate the models' baseline linguistic generation capabilities without the influence of few-shot tuning. Each of the 200 original articles was fed into the models via API or web interface with a standardized, neutral prompt: *„Генерирай новина, подобна на тази:“* (Generate a news article

similar to this one:"). No instructions regarding tone, length, or structural formatting were provided. This methodology isolates the models' intrinsic stylistic biases and their default interpretations of Bulgarian journalistic norms. The process yielded a total generated corpus of 1,200 AI-authored texts.

3.3 Analytical Tools and Evaluation Metrics

The generated texts were evaluated against the original human-written articles using a suite of advanced natural language processing tools and statistical metrics:

Morphosyntactic Annotation Pipeline: The core linguistic processing was executed using the CLASSLA-Stanza pipeline (Ljubešić et al., 2024). Unlike general-purpose NLP tools, CLASSLA is specifically pre-trained on South Slavic datasets, ensuring high accuracy for Bulgarian morphological disambiguation. The pipeline performed tokenization, multi-word expression grouping, lemmatization, and Part-of-Speech (POS) tagging based on the MULTEXT-East v6 morphosyntactic specifications. Furthermore, universal dependency parsing was applied to analyze the syntactic relationships within the complex sentences typical of journalistic discourse.

Semantic and Stylometric Metrics: To measure how well the models retained the informational core of the original articles, we applied both lexical and dense semantic evaluations. First, we utilized Term Frequency-Inverse Document Frequency (TF-IDF), fitted across the combined vocabulary space of all corpora. The Inverse Document Frequency was computed with smoothing to prevent zero divisions:

$$\text{idf}(t) = \ln\left(\frac{1+n}{1+\text{df}(t)}\right) + 1$$

where n is the total number of documents and $\text{df}(t)$ is the number of documents containing term t. The resulting vectors were strictly L2-normalized. To balance this shallow, purely lexical statistic, we simultaneously applied Sentence-BERT (SBERT) (Reimers and Gurevych, 2019). Utilizing a Siamese neural network architecture with a pre-trained transformer, SBERT captures deep semantic relationships and contextual embeddings, allowing us to measure the preservation of underlying meaning via cosine similarity even when models hallucinate or aggressively change the vocabulary. For stylometry, Burrows' Delta (Evert et al., 2015) was utilized to measure stylistic distance. Originally developed for authorship attribution, Delta calculates the z-scores of the most frequent words to determine how closely a generated text mimics the baseline stylistic fingerprint of the human journalistic corpus.

Statistical Significance Testing: To ensure that observed differences between human and AI writing were not merely artifacts of sampling variation, Student's T-test and the Wilcoxon Signed-Rank Test were employed across all morphosyntactic variables.

4 Results and Discussion

The evaluation was conducted across five primary dimensions: semantic similarity, stylistic characteristics, n-gram distribution, morphosyntactic features, and qualitative error analysis.

4.1 Semantic and Thematic Similarity

To determine how accurately the models preserved the informational core of the original prompts, we utilized TF-IDF and SBERT.

As visualized in Figure 1, BgGPT demonstrated the highest average semantic retention (83.3% via SBERT and 73.6% via TF-IDF). This indicates that the model fine-tuned for Bulgarian is substantially better at maintaining factual integrity. Conversely, models like Gemma and ChatGPT showed much lower overlap, frequently diverging from the source material.

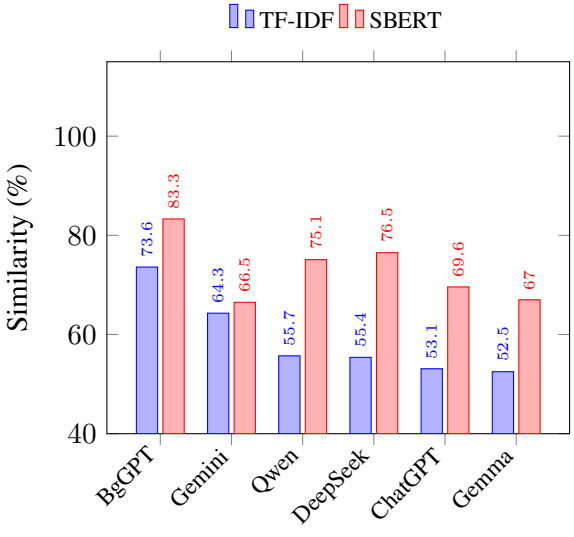


Figure 1: Average Semantic Similarity across models.

A supplementary K-Means clustering analysis confirmed these findings. When mapped into 17 thematic clusters based on n-gram frequencies, the texts generated by BgGPT remained in the exact same thematic cluster as their human-written source prompts in 196 out of 200 cases (shifting the cluster of only 4 texts). In contrast, Gemma and ChatGPT altered the thematic classification from the original text in 25 and 24 instances, respectively.

4.2 Stylometric and Structural Analysis

Stylistic distance was calculated using Burrows' Delta (Figure 2). A lower coefficient indicates closer stylistic proximity to the human baseline. BgGPT proved to be the most stylistically similar to the human reference corpus.

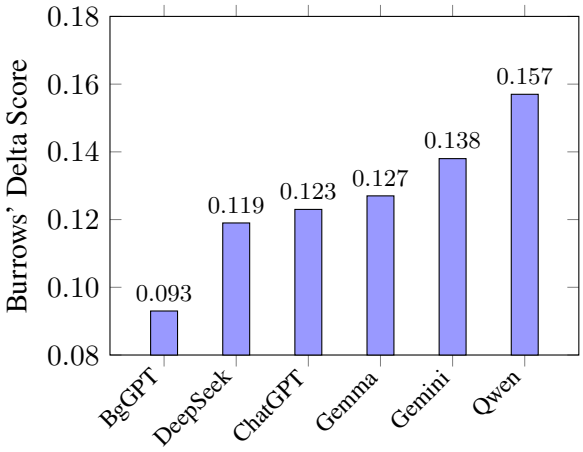


Figure 2: Stylistic distance from human texts using Burrows' Delta. Lower is closer to human style.

Structurally, human texts comprised 95.37% declarative, 2.63% interrogative, and 2.01% exclamatory sentences. BgGPT mirrored this closely (96.96% declarative). In contrast, Qwen and DeepSeek over-generated interrogative sentences (over 3%), while Gemma and Gemini produced a disproportionate amount of exclamatory sentences, making their outputs feel artificially dramatic rather than objective.

To evaluate the models' tendency to artificially expand text length, we analyzed the statistical distribution of corpus sizes (Table 2). Authentic human reporting exhibited significant length variance, averaging 364.12 words per document but ranging from short 7-word news flashes to extensive 1,882-word features (Standard Deviation = 318.32). In contrast, models like Qwen3 and Gemini 2.5 Flash demonstrated a clear tendency toward length inflation, generating maximums of 2,110 and 1,736 words, respectively, alongside much higher averages. Conversely, ChatGPT and DeepSeek V3 struggled to replicate the natural variance of human texts, severely compressing their outputs to a "safe" medium length (narrow standard deviations of 132.59 and 127.01, with maximum lengths strictly capped under 800 words). BgGPT achieved the most balanced approximation of the human baseline, avoiding excessive verbosity while maintaining a respectable variance.

Corpus	Avg Words	Std Words	Avg Sent.	Std Sent.
Human Baseline	364.12	318.32	16.38	15.75
BgGPT	327.95	163.85	15.06	9.81
ChatGPT	346.98	132.59	15.50	9.08
Gemini 2.5 Flash	433.44	276.94	18.50	14.47
Gemma 3 27B	293.02	134.55	13.90	8.72
DeepSeek V3	323.91	127.01	16.22	7.88
Qwen3 235B	475.74	245.41	20.88	22.01

Table 2: Distribution of document lengths (word and sentence counts) across the original human corpus and AI-generated models.

4.3 Lexical Richness and N-gram Distribution

To understand lexical preferences, we extracted the top n-grams. The analysis revealed that AI models exhibit a strong preference for specific formal structures that humans use more sparingly.

As detailed in Table 3, which presents the raw aggregate bigram frequencies across the 200 documents in each corpus, the construction *„да се“* (to be / to do) was uniformly the most frequent bigram across all AI models, utilized to string together complex verb predicates. Furthermore, phrases like *„по време“* (during) and *„за да“* (in order to) were significantly overrepresented in the AI corpora compared to the original human texts.

Bigram	Original Freq.	AI Avg. Freq.
да се (to be)	1,450	2,890
по време (during)	210	680
за да (in order to)	180	495
ще бъде (will be)	310	580
може да (can / may)	240	520

Table 3: Comparison of top structural bigrams between the original corpus and the AI-generated average.

Additionally, DeepSeek generated the highest number of unique lemmas (TTR: 0.662), while Gemini generated the least (TTR: 0.582). Human journalists utilized a lower proportion of content words (0.605) compared to the models (e.g., Qwen3 at 0.634), relying more heavily on

functional vocabulary to create organically cohesive narratives.

4.4 Part-of-Speech (POS) Distribution

A macro-level analysis of the Part-of-Speech distribution revealed that while language models produce grammatically correct Bulgarian, they achieve it using a significantly different morphological recipe than human authors.

Table 4 outlines the proportional distribution of the primary word classes. LLMs exhibit a pronounced tendency toward nominalization — relying heavily on nouns and adjectives to convey meaning. Qwen3 and Gemma 3 27B, for example, generated texts where nearly a third of all tokens were nouns. In contrast, human journalists utilized a higher proportion of verbs, prepositions, and conjunctions. This indicates that authentic Bulgarian journalistic writing is inherently more dynamic and action-oriented, driven by verbal predicates, whereas AI-generated texts are more static, descriptive, and reliant on heavily modified noun phrases.

Model	Noun	Verb	Adj	Pron	Prep	Conj
Human Baseline	26.9	11.0	11.6	5.4	18.4	4.2
BgGPT	28.5	10.6	12.4	4.5	18.6	4.6
ChatGPT	29.5	9.9	13.9	4.4	19.2	4.9
Gemini 2.5	28.7	9.6	14.7	4.6	18.4	5.0
Gemma 3	29.7	10.5	13.3	4.3	19.4	4.8
DeepSeek	28.9	10.5	13.9	4.3	18.0	4.3
Qwen3	30.6	9.2	14.8	3.6	19.8	3.7

Table 4: Average proportional distribution (%) of primary Part-of-Speech categories.

Specific Morphological Categories

Beyond basic POS distribution, the annotation pipeline revealed systematic biases in how LLMs handle specific Bulgarian morphological categories.

In Bulgarian, definiteness (*определеност*) is expressed via postpositive articles. Across all generated texts, models exhibited an overwhelming reliance on indefinite forms compared to definite forms (a ratio of approximately 3:1). While human journalists use indefinite forms to introduce new information, they frequently employ definite articles to refer back to established entities, creating cohesion. The models' over-reliance on indefinite forms contributes to a narrative style that feels disjointed, as if every entity is being introduced for the first time.

Furthermore, regarding verb aspect (*вид на глагола*), models like Qwen3 and Gemma over-generated perfective verbs (*свършен вид*) compared to the human baseline. Authentic news reporting often utilizes imperfective verbs (*несвършен вид*) to describe ongoing situations, background context, or habitual actions. The AI models' preference for perfective verbs results in an overly dynamic, action-heavy text that rushes through the narrative timeline, stripping the news report of its necessary descriptive background.

Syntactic dependency parsing showed that models like ChatGPT and Gemma frequently utilized subordinating conjunctions to string together overly complex sentences, whereas human authors preferred shorter independent clauses linked by coordinating conjunctions, optimizing for readability.

4.5 Syntactic Complexity and Dependency Parsing

While part-of-speech frequencies provide a macro-level view of vocabulary usage, the internal architecture of the generated sentences reveals deeper divergences. Utilizing the dependency parsing capabilities of the CLASSLA pipeline, which draws upon the structural frameworks established by resources like the BulTreeBank (Simov et al., 2002), but is trained on significantly larger datasets sourced from the Bulgarian web, we measured the average tree depth of the generated sentences.

Authentic journalistic writing in Bulgarian relies on a high degree of information condensation, often utilizing verbal nouns (*отглаголни съществителни*) and passive constructions to minimize clauses. The dependency trees for the human texts were relatively shallow but dense, averaging 4.2 nodes in depth.

In contrast, the AI models — particularly Qwen3 and Gemma — produced significantly deeper syntactic trees (averaging 5.8 nodes). This increased depth was driven by the models' over-reliance on explicit subordinate clauses introduced by relative pronouns (*който, която, което*) rather than utilizing participles or nominalizations. While grammatically flawless, this deep, highly subordinated syntax is characteristic of academic or explanatory discourse rather than the fast-paced, condensed nature of news reporting. BgGPT proved to be the exception, mirroring the shallower

dependency depth of the original texts and successfully utilizing participle constructions to condense information.

4.6 Qualitative Error Analysis and Hallucinations

Beyond the quantitative metrics extracted via the automated NLP pipeline, a manual qualitative review of the generated corpora exposed specific vulnerabilities in how generalized multilingual models handle facts and the strict stylistic conventions of Bulgarian journalism.

One of the most persistent issues in Natural Language Generation is "hallucination" — the generation of text that is grammatically fluent but factually incorrect or unfaithful to the source material, necessitating robust automated fact-checking mechanisms (Ji et al., 2023; Nakov et al., 2021). In our dataset, while the specialized BgGPT model demonstrated high factual fidelity, generalized models like Qwen3 frequently exhibited severe extrinsic hallucination. For example, when prompted with an article detailing the program of the "Apollonia" arts festival, the human text listed real performers and actual Bulgarian film premieres. BgGPT successfully retained these factual entities. In contrast, Qwen3 suffered from profound semantic drift, artificially inflating the article by fabricating non-existent performers (e.g., *„Боби Жежов и Off the edge band"*) and completely fictional movie premieres (e.g., *„Колелото на времето"*). This behavior highlights an algorithmic bias toward fulfilling the structural template of a comprehensive news report over factual grounding.

Furthermore, a critical linguistic divergence observed during manual review occurred in the treatment of the Bulgarian renarrative forms (*преизказни форми*). In Bulgarian publicistic texts, this is a mandatory grammatical category used to report unverified information. Authentic human texts systematically employ explicit renarrative verb forms to maintain necessary journalistic distancing (e.g., the renarrative perfect or pluperfect form „бил избягал"). Interestingly, our analysis revealed no significant difference in the overall use of these renarrative forms between human and AI-generated texts. However, the AI models demonstrated a significantly wider spread of the corresponding indicative forms — specifically the perfect tense (e.g., „е избягал") and the pluperfect tense (e.g., „беше избягал") — compared to human authors.

The AI models, however, demonstrated a severe qualitative deficit in activating this category. Generalized models almost exclusively defaulted to forms utilizing the present tense auxiliary verb (e.g., *„е избягал"*). As noted in theoretical Bulgarian grammar, such forms are morphologically ambiguous and can be classified as either perfect indicative or aorist conclusive depending on the specific theoretical framework and context. Regardless of the classification, these forms lack the explicit distancing provided by true renarrative markers. By presenting unverified allegations through these direct or deductive forms, the models strip the text of its required journalistic objectivity. BgGPT handled this constraint substantially better, successfully generating renarrative forms in appropriate contexts, reaffirming the necessity of language-specific pre-training.

5 Conclusion and Future Work

This study presented a comprehensive comparative evaluation of Bulgarian news texts generated by six leading large language models against authentic human-written articles. By employing a multi-level analytical framework encompassing semantic, stylometric, and morphosyntactic dimensions, we identified systematic differences that characterize machine-generated journalistic discourse in a relatively low-resource language.

Our findings indicate that while contemporary LLMs are highly capable of generating coherent Bulgarian text, their outputs often diverge from the stylistic and structural conventions of professional journalism. Notably, BgGPT — a model specifically adapted for the Bulgarian language and cultural context — demonstrated superior performance. It most accurately preserved the semantic core of the prompts, maintained thematic consistency, and closely mirrored the concise syntactic structure and style of human journalists.

In contrast, larger multilingual models such as Qwen3 and Gemini exhibited a tendency to artificially expand text length, over-generate interrogative and exclamatory sentences, and inflate lexical diversity. This resulted in a style that felt more descriptive or dramatic than objective reporting. Morphosyntactic analysis further revealed that AI models systematically differ from human writers in their grammatical choices. The generated texts relied more heavily on content

words (nouns and adjectives), agreeing modifiers, and specific pronoun categories, whereas human journalists utilized a higher proportion of functional words to create cohesive narratives.

As LLMs continue to evolve, distinguishing between human and machine-generated content will become increasingly challenging. The results of this study underscore the necessity for advanced, language-specific linguistic analyses to detect AI "fingerprints". Future research should focus on expanding the dataset to include diverse media genres and exploring targeted fine-tuning strategies. Additionally, while this study utilized a strict zero-shot strategy to isolate intrinsic stylistic biases, subsequent evaluations should compare these results against few-shot prompting to observe how models adapt to explicit in-context examples. Specifically, optimizing models to balance syntactic complexity, functional vocabulary usage, and the proper application of the renarrative forms will be crucial for generating highly authentic, genre-compliant texts in Bulgarian.

References

Academic Grammar. 1983. *Gramatika na savremenniya balgarski knizhoven ezik. Tom 2. Morfologiya [Grammar of the Contemporary Bulgarian Standard Language. Volume 2. Morphology]*. Izdatelstvo na BAN, Sofia. (in Bulgarian).

João Pedro Baptista, Rubén Rivas-de Roca, Anabela Gradim, and Concha Pérez-Curiel. 2025. Human-Made News vs AI-Generated News: A Comparison of Portuguese and Spanish Journalism Students' Evaluations. *Humanities and Social Sciences Communications*, 12(1):1–9.

DeepSeek-AI. 2024. DeepSeek-V3 Technical Report. *arXiv preprint arXiv:2412.19437*.

Stefan Evert, Thomas Proisl, Fotis Jannidis, Isabella Reger, Steffen Pielström, Christof Schöch, and Thorsten Vitt. 2015. Towards a Better Understanding of Burrows's Delta in Literary Authorship Attribution. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 79–88.

Adam Fauzi. 2023. Bulgarian Articles with Keywords. https://www.kaggle.com/datasets/auhide/bulgarian-articles-with-keywords/data.

Gemini Team, Google. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *Google DeepMind Technical Report*.

Gemma Team et al. 2025. Gemma 3 Technical Report. *arXiv preprint arXiv:2503.19786*.

George Georgiou. 2024. Differentiating Between Human-Written and AI-Generated Texts Using Linguistic Features Automatically Extracted from an Online Computational Tool. *arXiv preprint arXiv:2407.03646*.

INSAIT. 2024. INSAIT Releases State-of-the-Art Language Models for Bulgarian Setting a New Standard for Open National LLMs. https://models.bggpt.ai/blog/bggpt-2-release-en/.

INSAIT. 2026. Announcing BgGPT 3.0: State-of-the-Art Multimodal LLMs for Bulgarian. https://models.bggpt.ai/blog/bggpt-3-release-en/.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12):1–38.

Svetla Koeva, Ivelina Stoyanova, and Jordan Konstantinov Kralev. 2025. IfGPT: A Dataset in Bulgarian for Large Language Models. In *Proceedings of the First Workshop on Advancing NLP for Low-Resource Languages*, pages 65–75. INCOMA Ltd.

Elena Krejčová. 2014. *Příručka pro výuku bulharské stylistiky [Manual for Teaching Bulgarian Stylistics]*. Masarykova univerzita, Brno. (in Bulgarian and Czech).

Nikola Ljubešić, Luka Terčon, and Kaja Dobrovoljc. 2024. CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages. In *Proceedings of the Conference on Language Technologies & Digital Humanities (JTDH 2024)*, pages 235–258.

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2024. Contrasting Linguistic Patterns in Human and LLM-Generated News Text. *Artificial Intelligence Review*, 57(10):265.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated Fact-Checking for Assisting Human Fact-Checkers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4551–4558. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Chidimma Opara. 2024. StyloAI: Distinguishing AI-Generated Content with Stylometric Analysis. In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*, pages 105–114, Cham. Springer Nature Switzerland.

Nikolay Paev, Kiril Simov, and Petya Osenova. 2025. Word Sense Disambiguation with Large Language

Models: Casing Bulgarian. In *Proceedings of the 13th Global Wordnet Conference*, pages 171–178, Pavia, Italy. Global Wordnet Association.

Qwen Team. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*.

Nils Reimers and Iryna Gurevych. 2019. 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)*, pages 3982–3992.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can AI-Generated Text be Reliably Detected? *arXiv preprint arXiv:2303.11156*.

Kiril Simov, Petya Osenova, Milena Slavcheva, Sia Kolkovska, Elisaveta Balabanova, Dimitar Doikoff, Krassimira Ivanova, Alexander Simov, and Milen Kouylekov. 2002. Building a Linguistically Interpreted Corpus of Bulgarian: the BulTreeBank. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Spain. European Language Resources Association (ELRA).

Kiril Simov, Nikolay Paev, Petya Osenova, and Stefan Marinov. 2025. Bulgarian Event Extraction with LLMs. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 1163–1171, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Kiril Simov, Zdravko Peev, Milen Kouylekov, Alexander Simov, Marin Dimitrov, and Atanas Kiryakov. 2001. CLaRK-An XML-Based System for Corpora Development. In *Proceedings of the Corpus Linguistics 2001 conference*, volume 1, pages 553–560.

Irina Temnikova, Iva Marinova, Silvia Gargova, Ruslana Margova, and Ivan Koychev. 2023. Looking for Traces of Textual Deepfakes in Bulgarian on Social Media. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing (RANLP 2023)*, pages 1151–1161.

Towards a new Grammar Checker for Bulgarian

Svetla Koeva
Institute for Bulgarian Language,
Bulgarian Academy of Sciences
52 Shipchenski prohod Blvd. Bldg. 17, Sofia 1113, Bulgaria
svetla@dcl.bas.bg

Abstract

This paper presents the dataset **Massive Multitask Language Understanding of Bulgarian Grammar** (**BulMMLU:gr**), modelled on the **MMLU** (Massive Multitask Language Understanding) benchmark. Its aim is to test the ability of AI grammar-checking applications, as well as different LLMs, to detect grammatical errors in Bulgarian and thus provide a basis for analysing the need for a specialised LLM-based tool to detect such errors and suggest corrections. Experiments conducted with two grammar-checking applications and five LLMs reveal limitations that highlight the need for a new, dedicated Bulgarian grammar-checking system. BulMMLU:gr can be further extended and used both for evaluation and for fine-tuning LLMs for grammatical error detection and correction suggestions.

Keywords: grammatical error detection and correction suggestions; grammar-checking applications; Large Language Models

1 Introduction

Large Language Models (LLMs) can edit and transform texts by paraphrasing, shortening, and simplifying them, with varying accuracy depending on the model's purpose, size, and language support. In addition, several artificial intelligence-based applications offer tools for improving the grammatical quality of written text, including grammar error detection and correction suggestions.

The **aims of this study** are: (i) to examine the extent to which applications with integrated grammar checkers that support Bulgarian, and selected LLMs, can correctly identify specific types of spelling, grammatical, and punctuation errors;[1] and (ii) to analyse the current state of grammar checking for Bulgarian and to identify future directions for developing a new AI-based grammar service dedicated to Bulgarian.

For this purpose, a dataset called **Massive Multitask Language Understanding of Bulgarian Grammar** (**BulMMLU:gr**), based on the **Massive Multitask Language Understanding** (**MMLU**) benchmark (Hendrycks et al., 2021), has been created. It is used to evaluate the ability of specialised AI-based grammar-checking applications and a set of LLMs (both proprietary and open-weight models) to detect grammatical errors in Bulgarian. This empirical evaluation is particularly important because very few publications describe the pre-training and fine-tuning procedures of state-of-the-art LLMs, and even fewer discuss specialised grammar-checking applications.

The main contributions of this work are as follows:

- **Development of the BulMMLU:gr** benchmark for Bulgarian, including questions designed to cover six types of grammatical, spelling, and punctuation errors in Bulgarian. The dataset has a standardised format, following the MMLU format.
- **Accuracy comparison** of two specialised Bulgarian grammar-checking applications and five LLMs (including proprietary and open-weight models) on the BulMMLU:gr benchmark.

The analysis indicates that proprietary LLMs and one existing grammar service demonstrate reasonably good, though imperfect, performance in detecting errors in Bulgarian grammar. However, there is currently no widely available open-source online service providing comprehensive Bulgarian grammar checking, nor any open-weight LLM that achieves comparably high accuracy.

[1] In this study, we count as grammatical not only purely grammatical errors but also spelling and punctuation errors that are related to grammar.

Therefore, it is necessary to expand the BulMMLU:gr dataset to cover a broader – ideally comprehensive – range of Bulgarian grammatical phenomena, and to use it for both testing and fine-tuning suitable language models to create a reliable tool for grammar checking in Bulgarian.

2 Related work

Traditional rule-based and statistical grammar checkers have gradually evolved into AI-based applications. They now offer a range of features that help users improve style and word choice, correct grammatical errors, and make content more readable and understandable. LLMs can naturally improve grammar, style, and fluency; however, they often alter the original text and, in some cases, the author's intended meaning may be lost, transformed, or even combined with incorrect statements. In contrast, AI-based grammar checkers improve clarity, readability, and flow by explicitly presenting each suggested edit and allowing authors to decide whether to accept or reject it.

Two essential elements of grammar checking have been studied for many years: grammatical error detection and grammatical error correction (Bryant et al., 2019; Rothe et al., 2021). Research on grammatical error detection and correction for English is well established, as are many other NLP tasks. In addition, a well-defined error taxonomy and standardised evaluation mechanism have been developed for English (Bryant et al., 2017). Error classification methods and approaches for transforming erroneous sequences into correct ones, based on the nature of the error, have also been proposed (Omelianchuk et al., 2020).

Recently, several publications have proposed approaches to grammatical error correction for languages other than English, including LLM-based methods for Turkish (Kara et al., 2023) and Chinese (Tian and Zhou, 2025), as well as synthetic data generation based on the Mask-Translate-and-Fill framework for low-resource Indian languages (Sharma and Bhattacharyya, 2025). For Bulgarian, mT5 was fine-tuned specifically for grammatical error correction (Klouchek and Batista-Navarro, 2024). The training data includes errors involving definite articles, pronouns, agreement, and verb morphology from a synthetic Bulgarian error-correction dataset. The fine-tuned model reportedly achieves an F0.5 score of 68.18% and outperforms the evaluated LLM baselines.

One of the most widely used modern approaches to training grammatical error correction models is to create synthetic parallel datasets containing pairs of erroneous and corrected sentences (Bondarenko et al., 2023). For example, RoLegalGEC, a synthetic Romanian-language parallel dataset for grammatical error detection and correction in the legal domain, contains 350,000 synthetically generated text pairs with error annotations (Timpuriu et al., 2026). The dataset was used to train and evaluate several grammatical error detection and correction models, including knowledge-distilled Transformer models, sequence-tagging architectures for error detection, and pre-trained sequence-to-sequence Transformer models for grammatical error correction.

Another example is OmniGEC, a collection of multilingual silver-standard datasets for grammatical error correction covering eleven languages: Czech, English, Estonian, German, Greek, Icelandic, Italian, Latvian, Slovene, Swedish, and Ukrainian (Kovalchuk et al., 2025). These datasets were developed to help adapt English grammatical error correction methods to multilingual settings. The source texts are mainly from Wikipedia and Reddit in the eleven target languages, together with an additional Ukrainian-only dataset sourced from social media. The Wikipedia edits were derived from human corrections, while the Reddit and Ukrainian-only data were automatically corrected using the GPT-4o-mini model. The quality of the corrections was evaluated both automatically and manually, and the data were further used to fine-tune two open-weight large language models, Aya-Expanse (8B) and Gemma-3 (12B), which were reported to achieve state-of-the-art results for paragraph-level multilingual grammatical error correction.

Another approach combines synthetic and gold data by either (i) fine-tuning a large language model in two stages – first on synthetic data and then on gold data – or (ii) training a small sequence-to-sequence (seq2seq) Transformer model that is first pre-trained on synthetic data and subsequently fine-tuned on gold data (Palma Gomez et al., 2023).

In summary, recent approaches to grammatical error detection and correction suggestions mainly rely on supervised fine-tuning using parallel corpora of erroneous and corrected sentence pairs. In this framework, grammatical error correction is typically formulated either as a sequence-to-sequence

task, in which models directly generate corrected text, or as a sequence-tagging task, in which models predict token-level edit operations (Omelianchuk et al., 2020). Owing to the scarcity of high-quality annotated data, synthetic data generation is widely employed. Errors are typically introduced using rule-based methods, back-translation, or, increasingly, LLM-based generation, enabling the creation of large-scale synthetic training datasets (Stahlberg and Kumar, 2021).

To further improve performance, many systems adopt a two-stage training strategy in which models are first pre-trained on large synthetic datasets and then fine-tuned on smaller, high-quality, human-annotated corpora (Ingólfsdóttir et al., 2023). The BulMMLU:gr dataset can be readily converted into a gold-standard fine-tuning dataset, as it contains erroneous sentences together with their corresponding corrections for the targeted error types. Its usefulness for this purpose could be further enhanced by expanding the dataset to cover a broader range of grammatical error types. As a proof of concept, Gemma 3 27B was fine-tuned using a version of BulMMLU:gr redesigned as a fine-tuning dataset.[2]

3 Description of the BulMMLU:gr dataset

The **Bulgarian Massive Multitask Language Understanding benchmark** (**MMLU-BG**) was developed to evaluate whether an LLM possesses general knowledge beyond simple text prediction in Bulgarian (Koeva et al., 2026). The original **MMLU** (**Massive Multitask Language Understanding**) (Hendrycks et al., 2021) covers 57 diverse subjects across multiple domains and difficulty levels. The questions are divided into a few-shot development set, a validation set for selecting hyperparameters, and a test set. Although **MMLU-BG** is well suited to assessing the general knowledge of LLMs, as it is a human translation of the original MMLU into Bulgarian, it does not reflect features specific to Bulgarian history, culture, and language.

In contrast, the newly developed dataset, **Massive Multitask Language Understanding for Bulgaria and Bulgarian** (**BulMMLU**), evaluates LLMs' knowledge of Bulgarian history, literature, economics, and the Bulgarian language. Its component, **Massive Multitask Language Understanding of Bulgarian Grammar** (**BulMMLU:gr**), provides a benchmark for evaluating specific aspects of Bulgarian grammar, spelling, punctuation, and context-dependent usage.

Six main types of grammatical error are included in the dataset for evaluation, selected according to the following principles:

- To provide diverse cases for evaluation, ranging from those that can be resolved by semantic analysis of the context and grammar rules to purely spelling-related errors.
- To test language-specific grammar, spelling, and punctuation rules.
- To evaluate correct grammar and spelling rules that depend on sentence analysis or external knowledge.

Type 1: Incorrect use of a paronym from a pair differing by only one letter. For example, in the Bulgarian sentence **На 1 януари 2026 г. еврото ще влезе в обръщение за парични плащания в страната*, 'On 1 January 2026, the euro will enter into circulation for cash payments in the country', the word *обръщение*, 'a word or expression with which the speaker addresses the listener', is incorrectly used instead of the word *обращение*, 'the process of use or exchange of money and goods; turnover'. The appropriate word can be selected only after analysing the syntactic and semantic structure of the sentence in which the target word is used and cannot be determined by applying rule-based grammar or spelling rules alone.

Type 2: Incorrect comma placement involving the adverb *така* 'in such a way', the subordinating conjunction *че* 'that', and the compound subordinating conjunction *така че* 'so'. For example: **Направете, така че законът да бъде приет* 'Do it in such a way that the law is passed'; **Не съм поканена на рождения ден така, че няма да дойда* 'I am not invited to the birthday party, so I will not come'. The correct punctuation depends on whether *така* functions as an adverb followed by the conjunction *че*, or whether *така че* functions as a compound conjunction. Correctly identifying these errors requires syntactic and semantic analysis of the sentence to determine the grammatical function of these expressions.

Type 3: Incorrect fused or separate spelling of an adverb and an adjective. Combinations of an adverb and an adjective (including adjectival participles) are written separately when they do not denote a single concept, for example, *особено*

[2] https://github.com/DCL-IBL/BG-LLMs/tree/feat/spelling/gemma3 (accessed on 15.07.26)

опасен ‘especially dangerous’ and *прясно боядисан* ‘freshly painted’. They are written as a single word when they form compound adjectives that conventionally denote a single concept, for example, *светлозелен* ‘light green’ and *световноизвестен* ‘world-famous’. In these cases, rearrangement of the components is not possible (e.g. **зелен светло*, **известен световно*). This type of error is difficult to detect because not all compound adjectives written as a single word are included in orthographic dictionaries,[3] and the choice between fused and separate spelling depends on whether the components denote a single lexicalised concept or form a phrase. Consequently, such errors are hardly detectable by rule-based grammar checkers.

Type 4: Incorrect agreement between an adjective and an abbreviation of a multi-word proper name. The adjective agrees in gender and number with the head noun of the multi-word proper name, even when the expression is abbreviated. For example, **новото ДДС* (данък добавена стойност, ‘Value Added Tax’, VAT) should be replaced with *новият ДДС*, because the noun *данък* is masculine. Detecting this type of agreement error requires knowledge of the multi-word expression represented by the abbreviation, including the grammatical gender and number of its head noun.

Type 5: Incorrect fused, semi-fused, or separate spelling of two nouns. Compound nouns in which one component is of foreign origin and not used as an independent word are written as one word. If the first component of the compound noun is of foreign origin and also used as an independent word, both fused and separate spellings are allowed. For example, the fused spelling *геймконзола* (‘gaming console’), where *гейм* denotes ‘relating to computer games and video games’, is the only prescribed form, whereas both *еврозона* and *евро зона* (‘euro zone’) are allowed because *евро* means ‘the monetary unit of the euro’ and is also used independently. Some fused words are listed in modern orthographic dictionaries, but not all new forms are included, as new words regularly enter the language. Thus, detecting this spelling error requires knowledge of which words are of foreign origin and not used separately in Bulgarian.

Type 6: Incorrect use of plural forms of masculine and neuter nouns. These nouns are of foreign origin and are either relatively rare or have recently entered the language, so their forms are not yet well established. For example, **две уиски* instead of *две уискита* ‘two whiskeys’, **много билборди* instead of *много билбордове* ‘many billboards’. The words whose plural forms might be mistaken are of foreign origin and relatively new to the language. For some of them, the existence of a plural form and its correct ending is already recorded in modern orthographic dictionaries, while others are still not included in the dictionaries.

The **BulMMLU:gr** (**Massive Multitask Language Understanding of Bulgarian Grammar**) dataset provides a new resource for evaluating state-of-the-art LLMs and grammar-checking applications by comparing their performance across different aspects of Bulgarian grammar, spelling, and punctuation. **BulMMLU:gr** follows the structure and data organisation of the original MMLU dataset (Hendrycks et al., 2021). Like MMLU, BulMMLU:gr consists of three parts: Development, Validation, and Test, and is intended to facilitate few-shot experiments and the reliable evaluation of LLMs. Each part comprises 5%, 15%, and 80% of the entire dataset, respectively. The dataset questions are designed to evaluate grammar, spelling, and punctuation phenomena that depend on sentence analysis or external knowledge; each question is followed by four answers, only one of which is correct, and the correct answer is provided as a separate field. All questions contain one or two example sentences with an intentionally inserted error of one of the described types. In some questions, two elements of the sentence are targeted. The dataset was developed manually by one person, with some semi-automatic support for detecting and extracting sentences with targeted correct words or phrases and for spell-checking. Additional semi-automatic validation procedures were performed to detect issues involving Cyrillic and Latin characters and incorrect punctuation inserted at the end of answers. The dataset contains 118 questions covering all selected error types. The dataset is maintained in CSV format and can be easily converted into a fine-tuning dataset.

The **BulMMLU:gr** dataset is publicly available under the Creative Commons Attribution 4.0 International Licence (CC BY 4.0), facilitating the replication of experiments. It is distributed as part of the broader **Massive Multitask Language Understanding for Bulgaria and Bulgarian** (**BulMMLU**) benchmark, which evaluates LLMs’ knowledge of Bulgarian history, literature,

[3] https://beron.mon.bg (accessed on 15.07.26)

economics, and the Bulgarian language.[4]

Two types of experiments are conducted with the dataset. The first evaluates grammar-checking applications that support Bulgarian, while the second evaluates five LLMs that are either pre-trained on Bulgarian data or specifically trained and fine-tuned for Bulgarian.

4 Evaluation of grammar-checking applications

Contemporary grammar-checking applications such as Grammarly, QuillBot, and LanguageTool are AI-based, typically combining machine-learning methods with rule-based approaches to detect errors and suggest improvements to grammar, spelling, punctuation, and style, while providing real-time suggestions across browsers, mobile applications, and word processors. Modern tools go beyond correction by offering features such as paraphrasing, summarisation, and plagiarism and AI-generated content detection.

4.1 Brief description of grammar-checking applications

Here we present six publicly accessible grammar-checking applications.

Grammarly is an AI-based writing assistant that combines machine learning, natural language processing, and rule-based methods to provide grammar, spelling, punctuation, style, and fluency suggestions. It supports more than 20 languages, but not Bulgarian.[5] Grammarly is available as a web, desktop, and mobile application and integrates with tools such as Microsoft Word and Google Docs. Earlier work by Grammarly researchers introduced the GECToR grammatical error correction model, based on a Transformer encoder (Omelianchuk et al., 2020).

InstaText is an AI writing assistant that improves grammar, spelling, punctuation, clarity, and readability.[6] It integrates with popular writing platforms and supports editing in several major languages, while multilingual assistance extends to more than 40 languages, including Bulgarian.

LanguageTool is a grammar and style checker that combines rule-based and machine-learning techniques to detect grammatical, spelling, punctuation, and stylistic errors.[7] It also provides AI-based paraphrasing and supports 31 languages, although Bulgarian is not currently available.

Scribens is an online grammar and spell checker that combines rule-based and AI techniques to detect spelling, grammatical, and punctuation errors.[8] It supports more than 40 languages, including Bulgarian, and is available as both a free and a premium service.

QuillBot is an AI writing assistant offering paraphrasing, grammar-checking, summarisation, translation, plagiarism detection, and citation generation.[9] It integrates with Google Docs and Microsoft Word and supports several major European languages, but not Bulgarian.

Scribbr is an academic writing platform that provides plagiarism detection, citation generation, grammar-checking, and proofreading services.[10] Its Grammar Checker is powered by QuillBot and supports several major languages, but not Bulgarian.

In summary, there are various applications for checking and correcting grammar, which commonly provide a range of services, including grammar and style checking, improving fluency, text simplification, paraphrasing in different words or styles, plagiarism detection, and text summarisation. Most of these applications offer premium modes that use more advanced algorithms, perform better, allow checking larger volumes of text, and provide additional services. English is generally supported, and many applications also support widely used languages such as French, German, and Spanish. Less frequently, applications support other languages; among the six applications surveyed, only two support Bulgarian. Most applications also offer various options for use and integration with browsers, word processing programmes, and other platforms.

Table 1 presents a brief comparison of six AI-based applications that offer grammar-checking services. The comparison includes the number of supported languages, whether Bulgarian is among them, and whether the application supports various types of use: online writing or copying texts, integration in browsers and other online services, as well as in word processing tools, indicated with +. Support for Bulgarian mostly relies on multilingual

[4] https://huggingface.co/datasets/DCL-IBL/BulMMLU-Dataset (accessed on 15.07.26)

[5] https://support.grammarly.com/ (accessed on 15.07.26)

[6] https://instatext.io (accessed on 15.07.26)

[7] https://languagetool.org/ (accessed on 15.07.26)

[8] https://www.scribens.com/ (accessed on 15.07.26)

[9] https://quillbot.com/ (accessed on 15.07.26)

[10] https://www.scribbr.com/ (accessed on 15.07.26)

Tool	Languages	BG	Use
Grammarly	24	No	Online+
InstaText	40+	Yes	Online+
LanguageTool	31	No	Online+
QuillBot	5	No	Online+
Scribbr	5	No	Online+
Scribens	40+	Yes	Online+

Table 1: Grammar-checking applications

applications, which cover more than 40 languages but often provide only limited support for Bulgarian.

4.2 Evaluation, results and analysis

Only two grammar-checking applications that support Bulgarian are evaluated with BulMMLU:gr: **InstaText** and **Scribens**. For the evaluation, the example sentences from the dataset questions are extracted and submitted to each application. The results are recorded and analysed. For Scribens, however, manual inspection of the suggested replacements is required, as the application displays them only after the user clicks on the highlighted error. Four outcomes are distinguished:

1. **Correct error detection with a correct correction suggestion** – the application correctly identifies the error and proposes the intended correction.

2. **Correct error detection with an incorrect correction suggestion** – the application correctly identifies the location of the error but proposes an incorrect replacement.

3. **Failure to detect an error** – the application does not identify an existing error.

4. **False error detection** (overgeneration) – the application incorrectly flags a correct language unit as an error.

Table 2 presents the evaluation results for the two grammar-checking applications that support Bulgarian, InstaText and Scribens. The error detection rate (Det.) is calculated as the percentage of evaluated sentences that fall into cases 1 and 2, since the error is successfully detected regardless of the quality of the suggested correction. The correct correction rate (Sugg.) is calculated as the percentage of evaluated sentences that fall into case 1 only, as these are the instances in which both the error detection and the proposed correction are correct. Case 3 represents a failure of error detection and therefore lowers the error detection rate. Case 4 corresponds to overgeneration (a false positive) and is reported separately; it is not included in the calculation of either the error detection rate or the correct correction rate.

Appl	InstaText	InstaText	Scribens	Scribens
Mode	**Det.**	**Sugg.**	**Det.**	**Sugg.**
Type 1	0.65	0.63	0.88	0.82
Type 2	0.72	0.72	0.78	0.65
Type 3	0.51	0.48	0.89	0.84
Type 4	0.62	0.62	**0.99**	**0.99**
Type 5	0.45	0.42	0.90	0.78
Type 6	0.60	0.58	0.71	0.56
Total	0.59	0.57	**0.85**	**0.77**

Table 2: Detection and suggestion rates of grammar-checking applications

The results showed that, at the time of the experiment (July 2026), Scribens outperformed InstaText in detecting most error types, performing best with Type 4 errors (incorrect agreement between an adjective and an abbreviation of a multi-word proper name). However, both applications struggled to detect the remaining error types. InstaText achieved its lowest score on Type 5 (incorrect fused, semi-fused, or separate spelling of two nouns), whereas Scribens performed worst on Type 6 (incorrect use of plural forms of masculine and neuter nouns). For example, in the sentence:

В киносалона на ***гейм клуба*** *прожектираха* ***кино документалистика*** *за* ***гейм индустрията*** *и нейното влияние върху младите* 'In the game club's cinema, they screened a documentary about the game industry and its influence on young people',

both applications treat the incorrect separate spellings of **гейм клуб* and **гейм индустрия* as correct. Moreover, instead of detecting **кино документалистика* as an error and suggesting that it be written as a single word, both applications recommend replacing it with a different phrase. InstaText suggests *documentary film*, while Scribens provides the following explanation: *The phrase is stylistically incorrect. 'Documentary cinema' is the established term, whereas 'cinema documentary' is a literal and incorrect translation from a foreign language (a calque).*

In principle, Scribens' error explanation feature

often indicates that an error has been detected but not recognised as the correct error type. Consequently, it does not suggest the appropriate correction. This distinction is not taken into account in the error analysis. For example, the explanation provided for the plural form of the word *капучино* 'cappuccino' is incorrect: *The plural of капучино in Bulgarian is formed with the ending -a (капучина).*

These observations raise questions that cannot be answered because the training data, fine-tuning procedures, and underlying rules of both grammar-checking applications are not publicly documented. Consequently, it is only possible to hypothesise about the reasons for their behaviour. One possible explanation is that InstaText appears to rely primarily on AI-based methods, whereas Scribens appears to rely more heavily on rule-based error detection and correction.

Overall, although Scribens achieves relatively good error detection performance across the evaluated error types, neither application consistently provides the intended corrections. These findings suggest that both applications still have notable limitations in handling Bulgarian-specific grammar, spelling, and punctuation errors.

5 Evaluation of LLM performance in grammar checking

LLMs are used in grammar-checking applications because they learn language patterns from large-scale datasets, enabling them to detect and correct errors in a context-aware manner and potentially offering advantages over traditional rule-based systems.

5.1 Brief description of LLMs selected for the experiment

We present five LLMs selected to evaluate the performance of major proprietary models and open-weight models.

Gemma 4 31B IT is an open-weight multimodal LLM developed by Google DeepMind and released in 2026.Built on research related to the Gemini 3 series, it supports more than 140 languages (including Bulgarian), image inputs, and a 256K-token context window. The model is released under the Apache 2.0 licence and is publicly available on Hugging Face and Kaggle (Google DeepMind, 2026b).

BgGPT 3.0 27B is an instruction-tuned LLM developed by INSAIT and released in 2026.Based on the Gemma 3 architecture, it is specifically designed for Bulgarian-language understanding and generation, supporting Bulgarian and English, multimodal input, and a 131K-token context window. The model is freely available through downloadable weights and an API (INSAIT, 2026).

Claude Sonnet 4.6 is an instruction-tuned LLM developed by Anthropic and released in 2026.It is designed for reasoning, coding, and agentic workflows, supports a context window of up to one million tokens, and is available through the Anthropic API (Anthropic, 2026).

Gemini 3.5 Flash is a multimodal LLM developed by Google DeepMind and released in 2026.It is optimised for low-latency inference while maintaining strong reasoning and coding performance. The model supports text, image, audio, video, and PDF inputs, a one-million-token context window, and is available through the Gemini API, Vertex AI, Google AI Studio, and the Gemini application (Google DeepMind, 2026a).

GPT-5.5 is a general-purpose multimodal LLM developed by OpenAI and released in 2026.It is designed for reasoning, instruction following, coding, and agentic tasks. GPT-5.5 supports a 1.05 million-token context window and incorporates improvements in reasoning, coding, and tool use over previous GPT-5 models. The model is available through the OpenAI API (OpenAI, 2026).

Experiments conducted with different model families, and with different releases within the same model series, demonstrate a high degree of variability in performance. This suggests that solutions to specific tasks are influenced by multiple factors, many of which are not fully transparent or controllable. Consequently, our analysis focuses on identifying broader performance tendencies and patterns rather than drawing conclusions from the results of any single model or model release.

5.2 Evaluation, results and analysis

The BulMMLU:gr dataset is designed as a single prompt covering all six error types. The prompt asks the model to provide the correct answer to a given question by choosing from four options. No example is shown. The model is instructed to answer only with one of the letters designating each possible answer: A), B), C) or D). Three of the tested models produced outputs exactly as instructed, while two returned the letter indicating the correct answer together with an explanation, from which the selected option was extracted.

We evaluate model performance using accuracy, comparing each model prediction to the gold-standard answer and scoring it as correct or incorrect depending on whether it matches. The evaluation metric is binary accuracy, calculated as the proportion of instances in which the model's decision matches the gold standard out of the total number of answers. The results for five models, distributed among the six error types, are presented in Table 3.

Table 3: Accuracy of the five evaluated models

Type	**N**	**M1**	**M2**	**M3**	**M4**	**M5**
Type 1	20	0.80	0.25	0.80	**0.90**	0.85
Type 2	21	0.48	0.43	**0.90**	0.48	0.10
Type 3	21	0.10	0.24	**0.71**	0.19	0.19
Type 4	16	0.63	0.31	**0.75**	0.56	0.50
Type 5	20	0.15	0.25	**0.85**	0.20	0.30
Type 6	20	0.45	0.25	**0.55**	0.40	0.15
Total	118	0.42	0.29	**0.76**	0.45	0.34

The accuracy results reveal substantial performance differences across models and question types. Among the evaluated models, M3 achieves the highest overall accuracy (0.76), outperforming all others by a wide margin. The remaining models show markedly lower overall performance, with M4 achieving 0.45, M1 – 0.42, M5 – 0.34, and M2 – 0.29, and they exhibit more pronounced weaknesses for specific error types.

The results also show that model performance strongly depends on the type of question being addressed. No single model achieves the highest accuracy across all question types. For Type 1 questions, M4 performs best with an accuracy of 0.90, followed by M5 (0.85) and M1/M3 (0.80). In contrast, M4 performs substantially worse on Type 3 and Type 5 questions, indicating that its strengths may be limited to particular reasoning patterns or task formulations. M5 shows a similar pattern of narrow strength: it is the second-best model on Type 1 (0.85), but its accuracy drops sharply on Type 2 (0.10) and remains low for Types 3, 5, and 6, indicating that its performance is highly dependent on question type. The models are not specifically pre-trained or fine-tuned for Bulgarian grammar, spelling, or punctuation, and the strong performance of a particular model on a specific task may be due to several factors, such as training data, model architecture, Bulgarian coverage, and so on.

M3 shows a comparatively more consistent performance pattern across the evaluated categories. It achieves the best results for five of the six question types (Types 2–6), with Type 1 being the only category where another model (M4) performs better. M3's strongest results are on Type 5 (0.85) and Type 2 (0.90). M1, by contrast, is highly inconsistent: it performs reasonably well on Type 1 (0.80, on par with M3) but drops sharply on Type 3 (0.10), and remains weak on Type 5 (0.15) and Type 6 (0.45). This indicates that M1 lacks general robustness across error types. As a model with open weights, it nonetheless remains a good candidate for further fine-tuning to detect errors in Bulgarian grammar, spelling, and punctuation.

The lowest overall performance is observed for M2, which remains below 0.50 accuracy for all question types and achieves only 0.29 overall accuracy. This consistent underperformance suggests that the model may have difficulties with the specific reasoning requirements represented in the evaluation set. In comparison, M4 and M5 show intermediate performance but exhibit high variability between categories, indicating sensitivity to task formulation.

The observed variability confirms that model selection has a significant impact on task outcomes. Differences between models are not limited to average accuracy but also appear in their relative strengths across question types. Therefore, evaluating only aggregate performance may conceal important model-specific capabilities and limitations. These findings support the use of multiple evaluation categories and suggest that model choice should take into account the characteristics of the target task rather than relying solely on general performance rankings.

The McNemar test is a statistical test used to compare two classifiers on the same set of instances. Unlike tests based on independent samples, it evaluates paired predictions by focusing on cases where the two models disagree: instances correctly classified by one model but incorrectly classified by the other (McNemar, 1947). The test assesses whether the difference in error distributions between two models is statistically significant. In this work, we use the exact binomial form of the test rather than the chi-squared approximation, computing the p-value directly from the binomial distribution with $n = b + c$ trials and success probability 0.5. A low p-value indicates that the observed performance difference is unlikely to arise from random variation,

suggesting that one model achieves a statistically different level of accuracy compared with the other.

The pairwise McNemar test results for the complete set of questions (Table 4) show statistically significant differences between most model pairs. M3 exhibits the most distinctive performance profile, with statistically significant differences compared with every other model ($p < 0.0001$ in all comparisons). This confirms that the superior overall accuracy of M3 is not only a numerical improvement but also represents a statistically significant change in prediction behaviour compared with the remaining models.

The comparisons among M1, M2, M4, and M5 reveal a more mixed pattern. M1 differs significantly from M2 ($p = 0.0328$), but no significant difference is observed between M1 and M4 or M5 ($p = 0.6476$ and $p = 0.0872$, respectively). Similarly, M2 performs significantly differently from M4 ($p = 0.0094$), but its difference from M5 is not statistically significant ($p = 0.4709$). These results indicate that although the models have different average accuracies, some performance differences may not be sufficiently large or consistent across individual questions to reach statistical significance.

The results by question type (Table 5) provide additional insight into the source of these differences. M3 shows statistically significant differences compared with the other models across several question categories, particularly Types 2, 3, and 5. For Type 2 questions, M3 shows statistically significant differences in favour of M3 over all competing models ($p \leq 0.0063$), indicating a strong advantage in this category. Similarly, Type 3 and Type 5 questions show significant differences between M3 and all other models, suggesting that these categories contribute substantially to the overall performance gap.

Given the number of pairwise comparisons performed, results should be interpreted with awareness that no correction for multiple testing (e.g., Bonferroni) was applied; borderline p-values (e.g., 0.0328, 0.0391, 0.0386) would not remain significant under such correction.

Overall, the McNemar analysis confirms that M3 provides a statistically significant improvement over the other evaluated models for the majority of the benchmark. However, the category-level analysis shows that model superiority is task-dependent: some question types do not show significant differences, indicating that performance advantages arise primarily from specific capabilities rather than universal dominance across all tasks.

6 Conclusion and future work

Of the applications surveyed, only two supported Bulgarian and were included in the evaluation, specifically for tasks involving grammatical error detection and correction suggestions. Although large language models demonstrate strong capabilities in improving text fluency and correcting various types of errors, the evaluated LLM interfaces did not offer the same explicit correction-review workflow as grammar-checking applications, which allow users to review, select, and accept or reject individual corrections. This limitation highlights a potential direction for future research and development: creating a dedicated Bulgarian grammar-checking system that combines automatic error detection with user-controlled correction suggestions.

A key objective for future work is to develop a systematic description and classification of the most frequent errors in Bulgarian, covering syntax, punctuation, and spelling. This resource could support the extension of the BulMMLU:gr dataset by incorporating examples of incorrect sentences paired with the corresponding corrected versions. Further experiments will determine whether a synthetic dataset, symbolic approaches to data processing, or both are needed to achieve sufficient coverage of the different error types.

Future investigations may also explore the adaptation and fine-tuning of open-weight large language models for Bulgarian grammatical error detection and correction tasks. Potential candidate models include Gemma 3 27B and Gemma 4 27B, which could serve as foundations for developing specialised models capable of providing accurate, transparent, and user-controllable grammar correction assistance for Bulgarian. At present, the possibilities of quantising and running large open-weight language models (such as DeepSeek-V3, V4 Pro, or Kimi 3), or using pre-compressed or pruned community weights, remain largely unexplored.

7 Acknowledgments

The present study is carried out within the project Infrastructure for Fine-tuning Pretrained Large Language Models, Grant Agreement No. ПВУ – 55 from 12.12.2024 /BG-RRP-2.017-0030-C01/.

References

Anthropic. 2026. Claude Sonnet 4.6: Frontier performance for scale. Accessed: 15 July 2026.

Maksym Bondarenko, Artem Yushko, Andrii Shportko, and Andrii Fedorych. 2023. Comparative Study of Models Trained on Synthetic Data for Ukrainian Grammatical Error Correction. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 103–113, Dubrovnik, Croatia. Association for Computational Linguistics.

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 Shared Task on Grammatical Error Correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy. Association for Computational Linguistics.

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada. Association for Computational Linguistics.

Google DeepMind. 2026a. Gemini 3.5 Flash Model Card. Google AI.

Google DeepMind. 2026b. Gemma 4 Model Card. Accessed: 15 July 2026.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, pages 1–27. OpenReview.net.

Svanhvít Lilja Ingólfsdóttir, Petur Ragnarsson, Haukur Jónsson, Haukur Simonarson, Vilhjalmur Thorsteinsson, and Vésteinn Snæbjarnarson. 2023. Byte-Level Grammatical Error Correction Using Synthetic and Curated Corpora. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7299–7316, Toronto, Canada. Association for Computational Linguistics.

INSAIT. 2026. BgGPT 3.0: A new generation of Bulgarian AI models. Accessed: 15 July 2026.

Atakan Kara, Farrin Marouf Sofian, Andrew Bond, and Gözde Şahin. 2023. GECTurk: Grammatical Error Correction and Detection Dataset for Turkish. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 278–290, Nusa Dua, Bali. Association for Computational Linguistics.

Bozhidar Klouchek and Riza Batista-Navarro. 2024. Bulgarian Grammar Error Correction with Data Augmentation and Machine Translation Techniques. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pages 365–376, Trento. Association for Computational Linguistics.

Svetla Koeva, Ivelina Stoyanova, Dimiter Georgiev, Svetlozara Leseva, Valentina Stefanova, Maria Todorova, Tsvetana Ivanova Dimitrova, Hristina Kukova, Mihaela Moskova, and Tinko Tinchev. 2026. Bulgarian Massive Multitask Language Understanding Benchmark. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 4658–4672, Palma de Mallorca, Spain. European Language Resources Association (ELRA).

Roman Kovalchuk, Mariana Romanyshyn, and Petro Ivaniuk. 2025. Introducing OmniGEC: A Silver Multilingual Dataset for Grammatical Error Correction. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, pages 162–178, Vienna, Austria (online). Association for Computational Linguistics.

Quinn McNemar. 1947. Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages. *Psychometrika*, 12:153–157.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. GECToR – Grammatical Error Correction: Tag, Not Rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–170, Seattle, WA, USA – Online. Association for Computational Linguistics.

OpenAI. 2026. Introducing GPT-5.5. OpenAI Technical Report and API Documentation. Accessed: 15 July 2026.

Frank Palma Gomez, Alla Rozovskaya, and Dan Roth. 2023. A Low-Resource Approach to the Grammatical Error Correction of Ukrainian. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 114–120, Dubrovnik, Croatia. Association for Computational Linguistics.

Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2021. A Simple Recipe for Multilingual Grammatical Error Correction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 702–707, Online. Association for Computational Linguistics.

Ujjwal Sharma and Pushpak Bhattacharyya. 2025. IndiGEC: Multilingual Grammar Error Correction for Low-Resource Indian Languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22382–22396, Suzhou, China. Association for Computational Linguistics.

Felix Stahlberg and Shankar Kumar. 2021. Synthetic Data Generation for Grammatical Error Correction

with Tagged Corruption Models. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 37–47, Online. Association for Computational Linguistics.

Wei Tian and Yuhao Zhou. 2025. Chinese Error Corrector 3-4B: State-of-the-Art Chinese Spelling and Grammar Corrector. *arXiv:2511.17562*.

Mircea Timpuriu, Mihaela-Claudia Cercel, and Dumitru-Clementin Cercel. 2026. RoLegalGEC: Legal Domain Grammatical Error Detection and Correction Dataset for Romanian. *arXiv:2604.19593*.

Appendix A McNemar Test Results

Model	M1	M2	M3	M4	M5
M1	—	0.0328*	<0.0001*	0.6476	0.0872
M2	0.0328*	—	<0.0001*	0.0094*	0.4709
M3	<0.0001*	<0.0001*	—	<0.0001*	<0.0001*
M4	0.6476	0.0094*	<0.0001*	—	0.0146*
M5	0.0872	0.4709	<0.0001*	0.0146*	—

* Statistically significant difference at a significance level of $p < 0.05$.

Table 4: p-values from the McNemar test for pairwise comparisons between the models (total of 118 questions)

Question type	vs M1	vs M2	vs M4	vs M5
Type 1	1.0000	0.0034*	0.6875	1.0000
Type 2	0.0039*	0.0063*	0.0039*	<0.0001*
Type 3	0.0010*	0.0063*	0.0034*	0.0034*
Type 4	0.6875	0.0391*	0.5078	0.3438
Type 5	0.0005*	0.0005*	0.0010*	0.0074*
Type 6	0.7744	0.1796	0.5811	0.0386*

*Statistically significant difference compared to Model 3 at p < 0.05.

Table 5: p-values from the McNemar test: Comparison of Model 3 with the other models across question types

Corpus Building and LLM Benchmarking for Language Understanding

Building a cross-lingual corpus of BA theses: Rationale and preliminary observations

Karla Csuros
West University of Timisoara, Romania
karla.csuros@e-uvt.ro

Madalina Chitez
West University of Timisoara, Romania
madalina.chitez@e-uvt.ro

Roxana Rogobete
West University of Timisoara, Romania
roxana.rogobete@e-uvt.ro

Petya Osenova
Institute of Information and Communication Technologies
Bulgarian Academy of Sciences, Bulgaria
petya@bultreebank.org

Abstract

In recent years, the impact of AI has increased in all areas of our lives, including academic writing. While the writing of MA and doctoral theses had been researched extensively even prior to AI, BA theses were not a focus of analysis. However, BA theses are the first serious academic texts students produce, in which they must demonstrate their ability to logically construct academic texts, posit research questions, critically review related works, classify and analyze data, and make observations and conclusions. The growing use of these tools makes it urgent to establish empirical baselines for what student writing looks like in specific languages and genres, against which future change can be measured. In this paper, we report on the creation of an initial cross-lingual corpus of theses from Bulgarian and Romanian BA students and present some quantitative and qualitative results on the usage of hedges in both corpora. We are interested in seeing how certain or uncertain students are with respect to the various parts of academic texts.

Keywords: academic writing, BA theses, cross-lingual corpus, AI in education, corpus linguistics.

1 Introduction

Academic writing research has devoted substantial attention to MA dissertations and doctoral theses (see, for instance, Charles, 2007; Thompson, 2005), yet BA theses, as the first extended academic texts students produce, remain comparatively under-studied, particularly outside English-language contexts. This gap has acquired new urgency with the widespread availability of generative AI (genAI) tools. If we wish to understand how AI adoption is reshaping student writing, we need empirical baselines for what student writing looks like in specific languages, genres, and institutional settings.

This paper contributes to that effort by reporting on the construction of a pilot cross-lingual corpus of 36 BA theses (18 Romanian, 18 Bulgarian) from the 2025 cohort, formed of students who had access to large language models (LLMs) throughout their undergraduate studies. As a first analytical lens, we examine hedging: the linguistic devices through which writers signal uncertainty, attribute claims, and modulate the strength of their assertions. Hedging is a particularly sensitive indicator of epistemological development, as it reflects how students position themselves relative to knowledge claims in their discipline. To this end, we address two research questions:

(RQ1) What hedging profiles characterise BA writing in Romanian and Bulgarian, and how do they compare cross-linguistically?

(RQ2) What can these profiles tell us about the rhetorical strategies available to novice academic writers in each language?

2 Related work

2.1 Student BA thesis writing

Corpus-based approaches to academic writing have developed substantially since the shift from text-oriented to process-centred pedagogy in the 1970s, yet most corpus resources and pedagogical applications target MA and doctoral writing (Chitez et al., 2015; Flowerdew, 2015). Platforms such as *The-*

sis Writer provide real-time formulation support for graduate students through POS-informed pattern searches and discipline-specific phrase lists (Chitez et al., 2015), while workshop-based approaches use broad reference corpora to teach prototypical move structures at the postgraduate level (Flowerdew, 2015). These initiatives have demonstrated the value of Data-Driven Learning (DDL) for developing metadiscursive awareness, but their focus remains overwhelmingly on graduate genres. BA theses, by contrast, have received limited corpus-based attention, a gap that is particularly consequential for non-Anglophone academic traditions. Baniceru et al. (2012) offer one of the first studies of Romanian BA writing, finding that students successfully adopt descriptive rhetorical moves (introducing the topic, presenting structure) but struggle with reflective ones (summarising prior research, identifying gaps). This observation can be situated within what Doroholschi et al. (2018) explore as a broader Central and Eastern European context where academic writing cultures shaped by French-influenced, author-oriented traditions are undergoing a sometimes mechanical transition toward Anglo-Saxon reader-oriented conventions. The present study contributes to filling this gap by building a cross-lingual BA thesis corpus and applying corpus methods, specifically lexicon-based hedge identification, to a genre and pair of languages for which no comparable quantitative data has previously been reported.

More recent research has benefited from the development of the Corpus of Romanian Academic Genres (ROGER), a bilingual learner corpus containing undergraduate academic texts written in Romanian (L1) and English (L2). The ROGER corpus (see Chitez 2022) includes BA-level research papers and theses across multiple disciplines (Economics, Humanities, Political Sciences, Computer Science, etc.) and allows, together with EXPRES corpus (Corpus of Expert Writing in Romanian and English; Chitez et al. (2022), comparisons between novice student writing and expert academic discourse. Corpus-based analyses using ROGER have documented limited phraseological diversity, underuse of stance markers, and restricted metadiscursive resources in Romanian student writing, particularly at the undergraduate level (Oravițan et al. 2022, Dincă et al.2022).

2.2 Hedging in academic discourse

Hedging is a fundamental communicative strategy through which academic writers balance "conviction with caution" in what Hyland (1998) characterises as the social negotiation of knowledge. Rather than signalling weakness, hedges serve both epistemic and interpersonal functions, encoding the writer's degree of commitment to a proposition while maintaining collegial deference toward the reader (Hyland, 1998). Unhedged claims risk appearing as overstatements that threaten the writer's standing within a disciplinary community.

Several classification systems have been proposed for these devices. Prince et al. (1982) distinguish between approximators (affecting propositional content) and shields (affecting speaker commitment). Salager-Meyer (1994) and Crompton (1997) offer pragmatically grounded typologies, while Caffi (1999) proposes a tripartite model of mitigation. Hyland (1998) further distinguishes standard hedges, which qualify authorial commitment, from attribute hedges, which manage "permissible imprecision" without necessarily signalling uncertainty. Thompson and Yiyun (1991) complement this with a multi-layered taxonomy of reporting verbs, distinguishing author acts, writer acts, and stance categories, a system which is particularly relevant to the categories used in the present study.

Cross-linguistically, Vassileva (2001) demonstrates that hedging conventions differ substantially between academic traditions, while Vold (2006) shows that even typologically related languages can produce divergent hedging profiles in comparable academic genres. These findings motivate the present cross-lingual design for investigating hedge use in student BA thesis writing.

3 Corpus and methodology

3.1 TRAI cross-lingual BA thesis corpus

The pilot corpus compiled for this study belongs to the TRAI (Exploring the Triad of Academic Writing, Critical Thinking and AI Literacy: Interdisciplinary Perspectives on Undergraduate Thesis Writing) project and comprises 36 BA theses in two languages: 18 Romanian theses from the West University of Timișoara (412,654 words), and 18 Bulgarian theses from Sofia University St. Kliment Ohridski (369,721 words). The pairing of Romanian and Bulgarian is motivated first practically: both subcorpora are compiled within the TRAI

project, a partnership between the two institutions, which provides access to comparable BA theses from philology programmes in each country. Typologically, the pair also sets a Romance language against a Slavic one while holding constant their shared membership in the Balkan area, allowing hedging strategies to be compared across genetically distinct but areally linked systems. All theses were submitted by cohorts with access to generative AI tools throughout their undergraduate studies, situating the corpus in the post-LLM period. Because the design includes no pre-LLM comparison cohort, we make no claim that the profiles reported here reflect AI influence; that question is left to future work (Section 5). Table 1 summarises the composition and size of the two subcorpora alongside the lexicon specifications and aggregate hedging rates.

The two subcorpora were compiled as structured XML using a custom Python pipeline that identifies front matter, table of contents, body sections, and bibliography through heading-style detection and keyword matching. Both subcorpora share the same target structure: an XML document with elements segmenting each thesis into title page, table of contents, introduction, body, conclusion, and bibliography. For hedge analysis, only the content sections are retained; front matter and bibliography are excluded.

3.2 Cross-lingual hedge identification

Hedging devices are identified using a lexicon-based approach that draws on established hedging taxonomies (Caffi, 1999; Crompton, 1997; Hyland, 1998, 2005; Prince et al., 1982; Salager-Meyer, 1994; Thompson and Yiyun, 1991), with cross-linguistic adaptations informed by Vassileva (2001) and Vold (2006). Six categories are used for this pilot study: MODAL (epistemic modal verbs and particles), EPISTEMIC (lexical verbs encoding authorial stance), APPROXIMATOR (degree, quantity, and frequency modifiers), SHIELD (impersonal and distancing constructions), ADVERB (epistemic adverbs, concessives, and boosters), and DEFERENCE (attribution and perspectivisation markers).

For each language, a dedicated lexicon of regular expression patterns was developed: 132 patterns for Romanian and 159 for Bulgarian, with each pattern mapped to a canonical form and category. Each language's lexicon was compiled by a single expert linguist working from a shared English taxonomy: the English categories provided a common scaffold, which was then translated and enriched to fit the morphosyntactic resources of the target language. For instance, Romanian gerund forms (*considerând, observând*) and Bulgarian impersonal reflexive constructions (смята се, предполага се) have no direct English equivalents, but serve analogous hedging functions. Because the two experts worked from the same category definitions, the lexicons remain comparable across languages, and because a single annotator built each, there is no within-language disagreement to reconcile; every pattern is mapped to a canonical form and category and is released with the corpus for inspection. We note as a limitation that formal inter-annotator agreement and manual validation of the automatic annotation output were not carried out for this pilot; both, alongside the morphological analysis flagged below, are priorities for the next stage.

Pattern matching operates on the raw running text within each XML file, with overlapping matches resolved through a longest-match-first deduplication strategy. This surface-level approach prioritises transparency and reproducibility over the precision gains that POS tagging or lemmatisation could offer. Known limitations include the polysemy of high-frequency items, such as Romanian *poate* (modal verb "can" vs. adverb "perhaps") and Bulgarian около (approximator "approximately" vs. spatial preposition "around"), and the exclusion of the Bulgarian renarrative forms (преизказни форми), a grammaticalised evidential with no Romanian equivalent whose forms are homophonous with past active participles and therefore undetectable by regex. Both limitations are flagged for future work involving morphological analysis.

4 Preliminary results

Across the combined corpus of 36 theses (≈782,000 words), we identified 5,156 hedge instances in Romanian (mean 12.5/1000w, SD = 2.8) and 4,986 in Bulgarian (mean 14.18/1000w, SD = 3.71). Bulgarian theses thus hedge at a moderately higher overall rate, but the more revealing finding lies in how the two student populations hedge in the distribution across categories rather than in the aggregate frequency. Figure 1 visualises the distribution of hedges across the six categories, revealing markedly different profiles despite comparable overall rates.

	Romanian	Bulgarian
Institution	West Univ. Timisoara	Sofia Univ. St. Kliment Ohridski
Specialisation	Romanian Philology	Slavic Philology
Year	2025	2025
Theses	18	18
Total words	412,654	369,721
Mean words/thesis (SD)	22,925 (7,049)	20,540 (8,559)
Total hedges	5,156	4,986
Mean hedges/1000w (SD)	12.5 (2.80)	14.18 (3.71)
Lexicon patterns	132	159

Table 1: Pilot TRAI cross-lingual BA thesis corpus

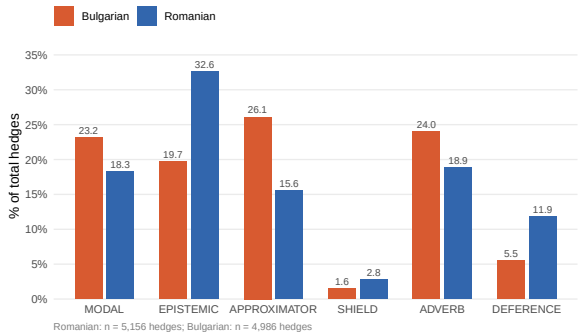


Figure 1: Hedging category profiles: Romanian vs. Bulgarian BA theses (% of total hedges, 18 theses per language).

To test whether these differences are statistically reliable, we treated each thesis as a single observation (n = 18 per language) and compared per-thesis hedging rates with Mann-Whitney U tests, applying Holm correction across the six categories and reporting Cliff's delta (δ) as an effect size. The aggregate hedging rate does not differ significantly between the two subcorpora (Romanian 12.5 vs. Bulgarian 14.18 per 1000 words; U = 123, p = .22), confirming that the contrast lies in the distribution rather than the overall frequency. Three categories differ significantly after correction, all with large effect sizes: Romanian shows higher rates of EPISTEMIC verbs (p = .04, δ = 0.51) and DEFERENCE markers (p = .002, δ = 0.69), while Bulgarian shows markedly higher use of APPROXIMATOR (p < .001, δ = 0.77). MODAL and ADVERB use trend higher in Bulgarian but do not survive correction (p = .08 and .13), and SHIELD use is low in both subcorpora and does not differ significantly (p = .13).

Romanian students rely heavily on EPISTEMIC verbs, which account for 32% of all hedges (4.11/1000w). The most frequent epistemic forms (*observă*, *consideră*, *sugerează*, *propune*) are reporting verbs through which writers attribute claims to cited scholars or signal their own interpretive stance. This pattern is reinforced by the DEFERENCE category, where Romanian rates are nearly double those of Bulgarian (1.45 vs. 0.77/1000w), driven by high-frequency attribution markers such as din *punct de vedere*, *după cum*, and *conform*. Together, EPISTEMIC and DEFERENCE account for over 44% of Romanian hedging, suggesting a rhetorical strategy oriented around positioning claims relative to external authorities. For instance, a typical Romanian hedging sequence reads: "Această emoție *pare a fi* apogeul existenței" ("This emotion *appears to be* the apex of existence"), where the evidential *pare a fi* simultaneously claims and distances. Similarly, attribution to authorities is pervasive: "Pe de altă parte, *conform lui* Albert Camus, individul ce recurge..." ("On the other hand, *according to* Albert Camus, the individual who resorts to..."), combining DEFERENCE (*conform*) with a perspectival frame that defers the claim.

The Bulgarian profile is structurally different. Hedging is more evenly distributed across APPROXIMATOR (26.1%, 3.75/1000w), ADVERB (24%, 3.34/1000w), and MODAL (23.2%, 3.35/1000w). The most frequent forms (често, обаче, може да + V, може би, почти) are degree modifiers, possibility markers, and concessive connectors. Bulgarian students appear to hedge primarily by softening the strength or scope of their propositions rather than by delegating epistemic responsibility to cited sources. A representative Bulgarian passage illustrates this strategy: "Оптимизмът може да се разглежда като позитивна черта" ("Optimism *can be* viewed as a positive trait"), where a MODAL (може да) combines with a reflexive impersonal (се разглежда) to double-hedge a proposition that an English or Romanian

student might assert directly. Elsewhere, frequency-based softening is common: "като често пренебрегват неблагоприятните обстоятелства" ("as they *often* disregard unfavourable circumstances"), where често ("often") limits the scope of the generalisation.

SHIELD devices (impersonal constructions such as *se poate afirma* or може да се каже) are rare in both subcorpora (2.8% in Romanian, 1.6% in Bulgarian), and, as the tests above show, do not differ significantly between them. The few SHIELD instances that do occur tend to be formulaic: Romanian "Generația 60 *se poate spune că* a urmat..." ("The 60s generation, *it can be said*, followed...") and Bulgarian "може да се допусне, че социалното време..." ("*it can be assumed that* social time...") both use fixed constructions that suggest imitation of a template rather than productive mastery of impersonal distancing.

These observations should be interpreted with the caveats noted earlier: the lexicon-based approach cannot fully resolve polysemy, and the Bulgarian renarrative mood, a grammaticalised evidential with no Romanian equivalent, remains outside the scope of automatic detection.

5 Discussion and next steps

The divergent hedging profiles suggest distinct rhetorical orientations at the BA level: Romanian students manage uncertainty primarily through attribution, positioning claims relative to cited authorities, while Bulgarian students distribute hedging more evenly across modal, approximative, and adverbial resources, softening propositions lexically rather than delegating epistemic responsibility. SHIELD devices are rare in both subcorpora, and the tests confirm the difference between languages is not significant. We resist reading this as evidence of a developmental stage, however: that interpretation would require comparison with expert or graduate writing in the same languages, which this pilot does not include. We therefore offer it as a hypothesis for future work, in which BA profiles would be compared against MA, doctoral, or expert reference corpora (for Romanian, for instance, EXPRES). These patterns are consistent with the novice writing profiles described by Thompson and Yiyun (1991) and with the difficulties in reflective rhetorical moves observed in Romanian BA writing by Baniceru et al. (2012).

A fuller account of why the two groups differ is beyond a pilot, but one plausible factor is writing tradition. The Romanian attribution-heavy profile in particular is consistent with the author-oriented tradition Doroholschi et al. (2018) describe for the region, in which claims are habitually anchored to named authorities. Social and pedagogical background are variables worth incorporating in future comparison.

Because the present corpus captures a post-LLM cohort, the hedging profiles reported here cannot be attributed to AI in the absence of a pre-LLM baseline; disentangling any AI effect would require comparison across cohort years, which we leave to future work. More immediate next steps within the TRAI project include adding Swiss German BA theses as a third comparison language, integrating POS tagging to resolve polysemy in high-frequency items, and expanding the corpus beyond philology programmes to test whether the patterns observed here are discipline-specific or generalisable.

Acknowledgements

This work was supported by the Swiss National Science Foundation (SNSF) under the Multilateral Academic Projects (MAPS) grant (No. IZ11Z0_230816), awarded for the project “Exploring the Triad of Academic Writing, Critical Thinking and AI Literacy: Interdisciplinary Perspectives on Undergraduate Thesis Writing” (TRAI), co-financed by the Romanian Ministry of Education and Research through UEFISCDI (Grant No. 15ROCH /01/07/2025) and by the Bulgarian Ministry of Education and Science (Reg N D03-34/30.07.2025).

The author from IICT-BAS gratefully acknowledges the financial and institutional support provided by the Swiss-Bulgarian Research Programme in the framework of the second Swiss Contribution to selected EU member states.

References

Cristina Ana Baniceru, Mirela Ioana Borchin, Claudia Ioana Doroholschi, and Dumitru Tucan. 2012. Academic writing in Romania: a contrastive analysis of BA thesis introductions in Romanian and English. *CICCRE*, 1.

Claudia Caffi. 1999. On mitigation. *Journal of Pragmatics*, 31(7):881–909.

Maggie Charles. 2007. Reconciling top-down and bottom-up approaches to graduate writing: Using

a corpus to teach rhetorical functions. *Journal of English for Academic Purposes*, 6(4):289–302.

Mădălina Chitez. 2022. Corpus of Romanian Academic Genres (ROGER). Slovenian Language Resource Repository CLARIN.SI.

Madalina Chitez, Christian Rapp, and Otto Kruse. 2015. Corpus-supported academic writing : how can technology help? *Critical CALL : Proceedings of the 2015 EUROCALL Conference, Padova, Italy*.

Mădălina Chitez, Roxana Rogobete, Valentina Mureşan, and Andreea Dincă. 2022. Corpus of Expert Writing in Romanian and English (EXPRES). West University of Timişoara.

Peter Crompton. 1997. Hedging in academic writing: Some theoretical problems. *English for Specific Purposes*, 16(4):271–287.

Andreea Dincă, Mădălina Chitez, Loredana Bercuci, Alexandru Oraviţan, and Roxana Rogobete. 2022. Using bilingual novice and expert corpora to teach academic writing at the university: The case of ROGER and EXPRES. Paper presented at the 15th Teaching and Learning Corpora Conference (TaLC 2022). Conference presentation, 13–15 July 2022.

Claudia Ioana Doroholschi, Dumitru Tucan, Mădălina Chitez, and Otto Kruse. 2018. *Introduction: Understanding Academic Writing in the Context of Central and Eastern European Higher Education*, pages 1–12. Springer International Publishing, Cham.

Lynne Flowerdew. 2015. Using corpus-based research and online academic corpora to inform writing of the discussion section of a thesis. *Journal of English for Academic Purposes*, 20:58–68.

Ken Hyland. 1998. Boosting, hedging and the negotiation of academic knowledge. *Text & Talk*, 18(3):349–382.

Ken Hyland. 2005. Stance and engagement: a model of interaction in academic discourse. *Discourse Stud.*, 7(2):173–192.

Alexandru Oraviţan, Mădălina Chitez, Loredana Bercuci, and Roxana Rogobete. 2022. Using the bilingual Corpus of Romanian Academic Genres (ROGER) platform to improve students' academic writing. In *Intelligent CALL, granular systems and learner data: Short papers from EUROCALL 2022*, pages 315–321. Research-publishing.net.

Ellen F Prince, Joel Frader, Charles Bosk, et al. 1982. On hedging in physician-physician discourse. *Linguistics and the Professions*, 8(1):83–97.

Françoise Salager-Meyer. 1994. Hedges and textual communicative function in medical English written discourse. *English for Specific Purposes*, 13(2):149–170.

Geoff Thompson and Ye Yiyun. 1991. Evaluation in the reporting verbs used in academic papers. *Applied Linguistics*, 12(4):365–382.

Paul Thompson. 2005. Points of focus and position: Intertextual reference in PhD theses. *Journal of English for Academic Purposes*, 4(4):307–323.

Irena Vassileva. 2001. Commitment and detachment in English and Bulgarian academic writing. *English for Specific Purposes*, 20(1):83–102.

Eva Thue Vold. 2006. Epistemic modality markers in research articles: a cross-linguistic and cross-disciplinary study. *International Journal of Applied Linguistics*, 16(1):61–87.

Appendix A Pilot cross-lingual hedge lexicon sample

Table 2 presents the core items from each of the six pilot hedging categories used in the cross-lingual analysis.

Category	English gloss	Romanian	Bulgarian
MODAL	may/can + V	*a putea* + V	може да + V
	perhaps	*poate* (adv.)	може би
	could (conditional)	*ar putea*	би могъл/а/о/и
	should + V	*ar trebui*	трябва да
	cannot exclude	*nu se poate exclude*	не може да се изключи
EPISTEMIC	seem, appear	*a părea*	изглеждам
	consider	*a considera*	смятам; считам
	assume, suppose	*a presupune*	предполагам; допускам
	believe	*a crede*	склонен съм
	suggest	*a sugera*	подсказвам; насочвам към
	observe, note	*a observa; a remarca*	наблюдавам; отбелязвам
	find, establish	*a constata; a rezulta*	установявам; разглеждам
	propose, interpret	*a propune; a interpreta*	предлагам; тълкувам
APPROXIMATOR	approximately	*aproximativ*	приблизително; около
	almost	*aproape*	почти
	relatively	*relativ*	относително
	to a certain extent	*într-o oarecare măsură*	в известна степен
	often, sometimes	*adesea; uneori*	често; понякога
	rather	*mai degrabă*	по-скоро
SHIELD	it can be said	*se poate afirma/spune*	може да се каже
	it can be assumed	*se poate presupune*	може да се допусне
	it is possible	*este posibil (că)*	възможно е (да)
	we could say	*am putea spune*	бихме могли да кажем
	remains open	*rămâne de văzut*	остава открит/спорен
ADVERB	probably	*probabil*	вероятно
	possibly	*posibil; eventual*	възможно; евентуално
	apparently	*aparent*	привидно
	however	*însă; totuși*	обаче; въпреки това
	still	*încă*	все още; все пак
	undoubtedly (B)	*fără îndoială*	несъмнено
DEFERENCE	according to	*conform; potrivit*	според; съгласно
	in the opinion of	*în opinia; în viziunea*	по мнението на
	from the perspective of	*din punct de vedere*	от гледна точка на
	as X states	*după cum*	както посочва/отбелязва
	in X's work	*în lucrarea/studiul*	в своя/неговия труд

Table 2: Core items from pilot cross-lingual hedge lexicon by category

Reasoning-BG: A Hybrid Benchmark for Long-Text Understanding and Reasoning in Bulgarian

Maria Todorova, Valentina Stefanova, Tsvetana Dimitrova, Mihaela Moskova, Hristina Kukova, Svetlozara Leseva, Ivelina Stoyanova, Dimitar Georgiev, Svetla Koeva

Department of Computational Linguistics, Institute for Bulgarian Language, Bulgarian Academy of Sciences

{maria,valentina,cvetana,mmoskova,hristina}@dcl.bas.bg
{zarka,iva,dgeorgiev,svetla}@dcl.bas.bg

Abstract

The paper introduces the Bulgarian benchmark **Reasoning-BG**, designed to evaluate how accurately large language models (LLMs) solve Bulgarian long-text comprehension tasks involving different comprehension demands. It was developed using a hybrid methodology that combines automated generation of four-option multiple-choice questions by the Gemma-3-27B-IT model, applied to 232 Bulgarian texts, followed by validation, correct-answer determination, and editing by human experts.

The question design, which addresses relatively long texts, is inspired by the Programme for International Student Assessment (PISA) framework, which targets reading-comprehension tasks such as retrieving information, interpreting and integrating ideas, and evaluating and reflecting on content. The benchmark provides a framework for evaluating model-driven, cognitive-like processes, shifting the focus from factual recall to the analysis of complex interpretative and logical performance in Bulgarian.

Keywords: benchmark, reasoning, LLMs, Bulgarian

1 Introduction

The increasing use of large language models (LLMs) to generate (educational) content and assessment tasks has created a need for benchmarks that can evaluate AI systems on their understanding and reasoning over texts with different levels of complexity.

This need is particularly acute for Bulgarian, which currently lacks specialised benchmarks designed to assess LLM understanding and reasoning. Some recent efforts to develop Bulgarian evaluation datasets have been made (see Section 3), but these resources do not specifically target the cognitive processes involved in understanding extended texts. Traditional benchmarks often prioritise factual accuracy or simple logical predictions, and fail to provide the depth required to measure complex comprehension effectively.

This paper addresses a significant gap in Bulgarian NLP resources by introducing **Reasoning-BG**, a benchmark comprising 232 relatively long texts covering a wide range of topics. Each text is accompanied by 10 multiple-choice questions aligned with a cognitive framework that targets key comprehension processes such as information retrieval, interpretation, and evaluation. A distinguishing characteristic of Reasoning-BG is its strong emphasis on content quality: every text and all questions have been carefully crafted and refined by experts to ensure grammatical accuracy, natural word order, and stylistic consistency that faithfully reflects the linguistic characteristics of Bulgarian.

The present study makes both methodological and empirical contributions by advancing the evaluation of text understanding, reasoning, and higher-order cognitive-like processes involved in long-text comprehension by large language models.

- First, a benchmark for LLM evaluation, Reasoning-BG, was developed that is explicitly aligned with the cognitive processes underpinning deep human text comprehension, including the retrieval and integration of information, the establishment of connections between distant textual segments, the construction of local and global inferences, the tracking of causal relationships, and the formation of a coherent mental representation of the text.
- Second, a hybrid human–LLM methodology for benchmark creation is used, combining expert human judgement with the efficiency of large language models. This approach facilitates the creation and validation of high-quality benchmarks while preserving human

oversight of the interpretation and reliability of the content.

- Third, unlike benchmarks primarily designed for fact retrieval or the assessment of short-text understanding, the proposed benchmark is specifically intended to evaluate long-text inferences. It therefore targets more complex levels of comprehension, including the integration of distributed information, the construction of a global text representation, and the generation of inferences across extended textual contexts. In this respect, the study contributes to the development of cognitively grounded approaches to evaluating understanding and reasoning and provides a foundation for future research.

The paper is organised as follows: Section 2 presents the goals and motivation of the research, followed by a review of related work in Section 3. Section 4 details the dataset design, including the selection of the text collection and the automated question generation process using LLMs, as well as the two-stage human review and validation process used to ensure logical consistency and linguistic adequacy. Section 5 outlines the dataset structure, the distribution of texts across domains and text types, and the distribution of questions by category. The experiments evaluating LLMs with the ReasoningBG benchmark are reported in Section 6, followed by a brief discussion of the results in Section 7 and a summary of the findings and directions for future work in Section 8.

2 Goals and motivation

Reading literacy has traditionally been conceptualised as a measure of human comprehension, with its most influential operationalisation provided by the Programme for International Student Assessment (PISA) (OECD, 2019). Rather than measuring factual recall alone, the PISA framework evaluates progressively more complex comprehension processes, including retrieving information, interpreting and integrating ideas, and evaluating and reflecting on textual content (OECD, 2019). These processes closely correspond to cognitive theories of text comprehension, such as Kintsch's construction–integration model, in which readers incrementally construct coherent mental representations by integrating explicit information with inference and prior context (Kintsch, 1998).

Although PISA was developed to assess human reading proficiency, its cognitive organisation is not specific to human learners. The three broad reading-process dimensions describe increasingly demanding forms of reasoning over text and therefore provide a shared conceptual framework for comparing human readers and large language models. Instead of treating reading comprehension as simple question answering, the PISA taxonomy distinguishes between different levels of reasoning, from locating explicitly stated information to integrating evidence distributed across a document and evaluating competing interpretations.

The PISA categories are used here as task design and analysis categories rather than as claims about the cognitive mechanisms employed by LLMs. In particular, successful performance on an inference-oriented item does not establish that a model performs inference in the same way as a human reader. The benchmark therefore evaluates observable task performance under different comprehension demands, while leaving the underlying mechanisms of model behaviour as an open empirical question.

The benchmark organises questions according to progressively more demanding comprehension processes, enabling systematic comparison across language models and, potentially, with human readers using the same cognitive framework. Rather than serving as a direct replication of the PISA assessment, the benchmark transfers its theoretical organisation to the evaluation of LLM reasoning, providing a common reference point for analysing similarities and differences between human and artificial text comprehension.

While PISA operationalises the cognitive processes of retrieving information, interpreting and integrating, and evaluating and reflecting on content (OECD, 2019) to differentiate human proficiency levels, this benchmark adapts this hierarchical approach to evaluate LLM performance on text-comprehension tasks organised according to PISA-inspired cognitive categories in the context of Bulgarian.

The benchmark is designed to include tasks that go beyond direct information retrieval, such as interpretation, causal reasoning, comparison, conditional reasoning, and text-based inference. These labels describe the requirements imposed by the questions and should not be taken as evidence that the models employ formal logical reasoning or human-like cognitive processes.

By mapping AI performance onto pedagogically grounded categories, the benchmark provides a robust framework for investigating whether performance varies systematically across tasks that place different demands on text comprehension, or instead reflects similar surface-level or statistical strategies across different question types.

3 Related work

Recent advances in LLMs have demonstrated substantial progress in common-sense reasoning and mathematical tasks (Chen et al., 2024; Wang et al., 2026a,b). While established benchmarks such as SQuAD (Rajpurkar et al., 2016), BIG-bench (Srivastava et al., 2023), and HellaSwag (Zellers et al., 2019) are widely used to evaluate machine reading comprehension and logical prediction, they often lack systematic alignment with structured pedagogical frameworks such as PISA (OECD, 2019). There have been only limited efforts to create examination datasets and benchmarks, for example RACE (Lai et al., 2017), that resemble a standardised school examination setting suitable for use as an LLM benchmark.

Existing evaluation tools such as TruthfulQA (Lin et al., 2022) focus primarily on factual accuracy rather than deep comprehension of extended texts. A complementary line of work focuses on educational assessment. Resources such as PISA-Bench (Haller et al., 2025), semi-automatic scoring frameworks such as Omethi (Zehner et al., 2025), and AI-based scoring methodologies (Okubo et al., 2023; Šindelář et al., 2026) have been developed to address the complexities of large-scale educational assessments and reduce human bias in scoring. Unlike general NLP benchmarks, these resources are grounded in educational theory; however, they focus primarily on assessment methodology and scoring rather than on constructing native-language benchmarks for evaluating the reading comprehension abilities of large language models.

The evaluation of reading comprehension in LLMs has evolved considerably beyond simple fact extraction towards deeper forms of understanding. The focus on comprehension of longer texts has led to the development of the QuALITY benchmark (Pang et al., 2022), with multiple-choice questions over passages of roughly 5,000 tokens; NarrativeQA (Kočiský et al., 2018), with questions on full novels and film scripts; and SCROLLS and ZeroSCROLLS (Shaham et al., 2022, 2023), which aggregate summarisation, question answering, and natural language inference tasks across domains ranging from literature to legal texts. Alongside text length, there has been growing interest in multi-step and multi-document reasoning, for example the LongBench benchmark (Bai et al., 2024).

For the Bulgarian language, there have been several evaluation datasets, for example MMLU-bgeval, Winogrande-bgeval, and GSM8k-bgeval (Alexandrov et al., 2024), EXAMS (Hardalov et al., 2020), BgGLUE (Hardalov et al., 2023), and MMLU-X (Thellmann et al., 2024). While these resources are crucial for assessing general knowledge, common sense, and mathematical reasoning in a Bulgarian context, they are largely (automatically) translated and adapted from English or other language benchmarks and do not specifically target language understanding with respect to hierarchical cognitive processes (Kintsch, 1998; OECD, 2019).

An important recent addition is MMLU-BG (Koeva et al., 2026), a native Bulgarian benchmark for large-scale multitask language understanding. The benchmark presented in this paper builds on this line of research by shifting the evaluation focus from multitask knowledge assessment to document-level reasoning, grounded in well-established cognitive and educational frameworks. Together, these resources constitute important infrastructure for Bulgarian LLM evaluation. However, none is specifically designed to assess reading comprehension through hierarchical cognitive processes over extended texts.

Recent advances in LLMs have also enabled the scalable generation of educational reading-comprehension datasets. Zou et al. (2025) introduce DocBench, a benchmark for end-to-end document-reading systems constructed through a hybrid pipeline combining human-written and synthetically generated questions over multi-page documents. Säuberli and Clematide (2024) propose an evaluation protocol, including the *text informativity* metric, for assessing the quality of LLM-generated multiple-choice items, demonstrating that current language models can generate educational questions of acceptable quality for PISA-style assessment. At the same time, several studies have identified limitations of LLM-generated reading-comprehension data. Neeman et al. (2024) show that language models frequently rely on internalised parametric knowledge rather than on the supplied document when the two conflict, po-

tentially compromising the validity of reading-comprehension evaluation. Similarly, Jain et al. (2025) demonstrate that although GPT-4o and o1 can estimate question difficulty in a manner correlated with Item Response Theory parameters, their performance remains insufficient for fully automated assessment design.

These findings highlight the limitations of fully automatic benchmark construction. LLM-generated questions may suffer from factual inconsistencies, ambiguity, and weak cognitive validity, allowing models to answer correctly through superficial lexical matching rather than genuine comprehension (Lin et al., 2022; Srivastava et al., 2023). Consequently, expert human validation remains an essential component of educational benchmark development (Zehner et al., 2025). Hybrid LLM–human pipelines combine the scalability of language models with the methodological control provided by domain experts, ensuring both linguistic quality and pedagogical validity (Haller et al., 2025).

Taken together, existing resources address complementary aspects of reading comprehension, educational assessment, multilingual evaluation, and automatic dataset construction. However, to our knowledge, there is currently no benchmark that simultaneously provides (i) a native Bulgarian resource, (ii) document-level reading comprehension over extended texts, (iii) explicit alignment with the cognitive processes underlying the PISA framework, and (iv) a hybrid LLM–human annotation methodology designed specifically for evaluating the reading-comprehension capabilities of large language models. The benchmark presented in this work is intended to fill this gap.

4 Benchmark design

4.1 Collection of texts

To facilitate the automated generation of assessment items, a specialised dataset of 232 texts was compiled. The collection focuses exclusively on text data drawn from a single source continuous text, composed of sentences organised into paragraphs. Materials containing graphics, tables or diagrams were excluded.

The initial phase of data collection yielded over 300 candidate texts, with an average length of 3,000–4,000 words. To ensure data integrity and avoid prior exposure to LLMs, only offline materials not indexed in public digital repositories were selected.

All sources were converted into a machine-readable format through scanning, OCR processing and manual correction to remove noise and ensure accuracy. The texts were orthographically normalised for consistent spelling and punctuation and stored in plain-text format for computational use. Sections that were overly long or complex were shortened while preserving semantic integrity. After manual verification, the corpus was refined to 232 texts. Approximately one third of the initial samples were excluded due to:

1. Lexical complexity: Texts containing an excessive density of specialised technical terminology were omitted.

2. Structural and technical constraints: A subset of materials was excluded due to high technical complexity in the OCR recovery process or structural irregularities that compromised standardisation of the dataset.

3. Length optimisation: Texts that significantly deviated from the average word count or were deemed excessively long for standard assessment windows were removed to ensure homogeneity of the dataset.

4.2 Text typology and classification

To ensure diversity in discourse characteristics while maintaining a coherent domain, the dataset was organised according to the primary communicative purpose and source of the texts. The categorisation reflects the intended use of the documents rather than a predefined educational taxonomy and supports the construction of questions targeting different comprehension skills.

The corpus comprises two broad groups:

- Informational texts: Sourced from scientific articles and biographies, these texts are specifically designed to support questions targeting factual accuracy and temporal reasoning.

- Popular science texts: Derived from magazines and encyclopaedias, these connected prose pieces facilitate questions focused on cause-and-effect relationships and explanatory understanding.

Although this functional categorisation guided corpus construction, it does not directly correspond to internationally established reading assessment

frameworks. Therefore, for analysis and comparison purposes, all texts were also mapped to some of the general PISA text categories. The PISA framework distinguishes texts according to their dominant rhetorical organisation rather than their source or topic, namely *Description*, *Narration*, *Exposition*, *Argumentation*, *Instruction*, *Transaction*. This secondary classification enables comparisons with previous reading comprehension studies and provides a standardised description of the discourse structures represented in the corpus.

As the present benchmark focuses exclusively on continuous prose intended for complex reading comprehension, texts whose primary function is procedural communication (*Instruction*) or interpersonal exchange (*Transaction*) were excluded from the dataset. The remaining texts were automatically mapped to the four applicable categories using an approximate classification based on the dominant rhetorical structure. The resulting distribution is: *Exposition* (36.32%), *Narration* (29.85%), *Description* (27.36%) and *Argumentation* (6.47%).

The assignment of texts to four categories reflects the dominant discourse function of each document. Many authentic texts contain mixed structures, so this categorisation should be regarded as an approximate mapping.

4.3 Generation of questions using large language models

The test questions were generated automatically using **Gemma-3-27B-IT**[1] (Google DeepMind, 2025), a high-performance, instruction-tuned 27B Transformer from the third-generation Gemma family. The model supports long-context processing, multilingual input, and hybrid-attention mechanisms, which enable efficient handling of extended texts. Its instruction-following capabilities and ability to generate coherent, context-aware outputs over long documents make it particularly suitable for automatic question generation. The model also provides structured outputs suitable for multi-step workflows. Despite its scale, the model is optimised for efficient deployment on modern GPUs and is released under the Gemma Terms of Use, which allow commercial application under defined safety constraints.

[1] https://huggingface.co/google/gemma-3-2B-it

4.4 Prompt design

Two prompts were used for the creation of multiple choice questions. Both prompts instruct the model to generate 15 questions with four possible answers, labeled A), B), C), D), with only one correct answer. The answers should not be explicitly stated in the text but should be logically derived from it.

The first prompt instructs the model to generate questions whose answers are not stated directly in the text, but can be logically inferred from it. The questions and answers generated by this type of prompt required minor editorial changes, such as improving the wording, refining misleading answers, removing ambiguities, or addressing instances where more than one answer was correct.

In the second prompt, the instruction is expanded to explicitly include rules from propositional logic (e.g. “modus ponens”, “modus tollens”, hypothetical syllogism, constructive dilemma, etc.), with the expectation that the model will generate questions requiring formal logical reasoning. In practice, however, the generated questions often adopt a logical “If … then …” structure and frequently introduce premises, conditions, or conclusions that are not supported by the original text. Many of these questions are based on hallucinations rather than valid logical inferences from the source text. As a result, only a limited number of questions generated by the second prompt were selected.

The generation process yielded two sets of comprehension questions for each text, resulting from Prompt 1 and Prompt 2.

4.5 Human review and validation

The automatic creation of the question set is followed by a two-stage human review process to ensure consistency, quality, and alignment with the source texts.

4.5.1 Stage 1: Question selection and editing

In the first stage, annotators review the automatically generated questions and select 10 items per text from an initial pool of 30, covering both general types of generated questions. Annotators are instructed to combine the results from the two prompt-generated sets and to choose 10 questions. The goal is to select difficult questions whose answers follow logically from the text, with only one correct option among the proposed answers.

The process involves careful rereading of the source texts to verify factual consistency, linguistic adequacy, and the relevance of questions and

answers; questions are edited, adapted, or rejected where necessary.

During evaluation and selection, preference is given to questions that (i) require drawing actual conclusions from the text, (ii) contain exactly one correct answer that can be derived from the source text, and (iii) demonstrate an appropriate level of difficulty. Since the first prompt consistently generates a higher proportion of questions meeting these criteria, most of the final questions originate from it, while only a small number of questions generated by the second prompt are retained after careful review and, where necessary, minor adjustments.

The automatically generated questions are edited to ensure clarity, accuracy, and readability. This includes correcting grammatical errors, spelling, and word order; improving lexical choice; and removing redundant or irrelevant elements. Questions are rejected if they are not inferable from the text, are excessively narrow or overly broad, duplicate the focus of other questions, or contain answers explicitly stated in the text without requiring interpretation.

4.5.2 Stage 2: Cross-annotation and refinement

In the second stage, the selected question–answer sets and corresponding texts are exchanged among annotators for further proofreading and refinement. The process includes inter-annotator agreement discussions to ensure consistent interpretation and evaluation criteria. Across both stages, all materials are reviewed and revised against the following validation criteria: clarity, correctness, alignment with the source text, absence of literal overlap with the answer, and appropriate cognitive level.

The four answer options for each question are revised to ensure that each question contained only one clearly correct answer, grounded in the source text. The editing process includes:

- resolving cases where multiple options appear partially correct,
- adding a correct answer when it is missing,
- balancing answer length and structure to prevent bias (e.g. avoiding overly detailed correct answers).

In some cases, annotators fully develop both the questions and all four answer options to meet the required quality standards. The incorrect answers are edited to include implicit information and partially true statements, testing models' ability to distinguish between accurate, misleading, and incomplete information.

5 Dataset

5.1 Classification

The overall observation from comparing the questions generated by the two prompts is that those derived from Prompt 1 are primarily grounded in explicit facts from the text and are more directly applicable and suitable for operational use. In contrast, the questions resulting from Prompt 2 require substantial restructuring and refinement, as they often introduce overly elaborate formulations or abstract criteria that are not immediately aligned with the intended assessment format.

To facilitate systematic analysis, the questions were broadly grouped into PISA-inspired task categories developed for this benchmark (Table 1).

The questions were classified into seven categories using a rule-based scheme applied to the question text in a fixed order, with Information retrieval questions serving as the default when no other pattern matches.

- **Information retrieval** — the default category for questions that retrieve information explicitly stated in the text, typically beginning with question words and not matching any higher-priority pattern.
- **Vocabulary and interpretation** — questions that ask about the meaning of a word, phrase, or abstract concept as used in the text.
- **Causal** — questions about reasons, causes, effects, or consequences.
- **Significance** — questions about the role, purpose, importance, or function of something.
- **Comparisons and relations** — questions that ask the reader to compare, contrast, or identify relationships between entities or concepts.
- **Conditional** — hypothetical questions that present a counterfactual or "if X, then what" scenario.
- **Inference** — questions that require reaching a conclusion not explicitly stated in the text, or evaluating a claim against the text.

Category	Features	Subcategories
Interpretative QA	who, when, where	Information retrieval, Vocabulary and interpretation
Abstractive QA	why, how, how do X and Y differ	Causal, Significance, Comparisons and relations
Inferential QA	if... then	Conditional, Inference

Table 1: Classification of question types by categories

These categories are vague and therefore assigned only approximately, and they do not strictly adhere to the formal PISA definitions. In particular, the feature markers shown in Table 1 served as heuristic cues during rule-based classification and do not exhaustively characterise each subcategory; inference questions, for instance, do not need to exhibit conditional syntax to qualify. The PISA-inspired categories should therefore be interpreted as properties of the assessment tasks rather than as direct measurements of internal cognitive processes.

5.2 Dataset structure

The dataset consists of 232 texts paired with a set of questions. The texts are stored in plain text (.txt) format, and the questions are in tab separated values (.tsv) format; each question contains: the question itself, the answers (A, B, C, D), and the correct answer (marked by the corresponding letter).

The following metadata presented in JSON format are stored for each text: **filenameText**, **filenameQuestions**, **title**, **author**, **source**, **domain**, **keywords**, **numberPar**, **numberSent**, and **numberWords** (number of paragraphs, sentences, and words, respectively), **validationStatus**.

Table 2 shows the statistics for the dataset and the texts' distribution across domains. The distribution across question categories is presented in Table 3.

6 Experiments

6.1 Description of selected LLMs

The LLMs used in the experiments span a range of model sizes and architectures, enabling meaningful comparison across models, particularly those that explicitly include or target Bulgarian during training.

Domain	#Texts	#Sents	#Words
Natural Sciences	60	2,535	40,462
Science & Technology	42	1,511	26,418
Literary Studies	38	1,292	26,658
History	37	1,279	25,312
Geography	23	1,108	19,816
Psychology	13	474	8,459
Culture	7	313	5,780
Medicine	7	253	4,347
Philosophy	5	206	4,103
Total	**232**	**8,971**	**161,355**

Table 2: Domain distribution of the dataset

Question Category	N	%	Avg/text
Information retrieval	1,057	45.6	4.56
Vocabulary and interpretation	64	2.8	0.28
Interpretative	1,121	48.3	4.83
Causal	255	11.0	1.10
Significance	104	4.5	0.45
Comparisons and relations	223	9.6	0.96
Abstractive	582	25.1	2.51
Conditional	114	4.9	0.49
Inference	503	21.7	2.17
Inferential	617	26.6	2.66
Total	**2,320**	**100.0**	**10.00**

Table 3: Distribution of the question according to question categories

Gemma-4-31B-IT[2] is an instruction-tuned large language model (released in 2026) with 31 billion parameters and a 256K token context window, supporting over 140 languages, including but not targeting Bulgarian.

BgGPT-Gemma-2-9B-IT-v1.0[3] is a Bulgarian-targeted instruction-tuned LLM (released in 2024) with 9 billion parameters, developed by INSAIT and continually pretrained on approximately 85 billion Bulgarian tokens on top of Google's Gemma 2 base model.

[2] https://huggingface.co/google/gemma-4-31b-it

[3] https://huggingface.co/INSAIT-Institute/BgGPT-Gemma-2-9B-IT-v1.0

BgGPT-Gemma-3-27B-IT[4] is a Bulgarian-targeted instruction-tuned LLM (released in 2025) with 27 billion parameters and a 131K token context window, developed by INSAIT through continual pretraining on Bulgarian data on top of Google's Gemma 3 base model.

Qwen2.5-14B-Instruct[5] is an instruction-tuned LLM (released in 2024) with 14 billion parameters and a 128K-token context window, developed by Alibaba Cloud and supporting more than 29 languages, without mentioning Bulgarian in its training data.

6.2 Evaluation environment and metrics

The evaluation is conducted on an NVIDIA A100 GPU with 80 GB of VRAM. The temperature is fixed at 0.7 across all models. The software is written in Python and is publicly available.[6] The language models are all programmatically downloaded from HuggingFace.

Model performance is evaluated by comparing each prediction with the manually determined answer in the dataset. The evaluation metric is binary accuracy, calculated as the proportion of instances in which the model's decision matches the human answer, out of the total number of answers.

6.3 Results

Testing large language models on the benchmark reveals pronounced performance differences across models.

Gemma-3-27B-IT, which is used for the generation of the dataset, achieves 91.5% accuracy across all text types. It is used in the analysis for comparative purposes.

The low scores of the newer **Gemma-4-31b-IT** are due to its failure to process the instruction format used in the evaluation experiments, which led to repeated response formatting failures (repetition of the question, repeatedly listing the answers, looping, hallucinations, etc.). This calls for a reconsideration of the experimental protocols in future applications.

BgGPT-Gemma-3-27B-IT also achieves high accuracy (87.84%), representing a negligible improvement on the previous, smaller version, **BgGPT-Gemma-2-9B-IT**, which scored 86.33%. The lower score of BgGPT-Gemma-3 relative to Gemma-3 motivates further controlled experiments investigating the effect of Bulgarian continual pretraining.

The strongest performer is **Qwen2.5-14B-Instruct** with 93.22%, despite having fewer parameters and only including Bulgarian as part of its multilingual support. Qwen2.5 was trained with an emphasis on instruction following and reasoning (Qwen Team, 2025), and its competitive performance on the multiple-choice task, despite the absence of Bulgarian-targeted training, may reflect strong general instruction-following calibration rather than Bulgarian language understanding.

It is worth noting that model performance is quite consistent across text types. Most models perform best on argumentation texts, except for BgGPT-Gemma-2-9B-IT-v1.0, which scores slightly better on exposition texts.

7 Discussion

The highest scores are achieved by the multilingual model Qwen2.5.The observed performance differences are consistent with the possibility that multilingual pre-training contributes to performance, but the present comparison does not permit causal attribution because model architecture, size, pretraining data, instruction tuning and language-specific training vary simultaneously.

Performance is broadly consistent across text types, with some models showing their strongest results on Argumentation texts and others showing their weakest on Narration. This may reflect the nature of the question types associated with each genre: Argumentation texts prompt more Inference questions, which require the model to evaluate claims against the text – a task that aligns well with the instruction-following training of LLMs. Narration texts, by contrast, tend towards information retrieval and causal questions anchored in specific narrative details, where a model lacking deep familiarity with Bulgarian cultural and historical context may be at a disadvantage.

Across question categories, variation within models is small, suggesting that the distinctions in cognitive demand encoded in the taxonomy (Causal, Inference, etc.) do not translate straightforwardly into differing levels of difficulty for LLMs. This contrasts with findings from human reading comprehension studies, where question type is a re-

[4] https://huggingface.co/INSAIT-Institute/BgGPT-Gemma-3-27B-IT

[5] https://huggingface.co/Qwen/Qwen2.5-14B-Instruct

[6] https://github.com/DCL-IBL/BG-Benchmark-Evaluation

Model	Type of text				
	Overall	Argumentation	Description	Exposition	Narration
Gemma-4-31B-IT	33.37	36.15	36.00	32.33	31.55
BgGPT-Gemma-2-9B-IT	86.33	85.38	85.64	87.81	85.34
BgGPT-Gemma-3-27B-IT	87.84	92.31	88.55	87.67	86.38
Qwen2.5-14B-Instruct	**93.22**	**96.15**	**93.27**	**93.56**	**92.07**

Table 4: Models' accuracy by text type. The highest observed accuracy for each text type (vertical maximum) is shown in bold. The results for the text type with the best performance for each model (horizontal maximum) are underlined.

Model	Question type						
	Fact	Voc&Int	Cause	Signif	Comp&Rel	Cond	Infer
Gemma-4-31B-IT	34.2	34.9	30.6	31.9	35.2	29.6	33.3
BgGPT-Gemma-2-9B-IT	87.0	86.0	88.6	89.4	83.2	82.4	85.6
BgGPT-Gemma-3-27B-IT	87.5	93.0	88.6	87.2	90.3	89.8	86.1
Qwen2.5-14B-Instruct	**93.8**	**95.3**	**94.1**	**90.4**	**91.8**	**91.7**	**92.9**

Table 5: Accuracy per question category. The model performing best on each question type (vertical highest) is marked in bold. The question category on which each model achieves its highest accuracy (horizontal highest) is underlined.

liable predictor of difficulty (OECD, 2019; Day and suk Park, 2005). One possible explanation is that the models may rely on broadly similar patterns of text–question–option association across different question types, rather than showing clearly distinguishable performance patterns across the reasoning demands represented in the benchmark. However, the present results do not allow us to determine the underlying mechanisms responsible for these patterns.

The dataset is available on GitHub: https://github.com/DCL-IBL/Reasoning-BG-Dataset and HuggingFace: https://huggingface.co/datasets/DCL-IBL/Reasoning-BG-Dataset.

8 Conclusions and future work

Reasoning-BG is a benchmark for measuring how accurately LLMs solve Bulgarian long-text comprehension tasks designed around different, PISA-inspired comprehension demands. It addresses a critical gap in Bulgarian AI benchmarks by shifting the focus from simple factual recall to more demanding forms of text comprehension, including interpretation, integration, and inference.

The main contribution of this work is the development of a human-validated, LLM-assisted dataset through a hybrid human–AI pipeline that combines the scalability of large language models such as Gemma-3-27B-IT with two-stage evaluation by human experts.

This dataset provides a foundation for future research on human–AI comparison and educational assessment tools in the Bulgarian context. There is considerable potential to expand it by adding more texts and designing increasingly complex reasoning tasks.

In addition, the dataset could be used to train LLMs to compile resources for teaching reading literacy and exam preparation.

Error analysis can, on the one hand, help improve dataset quality, particularly by comparing incorrect outputs from multiple models, and, on the other hand, be used to explore differences in the answer-generation behaviour of diverse models and in their approaches to different types of questions. In future work, more complex reasoning tasks can also be designed to test the capabilities of LLMs.

Acknowledgments

The present study is carried out within the framework of the project Infrastructure for Fine-tuning Pre-trained Large Language Models, Grant Agreement No. ПВУ–55 from 12.12.2024 /BG-RRP-2.017-0030-C01/.

References

Anton Alexandrov, Veselin Raychev, Dimitar I. Dimitrov, Ce Zhang, Martin Vechev, and Kristina Toutanova. 2024. BgGPT 1.0: Extending English-centric LLMs to other languages. ArXiv: 2412.10893.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3119–3137, Bangkok, Thailand. Association for Computational Linguistics.

Ruirui Chen, Weifeng Jiang, Chengwei Qin, Ishaan Singh Rawal, Cheston Tan, Dongkyu Choi, Bo Xiong, and Bo Ai. 2024. LLM-based multi-hop question answering with knowledge graph integration in evolving environments. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14438–14451, Miami, Florida, USA. Association for Computational Linguistics.

Richard R. Day and Jeong suk Park. 2005. Developing reading comprehension questions. *Reading in a Foreign Language*, 17(1):60–73.

Google DeepMind. 2025. Gemma 3 27B-IT. `https://huggingface.co/google/gemma-3-27b-it`.

Patrick Haller, Fabio Barth, Jonas Golde, Georg Rehm, and Alan Akbik. 2025. PISA-Bench: The PISA Index as a multilingual and multimodal metric for the evaluation of vision-language models. ArXiv: 2510.24792.

Momchil Hardalov, Pepa Atanasova, Todor Mihaylov, Galia Angelova, Kiril Simov, Petya Osenova, Veselin Stoyanov, Ivan Koychev, Preslav Nakov, and Dragomir Radev. 2023. bgGLUE: A Bulgarian General Language Understanding Evaluation Benchmark. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8733–8759, Toronto, Canada. Association for Computational Linguistics.

Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2020. EXAMS: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5427–5444. Association for Computational Linguistics.

Yoshee Jain, John Hollander, Amber He, Sunny Tang, Liang Zhang, and John Sabatini. 2025. Exploring the potential of large language models for estimating the reading comprehension question difficulty. In *Adaptive Instructional Systems*, pages 202–213. Springer Nature Switzerland.

Walter Kintsch. 1998. *Comprehension: A Paradigm for Cognition*. Cambridge University Press.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Svetla Koeva, Ivelina Stoyanova, Dimiter Georgiev, Svetlozara Leseva, Valentina Stefanova, Maria Todorova, Tsvetana Dimitrova, Hristina Kukova, Mihaela Moskova, and Tinko Tinchev. 2026. Bulgarian Massive Multitask Language Understanding Benchmark. In *Proceedings of the 2026 Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Ori Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2024. LLMs' reading comprehension is affected by parametric knowledge and struggles with hypothetical statements. ArXiv: 2404.06283.

OECD. 2019. *PISA 2018 Assessment and Analytical Framework: Reading, Mathematics and Science*. OECD Publishing, Paris, France. ISBN: 978-92-64-94031-4.

Tomoya Okubo, Wayne Houlden, Paul Montuoro, Nate Reinertsen, Chi Sum Tse, and Tanja Bastianić. 2023. *AI Scoring for International Large-Scale Assessments Using a Deep Learning Model and Multilingual Data*, volume 287 of *OECD Education Working Papers Series*. OECD Publishing, Paris, France.

Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, and Samuel Bowman. 2022. QuALITY: Question answering with long input texts, yes! In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5336–5358, Seattle, United States. Association for Computational Linguistics.

Qwen Team. 2025. Qwen2.5 technical report. ArXiv: 2412.15115.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Andreas Säuberli and Simon Clematide. 2024. Automatic generation and evaluation of reading comprehension test items with large language models. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, Mexico City, Mexico. Association for Computational Linguistics. ArXiv: 2404.07720.

Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. 2023. ZeroSCROLLS: A zero-shot benchmark for long text understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7977–7989, Singapore. Association for Computational Linguistics.

Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. 2022. SCROLLS: Standardised comparison over long language sequences. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 12007–12021, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pavel Šindelář, Dávid Slivka, Christopher Bouma, Filip Prášil, and Ondřej Bojar. 2026. Training data generation for context-dependent rubric-based short answer grading. ArXiv: 2603.28537.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, and et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. ArXiv: 2206.04615.

Klaudia Thellmann, Bernhard Stadler, Michael Fromm, Jasper Schulze Buschhoff, Alex Jude, Fabio Barth, Johannes Leveling, Nicolas Flores-Herr, Joachim Köhler, René Jäkel, and Mehdi Ali. 2024. Towards cross-lingual LLM evaluation for European languages. ArXiv: 2410.08928.

Peng-Yuan Wang, Tian-Shuo Liu, Chenyang Wang, Ziniu Li, Yidi Wang, Shu Yan, Chengxing Jia, Xu-Hui Liu, Xinwei Chen, Jiacheng Xu, and Yang Yu. 2026a. A survey on large language models for mathematical reasoning. *ACM Comput. Surv.*, 58(8).

Yi Wang, Yang Hu, and Jiajun Nie. 2026b. Knowledge graph-based chain-of-thought reasoning framework for complex questions. In *Proceedings of the International Conference on Artificial Intelligence and Advanced Autonomous Systems (AIAAS 2026)*, page 47–54, New York, NY, USA. Association for Computing Machinery.

Fabian Zehner, Hyo Jeong Shin, Emily Kerzabi, Andrea Horbach, Sebastian Gombert, Frank Goldhammer, Torsten Zesch, and Nico Andersen. 2025. Down the cascades of Omethi: Hierarchical automatic scoring in large-scale assessments. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 660–671, Vienna, Austria. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Anni Zou, Wenhao Yu, Hongming Zhang, Kaixin Ma, Deng Cai, Zhuosheng Zhang, Hai Zhao, and Dong Yu. 2025. DocBench: A benchmark for evaluating LLM-based document reading systems. In *Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing*, pages 359–373, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Detecting Pseudoscientific Narratives in Bulgarian Science Communication: Nationalistic Mystification, Pseudo-Expertise and Cultural Blind Spots in AI-Assisted Analysis

Ruslana Margova
GATE, Sofia University "St. Kliment Ohridski"
5 James Bourchier Blvd., Sofia, Bulgaria
ruslana.margova@gate-ai.eu

Abstract

Science-related communication increasingly could serve not only as a channel for disseminating scientific knowledge but also as a vehicle for pseudoscientific and misinformation narratives. This study investigates the prevalence and thematic structure of pseudoscientific content in Bulgarian science-related media discourse through a hybrid computational approach that combines lexicon-based classification, statistical lexical analysis, and thematic grouping assisted by the Large Language Model (LLM). Two datasets were analysed: a large dataset of 17,833 Bulgarian social media posts collected through CrowdTangle and a smaller of 498 science-related news items retrieved via PyGoogleNews. Using TF-IDF extraction, n-gram analysis, domain-specific lexical resources, and thematic categorisation assisted by ChatGPT-4o, we identified the dominant pseudoscientific narratives circulating in the two communication environments. The results indicate that 17.68% of the analysed social media posts contain lexical markers associated with pseudoscientific or misinformation-related content. The dominant narrative category, *"Bulgaria Above All – The Most Ancient"*, accounts for more than 55% of the identified pseudoscientific content, suggesting a strong association between pseudoscientific discourse and identity-based narratives. The analysis reveals the recurrent appropriation of scientific terminology within astrological and esoteric contexts as pseudo-expertise. A comparison between social media and news-media corpora demonstrates substantial differences in the dominant forms of pseudoscientific communication. The comparison between ChatGPT-4o and the lexicon-based approach shows considerable overlap in the identification of major thematic categories, while also revealing limitations of general-purpose LLMs in detecting culturally specific narratives. The findings highlight the complementary value of LLM-assisted and domain-specific NLP methods for misinformation research and provide a foundation for future media-literacy and prebunking interventions in the Bulgarian context.

Keywords: disinformation, science, communication, Bulgarian media, inoculation, pseudoscience, news

1 Motivation and research questions

Science communication plays an increasingly important role in contemporary society, both as a mechanism for disseminating scientific knowledge and as a means of fostering public trust in research. At the same time, the growing visibility of science-related content in digital media has created opportunities for the spread of pseudoscientific claims that imitate the language and authority of science while lacking empirical support. This study does not focus on problems of scientific misconduct within research itself, which are partially addressed through open-science initiatives and the FAIR[1] (Findable, Accessible, Interoperable, and Reusable) principles. Rather, it examines how science is represented in the media environment and to what extent pseudoscientific narratives are embedded within science-related communication. Here, pseudoscientific communication refers to science-related content that lacks empirical support while imitating the language, terminology, or authority of science (Ramalho and Marinho, 2026),(Frahm, 2026). This understanding is contrasted with evidence-based science communication grounded in verifiable information and established scientific methods. Thus, pseudoscientific elements are understood as lexical, semantic, or narrative features that present claims lacking empirical support while adopting scientific terminology,

[1] https://www.go-fair.org/fair-principles/

authority, or explanatory frameworks. Such elements may include references to esoteric concepts, alternative medical claims, pseudo-historical interpretations, conspiracy narratives, or other forms of content that imitate scientific discourse without adhering to established scientific standards. We address three research questions: (1) what proportion of science-related media content contains pseudoscientific elements; (2) what thematic patterns characterise such content; and (3) to what extent can a Large Language Model (ChatGPT-4o) reproduce or complement human-defined thematic classifications. To answer these questions, two datasets are analysed. The first is collected through CrowdTangle and consists primarily of social-media content, while the second corpus is gathered through PyGoogleNews and contains science-related news articles from Bulgarian media sources. Following data preprocessing, a hybrid analytical framework was applied, combining lexical extraction, rule-based classification, LLM-assisted thematic clustering, and human validation.

2 Related work

Popper emphasised the criterion of falsifiability to distinguish science from non-science, using astrology and psychoanalysis as examples of pseudoscience and Einstein's theory of relativity as an example of science (Popper, 2005). We will accept that in our case, this is a statement that claims to be scientific or factual, but without evidence. Without going into the definitions of disinformation and misinformation, we accept those of the European Commission[2]. The question about definitions in the field of information integrity, including the definition of pseudoscience is on the top (Hansson, 2013; Dahms, 2022). Some researchers try to show how the study of pseudoscience can teach us something about science, ourselves, and society, and the whole contemporary time (Tuboly, 2025). Others warn that by echoing disinformation, some media outlets, often willing accomplices in conflating denial and scepticism, amplify manufactured controversies and cast growing doubt upon scientific credibility (Schiele, 2020). Several Bulgarian scholars are engaged in the automatic detection of disinformation (Dinkov et al., 2019; Nakov, 2020; Nakov et al., 2021; Temnikova et al., 2023, 2025; Margova, 2025).

[2] https://commission.europa.eu/

3 Data

To ensure a comprehensive analysis of science-related narratives and pseudoscientific discourse in Bulgaria, we utilise two distinct datasets: (1) a CrowdTangle dataset and (2) a PyGoogleNews dataset. The primary and larger dataset was retrieved via CrowdTangle, a public insights platform developed by Meta. The CrowdTangle dataset spans a ten-year period (2014–2024) and consists of 20,000 public Facebook posts retrieved using the Bulgarian keyword *наука* ("science") from Bulgarian public pages, media outlets and groups. Following preprocessing, which included the removal of 2,167 duplicate entries and empty records, the final corpus comprised 17,833 unique posts. All subsequent analyses, including TF-IDF extraction, lexicon construction, rule-based classification, and LLM-assisted thematic grouping, were performed on this cleaned corpus. This dataset serves as the basis for analysing high-engagement public discourse and the circulation of pseudoscientific narratives on social media. The second dataset contains 498 science-related news items from the beginning of 2026 collected through PyGoogleNews, a Python wrapper for the Google News RSS feed. The tool was used to retrieve news articles and metadata from Bulgarian mainstream and digital news outlets, enabling a comparative analysis between social media and professional news content.

Both datasets underwent an identical preprocessing pipeline consisting of deduplication, URL removal, whitespace normalisation, and noise reduction. Additional filtering included the removal of stop words and highly frequent non-discriminative terms (e.g., science, many, Bulgaria) in order to improve the interpretability of the extracted lexical patterns. The code and derived lexical inventories are available in an online repository. The raw CrowdTangle data are not redistributed due to platform restrictions.

4 Methodology

The study employed a five-stage hybrid methodology combining statistical lexical extraction, rule-based lexical identification, LLM-assisted thematic grouping, expert qualitative validation, and comparative corpus analysis. The objective was not to develop a supervised

classifier, but to identify and analyse recurrent pseudoscientific narratives in Bulgarian science-related communication through complementary quantitative and qualitative methods.

First, TF-IDF (Term Frequency–Inverse Document Frequency) analysis was applied to the cleaned CrowdTangle corpus containing 17,833 unique posts. TF-IDF was selected because the objective was to identify statistically salient lexical markers that characterise recurrent themes in Bulgarian science-related online discourse while reducing the influence of highly frequent vocabulary. Unlike simple word-frequency counts, TF-IDF highlights lexical items that are statistically distinctive within the analysed corpus.

The analysis was implemented using a custom Python workflow incorporating text cleaning, linguistic normalisation through light rule-based suffix stripping, iterative stop-word filtering, and unigram–bigram extraction. The TF-IDF vectoriser was configured to retain the 100 highest-ranked lexical items and n-grams, providing a compact inventory of statistically salient lexical markers. This threshold was selected as a compromise between lexical coverage and interpretability during the subsequent qualitative analysis. The resulting ranked lexical inventory is presented in Appendix B. Because Bulgarian is a morphologically rich language, rule-based suffix stripping may occasionally introduce lexical ambiguity (e.g., the stem меди- may refer to both медицина “medicine” and медии “media”). To minimise such ambiguities, the extracted lexical inventory was subsequently reviewed during the expert validation stage.

Second, the ranked TF-IDF inventory served as the starting point for constructing a domain-specific pseudoscientific lexicon. The statistically derived lexical markers were subsequently expanded through qualitative analysis of Bulgarian online discourse in order to include semantically important markers that may occur infrequently but nevertheless characterise recurrent pseudoscientific narratives. Additional lexical roots and semantic markers were included when they appeared consistently across multiple pseudoscientific contexts and were assessed by the authors as representative of recurring misinformation themes. The resulting lexicon covers pseudoscientific, conspiratorial, esoteric, alternative-medical, and pseudo-historical discourse.

Third, the completed lexicon was employed in a rule-based identification procedure. Documents containing one or more lexical markers from the compiled lexicon were automatically classified as belonging to the pseudoscientific subset. The classification was non-exclusive; individual posts could be assigned to multiple thematic categories whenever they contained lexical markers associated with more than one narrative.

Because the proposed methodology relies on a rule-based lexicon rather than a supervised machine-learning classifier, standard evaluation metrics such as precision, recall, and F1-score were not computed, as no manually annotated gold-standard corpus was available. Instead, the resulting lexical inventory and thematic assignments were reviewed and refined by the authors through expert validation to ensure their semantic consistency and relevance. This procedure identified 3,152 posts (17.68% of the cleaned CrowdTangle corpus) containing one or more lexical markers associated with pseudoscientific discourse.

Fourth, the extracted lexical markers were organised into broader thematic categories using ChatGPT-4o as an exploratory thematic grouping tool. The model received the extracted lexical markers together with candidate thematic terms derived from the corpus and was instructed to identify the dominant themes using the Bulgarian prompt: „Класирай по основните теми“ (“Classify according to the main themes”).

No prompt engineering or parameter optimisation was performed because the objective was to evaluate whether a general-purpose Large Language Model could independently recover thematic structures comparable to those obtained through the rule-based lexical approach. The thematic groups produced by ChatGPT-4o were subsequently reviewed, consolidated where necessary, and interpreted by the authors. This procedure does not constitute probabilistic topic modelling (e.g., LDA, NMF, or BERTopic), but rather an LLM-assisted exploratory thematic grouping approach followed by expert qualitative interpretation.

Fifth, the thematic structure identified in the CrowdTangle corpus was compared with that obtained from the PyGoogleNews corpus in order to examine differences between pseudoscientific narratives circulating in Bulgarian social media and science-related reporting published by Bulgarian news outlets. In parallel, the thematic categories generated by ChatGPT-4o were compared with

those obtained through the rule-based lexical approach to assess the extent to which a general-purpose LLM could reproduce culturally specific patterns of pseudoscientific discourse.
The analytical workflow consisted of five consecutive stages:

- (1) TF-IDF extraction of statistically salient lexical markers from the cleaned CrowdTangle corpus;
- (2) construction and qualitative expansion of a domain-specific pseudoscientific lexicon;
- (3) rule-based identification of pseudoscientific content;
- (4) LLM-assisted thematic grouping followed by expert validation;
- (5) comparative analysis of both communication environments (CrowdTangle and PyGoogleNews) and thematic classifications (rule-based versus ChatGPT-4o).

ChatGPT-4o was accessed through the standard OpenAI interface in April 2026.

5 Results

5.1 Rule-Based Identification of Pseudoscientific Content

Applying the rule-based lexical identification procedure to the cleaned CrowdTangle corpus resulted in the identification of **3,152 posts** containing one or more lexical markers associated with pseudoscientific discourse. This corresponds to **17.68%** of the analysed corpus, while the remaining **82.32%** of the posts did not contain lexical indicators included in the domain-specific pseudoscientific lexicon.

The identified pseudoscientific subset was subsequently used for thematic analysis. Because individual posts frequently contained lexical markers associated with multiple narratives, thematic assignment was non-exclusive. Consequently, the percentages reported in Table 1 exceed 100%.

5.2 Dominant Thematic Categories

The thematic distribution of the identified pseudoscientific posts is presented in Table 1. The dominant category is *Bulgaria Above All – The Most Ancient*, accounting for 55.27% of the analysed pseudoscientific subset. This is followed by *Energy & Quantum Pseudoscience* (40.32%), *Health & Alternative Medicine* (34.49%), and *Conspiracies & Hidden Truth* (32.42%).

The predominance of the first category suggests a strong association between pseudoscientific discourse and narratives emphasising national identity, historical exceptionalism, and Bulgarian scientific achievements. Representative examples include headlines such as *"Bulgarian scientist makes world-shattering discovery in quantum physics, resolving a 100-year-old dispute between Einstein, Bohr, and Schrödinger"* and *"US scientists claim the 'Bulgarian Bacillus' could revolutionize mental health."* These examples illustrate how references to Bulgarian history, archaeology, or scientific achievement are combined with exaggerated or weakly substantiated claims.

The category *Energy & Quantum Pseudoscience* demonstrates the frequent appropriation of scientific terminology such as *quantum*, *energy*, and *frequency* within non-scientific explanatory frameworks. Likewise, the category *Health & Alternative Medicine* is characterised by emotionally oriented vocabulary associated with health, disease, healing, and everyday life.

Another notable observation is the recurrence of the lexical marker *horoscope* within both the *Energy & Quantum Pseudoscience* and *Cosmic Mysteries* categories, indicating a substantial overlap between astrological narratives and the use of scientific vocabulary.

5.3 Comparison Between Social Media and News Media

The thematic distribution observed in the CrowdTangle corpus differs substantially from that identified in the PyGoogleNews corpus. While the social media corpus is dominated by the narrative *Bulgaria Above All – The Most Ancient*, this category is almost absent from the news corpus. Instead, the PyGoogleNews dataset contains comparatively more content related to horoscopes, astrology, and sensationalised reporting on astronomy and space exploration.

The lexical profiles of the two datasets also differ considerably. Social media posts frequently employ emotionally charged vocabulary (e.g., *life*, *people*), whereas the news corpus generally exhibits a more neutral and informational style. Because the

Thematic Narrative	Share (%)	Representative Lexical Markers
Bulgaria Above All – Most Ancient	55.27%	education, Europe, energy, history, life
Energy & Quantum Pseudoscience	40.32%	life, horoscope, simple, more, "you can"
Health & Alternative Medicine	34.49%	medicine, health, diseases, heals, energy
Conspiracies & Hidden Truth	32.42%	people, history, facts, so much, then
Cosmic Mysteries	18.69%	life, people, horoscope, Earth, Sun
Ancient Knowledge & Archaeology	13.86%	history, Bulgarians, Bulgarian, life, horoscope
Uncategorized	8.66%	N/A

Table 1: Distribution of pseudoscientific narratives and their representative lexical markers. The 17.68% figure reflects posts identified as pseudoscientific through rule-based lexical classification. Percentages exceed 100% because individual posts may belong to multiple thematic categories simultaneously.

CrowdTangle corpus includes content produced both by professional media outlets and by other public actors, this comparison should be interpreted as a comparison between two communication environments rather than between professional journalism and social media alone.

The comparison indicates that pseudoscientific communication assumes different forms across the two datasets. Nationalist and identity-based narratives are considerably more prominent in social media, whereas mainstream news reporting is characterised by less conspiratorial and less emotionally framed pseudoscientific content.

5.4 ChatGPT-4o versus Rule-Based Classification

The thematic groupings produced by ChatGPT-4o show substantial agreement with the rule-based lexical analysis. Both approaches identify health-related misinformation, conspiracy narratives, emotionally framed content, and science-inspired esoteric discourse as major components of Bulgarian pseudoscientific communication.

An important difference, however, concerns the category *Bulgaria Above All – The Most Ancient*. Whereas this narrative emerged as the dominant category in the rule-based analysis, it was not identified by ChatGPT-4o as a major thematic cluster. This finding suggests that general-purpose Large Language Models may be less sensitive to culturally specific misinformation narratives than to globally recurring pseudoscientific themes.

LLM-assisted thematic grouping and rule-based lexical analysis should be regarded as complementary approaches. While ChatGPT-4o successfully identifies broad semantic themes, the domain-specific lexicon demonstrates greater sensitivity to culturally embedded narratives characteristic of the Bulgarian information environment.

6 Discussion

The findings demonstrate that pseudoscientific discourse in Bulgarian science-related communication is characterised not only by globally recurring themes such as alternative medicine, conspiracy narratives, and esoteric interpretations of science, but also by culturally specific narratives centred on Bulgarian history and national identity. In particular, the predominance of the category *Bulgaria Above All – The Most Ancient* suggests that pseudoscientific communication in Bulgaria is frequently intertwined with identity-based narratives that extend beyond the scientific domain itself.

The predominance of nationally framed pseudoscientific narratives also has practical implications for misinformation prevention. Recurrent rhetorical combinations linking Bulgarian identity, historical exceptionalism, and claims of revolutionary scientific discoveries may represent promising targets for prebunking interventions. From the perspective of inoculation theory, increasing public awareness of these recurring rhetorical patterns may reduce the persuasive potential of sensationalised or weakly substantiated claims before they become widely disseminated (Lewandowsky, 2021). Rather than responding to individual false claims after they have spread, such interventions could focus on recognising the rhetorical structures commonly employed in pseudoscientific communication.

More broadly, the observed differences between the CrowdTangle and PyGoogleNews corpora highlight the importance of maintaining researcher access to large-scale social media data. Comparative analyses of different communication environments provide valuable insights into how scientific information and pseudoscientific narratives evolve across platforms. As access to social media research infrastructures becomes increasingly restricted

following the discontinuation of CrowdTangle, opportunities to monitor misinformation ecosystems and evaluate countermeasures may become substantially more limited.

Taken together, these findings suggest that combining statistical lexical extraction, expert-curated lexical resources, and LLM-assisted thematic grouping provides a practical framework for analysing pseudoscientific discourse in low-resource languages. Rather than replacing expert knowledge, large language models appear to complement traditional NLP approaches by facilitating exploratory thematic analysis while domain-specific lexical resources remain essential for identifying culturally embedded misinformation narratives. The proposed workflow is intentionally transparent and reproducible, allowing future studies to adapt the lexical resources to other languages, scientific domains, or misinformation contexts while preserving the overall analytical framework.

7 Limitations

Several limitations should be considered when interpreting the results of this study.

First, a primary constraint is the **lack of real-time data** regarding the current evolution of the social media platform. Since Meta has deprecated CrowdTangle, researchers face significant challenges in tracking how scientific and pseudoscientific narratives evolve over time, limiting the ability to observe the most recent developments within the platform ecosystem.

Second, there is an inherent **structural hybridity in the CrowdTangle dataset**. Although the dataset includes posts published by professional media outlets, it is primarily composed of content from public pages and groups that do not necessarily follow journalistic standards. Consequently, the comparison between social media and mainstream news involves fundamentally different types of content producers, which may influence the observed differences in linguistic style and factual accuracy.

Third, the **disparity in dataset sizes** constitutes a methodological limitation. The news corpus collected through PyGoogleNews is considerably smaller than the CrowdTangle corpus, which may affect the granularity of the comparison and the statistical representativeness of the identified news-media trends.

Fourth, **stem-based lexical normalisation** in Bulgarian may occasionally introduce ambiguity because morphologically related stems can correspond to different semantic concepts. Although the subsequent expert review substantially reduced such ambiguities, the lexicon-based approach still requires human supervision when interpreting borderline cases.

Finally, although the observed **discrepancy between the lexicon-based approach and the LLM-assisted analysis** provides valuable insights into potential cultural blind spots in general-purpose AI models, the study relied on a single LLM (ChatGPT-4o). Future research should evaluate additional models (e.g., Claude, Gemini, or Llama) to determine whether similar patterns are observed across different architectures or whether some models demonstrate greater sensitivity to culturally specific narratives.

8 Ethics

The data utilised in this study, sourced via CrowdTangle, has been fully **anonymised** to ensure the privacy and protection of individual users. In alignment with the principles of Open Science and the Digital Services Act (DSA), the authors intend to make this dataset available for academic and research purposes, strictly following legal frameworks and platform terms of service. Once again, we underscore the critical necessity for **guaranteed and transparent data access** for the scientific community. Such access is indispensable for maintaining information integrity and developing effective strategies to safeguard the public discourse against the erosion of scientific truth.

9 Conclusions and Future Work

The findings provide important insights into the current state of science communication and pseudoscientific discourse in the Bulgarian media environment. A key observation is the substantial prevalence of pseudoscientific content, with approximately one-fifth of the analysed social media posts (17.68%) containing lexical markers associated with misinformation. Such content frequently relies on emotionally charged vocabulary, including terms such as life and people, which may increase the persuasive appeal of unsupported claims.

The analysis further reveals a strong association

between pseudoscientific narratives and identity-based themes. More specifically, the findings point to a process of nationalistic mystification, in which references to Bulgarian antiquity, cultural heritage, and national uniqueness are mobilised to legitimise pseudoscientific claims and reinforce identity-based narratives. The dominant category, "Bulgaria Above All – The Most Ancient", accounts for more than 55% of the analysed pseudoscientific content, suggesting that references to national history, heritage, and collective identity play an important role in the construction and dissemination of certain misinformation narratives. In addition, the recurrent co-occurrence of scientific and astrological terminology—particularly in categories related to cosmic mysteries and energy—illustrates how scientific language can be appropriated to enhance the perceived credibility of esoteric claims.

The comparison between the CrowdTangle and PyGoogleNews corpora demonstrates that different communication environments facilitate different forms of pseudoscientific discourse. While social media contains a higher concentration of emotionally framed, identity-based, and conspiratorial narratives, the news corpus is characterized by more conventional forms of pseudoscientific content, including horoscopes, alternative medicine, and sensationalized scientific reporting.

From a methodological perspective, the study highlights both the potential and the limitations of Large Language Models for thematic analysis. Although ChatGPT-4o successfully identified several major pseudoscientific themes, it did not detect the dominant nationalist narrative identified through the lexicon-based approach. This discrepancy may indicate a cultural blind spot in general-purpose LLM-assisted analysis when dealing with highly localised narratives embedded in specific historical and sociocultural contexts. Thus, the general-purpose LLMs and domain-specific NLP methods should be regarded as complementary analytical tools, particularly when investigating culturally embedded forms of misinformation.

Future research should expand both the social media and news-media corpora, explore additional Large Language Models, and further investigate the relationship between pseudoscientific narratives, national identity, and science communication. More specifically, future studies could employ more targeted filtering of social media datasets, expand the news corpus through large-scale collection of Bulgarian regional and specialized science news sources, and evaluate the robustness of the observed patterns across different LLM architectures, including Claude, Gemini, and Llama.

The identified patterns of pseudo-expertise and rhetorical legitimization may also provide a useful foundation for future prebunking and media-literacy interventions aimed at strengthening public resilience against pseudoscientific misinformation. In particular, the recurrent appropriation of scientific concepts such as quantum, energy, or frequency within astrological and esoteric contexts provides identifiable rhetorical patterns that could be incorporated into inoculation-based communication strategies. By familiarizing audiences with such misleading semantic combinations and explaining their persuasive mechanisms, future interventions may help strengthen resistance to pseudoscientific narratives and reduce their persuasive impact.

Acknowledgements

This research is part of the GATE project funded by the Horizon 2020 WIDESPREAD-2018-2020, TEAMING Phase 2, the Program under grant agreement no. 857155, the Program "Research, Innovation and Digitalization for Smart Transformation" 2021-2027 (PRIDST) under grant agreement №BG16RFPR002-1.014-0010-C01 and the project BROD (Bulgarian-Romanian Observatory of Digital Media), funded by the Digital Europe Program of the European Union under contract №101226153.

References

Hans-Joachim Dahms. 2022. Alternative facts, fake news, pseudoscience–new challenges for a scientific world-view. an essay. In *The Socio-Ethical Dimension of Knowledge: The Mission of Logical Empiricism*, pages 195–216. Springer.

Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2019. Detecting toxicity in news articles: Application to Bulgarian. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 247–258.

Gabriel Frahm. 2026. What is science? *Available at SSRN 6307218*.

Sven Ove Hansson. 2013. Defining pseudoscience and science. *Philosophy of pseudoscience: Reconsidering the demarcation problem*, pages 61–78.

Stephan Lewandowsky. 2021. Climate change disinformation and how to combat it. *Annual review of public health*, 42(1):1–21.

Ruslana Margova. 2025. The (possible) use of AI tools for processing texts in journalism in Bulgarian. *Computational Linguistics in Bulgaria*, 1(1):84–102.

Preslav Nakov. 2020. Can we spot the "fake news" before it was even written? *arXiv preprint arXiv:2008.04374*.

Preslav Nakov, Firoj Alam, Shaden Shaar, Giovanni Da San Martino, and Yifan Zhang. 2021. COVID-19 in Bulgarian social media: Factuality, harmfulness, propaganda, and framing. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 997–1009.

Karl Popper. 2005. *The logic of scientific discovery*. Routledge.

Tiago Ramalho and Sandra Marinho. 2026. Media effects in the use of pseudoscience: A systematic literature review (1900–2022). *Cultures of Science*, 9(2):125–144.

Alexandre Schiele. 2020. Pseudoscience as media effect. *Journal of Science Communication*, 19(2):L01.

Irina Temnikova, Silvia Gargova, Ruslana Margova, Veneta Kireva, Ivo Dzhumerov, Tsvetelina Stefanova, and Hristiana Krasteva. 2023. New Bulgarian resources for studying deception and detecting disinformation. *Human Language Technologies as a Challenge for Computer Science and Linguistics–2023*.

Irina Temnikova, Ruslana Margova, Stefan Minkov, Tsvetelina Stefanova, Nevena Grigorova, Silvia Gargova, and Venelin Kovatchev. 2025. Automatic detection of the Bulgarian evidential renarrative. *Computational Linguistics in Bulgaria*, page 61.

Adam Tamas Tuboly. 2025. On the value of pseudoscience and its philosophical study. *European Journal for Philosophy of Science*, 15(3):43.

Appendix A Lexicon Construction and Thematic Categorisation

The pseudoscientific lexicon was constructed using the cleaned CrowdTangle corpus containing 17,833 unique posts. Following duplicate removal and text preprocessing, TF-IDF (Term Frequency–Inverse Document Frequency) analysis was applied to identify statistically salient lexical markers. The resulting ranked lexical items served as the basis for the subsequent expert-guided construction of the domain-specific pseudoscientific lexicon.

Unlike simple word-frequency counts, TF-IDF assigns greater importance to lexical items that are distinctive within the corpus while reducing the influence of very common vocabulary. This approach facilitates the identification of lexical markers that are particularly informative for distinguishing recurrent themes in Bulgarian science-related online discourse.

The analysis was implemented using a custom Python workflow consisting of the following stages:

- tokenisation and text cleaning, including the removal of punctuation, URLs, HTML fragments, special characters, and other technical artefacts;
- linguistic normalisation through light suffix-based stemming to reduce morphological variation and group related Bulgarian lexical forms under common lexical roots;
- extraction of unigrams and bigrams to preserve both individual lexical items and frequently co-occurring expressions;
- iterative stop-word filtering using a custom Bulgarian stop-word inventory.

The stop-word inventory contained more than 800 entries and was developed through multiple rounds of iterative refinement. It included grammatical function words, highly frequent news vocabulary, temporal and quantitative expressions, technical artefacts, and other non-discriminative lexical items that did not contribute to thematic differentiation. Because Bulgarian is a morphologically rich language, stem-based normalisation may occasionally introduce lexical ambiguity (for example, the stem меди- may refer to both медицина ("medicine") and медии ("media"). Such cases were subsequently reviewed during the expert validation stage.

The TF-IDF vectoriser was configured to extract the 100 highest-ranked lexical items and n-grams, providing a concise inventory of statistically salient lexical markers. This threshold was selected as a compromise between lexical coverage and interpretability during the subsequent qualitative analysis. The complete ranked list is presented in Appendix B.

The ranked lexical inventory served as the empirical basis for the construction of the domain-specific pseudoscientific lexicon used in the

subsequent rule-based identification and thematic analysis presented in the main text. Because linguistic normalisation was performed through light suffix-based stemming, several entries are represented as lexical roots rather than complete word forms.

This procedure resulted in a broader inventory of lexical indicators covering pseudoscientific, conspiratorial, esoteric, alternative-medical, and pseudo-historical narratives. Additional lexical markers were included when they appeared consistently across multiple pseudoscientific contexts and were assessed by the authors as semantically representative of recurring misinformation themes.

The resulting lexical markers were organised into the following thematic clusters:

- Core Pseudoscientific Markers: energy (енерги-), radiation (радиац-), phenomenon (феномен), unexplained (необясним), mystery (мистерия), mysterious (мистериоз-), hypothesis (хипотеза), quantum (квант-), unknown (непознат).
- Esoteric and New Age Terms: vibration (вибраци-), frequency (честот-), astrology (астрологи-), secret (секрет-), cosmic (космическ-), mystical (мистич-), aliens (извънзем-), supernatural (свръхестеств-).
- Conspiracy and Alternative Narratives: conspiracy (конспир-), hidden (скрит-), truth (истин-), forbidden (забранен), suppressed (таен-), shocking (шокиращ), sensation (сензац-).
- Pseudo-Historical and Biological Markers: ancient (древ-), DNA (ДНК), anomaly (аномал-), miracle (чуд-).

Appendix B Ranked TF-IDF Lexical Markers

This appendix presents the 100 highest-ranked lexical items and n-grams identified through TF-IDF (Term Frequency–Inverse Document Frequency) analysis of the cleaned CrowdTangle corpus containing 17,833 unique posts retrieved using the Bulgarian keyword наука (“science”).

The lexical extraction procedure, including corpus preprocessing, linguistic normalisation, iterative stop-word filtering, unigram–bigram extraction, and TF-IDF ranking, is described in Appendix A.

The ranked lexical inventory served as the empirical basis for domain-specific pseudoscientific lexicon used in the subsequent rule-based identification and thematic analysis presented in the main text. Because linguistic normalisation was performed through light suffix-based stemming, several entries are represented as lexical roots rather than complete word forms.

Table 2: Top 100 TF-IDF lexical markers, ranks 1–50

Rank	Lexical Item	Weight	Rank	Lexical Item	Weight
1	кои	0.135389	26	душа	0.017550
2	най	0.108256	27	бъдеще	0.017278
3	кое	0.078467	28	духовн	0.016881
4	нов	0.056786	29	истинск	0.016308
5	све	0.043854	30	велик	0.016090
6	истин	0.042353	31	списани	0.015968
7	здрав	0.039206	32	свое	0.015884
8	технологи	0.036516	33	име	0.015793
9	къде	0.028469	34	политиче	0.014945
10	иноваци	0.028059	35	любов	0.014903
11	образование	0.027825	36	основ	0.014870
12	негов	0.027748	37	дух	0.014660
13	болест	0.025659	38	обучени	0.014648
14	развитие	0.023316	39	участи	0.014498
15	история	0.023010	40	лечени	0.014482
16	време	0.022343	41	ваксин	0.014146
17	може	0.022187	42	значени	0.014081
18	здраве	0.020766	43	управлен	0.013733
19	тайн	0.020081	44	късно	0.013696
20	световн	0.019425	45	отношени	0.013383
21	събитие	0.019043	46	универси	0.013114
22	космиче	0.018786	47	научен	0.013013
23	историче	0.018483	48	намер	0.012979
24	българско	0.018115	49	състояни	0.012967
25	наше	0.018095	50	бързо	0.012925

Table 3: Top 100 TF-IDF lexical markers, ranks 51–100

Rank	Lexical Item	Weight	Rank	Lexical Item	Weight
51	интелект	0.012875	76	вижда	0.011200
52	благодар	0.012678	77	лев	0.011120
53	път	0.012613	78	комисия	0.011018
54	иде	0.012542	79	внимани	0.010998
55	смърт	0.012459	80	точка	0.010984
56	далеч	0.012347	81	кмет	0.010835
57	бъд	0.012264	82	лични	0.010796
58	колег	0.012234	83	историчео	0.010576
59	реалност	0.012204	84	промяна	0.010486
60	създаване	0.012166	85	организа	0.010393
61	човешко	0.012089	86	институц	0.010367
62	мъдрост	0.012058	87	инвестици	0.010322
63	анализ	0.012040	88	мар	0.010163
64	партньор	0.012015	89	икономи	0.009934
65	две	0.011835	90	лун	0.009454
66	писменос	0.011798	91	ред	0.009355
67	физиче	0.011798	92	слънце	0.009285
68	смисъл	0.011762	93	гор	0.009116
69	услов	0.011527	94	майка	0.008925
70	заместни	0.011457	95	адрес	0.008869
71	знани	0.011403	96	хороскоп	0.008834
72	периода	0.011398	97	трак	0.008816
73	използва	0.011269	98	бога	0.008093
74	меди	0.011226	99	конферен	0.007761
75	ръководи	0.011148	100	спрямо	0.007516

PS-CHILDES-BG: a constituency treebank of Bulgarian morphemically tokenised child-directed speech

Mila Marcheva-Nash
Department of Computer Science & Technology,
University of Cambridge, UK
mmm67@cam.ac.uk

Weiwei Sun
Department of Computer Science & Technology,
University of Cambridge, UK
ws390@cam.ac.uk

Abstract

We present PS-CHILDES-BG, a constituency treebank of 2,434 Bulgarian morphemically tokenised child-directed speech sentences. We supply a morphemic tokenisation algorithm for Bulgarian, which uses UD parses as input. Additionally, we introduce a linguistically informed algorithm for Bulgarian which converts UD parses into constituency parses. These resources are intended to further the language acquisition research of Bulgarian via computational methods.

Keywords: constituency treebank, morphemic tokenisation, child-directed speech

1 Introduction

A central question in first language acquisition (FLA) is how children acquire functional categories (Dye et al., 2018). Functional categories are expressed by functional items and encode grammatical content (Baker, 2008), in contrast with lexical categories, which encode semantic content. Traditionally, FLA relies on labour-intensive methods, such as manually annotating child-directed speech (CDS) with syntactic features and descriptively analysing the resulting data. Computational methods can complement traditional approaches. For example, grammar and category induction experiments can be used to approximate the corresponding category or grammar acquisition processes. Seminal computational experiments for FLA include Carroll and Charniak (1992); Cartwright and Brent (1997); Buttery (2006); Parisien et al. (2008). Computational experiments allow for confounding factors to be controlled for and can provide a lower bound on the types of grammatical information that may be recoverable from linguistic signal alone.

In computational modelling of FLA, standard tokenisation only captures the free functional items, such as the article *the*, but ignores bound functional morphemes, such as the progressive aspect suffix *-ing*. Morphemic tokenisation allows both bound and free functional items to be represented, leading to a more complete account of the grammatical material available in the input. The motivation for morphemic tokenisation of CDS is also grounded in FLA literature. Functional items have a higher frequency than lexical items and occur in predictable positions (Lieven, 2010; Biberauer, 2019). The salience of functional items for distributional analysis makes morphemic tokenisation compatible with both generativist (Guasti, 2017) and usage-based (Lieven, 2016) accounts of FLA. Morphemic tokenisation of English CDS has recently been explored for functional category induction and grammar induction (Marcheva et al., 2025a,b; Marcheva-Nash et al., 2026b). The need for morphemic tokenisation is even stronger in morphologically richer languages, such as Bulgarian, where a larger number of functional items manifest as bound morphemes.

Caregivers are known to use a special register when communicating with infants, known as child-directed speech (CDS), which differs from adult-directed speech in prosody (higher pitch, exaggerated contours), lexical choice, and utterance length (Fernald, 1989; Soderstrom, 2007; Kuhl, 2006). The usage of CDS in computational language modelling is recommended in order to make the computational experiments more cognitively plausible, i.e. to more closely resemble the real-world learning problem. CDS corpora exist for very few languages: English (Pearl and Sprouse, 2013; Yang et al., 2025); Dutch (Odijk et al., 2018); Hebrew (Szubert et al., 2024); Japanese (Butler et al., 2022); Bulgarian UD-CHILDES-BG (Marcheva-Nash et al., 2026a). CDS corpora predominantly follow Universal Dependencies (UD; de Marneffe et al., 2021). However, constituency (phrase structure) parses can be more informative for models which operate over hierarchical phrase structure,

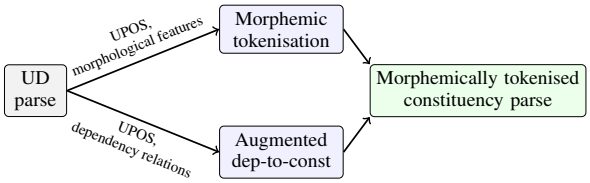


Figure 1: Pipeline for deriving a morphemically tokenised constituency parse from a UD parse.

such as grammar induction, because they represent the hierarchical structure of language directly (Carnie, 2021). A morphemically tokenised constituency parse resource currently exists only for English CDS (Pearl and Sprouse, 2013; Marcheva et al., 2025b). The algorithms and the treebank we contribute in this paper would allow for the current research based on morphemically tokenised CDS to be extended to Bulgarian.

We present a morphemic tokeniser for Bulgarian inflectional morphology, which we apply to Bulgarian CDS. We also develop a linguistically informed automatic conversion system from dependency to constituency parses. Both algorithms rely on UD parses as input, which in our case are manually corrected, but automatic parses could also be used. While the two algorithms can be used independently, we combine them, as seen in Figure 1, to produce a linguistically informative resource, PS-CHILDES-BG, a constituency treebank of 2,434 Bulgarian morphemically tokenised CDS sentences, of which 15% are manually verified. The existing resources are presented in section 2, the algorithms are described in more detail in section 3, an evaluation of the algorithms is presented in section 4. The morphemic tokeniser for Bulgarian, the modified dependency to constituency algorithm, as well as the resulting treebank are available on GitHub.[1]

2 Existing resources

Bulgarian NLP resources include the constituency treebank BulTreeBank (BTB; Simov et al., 2002; Simov and Osenova, 2003) and its dependency counterpart, UD-BTB (Osenova and Simov, 2017). Bulgarian also has a CHILDES (MacWhinney, 1992) resource: CHILDES-BG provides longitudinal developmental data (Popova and Popov, 2020). Parsing tools for Bulgarian include the Berkeley constituency parser (Petrov et al., 2006), the dependency parsers Stanza (Qi et al., 2020) and CLASSLA-Stanza (Terčon and Ljubešić, 2023; Ljubešić et al., 2024).

At the subword level, several resources are also relevant. BulStem (Nakov, 2003) is an inflectional stemmer for Bulgarian. Iliev et al. (2015) present a dictionary-based Bulgarian lemmatiser. However, neither of these specifically tokenises functional morphemes. MorphScore (Arnett et al., 2025; Arnett and Bergen, 2025) is a tool for evaluating morphemic tokenisation, which includes a Bulgarian dataset based on UD-BTB. More generally, work on morpheme segmentation is represented in SIGMORPHON, a shared task on morpheme segmentation (Batsuren et al., 2022), but it excludes Bulgarian. The most closely related resources for functional morpheme tokenisation are morphemic tokenisation of English CDS (Marcheva et al., 2025a) and the morphemically tokenised English CHILDES-TB (Marcheva et al., 2025b).

3 Pipeline

In this section we give an overview of the data, the morphemic tokenisation algorithm, the augmented dependency to constituency (dep-to-const) conversion algorithm, including detail on the morphemically tokenised version. The workflow is illustrated in Figure 1.

3.1 Data

The UD parses we use as input are the manually verified CDS portion of UD-CHILDES-BG (Marcheva-Nash et al., 2026a), a UD corpus based on CHILDES-BG (Popova and Popov, 2020), but any automatic UD parses can be used as input. The current system only supports projective parses, and there are 2,434 such CDS parses in UD-CHILDES-BG. The manually verified section of UD-CHILDES-BG includes sentences from the input of five children, stratified by age. The age range of the five children is between one and three years old, which encompasses the key stages for the emergence of functional categories (Brown, 1973; Popova, 2023).

3.2 Morphemic tokenisation

There are several challenges specific to morphemic tokenisation of Bulgarian, which are not encountered in the same task for English. Several competing segmentation choices are often possible. In such cases, it is important to select a single lin-

[1] https://github.com/milamarcheva/UD-CHILDES-BG

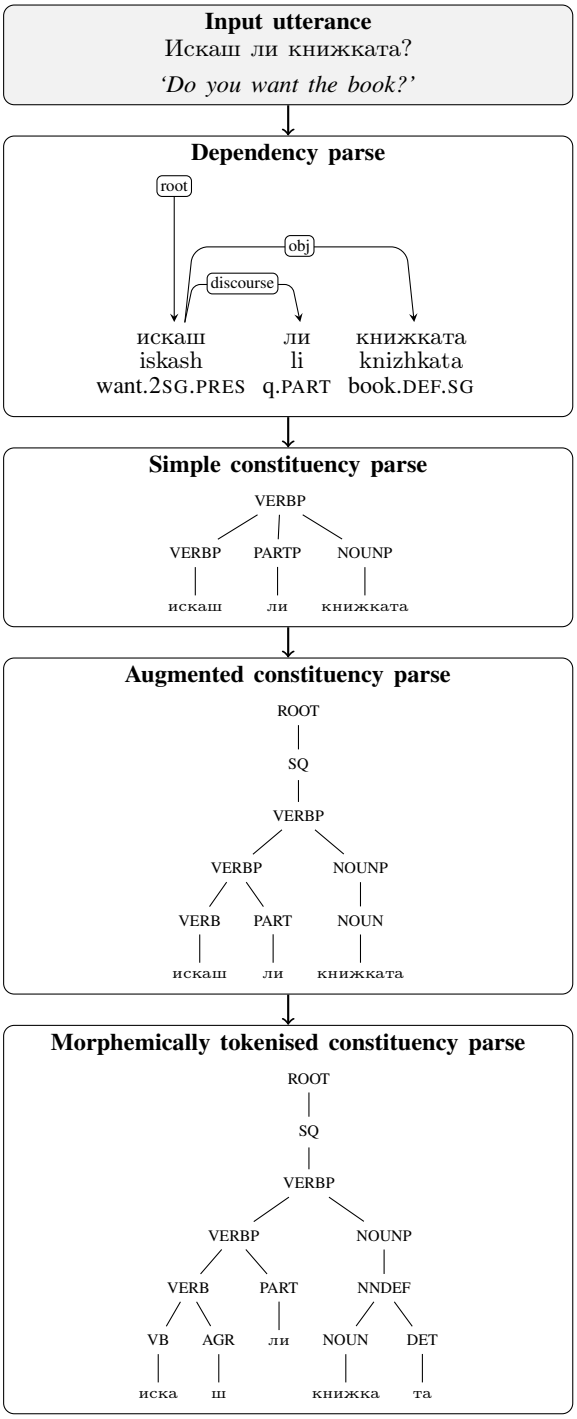


Figure 2: Different types of parses. The augmented and morphemically tokenised constituency parses are the result of the algorithms presented here.

guistically defensible implementation and apply it consistently throughout the corpus.

3.2.1 Stem vs. lemma

When the functional morpheme is identified and segmented, the remaining stem does not always correspond to the dictionary form. In some cases, the stem and the lemma are the same, such as in въпроси → въпрос и *(questions → question* + PL). In other cases, such as правим → прав им *(we do → do.STEM* + 1PL*)* and правят → прав ят *(they do → do.STEM* + 3PL*)*, the remaining stem is not a lemma (the lemma would be the first person form правя). In such cases, for PS-CHILDES-BG we choose to use the stem so that the different conjugations of the verb can still have the same base, as this is important for symbolic downstream applications. However, our morphemic tokeniser also provides functionality to instead use the lemma (which is part of the UD parse), if that is appropriate for another user's application.

3.2.2 Multiple functional morphemes

Sometimes in Bulgarian several functional morphemes can be strung together. Here, тротинетките → тротинетк и те *(the scooters → scooter* + PL + DEF.PL*)*, the functional morpheme for plural и and the functional morpheme for definite marking of a plural nominal те are both present. As the morphemes are separable, for PS-CHILDES-BG we segment them, because this would enable the morphemes to be learned separately in category and grammar induction experiments. However, the morphemic tokeniser includes an option to keep all functional morphemes as one suffix, if appropriate for other users' applications.

3.2.3 Gender

In Bulgarian, gender is one of the grammatical axes of agreement: a noun has a fixed grammatical gender and a nominal modifier of the noun, such as an adjective, needs to agree with its gender. In the regular case segmenting the suffix marking gender is easy as it is an additive suffix added to the masculine form: красив, красива, красиво *(beautiful.*M.SG*, beautiful.*F.SG*, beautiful.*N.SG*)*. However, often the stem of an adjective differs between masculine and neuter and feminine, e.g. малък, малка, малко *(small.*M.SG*, small.*F.SG*, small.*N.SG*)* or странен, странна, странно *(strange.*M.SG*, strange.*F.SG*, strange.*N.SG*)*. As there is neither a common lemma nor a common stem for all three genders in these cases, we decide to not tokenise gender in the current iteration of the system.

3.2.4 Verbal conjugation classes

Bulgarian verbal inflection is organised around conjugation-specific bases and endings. The three

conjugation classes are distinguished by the vowel of the third person singular present tense form: first-conjugation verbs have е, second-conjugation verbs have и, and third-conjugation verbs have an а/я-final base.[2] The morphemic tokeniser therefore does not apply a single uniform suffix rule to all verbs. Instead, it selects a segmentation rule on the basis of the lemma, UD morphological features, surface form, and, where relevant, conjugation class and tense. This allows the tokeniser to distinguish cases where a vowel belongs to the inflectional ending from cases where it belongs to the verbal base, while still keeping the lexical base stable across forms whenever possible.

3.3 Dependency to constituency conversion

Dependency to constituency (dep-to-const) conversion is commonly performed using the method of Collins et al. (1999). Here we use the implementation of Kando et al. (2022) as a starting point: the output of their algorithm is the "Simple constituency parse" in Figure 2. We perform a number of Bulgarian-specific modifications, some of which are described below, which aim to introduce more hierarchical structure through linguistic grounding to yield parses of the calibre of the "Augmented constituency parse" in Figure 2. More parses, both standardly and morphemically tokenised constituency parses are presented in Appendix A.

3.4 Hierarchical restructuring

To every parse we add a ROOT wrapper. We distinguish between pre-terminals, the UPOS tagset (e.g. NOUN) and non-terminals, the UPOS tagset + P (e.g. NOUNP), and where necessary introduce PTB-style non-terminals (e.g. SBAR for subordinate clauses). Discourse particles, vocatives, and interjections attach at the highest possible level to the clause as a whole, instead of being inserted inside the core predicate structure, such as in Figure 3. Within verbal projections, the verb should combine with its object first, such as in Figure 4. However, due to the relatively free word order of Bulgarian, an adjective may intervene in the surface order between verb and object, such as in Figure 5. Due to the restrictions of constituency grammar, the verb needs to form a constituent with the adverb and then with the object through a recursive VERBP rule.

For subordinated clauses introduced by the dependency relations `xcomp`, `ccomp`, and `advcl`, we use SBAR and subordinate them to the verb that is the head of the subordinate clause in the dependency parse, as seen in Figure 6. To distinguish interrogative clauses from non-question clauses, we use the PTB-style label SQ to introduce questions, as seen in Figure 7. Since the dependency parse does not reliably encode interrogative force, we use the presence of the question mark as the necessary condition for assigning SQ. Furthermore, within such clauses, Bulgarian interrogative *k*-words are projected as WH-type phrases according to their POS category, for example WHNP for pronominal and determiner-like forms such as кой (*who*) and WHADVP for adverbial forms such as къде (*where*).

3.5 Morphemically tokenised constituency parses

The morphemically tokenised constituency parses follow the same rules as above and additional rules are introduced to replace word-level terminals with subtrees comprising a word base and functional morphemes. For example, we use the pre-terminal AGR to mark verbal agreement morphology, DET to mark nominal definiteness morphology, and DIV to mark nominal count or plural number. Some examples are presented in Appendix A. We manually verify 15% of the total 2,434 parses in PS-CHILDES-BG.

4 Evaluation

We evaluate the morphemic tokeniser using MorphScore. In the original paper (Table 3; Arnett and Bergen, 2025), the reported accuracy score for Bulgarian is 0.584, whereas our morphemic tokeniser achieves 0.814. MorphScore is not designed specifically for functional morphology and also includes derivational morphology. MorphScore tokenises gender which we do not as explained in subsubsection 3.2.3, and this accounts for 55% of the mismatches with the MorphScore gold tokenisation, and some disagreements around verbal endings, which we tokenise as explained in subsubsection 3.2.4, account for 26.7% of the mismatches. As MorphScore is not primarily aimed at functional morpheme tokenisation, not achieving an accuracy of 1.0 is not necessarily a bad result.

[2] See the overview of Bulgarian verbal conjugation at https://www.slav.uni-sofia.bg/grammar/morf25.html and the overview of Bulgarian tense formation at https://www.slav.uni-sofia.bg/grammar/morf30.html.

Tokenisation →	STANDARD		MORPHEMIC
Metric ↓	Simple	Augm.	Augm.
Branching	2.74	1.51	1.54
Arity≥3	52.2%	0.24%	0.21%
Depth	2.35	4.94	5.18

Table 1: Tree structure evaluation using mean branching, percentage of nodes with arity≥3, and mean tree depth, for the simple dep-to-const conversion (standard tokenisation) and for the augmented dep-to-const conversion (standard and morphemic tokenisation).

Additionally, to demonstrate the quantitative difference between the parses resulting from the Collins algorithm (Collins et al., 1999; Kando et al., 2022) and the ones from our augmented algorithm we quantify several measures of tree flatness in Table 1. Our parses have a decreased branching factor, very few nodes with high arity, and an increased tree depth. As expected, the morphemically tokenised trees are even deeper, due to the way we introduce functional morphemes with subtrees. Overall, our augmented algorithm introduces hierarchy in the parses, making them more linguistically informed.

5 Conclusion

We present PS-CHILDES-BG, a constituency treebank of 2,434 Bulgarian CDS sentences, of which 15% are manually verified. The constituency parses are converted from UD parses, using a linguistically augmented dep-to-const algorithm. We also supply a morphemic tokenisation algorithm for Bulgarian, which can be used independently or in conjunction with the augmented dep-to-const algorithm to yield morphemically tokenised constituency parses for Bulgarian. In future work we will use PS-CHILDES-BG for FLA research via computational methods.

Limitations

A limitation of the resource, as also noted by a reviewer, is that the manual verification of the data is limited, due to how time-consuming manual verification is. We have increased it from 10% to 15% to address the reviewer's comment, but the issue is still worth considering in future iterations. Another limitation of the resource is that it relies on existing dependency parsing models which may not have the desired accuracy. However, the focus of this paper is to demonstrate how we can harness existing resources to create a silver-standard, if not gold-standard, resource which has potential to offer a different insight into computational modelling of acquisition than UD resources. Finally, the current verbal morphemic tokenisation does not label different verbal endings according to their function, such as person, number, tense, aspect, and mood, because of some fusional characteristics of Bulgarian morphology, which make the disentanglement of verbal morphology a problem for future work.

References

Catherine Arnett and Benjamin Bergen. 2025. Why do language models perform worse for morphologically complex languages? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6607–6623, Abu Dhabi, UAE. Association for Computational Linguistics.

Catherine Arnett, Marisa Hudspeth, and Brendan O'Connor. 2025. Evaluating Morphological Alignment of Tokenizers in 70 Languages. In *Proceedings of the ICML 2025 Tokenization Workshop (TokShop)*.

Mark C. Baker. 2008. The macroparameter in a microparametric world. In *The Limits of Syntactic Variation*, page 351–373. John Benjamins Publishing Company.

Khuyagbaatar Batsuren, Gábor Bella, Aryaman Arora, Viktor Martinovic, Kyle Gorman, Zdeněk Žabokrtský, Amarsanaa Ganbold, Šárka Dohnalová, Magda Ševčíková, Kateřina Pelegrinová, Fausto Giunchiglia, Ryan Cotterell, and Ekaterina Vylomova. 2022. The SIGMORPHON 2022 shared task on morpheme segmentation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 103–116, Seattle, Washington. Association for Computational Linguistics.

Theresa Biberauer. 2019. Children always go beyond the input: The maximise minimal means perspective. *Theoretical Linguistics*, 45(3–4):211–224.

Roger Brown. 1973. *A first language: The early stages*. Cambridge, MA: Harvard University Press.

Alastair Butler, Susanne Miyata, and Yumiko Kinjo. 2022. The Soyogo Treebank – a parsed corpus of child Japanese. https://soyogo.github.io.

Paula J. Buttery. 2006. Computational models for first language acquisition. Technical Report UCAM-CL-TR-675, University of Cambridge, Computer Laboratory.

Andrew Carnie. 2021. *Syntax*, 4 edition. Introducing Linguistics. Wiley-Blackwell, Hoboken, NJ.

Glenn Carroll and Eugene Charniak. 1992. Two experiments on learning probabilistic dependency

grammars from corpora. Technical Report CS-92-16, Dept. of Computer Science, Brown University, Providence, RI.

Timothy A. Cartwright and Michael R. Brent. 1997. Syntactic categorization in early language acquisition: formalizing the role of distributional analysis. *Cognition*, 63(2):121–170.

Michael Collins, Jan Hajic, Lance Ramshaw, and Christoph Tillmann. 1999. A statistical parser for Czech. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 505–512, College Park, Maryland, USA. Association for Computational Linguistics.

Cristina Dye, Yarden Kedar, and Barbara Lust. 2018. From lexical to functional categories: New foundations for the study of language development. *First Language*, 39(1):9–32.

Anne Fernald. 1989. Intonation and communicative intent in mothers' speech to infants: Is the melody the message? *Child Development*, 60(6):1497.

Maria Teresa Guasti. 2017. *Language acquisition*, 2 edition, chapter 1: Basic Concepts. Bradford Books, Cambridge, MA.

Grigor Iliev, Nadezhda Borisova, Elena Karashtranova, and Dafina Kostadinova. 2015. A publicly available cross-platform lemmatizer for Bulgarian.

Shunsuke Kando, Hiroshi Noji, and Yusuke Miyao. 2022. Multilingual syntax-aware language modeling through dependency tree conversion. In *Proceedings of the Sixth Workshop on Structured Prediction for NLP*, pages 1–10, Dublin, Ireland. Association for Computational Linguistics.

Patricia K. Kuhl. 2006. Is speech learning 'gated' by the social brain? *Developmental Science*, 10(1):110–120.

Elena Lieven. 2010. Input and first language acquisition: Evaluating the role of frequency. *Lingua*, 120(11):2546–2556.

Elena Lieven. 2016. Usage-based approaches to language development: Where do we go from here? *Language and Cognition*, 8(3):346–368.

Nikola Ljubešić, Luka Terčon, and Kaja Dobrovoljc. 2024. CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages. In *Conference on Language Technologies and Digital Humanities (JT-DH-2024)*, Ljubljana, Slovenia. Institute of Contemporary History.

Brian MacWhinney. 1992. The CHILDES project: tools for analyzing talk. *Child Language Teaching and Therapy*, 8(2):217–218.

Mila Marcheva, Theresa Biberauer, and Weiwei Sun. 2025a. Functional category induction with theory-neutral cognitive biases. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 47.

Mila Marcheva, Theresa Biberauer, and Weiwei Sun. 2025b. Profiling neural grammar induction on morphemically tokenised child-directed speech. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 47–54, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Mila Marcheva-Nash, Yasena Chantova, Tsvetina Kirilova, Ivelina Pavlova, Tsvetelina Stefanova, Yoana Vasileva, and Weiwei Sun. 2026a. UD-CHILDES-BG: A Dependency Treebank of Bulgarian Child and Child-Directed Speech. In *Proceedings of the 20th Linguistic Annotation Workshop*, San Diego, United States of America. To appear.

Mila Marcheva-Nash, Suchir Salhan, and Weiwei Sun. 2026b. A computational operationalisation of competing maturational theories of syntactic development via statistical grammar induction. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 48. To appear.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics*, page 1–54.

Preslav Nakov. 2003. BulStem: Design and evaluation of inflectional stemmer for Bulgarian. In *Proceedings of the Workshop on Balkan Language Resources and Tools*, Thessaloniki, Greece. 1st Balkan Conference in Informatics.

Jan Odijk, Alexis Dimitriadis, Martijn van der Klis, Marjo van Koppen, Meie Otten, and Remco van der Veen. 2018. The AnnCor CHILDES treebank. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Petya Osenova and Kiril Simov. 2017. Recent developments within BulTreeBank. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 129–137, Prague, Czech Republic.

Christopher Parisien, Afsaneh Fazly, and Suzanne Stevenson. 2008. An incremental Bayesian model for learning syntactic categories. In *CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 89–96, Manchester, England. Coling 2008 Organizing Committee.

Lisa S. Pearl and Jon Sprouse. 2013. Syntactic islands and learning biases: Combining experimental syntax and computational modeling to investigate the language acquisition problem. *Language Acquisition*.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for*

Computational Linguistics, pages 433–440, Sydney, Australia. Association for Computational Linguistics.

Velka Popova. 2023. The emergence of word classes in early Bulgarian language ontogenesis. a pilot corpus study. *Journal of Bulgarian Language*, 70(PRIL):290–308.

Velka Popova and Dimitar Popov. 2020. CHILDES Bulgarian labling corpus.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. 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*.

Kiril Simov and Petya Osenova. 2003. Practical annotation scheme for an HPSG treebank of Bulgarian. In *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-2003)*, pages 17–24.

Kiril Simov, Petya Osenova, Milena Slavcheva, Sia Kolkovska, Elisaveta Balabanova, Dimitar Doikoff, Krassimira Ivanova, Alexander Simov, and Milen Kouylekov. 2002. Building a linguistically interpreted corpus of Bulgarian: the BulTreeBank. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Melanie Soderstrom. 2007. Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review*, 27(4):501–532.

Ida Szubert, Omri Abend, Nathan Schneider, Samuel Gibbon, Louis Mahon, Sharon Goldwater, and Mark Steedman. 2024. Cross-linguistically consistent semantic and syntactic annotation of child-directed speech. *Language Resources and Evaluation*, 59(2):727–776.

Luka Terčon and Nikola Ljubešić. 2023. CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages.

Xiulin Yang, Zhuoxuan Ju, Lanni Bu, Zoey Liu, and Nathan Schneider. 2025. UD-English-CHILDES: A collected resource of gold and silver Universal Dependencies trees for child language interactions. In *Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025)*, pages 52–58, Ljubljana, Slovenia. Association for Computational Linguistics.

A Constituency parse examples

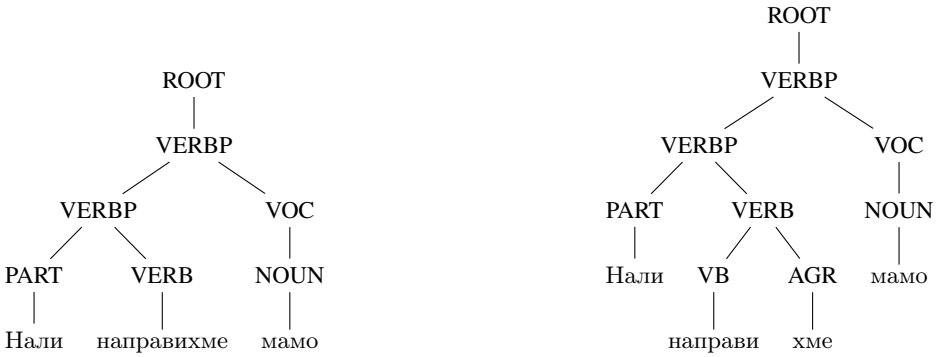


Figure 3: Нали направихме, мамо. ‘Right, we did it, *inverse-vocative-address*.’

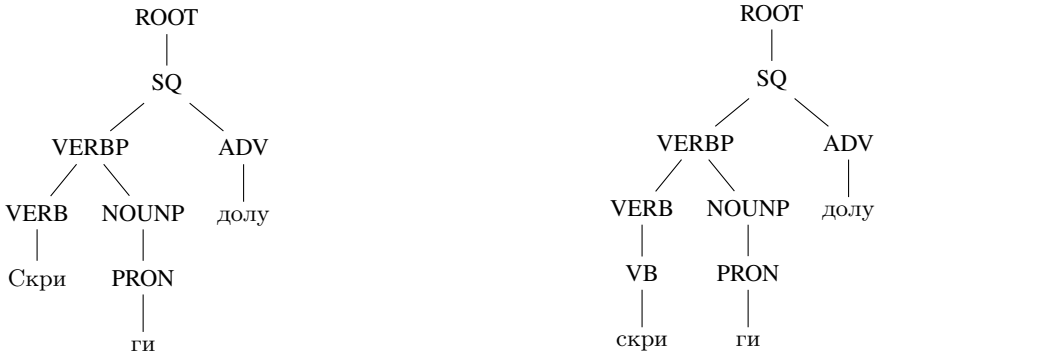


Figure 4: Скри ги долу. ‘He/She/It hid them under.’

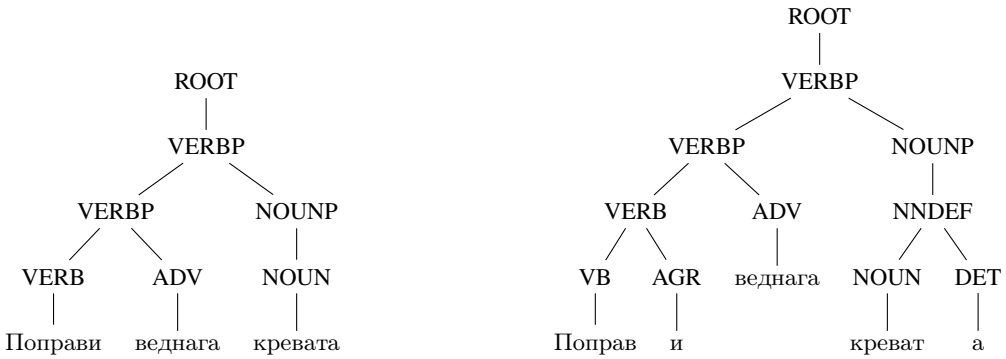


Figure 5: Поправи веднага кревата. ‘Fix the bed immediately.’

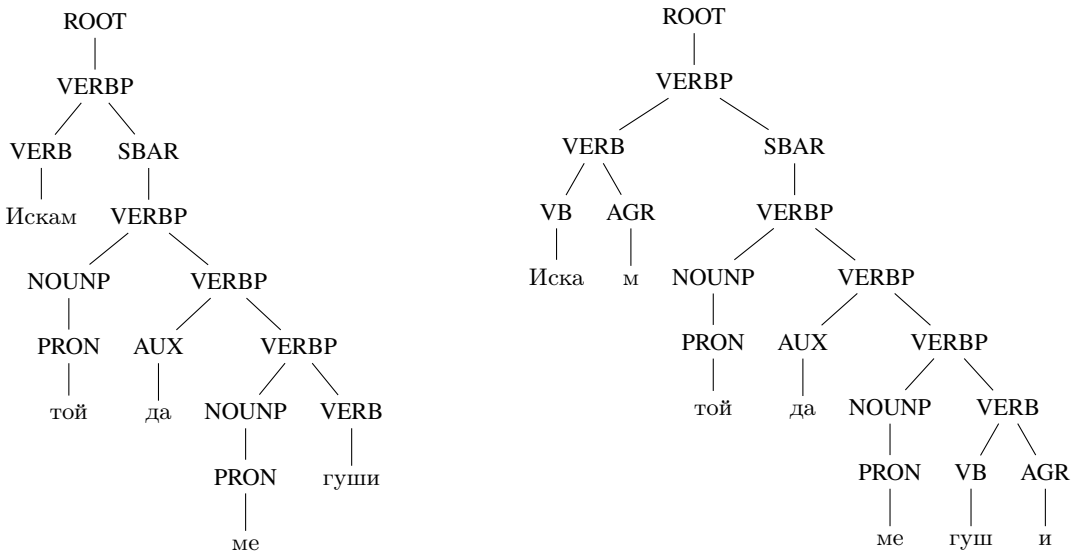


Figure 6: Искам той да ме гуши. ‘I want him to hug me.’

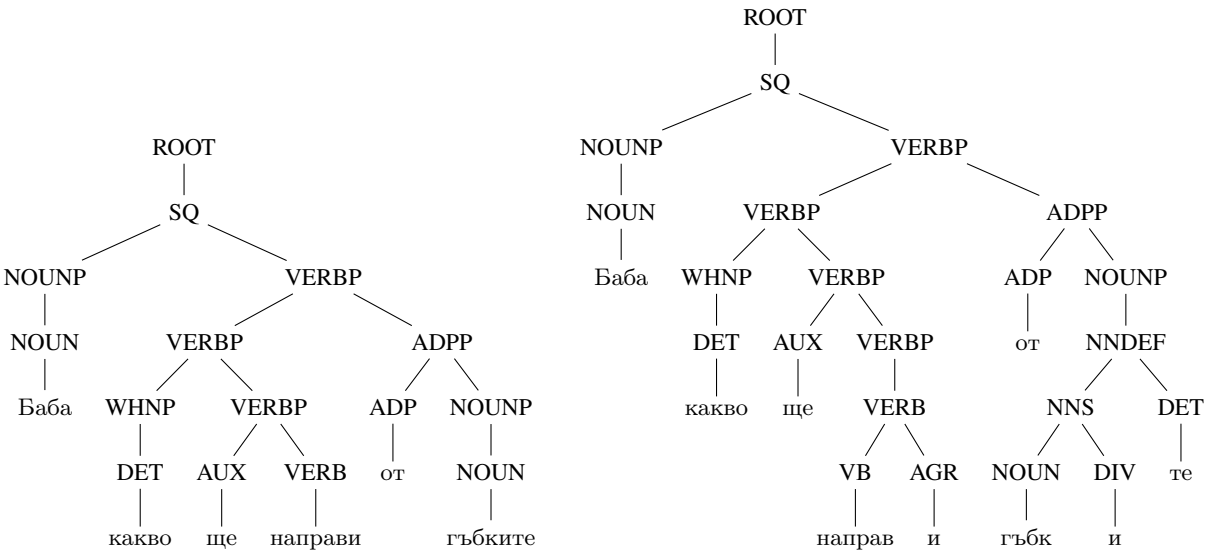


Figure 7: Баба какво ще направи от гъбките? ‘What will grandma make from the mushrooms?’

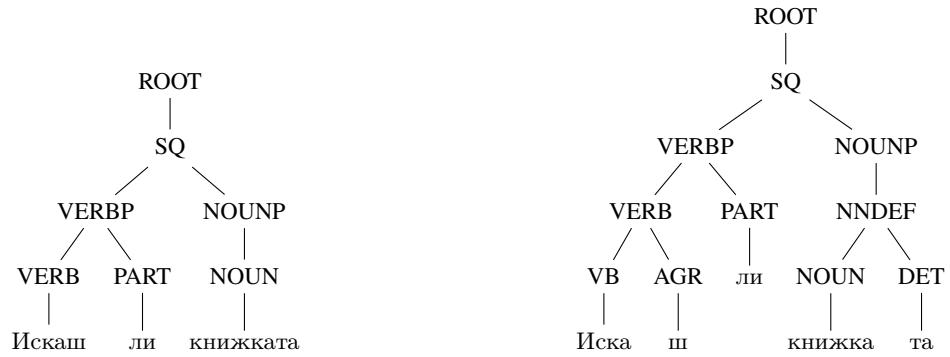


Figure 8: Искаш ли книжката? ‘Do you want the book?’

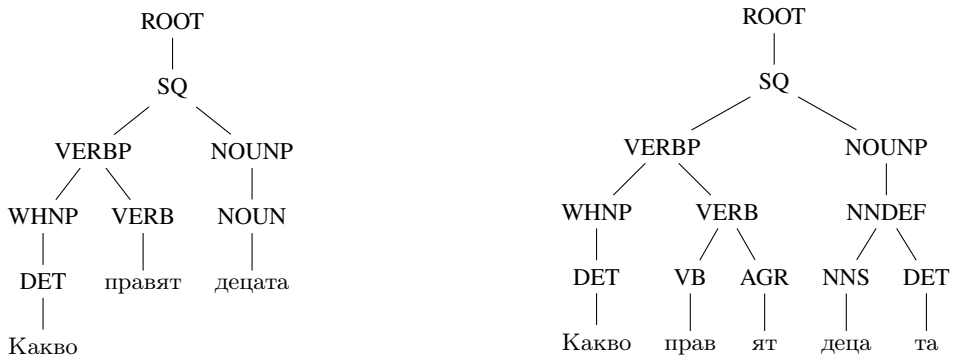


Figure 9: Какво правят децата? ‘What are the children doing?’

Special Session on Semantic Relations, Framenets and Ontologies

Toward a Serbian FrameNet: Resource Construction, Methodology, and Initial Annotation Results

Marina Bagi
The Institute for the Serbian Language of SASA
Serbia

Aleksandra Marković
The Institute for the Serbian Language of SASA
Serbia

Ranka Stanković
University of Belgrade – Faculty of Mining and Geology
Serbia

{marina.bagi|aleksandra.markovic}@isj.sanu.ac.rs ranka@rgf.rs

Abstract

This paper presents the initial steps toward developing the FrameNet lexical database for Serbian. FrameNet is a lexical database, primarily for English, based on Charles Fillmore's theory of Frame Semantics. Following the establishment of the Berkeley FrameNet, databases were developed for many other languages. There is growing interest in FrameNet and its applications in NLP, owing to its role in text structuring through role labeling. In addition to NLP, FrameNet has found applications across various research areas of applied linguistics.

Building FrameNet for Serbian is an important step in developing crucial resources for this underrepresented language. The paper reports on the construction of the Serbian database.

After translating the names of the English frames and frame elements into Serbian, we expanded the list with additional frames from the Brazilian database. We began our first annotations, and we present annotations of 162 sentences containing the lexemes *oštetiti* ('damage') and *uništiti* ('destroy'). As a result, we offer lexical unit reports for two frames nonexistent in the English FrameNet, specific to Serbian. We further discuss the main challenges and uncertainties encountered during the annotation process, as well as the solutions adopted. The development of the Serbian FrameNet contributes to the enrichment of multilingual FrameNet initiatives and provides a foundation for future research in semantic annotation, lexicography, and NLP applications for Serbian.

Keywords: FrameNet, Serbian language, database, destroying and damaging frames.

1 Introduction

In this paper, we present the initial phase of developing the Serbian FrameNet, a lexical database. FrameNet is based on Charles Fillmore's theory of Frame Semantics. Our work focuses on (i) adapting the FrameNet framework to Serbian, (ii) integrating existing lexical resources, and (iii) establishing an annotation workflow. We build upon the platform developed by the Brazilian FrameNet team (Torrent et al., 2025) and adapt it to the linguistic and resource-specific characteristics of Serbian.

The initial version of the resource includes translated frame and frame element inventories, aligned with the original FrameNet structure. The lexicon has been enriched with approximately 17,000 lexemes derived from the Leximirka lexical database, including their inflectional paradigms. In addition, a corpus of Serbian sentences has been integrated to support annotation and evaluation.

We have begun manual annotation of selected frames, focusing on the core semantic domains. Preliminary results indicate both the feasibility of direct frame transfer and the need for language-specific adaptations, particularly in cases of variation in argument structure and surface linguistic realization.

1.1 Theoretical background

FrameNet is an electronic lexicographic database of English, based on the theory of Frame Semantics – the cognitive-semantic approach to lexical semantic analysis, developed by Charles Fillmore and his collaborators. The research focus of this approach is on the meaning and use of words, and the idea is that lexical semantics should be described in relation to the frames to which words belong. Namely, the essential idea is that words of specific classes (verbs, nouns, adjectives, and adverbs) evoke frames – systems of concepts related in such a way that understanding any of these concepts requires understanding the entire system to which it belongs (Fillmore, 1982).

A semantic frame is a conceptual gestalt that could represent an event, state, attribute, relation,

or entity (Boas et al., 2024). Namely, Fillmore holds that word meanings are relativized to scenes (Fillmore, 1977) or frames, and that cultural and world knowledge motivates linguistic expressions and is embedded in them. That's why the FrameNet frame definition represents the schematic representation of the situation type that underlies the meaning of the word, together with the named roles of the participants or aspects of the situation (Fillmore et al., 2003).

The work on FrameNet, which began in 1997, can be considered an implementation of the theory of Frame Semantics, since it is organized around semantic frames evoked by different words. At the moment, it contains more than 10,000 lexical units which evoke frames (Boas et al., 2024). This database contains manually annotated sentences from real-world corpora and provides detailed, precise information on the semantic and syntactic properties of words. Namely, its importance is reflected in the fact that word meanings are explicitly related to the syntactic contexts in which they appear (e.g., FEs to their syntactic realizations), thereby enabling conclusions about semantico-syntactic interdependencies.

1.2 Related work

Linguistic researchers were drawn to such a detailed resource and motivated to adapt FrameNet to other languages. Over the last few years, similar lexicographic databases have been developed for Japanese, German, Spanish, Chinese, Brazilian Portuguese, Finnish, French, Greek, Hebrew, Italian, Latvian, Swedish, and other languages. Besides, the project *Multilingual FrameNet* has started, and interest in contrastive and typological research has increased, as has interest in creating frames for specific terminologies (Torrent et al., 2018; Boas et al., 2025).

Although primarily designed to improve and enrich lexicography with corpus-linguistic data based on semantic frame primitives (Fillmore and Atkins, 1994; Atkins et al., 2003a,b; Boas et al., 2024), FrameNet is being used across a wide range of research areas in linguistics, computational linguistics, language acquisition, natural language processing, and applied linguistics.

FrameNet was developed for Slavic languages, such as Bulgarian, BulFrameNet (Koeva, 2010), and research in this area has addressed various topics, such as (Stoyanova and Leseva, 2025). Initial results for Serbian frames were presented in (Marković et al., 2021; Dinić Marinković et al., 2024). The present research is part of a PhD thesis that examines the potential application of Frame Semantics theory to Serbian descriptive lexicography, with a focus on verbs of damage and destruction. Also, in (Bagi, 2025), a semantico-syntactic analysis of these verbs was presented. The frames evoked by these verbs were described, as well as the process of work on identifying and describing frames and frame elements. In the present research, we focus on sentence annotation involving the verbs damage and destroy.

In the remainder of the paper, we describe the methodology used to establish the database (Section 2.1) and the annotation process (Section 2.2). In Section 3, we present the annotation results, which we discuss in Section 4. We finish with the concluding remarks in Section 5.

2 Methods

2.1 Establishing the database

In preparing a Serbian FrameNet resource, we relied on several well-established linguistic resources and tailored them to the specific requirements of frame-semantic annotation. Given that Serbian is written in two scripts, Cyrillic and Latin, we standardized all resources in the Cyrillic script to ensure consistency across datasets and compatibility with existing language technology tools. This decision is also aligned with the predominant use of Cyrillic in formal linguistic resources for Serbian.

A central component of our work is the lexical resource, which contains 255,190 monolexemic entries formatted as ⟨`wordform`, `POS`, `lexeme`⟩ and covers 17,860 distinct lexemes. This lexicon was derived from Leximirka (Stanković et al., 2018; Lazić and Škorić, 2019), a comprehensive lexical database based on the electronic dictionaries (Krstev, 2008; Vitas and Krstev, 2012) developed over several decades by Cvetana Krstev and Duško Vitas. Among other features, Leximirka provides rich morphological information, including lemmas and their inflected forms, making it a suitable foundation for mapping lexical units to frames and supporting downstream annotation tasks.

The database import and configuration were carried out by Ely Matos and the development team. The database is currently hosted in a central repos-

itory[1], with plans to migrate it to a local Serbian installation.

POS	Lemmas	Word forms
V	3,708	112,119
A	4,067	82,756
N	8,887	58,514
PRON	67	863
ADV	625	625
NUM	34	93
PREP	77	77
ART	63	63
INTJ	29	29
CCON	28	28
SCON	23	23
Total	**17,508**	**254,670**

Table 1: Distribution of lexical categories (POS) sorted by frequency (word forms)

In addition to lexical resources, we constructed bilingual alignment lists to facilitate the transfer of frame-semantic knowledge from English to Serbian. The parallel list of frame names and frame elements includes manually translated and subsequently double-checked translations to ensure terminological accuracy and conceptual consistency. Initially, 1,224 English frame names were translated, along with an additional 209 frames identified during the process, resulting in 1,433 Serbian frame names. Furthermore, 1,251 distinct frame elements were translated, corresponding to 11,494 entries when accounting for their associated identifiers and variations.

To support empirical validation and annotation experiments, we also prepared a dataset of 600 Serbian sentences for frame annotation[2]. These sentences serve as an initial corpus for testing annotation guidelines, evaluating inter-annotator consistency[3], and exploring the applicability of frame-semantic structures in Serbian. The integration of these resources provides a solid foundation for developing a Serbian FrameNet, enabling both linguistic analysis and training computational models for semantic role labeling and related tasks.

[1] https://external.frame.net.br/

[2] We extracted these sentences from the Serbian referent corpora, mainly from SrpKor2021 (Vitas et al., 2025), available at: https://noske.jerteh.rs/. This corpus contains texts from different genres – publicistic, literary, and scientific – as well as from various internet domains.

[3] For this research, we had two annotators, but we didn't measure IAA; this could be a worthwhile area for future research.

2.2 Annotation

After establishing the Serbian FrameNet platform, we prepared a set of several hundred sentences containing the lexemes *oštetiti* (EN to damage) and *uništiti* (EN to destroy) and began manual annotation. In the Frame Semantics theory and in FrameNet, one of the central analytical elements is the *lexical unit*[4], which represents a word in one of its senses (Fillmore et al., 2003). In the typical case, every sense of a polysemous word belongs to a different semantic frame (Ruppenhofer et al., 2016). Since the SR lexemes *oštetiti* ('damage') and *uništiti* ('destroy') are polysemous, they are related to a number of frames: lexeme *oštetiti* with the frames `Oštećivanje (Damaging)`, `Uskraćivanje (Deprivation)`, `Izazvati_emociju (Cause_emotion)`, lexeme *uništiti* with the frames `Uništavanje (Destroying)`, `Dovesti_u_teško_stanje (Cause_to_be_in_a_difficult_state)`, `Pobediti_protivnika (Beat_opponent)`, `Izazvati_emociju (Cause_emotion)`, `Stimulisati_emociju (Stimulate_emotion)`, `Ubijanje (Killing)`. While working on manual annotation, we followed the guidelines in (Ruppenhofer et al., 2016). According to it, the annotation is primarily done relative to the target word (the word evoking the frame), and then the focus shifts to the target dependents, which are annotated as whole constituents.

3 Results

As mentioned above, following the guidelines in (Ruppenhofer et al., 2016), we annotated 162 sentences: 82 with the verb *to damage* and 80 with the verb *to destroy*. The FrameNet platform contains two annotation types: FE annotation (directed only to frame elements) and full-text annotation, which supports the so-called multilayer annotation related to the target word. We opted for the second type because it allows labeling frame elements (in further text, FEs) (Frame Element layer), their grammatical functions (Grammatical Function layer), and their grammatical realizations (Phrase Type layer).[5]

[4] In Serbian lingistics, the term sublexeme is used.

[5] For the moment, the layers' and labels' names are in English. We decided not to translate them. The possibility of translating them is open for a later time, and the database can

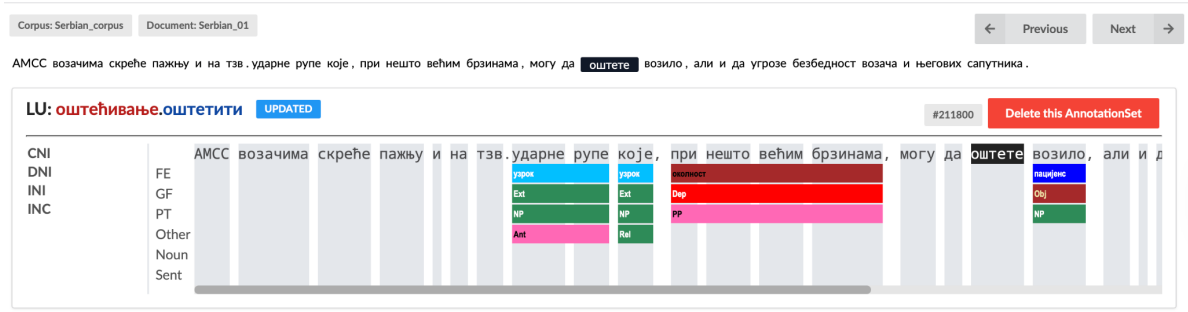


Figure 1: Annotation window

An example of the annotation window is shown in Figure 1.

3.1 The valence description

In order to briefly represent the results of the three-layer annotation of sentences with lexemes *damage* and *destroy*, we will represent a few Lexical Units reports, that is, syntactic realizations of FEs, and a part of the valence description for the frames non existing in the Berkley FrameNet. Those are frames `Uskraćivanje (Deprivation)`, which is evoked by a sublexeme *damage*, and frame `Dovesti_u_teško_stanje` (`Cause_to_be_in_a_difficult_state`), evoked by a sublexeme *destroy* (see Figure 2, 3, 4 and 5).

The `Deprivation` frame refers to a situation in which the AGENT affects the PATIENT in such a way that the Patient remains deprived of a STATE_OF_AFFAIRS they are expected to have, need, or want. The STATE_OF_AFFAIRS does not always have to be specified and may refer to the PATIENT's right to possess something. In this frame, sublexeme *damage* means 'to cause someone to be without something, to deprive someone of something').

As can be seen in Figure 4, the core elements of this frame are AGENT (less frequently, CAUSE), PATIENT, and STATE_OF_AFFAIRS, which is expressed by a PP with the preposition *za* ('for') and nominal in the accusative case. Other elements testified are CRITERION, EXPLANATION, ENABLING_FACTOR[6], REASON, CONTAINING_EVENT, TIME, MANNER, MEANS, DEGREE.

In frame `Cause_to_be_in_a_difficult_state`, the AGENT or CAUSE affects the PATIENT in such a way that the PATIENT (or some part of it) ends up in a completely non-canonical state, such that it no longer retains its original appearance, function, etc. The PATIENT loses its inherent characteristics and is nearly completely altered; the resulting state is extremely unfavorable. Sublexeme *to destroy* here doesn't mean 'to cause something to stop existing' but 'to cause an entity to end up in a severely compromised state'.

Therefore, core elements of this frame are AGENT, CAUSE, and PATIENT; besides, MALEFICIARY, EXPLANATION, CONTAINING_EVENT, FREQUENCY, TIME, INSTRUMENT, PLACE, MANNER, MEANS, and DEGREE appear.

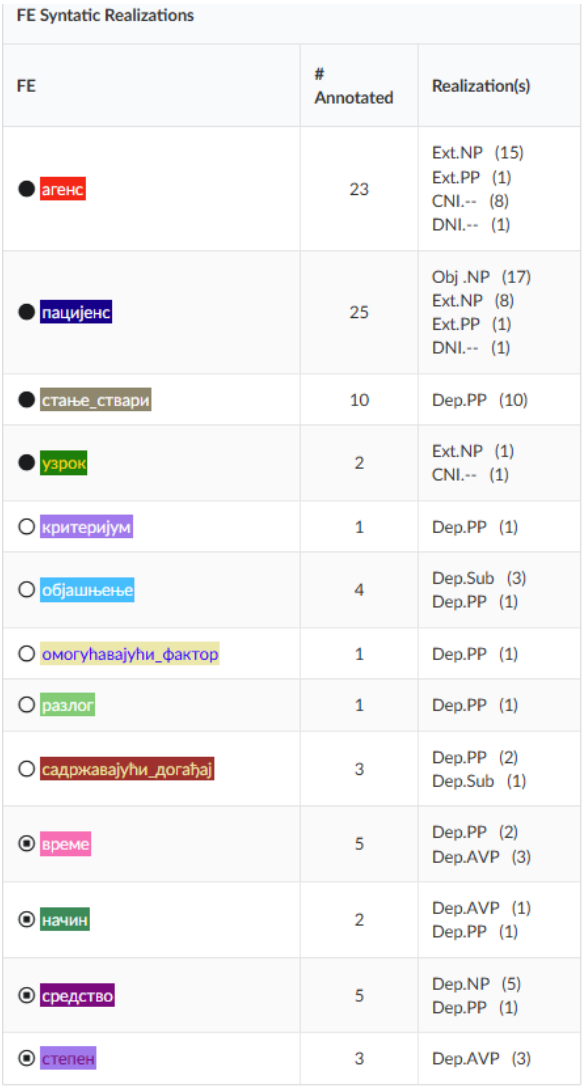
4 Discussion

During the annotation process, we encountered several challenges, some of which are inherent to Serbian. Firstly, they are connected with the null instantiation cases – cases in which conceptually salient frame elements don't appear as lexical or phrasal material in the sentence. It is well-known that Frame Semantics differ among the three types of null instantiation: DNI (Definite Null Instantiation), INI (Indefinite Null Instantiation), and CNI (Constructional Null Instantiation).

Related to the last case, the omission of the frame element licensed by the grammatical construction in which the target word appears (Ruppenhofer et al., 2016), such as the omission of subjects in imperative and passive sentences, posed several issues. Namely, as a pro-drop language, Serbian allows the omission of the subject of any predicate clause, since it is "contained" in the predicate form (the finite form of the verb, which differs for per-

be enriched with Serbian labels. For example, in the Phrase Type layer, under the label PP (preposition phrase), the most frequently used prepositions in Serbian should be entered into the database to make the information on the syntactic realizations of frame elements more complete.

[6]As far as we know, this FE doesn't exist in the English FrameNet. In Serbian, it is expressed by a PP *zahvaljujući* and a dative nominal (*zahvaljujući njemu* 'thanks to him'), and it refers to an element that doesn't directly cause a certain event, but indirectly enables or makes it easier.



FE Syntatic Realizations

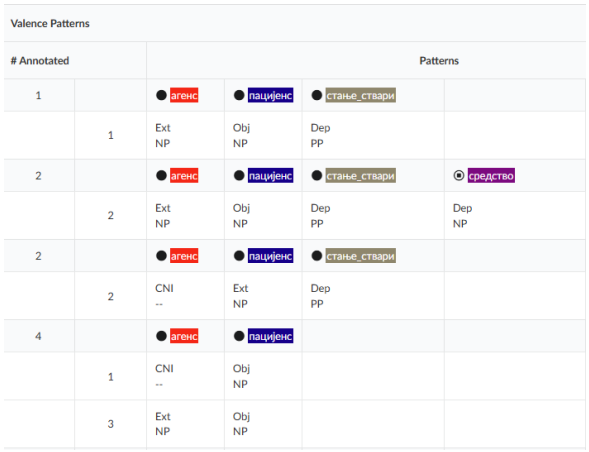
FE	# Annotated	Realization(s)
агенс	23	Ext.NP (15) Ext.PP (1) CNI.-- (8) DNI.-- (1)
пацијенс	25	Obj .NP (17) Ext.NP (8) Ext.PP (1) DNI.-- (1)
стање_ствари	10	Dep.PP (10)
узрок	2	Ext.NP (1) CNI.-- (1)
критеријум	1	Dep.PP (1)
објашњење	4	Dep.Sub (3) Dep.PP (1)
омогућавајући_фактор	1	Dep.PP (1)
разлог	1	Dep.PP (1)
садржавајући_догађај	3	Dep.PP (2) Dep.Sub (1)
време	5	Dep.PP (2) Dep.AVP (3)
начин	2	Dep.AVP (1) Dep.PP (1)
средство	5	Dep.NP (5) Dep.PP (1)
степен	3	Dep.AVP (3)

Figure 2: LU: deprivation.damage

son and number). For example, in the sentence *Uništio si me* (EN lit. (You) destroyed me, 'You broke me'), it is clear that the agent is *you*, since it is "embedded" in the finite, congruent form of the verb. Also, the word *you* is considered the grammatical subject of the respective sentence. What we want to say is that, in this example, in the situation denoted by the verb *destroy*, the agent exists but is not explicit. It is clear that these cases differ from deagentization[7] in which the agent is removed from the sentence's structure and cannot occupy the grammatical subject position,[8] as in passive sentences. In the sentence *Na dokovima je oštećen jedan teretni brod* ('A cargo ship was damaged at

[7]This term is used in the Serbian syntax, for backgrounding or removing an agent. A more usual term is *agent demotion*, like in (Matthies, 2022).

[8]Though sometimes present as an agentive complement or adjunct.



Valence Patterns

# Annotated		Patterns			
1		агенс	пацијенс	стање_ствари	
	1	Ext NP	Obj NP	Dep PP	
2		агенс	пацијенс	стање_ствари	средство
	2	Ext NP	Obj NP	Dep PP	Dep NP
2		агенс	пацијенс	стање_ствари	
	2	CNI --	Ext NP	Dep PP	
4		агенс	пацијенс		
	1	CNI --	Obj NP		
	3	Ext NP	Obj NP		

Figure 3: LU: deprivation.damage

the docks'), there is no agent, and it is not related to any person. In other words, the situation is described with an anonymized agent.

This was the reason we asked whether we should introduce a new label for the CNI in cases of agent omission in active sentences. One of the solutions was proposed in (Subirats, 2009), which introduces the label ECNI from the Spanish FrameNet. Typically, it is used to instantiate external arguments (External Constructional Null Instantiation), such as subjects. ECNI is not licensed lexically but constructionally, e.g., it depends on sentential constructions with predicative or support verbs, and it is subject to contextual and pragmatic limitations. If we implemented this solution, our database would have two labels for the two types of CNI: CNI in passive, and ECNI in active sentences.

During annotation, we encountered a few additional issues. As already mentioned, annotation is done relative to the target word: only those frame elements that are part of the maximal phrase headed by the target word are annotated (Ruppenhofer et al., 2016). But, sometimes it is not completely clear what is to be considered a scope of the target word (in our case, the predicate of a clause), especially in cases in which two clauses are in copulative coordination, or when there are two predicative verbs, the other additionally specifying the action denoted by the first one. This is the case in the example, where the action of destroying is specified by the construction *da uništi, odnosno zapali* (to destroy, more precisely, to burn). That's why we wonder whether the patient (*dokaz svoje pobede*, 'proof of his/her victory') falls under the maximal phrase of the target (Figure 6).

Further, since subjects can be out of the clause in which the target is, we wonder if it is neces-

FE Syntatic Realizations

FE	# Annotated	Realization(s)
агенс	9	CNI.-- (3) Ext.NP (7)
пацијенс	22	Obj .NP (22) DNI.-- (2)
узрок	13	Ext.NP (13) Ext.PP (1) Ext.VPing (1)
малефицијар	3	Dep.NP (3)
објашњење	1	Dep.Sub (1)
садржавајући_догађај	1	Dep.PP (1)
фреквенција	1	Dep.AVP (1)
време	3	Dep.NP (1) Dep.PP (1) Dep.AVP (1)
инструмент	4	Dep.NP (4)
место	2	Dep.PP (2)
начин	3	Dep.AVP (3)
средство	1	Dep.VPing (1)
степен	3	Dep.AVP (2) Dep.Sub (1)

Figure 4: LU: cause_to_be_in_a_difficult_state.destroy

sary to annotate the whole subject constituent. In other words, in our example: *Protiv Šarančića i trojice njegovih saradnika, koji su uhapšeni 2. juna u Beogradu, Posebno odeljenje Višeg suda u Beogradu vodi istragu zbog sumnje da su zloupotrebom službenog položaja oštetili Železnice Srbije za 1.24 miliona evra kupovinom vozova iz Švedske i Slovenije* ('The Special Department of the Higher Court in Belgrade is conducting an investigation against Šarančić and three of his associates, who were arrested on June 2 in Belgrade, on suspicion of having abused their official positions and caused damage to Serbian Railways amounting to €1.24 million through the purchase of trains from Sweden and Slovenia'), agent is in the preposition phrase (PP) *against* + nominal phrase (NP) in genitive case; we are not sure whether to annotate preposition *against* as part of the agent. If we annotate it, in the valence description of the lexical unit (sublexeme) *damage*, we will find that the agent is expressed by PP, which doesn't hold for the sit-

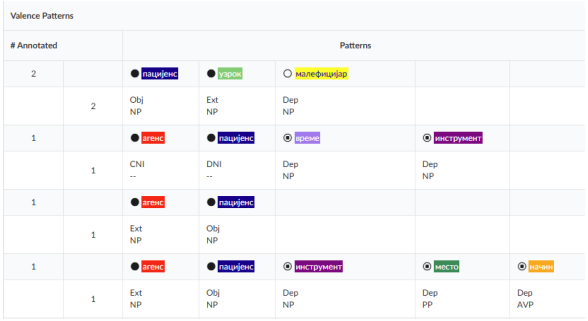


Figure 5: LU: cause_to_be_in_a_difficult_state.destroy

uation denoted by this sublexeme – that form is governed by the predicate of the other clause, the verb *to conduct* (to conduct an investigation against someone) (see Figure 7 and 8).[9].

We also wonder whether we do the right thing when the target is in a relative clause, and its antecedent is annotated as a whole PP. This is the case with an example: *Očekivali smo veću ažurnost sudskih organa po pitanju donošenja presude za huligane koji su spalili i uništili džamiju u Nišu* ('We expected the judicial authorities to act more promptly in issuing verdicts for the hooligans who set fire to and destroyed the mosque in Niš'), in which the antecedent is PP *for the hooligans* (see Figure 9[10].

In the rest of the section, we will briefly present typical dilemmas arising from cases in our corpus. One of them concerns the possibility of a double interpretation of some FEs. Namely, in some ex-

[9]According to Fillmore (Fillmore, 2007), Ext was chosen as the subject marker instead of, for example, Subj (Subject), because External Grammatical Function doesn't correspond only to the subject of a finite sentence. Also, the argument that participates in the grammatical construction with the target word will be marked as Ext, as in the example: *We persuaded [the general] to* ***release*** *the prisoners*, in which *the general* is the external argument of the verb *to release*. For this reason, we have decided to record Ext in such examples, even when it is not the subject of the clause containing the target word, but refers to it. It seems that these cases should be distinguished from cases of null instantiation, especially DNI or CNI, because they typically refer to instances when there is no subject in the sentence at all. In the case of DNI, the subject can be recovered from the preceding context, rather than from the sentence being annotated. Of course, Serbian allows the omission of overt subjects with finite verbs, but if two predicates share the same subject in the sentence, that is, if the subject is explicitly present, we decided to mark it as Ext.

[10]The answer to the questions we posed regarding the coreference of subjects and antecedents of relative clauses isn't straightforward, since it is not covered by (Ruppenhofer et al., 2016). These issues are relevant for the FrameNet, especially the latter, since it is necessary to annotate antecedent and relativizer on the fourth layer, if the target word occurs in the relative clause.



Figure 6: Maximal phrase of the target

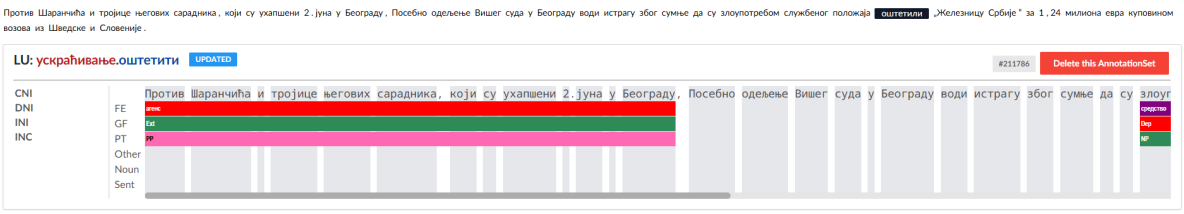


Figure 7: The subject constituent in PP(1)

amples, CONTAINING_EVENT and CAUSE could be used to annotate the same constituent.

E1 Godine 1984, Teri Volis je doživeo saobraćajni udes u kojem mu je centralni nervni sistem teško oštećen. ('In 1984, Terry Wallis suffered a traffic accident in which his central nervous system was severely damaged.')

E2 U oblasti severno od velikog turističkog centra Orlanda, u oluji je oštećeno ili uništeno više od 500 kuća, zgrada i crkava. ('In the area north of the major tourist center of Orlando, more than 500 houses, buildings, and churches were damaged or destroyed in the storm.')

In the example E1, saobraćajni udes[Ant] u kojem[Rel] ('traffic accident[Ant] in which[Rel]'), *traffic accident* could be interpreted as a CONTAINING_EVENT or as a CAUSE. But, if we reformulate the sentence into active voice, we couldn't tell that the traffic accident severely damaged the central nervous system, which was the reason for us to annotate this constituent as a CONTAINING_EVENT. A similar situation arises in sentence E2, where the FE *u oluji* ('in the storm') could be interpreted as either a CAUSE or a CONTAINING_EVENT. In an equivalent active sentence, we say that the storm destroyed more than 500 houses, buildings, and churches, which is why we annotated *in the storm* as a CAUSE.

Sometimes it is difficult to make a distinction between FEs CONTAINING_EVENT and TIME:

E3 *Kritičari su me jako oštetili kada nisam dobio nagradu za Prisluškivanje* ('The critics really hurt me when I was denied the award for Eavesdropping').

Namely, the constituent *kada nisam dobio nagradu za Prisluškivanje* ('when I was denied the award for Eavesdropping') could be considered an event that entails the action of damaging (CONTAINING_EVENT) or TIME the damaging is taking place. We consider that it is better to give preference to FE CONTAINING_EVENT, since the component of time, though present, doesn't dominate.

One of the most frequent issues we encountered concerns cases in which it is unclear whether a given construction should be considered part of a single FE or represents a separate FE. There are cases in which certain PP can be interpreted as a modifier in an NP, modifying a nominal constituent, or as a place modifier of a predicate. Here are some examples:

E4 Prema prvim procenama, **požar u Hilandaru** uništio je dve trećine manastirskog kompleksa. ('According to initial estimates, **the fire in Hilandar** destroyed two-thirds of the monastery complex.)

E5 Vesić je u prijavi naveo da je tokom noći izmedju 12. i 13. jula grupa nepoznatih lica sprejevima uništila **upravo rekonstruisanu fasadu zgrade u Pop Lukinoj ulici br. 12 na opštini Stari Grad u Beogradu** ('In the report, Vesić stated that during the night between July 12th and 13th, a group of unknown individuals used spray paint to destroy **the recently renovated facade of the building at 12 Pop Lukina street in the municipality of Stari grad in Belgrade'**).

E6 Albanska policija je saopštila da je uništila **7300 stabljika marihuane na severu zemlje**

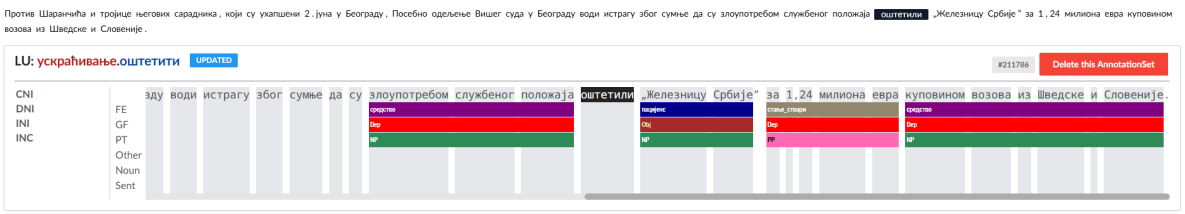


Figure 8: The subject constituent in PP(2)

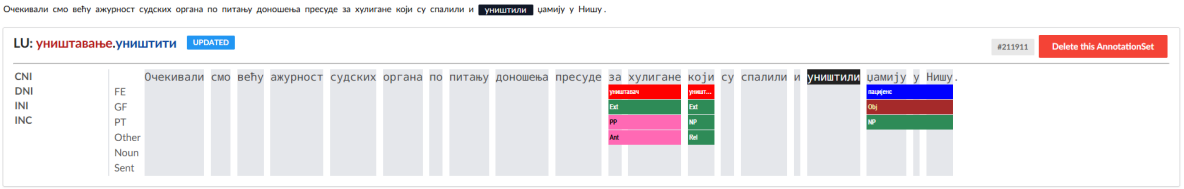


Figure 9: Antecedent of the subject in relative clause as PP

(The Albanian police announced that they had destroyed **7,300 marijuana plants in the north of the country**).

In example E4, PP *in Hilandar* could be considered a dependent modifier in NP, with the noun *fire* as a head (the fire that broke out in Hilandar), or it could be interpreted as a separate constituent, a place modifier. In that case, it would modify the action denoted by the verb *damage*, and it would be an answer to a question: where the fire destroyed two-thirds of the monastery complex? Though these constructions really are ambiguous, it seems that in this case, PP *in Hilandar* is the modifier of the noun *fire*.

A similar situation arises with PPs from examples E5 (at 12 Pop Lukina Street in the municipality of Stari grad in Belgrade) and E6 (in the north of the country), which could also be interpreted in two ways. We consider that they should be regarded as a separate constituent, or separate FE, PLACE, since they modify the predicate to destroy, and they indicate where the action of destroying is taking place.

Some FEs were difficult to differentiate. This holds for the relation between the elements NAČIN (MANNER) and OPIS_DOGADJAJA (EVENT_DESCRIPTION). Namely, it should be noted that the FrameNet element MANNER represents a broader category than what is considered an adverbial of manner in Serbian syntax. Under this category is understood the way an action is performed (i.e., the answer to a question about how something is done). In FrameNet, this FE in the frame `Destroying` relates to "any description of the process of destruction which is not covered by more specific FEs, including epistemic modification (probably, presumably, mysteriously), secondary effects (quietly, loudly), and general descriptions comparing events (the same way). It may also indicate salient characteristics of the DESTROYER that also affect the action (presumptuously, coldly, deliberately, eagerly, carefully)"[11]. Since it includes epistemic modifiers and general event descriptions, FE MANNER is hard to differentiate from EVENT_DESCRIPTION, extra-thematic FE, typically used for expressions describing the clause of the target word as a whole, representing a sentential modifier, like a word *oddly* from the next example: [Oddly], there was a very small earthquake in Miami today (Ruppenhofer et al., 2016).

There are also cases in which a modifier could be interpreted as both clausal and sentential:

E7 Načelnik odeljenja za strance u Ljubljani Marjan Kompara je u svojoj kancelariji **bukvalno** makazama razrezao, uništio moj pasoš (The head of the foreigners' department in Ljubljana, Marjan Kompara, **literally** cut up and destroyed my passport with scissors in his office).

Though *literally* is not between commas, it maybe could be considered as a sentential modifier, since the word *literally* represents a comment which relates to a whole complex sentence. Moreover, it cannot be said that a passport is literally being destroyed; it doesn't specify how this action is performed. This word expresses the speaker's amazement at the fact that the DESTROYER really cut and destroyed his or her passport. As for now, we left this FE as a MANNER, mostly because it is not between commas.

[11] https://framenet.icsi.berkeley.edu/frames/Destroying

5 Conclusions

In this paper, we tried to explain all the doubts we faced during the first phase of work on the Serbian FrameNate database: from translating frames and their elements, to more general dilemmas related to important terms (like constructionally-licensed null instantiation) and FEs (CONTAINING_EVENT and CAUSE), to specific sentences and some phenomena specific to the Serbian language.

Previous papers on the study of various phenomena in Serbian, based on Frame Semantics theory, as well as our analysis, show that Serbian can be analyzed within this framework, yielding a detailed semantico-syntactic description of the use of words evoking frames.

We assume that English frames can be adapted to Serbian, but they may need to be modified and, in some cases, additional frames typical of Serbian may need to be added.

As is well-known, the work on FrameNet is not linear, and some details will be revised and possibly transformed in future work on the database.

In order to complete the analysis of frames evoked by verbs *damage* and *destroy*, it is necessary to include other sublexemes that evoke these frames, as well as present the relations among the frames and FEs.

Since this lexicographic database is a unique resource, we consider it important to continue building FrameNet, especially given that Serbian is an underresourced language.

Acknowledgments

This research was supported by the Science Fund of the Republic of Serbia under Project No. 7276, "Text Embeddings – Serbian Language Applications" (TESLA) and the Ministry of Science of the Republic of Serbia, (451-03-33/2026-03/200174).

References

Sue Atkins, Charles J Fillmore, and Christopher R Johnson. 2003a. Lexicographic Relevance: Selecting Information from Corpus Evidence. *International Journal of lexicography*, 16(3):251–280.

Sue Atkins, Michael Rundell, and Hiroaki Sato. 2003b. The Contribution of FrameNet to Practical Lexicography. *International Journal of Lexicography*, 16(3):333–357.

Marina Bagi. 2025. Sintaksička i semantička analiza glagola "oštetiti" i "uništiti" iz ugla teorije semantike okvira. In *Proceedings of the International Conference South Slavic Languages in the Digital Environment JuDig: Thematic Collection of Papers*, volume 1, pages 181–207.

Hans C Boas, Josef Ruppenhofer, and Collin Baker. 2024. Framenet at 25. *International Journal of Lexicography*, 37(3):263–284.

Hans C Boas, Josef Ruppenhofer, and Collin F Baker. 2025. FrameNet at 25: Results and Applications. *International Journal of Lexicography*, 38(2):159–189.

Milica Dinić Marinković, Aleksandra Marković, and Ranka Stanković. 2024. FrameNet as a Tool for Tracking Semantic Development in FLA. Case of Lexical Unit "uspomena" (eng. memory.n) in School-age Children in Serbia. In *21st EURALEX International Congress - 2024*.

Charles J Fillmore. 1977. The Case for Case Reopened. In *Grammatical relations*, pages 59–81. Brill.

Charles J Fillmore. 1982. Frame Semantics. *Linguistics in the Morning Calm: Selected Papers from SICOL-1981*, pages 111–137.

Charles J. Fillmore. 2007. *Valency issues in FrameNet*, pages 129–162. De Gruyter Mouton, Berlin, New York.

Charles J Fillmore and Beryl TS Atkins. 1994. Starting where the Dictionaries Stop: The challenge of Corpus Lexicography. *Computational approaches to the lexicon*, pages 349–394.

Charles J Fillmore, Miriam RL Petruck, Josef Ruppenhofer, and Abby Wright. 2003. FrameNet in Action: The Case of Attaching. *International journal of lexicography*, 16(3):297–332.

Svetla Koeva. 2010. Lexicon and grammar in Bulgarian FrameNet. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Cvetana Krstev. 2008. *Processing of Serbian – Automata, Texts and Electronic Dictionaries*. University of Belgrade, Faculty of Philology, Belgrade.

Biljana Lazić and Mihailo Škorić. 2019. From DELA Based Dictionary to Leximirka Lexical Database. *INFOtheca: Journal of Information and Library Science*, 19:81–98.

Aleksandra Marković, Ranka Stanković, Natalija Tomić, and Olivera Kitanović. 2021. FrameNet Lexical Database: Presenting a Few Frames Within the Risk Domain. *Infotheca–Journal for Digital Humanities*, 21(1):7–33.

Jochen Matthies. 2022. Agent Demotion in German and Polish. *Linguistik online*, 115(3):91–118.

Josef Ruppenhofer, Michael Ellsworth, Myriam Schwarzer-Petruck, Christopher R Johnson, and Jan Scheffczyk. 2016. *FrameNet II: Extended Theory and Practice*. International Computer Science Institute, Berkeley, California.

Ranka Stanković, Cvetana Krstev, Biljana Lazić, and Mihailo Škorić. 2018. Electronic Dictionaries–from File System to Lemon Based Lexical Database. In *Proceedings of the 6th Workshop on Linked Data in Linguistics (LDL-2018): Towards Linguistic Data Science*.

Ivelina Stoyanova and Svetlozara Leseva. 2025. Towards Classifying Predicates Involving Change. In Jasmina Moskovljević Popović and Ranka Stanković, editors, *Proceedings of the International Conference South Slavic Languages in the Digital Environment JuDig : Thematic Collection of Papers*, volume 1 of *South Slavic Languages in the Digital Environment JuDig*, chapter 8, pages 129–146. University of Belgrade — Faculty of Philology, Belgrade. 8.

Carlos Subirats. 2009. Spanish FrameNet: A frame-semantic Analysis of the Spanish Lexicon. *Multilingual FrameNets in computational lexicography: Methods and applications*, pages 135–162.

Tiago T. Torrent, Michael Ellsworth, Collin F. Baker, and Ely Edison da Silva Matos. 2018. The Multilingual FrameNet Shared Annotation Task: A Preliminary Report. In *Proceedings of the LREC 2018 Workshop International FrameNet Workshop 2018: Multilingual FrameNets and Constructicons*, pages 62–68, Miyazaki, Japan. European Language Resources Association (ELRA).

Tiago Timponi Torrent, Ely Edison da Silva Matos, Alexandre Diniz da Costa, Maucha Andrade Gamonal, Simone Peron-Corrêa, and Vanessa Maria Ramos Lopes Paiva. 2025. A flexible tool for a qualia-enriched FrameNet: the FrameNet Brasil WebTool. *Language Resources and Evaluation*, 59(4):4529–4557.

Duško Vitas and Cvetana Krstev. 2012. Processing of Corpora of Serbian Using Electronic Dictionaries. *Prace Filologiczne*, LXIII:279–292.

Duško Vitas, Ranka Stanković, and Cvetana Krstev. 2025. O familiji korpusa savremenog srpskog jezika srpkor. In Jasmina Moskovljević Popović and Ranka Stanković, editors, *Proceedings of the International Conference South Slavic Languages in the Digital Environment JuDig*, volume 1, pages 3–27. university of Belgrade – Faculty of Philology.

A WordNet-Aligned Taxonomy of Logical Fallacies for Low-Resource NLP

Stanislav Penkov
GATE Institute
Sofia University "St. Kliment Ohridski"
stanislav.penkov@gate-ai.eu

Abstract

We present a compact taxonomy of eight logical fallacies designed for the analysis of manipulative reasoning and argumentative discourse, with a specific focus on low-resource settings. The taxonomy consolidates multiple public datasets into a nine-class training target (eight fallacies plus a background label *no fallacy*), with class distributions documented to support future stratified sampling or class-weighted modeling, and organizes these classes within a small parent–child hierarchy (relevance, presumption, causality, reasoning form) used for interpretability only (not as training labels), semantically grounding the target categories in WordNet 3.0 via synset IDs and offsets. From corpus evidence, we extract representative lexical cues using log-odds weighting with lemmatization and bigram detection, and link them to synsets to form a two-layer conceptual–lexical mapping. Crucially, the WordNet anchoring enables cross-lingual projection via Open Multilingual WordNet (OMW), supporting the creation of candidate seed lexicons and potential weak-supervision signals without prior in-language fallacy annotation. The framework thus provides a practical recipe for taxonomy-based lexical-semantic resource development, multilingual bootstrapping, and interpretable, fallacy-aware natural language processing (NLP).

Keywords: logical fallacies, WordNet, Open Multilingual WordNet, lexical-semantic resources, low-resource NLP, argumentation

1 Introduction

Logical fallacies — recurrent patterns of flawed reasoning that undermine the power of argumentation (Tindale, 2007) — are pervasive in public discourse. They are central to manipulative reasoning — a key mechanism through which information manipulation and propaganda operate in public discourse. Logical fallacy detection has thus emerged as an important component of explainable manipulative reasoning and argumentation analysis, where shallow linguistic cues often mask deeper reasoning flaws. However, this task is hindered by a fragmented resource ecosystem. Existing datasets adopt inconsistent label sets, complicating reuse and comparison across studies. Furthermore, prominent resources such as LOGIC (Jin et al., 2022), MAFALDA (Helwe et al., 2024), and CoCoLoFa (Yeh et al., 2024) focus almost exclusively on English and exhibit fragmented label schemes that hinder cross-lingual transfer. To address this, we propose a compact taxonomy of eight fallacy categories (with a separate *no fallacy* background label) semantically grounded in WordNet 3.0 (Princeton University, 2006), providing a stable schema that is portable across languages and resources. For low-resource languages such as Bulgarian — used here as a first test case — where annotated argumentation data is scarce, this semantically grounded taxonomy provides a crucial foundation for future resource building and for aligning large language model (LLM) outputs with interpretable reasoning categories. We will register the taxonomy, mappings, and normalized counts in the LRE Map to facilitate reuse and comparability across languages.

This work forms part of a broader research line on reasoning-aware NLP and manipulative rhetoric detection. Within this program, the present study contributes the linguistic and semantic infrastructure — a compact, WordNet-aligned taxonomy and normalized dataset — intended to support subsequent modules for fallacy detection and explainable argument analysis in low-resource languages.

Contributions. (i) A unified 8-class fallacy schema (plus a *no fallacy* background label) with transparent mappings from five widely used datasets. We selected the eight most stable, cross-source categories with clean WordNet anchors; rarer or source-

specific labels are retained in metadata only;

(ii) a WordNet 3.0-anchored conceptual layer with stable offsets, coupled with a data-driven lexical layer of representative cues;

(iii) a cross-lingual projection recipe that leverages Open Multilingual WordNet (OMW) to derive seed lexicons and weak supervision signals for low-resource languages without prior in-language fallacy annotation, illustrated through Bulgarian projection examples;

and (iv) a reproducible normalization pipeline that others can adopt to extend the resource.

2 Related Work

Research in argument mining and fallacy detection has proposed numerous taxonomies and datasets across diverse discourse domains, ranging from debate platforms (Habernal and Gurevych, 2017) and essays (Stab and Gurevych, 2017) to social-media and news discussions (Yeh et al., 2024). However, their label definitions and taxonomic structures vary substantially, limiting cross-dataset comparability and semantic reuse.

Early efforts such as *Argotario* (Habernal et al., 2017) gamified the creation of fallacy-annotated arguments through multilingual player interactions, introducing a scalable, crowd-driven approach to fallacy data collection. Related work by Habernal and Gurevych (2017) identified argumentative relations in user-generated web discourse, while Stab and Gurevych (2017) analyzed insufficiently supported claims in essays. Fallacy-specific expansions followed, such as *ad hominem* recognition — the identification of personal attacks directed at the speaker rather than the argument itself — (Habernal et al., 2018) and reasoning comprehension benchmarks (Wachsmuth et al., 2017).

A major step toward standardized evaluation came with Jin et al. (2022), who formalized the detection of logical fallacies as an NLP task and released the `LOGIC` and `LOGIC_CLIMATE` datasets (13 classes, ∼3.5k examples). Their structure-aware classifier masked semantically similar spans to highlight logical form over content, providing the first strong reasoning-based baseline. Sourati et al. (2023) extended this work via case-based reasoning (CBR), retrieving and adapting analogous examples enriched with goals, counterarguments, and structural annotations to improve cross-domain robustness. Lei and Huang (2024) further advanced structural modeling by introducing a Logical Structure Tree representation integrated into large language models through hard and soft prompts, achieving state-of-the-art results.

More recently, Yeh et al. (2024) combined LLM-assisted crowdsourcing and expert review to construct the largest English fallacy dataset (7,706 annotated news comments across eight categories). By integrating GPT-4 (OpenAI, 2023) as an in-task writing assistant, it demonstrated how hybrid human–AI pipelines can generate linguistically rich and contextually realistic fallacious discourse that closely resembles "in-the-wild" arguments. In parallel, Helwe et al. (2024) unified several existing corpora under a harmonized taxonomy, further establishing multi-domain evaluation standards.

Despite these advances, most resources standardize format rather than semantics. Few have linked fallacy or argument-type taxonomies to general lexical-semantic databases such as WordNet (Princeton University, 2006) and FrameNet (Baker et al., 1998). To our knowledge, fallacy categories have rarely been aligned with such lexical-semantic resources at the schema level, constraining semantic interoperability and multilingual reuse. Our work addresses this gap by pursuing taxonomy-driven, multilingual alignment between fallacy categories and general lexical resources, enabling cross-lingual representation within a unified conceptual framework.

3 Taxonomy Construction and Dataset Normalization

Recent work (e.g., Helwe et al. 2024) has reviewed extensive fallacy catalogues, but most lack semantic grounding for computational use. We prioritize interpretability and tractability. We consolidate the empirically represented fallacy types across five public datasets—Argotario, `LOGIC_EDU`, `LOGIC_CLIMATE`, MAFALDA, and CoCoLoFa—normalize them to a common JSONL schema, and merge into 12 531 examples before filtering, yielding a unified schema suitable for multilingual, reasoning-aware NLP.

Selection criteria. The eight fallacy categories were selected according to four practical criteria: (i) cross-source stability, i.e., the category or a close equivalent occurs in more than one existing resource or is clearly represented in a major source dataset; (ii) sufficient empirical support after normalization, so that the class can be meaningfully profiled using corpus-derived lexical evidence; (iii)

semantic anchorability, meaning that the category can be linked to one or more interpretable WordNet concepts without relying only on dataset-specific labels; and (iv) usefulness for low-resource transfer, where broad and recurrent reasoning patterns are preferable to fine-grained labels that cannot be reliably projected across languages. Rarer, highly source-specific, or theoretically ambiguous labels were therefore retained in metadata rather than included in the main training target. This design favours a compact and reusable schema over exhaustive coverage.

Parent–Child structure. Beyond the nine target classes used for modeling (eight fallacies plus `no_fallacy`), we attach a compact set of parent families that explain *why* the reasoning fails: *relevance* (redirection/irrelevant warrant), *presumption* (claim exceeds evidence), *causality* (unwarranted causal structure), and *reasoning form* (self-supporting loop). This hierarchy is used only for interpretability and error analysis and does not define the training label space. The intermediate label `causal_fallacy` consolidates source labels related to false causality and causal oversimplification for taxonomic analysis. It is assigned to the parent family `causality`, but is not included in the nine-class training target. The original labels and their intermediate mapping are retained in metadata rather than silently reassigned to another target class.

Consolidation rules (child → unified). We harmonize heterogeneous labels in lowercase with underscores. Table 1 shows the child → unified mappings used to normalize sources; arrows indicate the normalized unified label; not all unified labels are members of the nine-class target. Records mapped to `causal_fallacy` are excluded when filtering to the nine-class target but remain available in the normalized pre-filter corpus with their original and intermediate labels preserved in metadata.

Note on `appeal_to_worse_problems`. We group this label under `red_herring` (relevance) as it redirects attention to a different issue rather than supporting the claim; when realized as whataboutism, it functions as a relevance diversion.

Rationale (Popularity/Crowd). We fold `appeal_to_tradition` under `appeal_to_popularity` because its warrant is social acceptance via longevity ("it's always been done") rather than authority or evidence,

Family	**Child labels → Unified**
Popularity/Crowd	`ad_populum`, `appeal_to_majority`, `appeal_to_tradition` → `appeal_to_popularity`
Authority/Credibility	`fallacy_of_credibility`, `irrelevant_authority`, `appeal_to_false_authority`, `appeal_to_authority` → `appeal_to_authority`
Emotion	`appeal_to_fear`, `appeal_to_pity`, `appeal_to_anger`, `appeal_to_ridicule`, `appeal_to_positive_emotion`, `appeal_to_nature` → `appeal_to_emotion`
Relevance/Redirection	`fallacy_of_relevance`, `appeal_to_worse_problems`, `straw_man`, `fallacy_of_extension`, `red_herring` → `red_herring`
Generalization	`faulty_generalization`, `hasty_generalization` → `hasty_generalization`
Forced choice	`false_dilemma` → `false_dilemma`
Causality	`false_causality`, `causal_oversimplification` → `causal_fallacy`; `slippery_slope` kept as-is
Person-based	`guilt_by_association`, `tu_quoque`, `ad_hominem` → `ad_hominem`
Background & other	`none`/`no_fallacy`/`nothing` → `no_fallacy`; `circular_reasoning` retained in metadata (parent-only; not a training label); `intentional` excluded from 9-class training target

Table 1: Label consolidation map used to normalize heterogeneous source labels. Only records whose unified labels belong to the nine-class target, including `no_fallacy`, are retained during target filtering. Intermediate and parent-only labels are preserved for analysis and provenance.

keeping "popular acceptance ⇒ correctness" mechanisms in one class.

Target vs. parent-only. The nine-class modeling target comprises eight fallacies (ad_hominem, appeal_to_authority, appeal_to_emotion, appeal_to_popularity, false_dilemma, hasty_generalization, red_herring, slippery_slope) plus no_fallacy. Labels in the parent-only hierarchy (e.g., relevance, presumption, causality) and intermediate groupings (e.g., circular_reasoning, causal_fallacy) are used for analysis and are not members of the nine-class target. Child labels mapped to these parents are preserved in metadata.

Parents (unified → parent). *relevance:* red_herring, ad_hominem, appeal_to_authority, appeal_to_popularity, appeal_to_emotion; *presumption:* hasty_generalization, false_dilemma; *causality:* causal_fallacy, slippery_slope; *reasoning_form:* circular_reasoning; *background:* no_fallacy; *source_specific:* intentional. These parent families follow standard typologies in argumentation theory (Tindale, 2007); see also Helwe et al. (2024) for recent cross-corpus harmonization.

Source datasets. We integrate: *Argotario* (Habernal et al., 2017) (game-based; multilingual), *LOGIC_EDU* and *LOGIC_CLIMATE* (Jin et al., 2022) (curated 13-class schemes), *MAFALDA* (Helwe et al., 2024) (hierarchical spans), and *CoCoLoFa* (Yeh et al., 2024) (news comments; 8 classes). Each record follows {id, text, label, orig_label, source, domain, meta}. In LOGIC_EDU we adopt updated_label as canonical (preserving old_label in meta); MAFALDA is extracted at span level; Argotario retains consensus-voted rows.

3.1 Balancing and Stratified Splits

The merged corpus contains 12 531 normalized rows. Filtering to the nine target classes yields 11 111 usable examples. If used in downstream modeling, train/dev/test splits (80/10/10) can be produced and stratified by label_final and source; we focus here on the taxonomy and alignment.

Class	Count
no_fallacy	3248
appeal_to_popularity	1463
hasty_generalization	1265
red_herring	1163
appeal_to_emotion	1138
appeal_to_authority	836
false_dilemma	780
slippery_slope	695
ad_hominem	523

Table 2: Original class distribution (9-class target).

Source	Count (9-class)
CoCoLoFa	7,706
LOGIC_EDU	1,718
LOGIC_CLIMATE	813
MAFALDA	457
Argotario	417
Total	11,111

Table 3: Counts by source after filtering to the 9-class schema.

Domain	Count (9-class)
news_comments	7,706
education	1,718
climate	813
forum	457
game	417
Total	11,111

Table 4: Counts by domain after filtering to the 9-class schema.

3.2 Normalization Pipeline

All preprocessing is implemented in Python 3.10 using pandas and orjson. Modules: (1) per-dataset normalization to the canonical schema; (2) merge and mechanical label normalization; (3) consolidation to label_unified and parent; (4) filtering to the 9-class target; (5) integrity checks. Artifacts include mapping CSVs (child→unified, unified→parent), split manifests, and per-class count logs. All outputs are stored under data/interim/.

3.3 Rationale and Design Considerations

Normalization ensures semantic comparability across sources so that each class can be analyzed on equal conceptual and lexical footing. For lexical profiling, we compute log-odds ratios with informative Dirichlet priors (Monroe et al., 2008) over lemmatized unigrams and bigrams, identifying empirically distinctive cues per class. These cues form

the empirical lexical layer. A subset of the highest-ranked, semantically interpretable tokens is subsequently linked to WordNet 3.0 synsets (via domain-filtered mapping), creating the symbolic layer used in the taxonomy. For public tables we apply a minimal blocklist to remove non-scholarly profanity and single-instance named entities from display, while retaining the full ranked lists in the release artifacts for transparency. We compute log-odds with informative Dirichlet priors using the pooled background ($\alpha_k = 0.01 \times c_k^{\text{bg}}$; Monroe et al. 2008). Bigrams are selected with $PMI > 5$ and minimum count ≥ 5, then lemmatized. We use log-odds with informative Dirichlet priors because it is well suited for identifying class-distinctive terms under uneven class distributions, while reducing the instability of raw frequency ratios in small or imbalanced classes. Bigrams are included because several fallacy-related cues are relational or constructional rather than purely lexical, e.g., `vote_majority`, `expert_testimony`, or `easily_lead`. The extracted cues are not treated as universal markers of a fallacy; rather, they provide candidate lexical evidence that can be inspected, filtered, and linked to semantic anchors.

Methodological alternatives. Log-odds was selected over raw frequency and TF–IDF because the objective is to identify class-distinctive cues under unequal class sizes rather than document-level relevance. PMI is used only to identify recurrent bigram collocations. The head-based WordNet lookup is adopted as a transparent baseline; contextual word-sense disambiguation and embedding-based synset alignment may better address polysemy and should be compared systematically in future work.

4 WordNet Alignment

We enrich each fallacy class with two complementary semantic layers derived from the previous analysis. **(i) Conceptual anchors:** each class is linked to a primary WordNet synset representing the core reasoning concept (e.g., *false dilemma* → `dilemma.n.01`). **(ii) Lexical evidence:** we extract class-distinctive lemmatized unigrams and bigrams using log-odds with informative Dirichlet priors (Monroe et al., 2008). The ranked lists reflect corpus usage and therefore include both reasoning-related terms and corpus-specific artifacts.

Mapping policy. We map only a subset of semantically interpretable cues via automatic lookup with lexical-domain filtering, retaining `noun.cognition`, `noun.communication`, `noun.person`, `noun.act`, `noun.group`, `noun.quantity`, `noun.event`, `verb.communication`, `verb.cognition`, `verb.change` and excluding adjectives and attribute senses. For multiword cues (e.g., `vote_majority`), we perform a head-lemma lookup (*right head* by default: `majority`) and fall back to the left or the full bigram if no head match is found.

For polysemous lemmas (e.g., *authority*), we select one sense per cue by manual inspection of glosses and usage; ambiguous cases are omitted from the presented table but retained in release artifacts with an 'ambiguous' flag.

This produces a sparse candidate symbolic layer that requires filtering and human validation. In our presented run (after blocklisting table-only artifacts and excluding adjective/attribute synsets), the automatic mapping yielded 401 token–synset correspondences across the nine classes, with per-class distinct token counts: `appeal_to_authority=48`, `false_dilemma=46`, `slippery_slope=45`, `appeal_to_emotion=45`, `appeal_to_popularity=45`, `red_herring=44`, `no_fallacy=44`, `ad_hominem=42`, `hasty_generalization=42`. Counts reflect the filtered set of mappings reported here; the full mapping is released for reproducibility.

We keep the full empirical ranked lists for transparency and use the filtered synset layer for interpretability analyses. Table 5 summarizes representative cues for each class.

Preliminary quality audit of WordNet alignment. The present work does not claim a full downstream evaluation of fallacy detection in Bulgarian. Instead, we report a preliminary quality audit of the English cue-to-synset mapping used to construct the symbolic layer. We manually inspected a random sample of 50 cue→synset pairs from the automatically generated mapping. Of these, 26 (52%) were judged *acceptable* (correct or meaningfully related), 11 (22%) *borderline* (semantically plausible but overly broad/vague), and 13 (26%) *incorrect*. These results indicate that the heuristic linker is useful for generating candidate

Fallacy	Representative lexical cues (examples)
Ad Hominem	hypocrite, opponent_raise, move_town
Appeal to Authority	listen_expert, citation_expert, expert_testimony
Appeal to Emotion	mother_nature, power_line, natural_flow (*climate-domain artifacts*)
Appeal to Popularity	vote_majority, want_majority, human_tradition
False Dilemma	half_measure, meaningful_alternative, love_hate
Hasty Generalization	class_hard, college_good, philosophy_class
Red Herring	issue_poverty, cut_military, pale_comparison
Slippery Slope	easily_lead, censorship_start, lawlessness

Table 5: **Representative lexical cues.** Class-distinctive cues from the log-odds analysis (exported from `top_terms_per_class.presented.json`). Items may include corpus-specific artifacts; full ranked lists with provenance are released. Only an interpretable subset is used for WordNet alignment.

mappings, but not reliable enough to be used without filtering or human review. We therefore treat the WordNet-linked lexical layer as a bootstrapping resource rather than a finished annotation or classification system. A full evaluation will require downstream experiments, including Bulgarian seed lexicon validation, manual review of projected cues, and classification or retrieval experiments on Bulgarian argumentative texts.

The mappings should therefore not be interpreted as fully automatic fallacy annotations. Their value lies in narrowing the search space for annotation, prompting, and lexicon construction by producing inspectable candidate anchors. Human validation is part of the intended workflow: the automatic layer supports bootstrapping and prioritization, but is not sufficient as a standalone classification resource.

Cross-lingual projection. Anchored by WordNet synsets, we project to Bulgarian through Open Multilingual WordNet (OMW) (Bond and Foster, 2013) by retrieving aligned Bulgarian lemmas. We treat these lemmas strictly as seeds or weak constraints rather than inferred fallacy labels. Language-specific resources such as BulNet (Koeva, 2021) provide a complementary basis for future validation and extension. We use OMW offsets to collect aligned lemmas (not to infer labels) and treat them strictly as seeds or weak constraints (e.g., label propagation, pattern matching, lexicon-guided prompting) in low-resource settings. Table 7 illustrates how selected English WordNet anchors can be projected into Bulgarian seed cues. The projected lemmas are not used as gold labels or direct classifiers; they serve as candidate lexical material for weak supervision, lexicon-guided prompting, or manual annotation support.

Illustrative chain. *Text:* "If we allow this exception, it could **easily lead** to allowing everything."

The bigram `easily_lead` is extracted as class-distinctive lexical evidence for `slippery_slope`. At the conceptual layer, the class is associated with `consequence.n.01`, representing the anticipated outcome expressed by the argument. This is a class-level conceptual association, not a claim that the lemma `lead` is synonymous with `consequence`.

5 Discussion

The proposed two-layer taxonomy — combining conceptual WordNet anchors and lexical evidence — supports several research directions:

- **Taxonomy and resource development:** The present contribution is a taxonomy with lexical-semantic anchors, not a full ontology with axioms, formal relations, or constraints. The WordNet-linked nodes may nevertheless support future argumentation or reasoning ontologies by providing stable conceptual entry points for fallacy categories.
- **Cross-lingual mapping:** WordNet-based identifiers allow alignment with multilingual lexical resources such as BulNet (Koeva, 2021) for Bulgarian and related WordNets for other low-resource languages. This facilitates the creation of comparable reasoning datasets and seed lexicons in settings where manual fallacy annotation is limited or infeasible.
- **Interpretability:** The lexical layer clarifies how abstract reasoning flaws manifest linguistically, supporting transparent and explainable fallacy detection in manipulative or persuasive discourse.
- **Dataset integration:** The use of stable WordNet offsets (`offset--pos`) supports alignment where corresponding multilingual synset mappings are available to equivalent synsets in other languages, providing a bridge for future bilingual and multilingual reasoning corpora.

Beyond linguistic alignment, the taxonomy is designed to interface with cognitive and applied models of manipulative reasoning that distinguish between structural and content-based fallacies. By

Fallacy	Term	Synset	ID	Gloss (shortened)
Ad Hominem	hypocrite	hypocrite.n.01	10195593-n	professes beliefs not actually held
	liar	liar.n.01	10256537-n	a person who lies repeatedly
Appeal to Authority	authority	authority.n.01	05196582-n	power or right to give orders / decide
	expert	expert.n.01	09617867-n	person with special knowledge
	testimony	testimony.n.01	07184149-n	a solemn statement to establish a fact
Appeal to Emotion	pity	pity.n.01	07517737-n	a feeling of sympathy and sorrow
	fear	fear.n.01	07519253-n	an emotion experienced in anticipation of danger
Appeal to Popularity	majority	majority.n.02	13581067-n	more than half of the votes
	tradition	tradition.n.02	06685456-n	cultural continuity transmitted across generations
False Dilemma	dilemma	dilemma.n.01	05907857-n	uncertainty between two unfavorable options
	alternative	alternative.n.01	05790944-n	one of several possibilities
Hasty Generalization	generalization	generalization.n.01	05774415-n	reasoning from detailed facts to general principles
	assumption	assumption.n.02	05924488-n	a hypothesis that is taken for granted
Red Herring	diversion	diversion.n.02	00350380-n	turning aside of attention or concern
	irrelevance	irrelevance.n.01	04720284-n	lack of relation to the matter at hand
Slippery Slope	escalation	escalation.n.01	00290257-n	increase to a higher level of intensity
	consequence	consequence.n.01	11410625-n	outcome of an event or action
Background (`no_fallacy`)	balance	balance.n.01	14002279-n	state of equilibrium
	neutrality	neutrality.n.01	01240850-n	nonparticipation in a dispute

Table 6: Conservative WordNet 3.0 anchors per class after domain filtering. Shown items are semantically central, not exhaustive.

Fallacy	Synset	BG seed
appeal_to_popularity	majority.n.02	*mnozinstvo*
appeal_to_emotion	fear.n.01	*strah*
appeal_to_authority	expert.n.01	*ekspert*
slippery_slope	consequence.n.01	*sledstvie*

Table 7: Illustrative Bulgarian seed cues retrieved through OMW-aligned synsets. These are candidate cues for bootstrapping, not validated Bulgarian fallacy labels.

grounding each category in explicit WordNet concepts, these theoretical distinctions can be operationalized in computational form, facilitating integration with broader studies of argument quality and reasoning transparency.

The taxonomy further offers a structured target space for supervised and hybrid model training. Because each class includes both conceptual and lexical anchors, it supports multi-task learning setups where language models predict not only fallacy labels but also their semantic relations. This approach is particularly beneficial for low-resource languages such as Bulgarian, where OMW-based projection can provide initial seed cues for annotation, retrieval, and weak supervision, while full downstream validation remains necessary.

6 Reproducibility and Tools

All preprocessing and alignment were conducted in Python 3.10 within an isolated virtual environment. Key libraries include `spaCy 3.8.7` for tokenization and lemmatization, `NLTK 3.9` with Princeton WordNet 3.0 for English synset access, and `omw-1.4` for multilingual lemma retrieval. All operations use deterministic random seeds to ensure replicability.

The project follows a reproducible structure:

- `data/` — raw and normalized datasets
- `splits/` — stratified train/dev/test splits (not pre-balanced)
- `taxonomy_paper/` — scripts for lexical cue extraction and WordNet lookup
- `cache_json/` — hashed intermediate results
- `analysis/` — evaluation and inspection notebooks

All scripts and configuration files are implemented in Python and documented for reusability. The repository is currently private and will be released under a **CC BY 4.0** license upon publication, permitting reuse with attribution. All mapping files and scripts are versioned; random seeds are fixed for reproducibility.

7 Conclusion

We present a compact, WordNet-aligned taxonomy of logical fallacies designed as a semantic backbone for reasoning-aware NLP. The framework

unifies conceptual definitions with corpus-derived lexical cues, supporting taxonomy extension, multilingual resource development, and interpretable model evaluation. Future work will include release of the taxonomy and mappings, systematic Bulgarian seed-lexicon validation, and integration within hybrid fallacy-detection pipelines. The release includes both the parent–child hierarchy for interpretability and the 9-class training target (the eight fallacies plus `no_fallacy` as a background label) for practical modeling.

The resource is intended to interoperate with broader research on argumentative reasoning and manipulative discourse. Its WordNet-based structure enables seamless inclusion in forthcoming multilingual systems and cognitively informed studies, ensuring that linguistic, semantic, and reasoning layers evolve within a unified, transparent framework.

8 Ethics Statement and Limitations

This work analyzes reasoning structures in text, not individuals. Fallacy labels may reflect culturally specific reasoning norms; caution is required when transferring them across languages or social contexts. WordNet mappings for rhetorical phenomena are necessarily approximate, and the current manually evaluated alignment is limited to English. Future iterations will expand coverage and refine sense disambiguation. Projected lexicons may carry sense drift across languages; we recommend human review of seed lists before deployment. In particular, the Bulgarian examples in this paper demonstrate the projection mechanism, but they do not constitute a validated Bulgarian fallacy dataset or a downstream Bulgarian classifier. Their intended role is to support future annotation, retrieval, and weak supervision experiments.

This study forms part of a broader research effort on reasoning and manipulation across languages. The taxonomy was designed to remain conceptually transparent while avoiding culturally prescriptive definitions of "fallacious" reasoning (Tindale, 2007). Ongoing collaborative work will further examine how cognitive and cultural factors influence the perception and annotation of fallacies in diverse linguistic settings.

Lexical cue artifacts. The log-odds procedure can surface topic-specific or named-entity cues (e.g., items highly specific to a single example). We retain these in the released lists for transparency, but they may not generalize and should be interpreted as corpus artifacts rather than universal markers of a fallacy's form.

Acknowledgments

This research is part of the GATE project funded by the Horizon 2020 WIDESPREAD-2018-2020 TEAMING Phase 2 Programme under grant agreement no. 857155, the Programme "Research, Innovation and Digitalization for Smart Transformation" 2021–2027 (PRIDST) under grant agreement no. BG16RFPR002-1.014-0010-C01, and the project BROD (Bulgarian-Romanian Observatory of Digital Media), funded by the Digital Europe Programme of the European Union under contract number 101226153.

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley FrameNet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pages 86–90, Montreal, Quebec, Canada.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362, Sofia, Bulgaria. Association for Computational Linguistics.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179.

Ivan Habernal, Raffael Hannemann, Christian Pollak, Christopher Klamm, Patrick Pauli, and Iryna Gurevych. 2017. Argotario: Computational argumentation meets serious games. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 7–12, Copenhagen, Denmark. Association for Computational Linguistics.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 386–396, New Orleans, Louisiana. Association for Computational Linguistics.

Chadi Helwe, Tom Calamai, Pierre-Henri Paris, Chloé Clavel, and Fabian Suchanek. 2024. MAFALDA: A benchmark and comprehensive study of fallacy detection and classification. In *Proceedings of the 2024*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 4810–4845, Mexico City, Mexico. Association for Computational Linguistics.

Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schölkopf. 2022. Logical fallacy detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7180–7198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Svetla Koeva. 2021. The bulgarian WordNet: Structure and specific features. *Papers of the Bulgarian Academy of Sciences: Humanities and Social Sciences*, 8(1):47–70.

Yuanyuan Lei and Ruihong Huang. 2024. Boosting logical fallacy reasoning in LLMs via logical structure tree. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13157–13173, Miami, Florida, USA. Association for Computational Linguistics.

Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. 2008. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Princeton University. 2006. WordNet 3.0. `https://wordnet.princeton.edu/`. Accessed: 2025-10-20.

Zhivar Sourati, Filip Ilievski, Hông-Ân Sandlin, and Alain Mermoud. 2023. Case-based reasoning with language models for classification of logical fallacies. *arXiv preprint arXiv:2301.11879*.

Christian Stab and Iryna Gurevych. 2017. Recognizing insufficiently supported arguments in argumentative essays. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 980–990, Valencia, Spain. Association for Computational Linguistics.

Christopher W. Tindale. 2007. *Fallacies and Argument Appraisal*. Cambridge University Press, Cambridge, UK.

Henning Wachsmuth, Martin Potthast, Khalid Al-Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. 2017. Building an argument search engine for the web. In *Proceedings of the 4th Workshop on Argument Mining*, pages 49–59, Copenhagen, Denmark. Association for Computational Linguistics.

Min-Hsuan Yeh, Ruyuan Wan, and Ting-Hao Kenneth Huang. 2024. CoCoLoFa: A dataset of news comments with common logical fallacies written by LLM-assisted crowds. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 660–677, Miami, Florida, USA. Association for Computational Linguistics.

Relation Extraction from Text of Various Domains: A Serbian Case Study

Ranka Stanković
University of Belgrade
Faculty of Mining and Geol.
Belgrade, Serbia
ranka@rgf.bg.ac.rs

Cvetana Krstev
Language Resources and
Technologies Society–JeRTeh
Belgrade, Serbia
cvetana@jerteh.rs

Milica Ikonić Nešić, Miloš Utvić
University of Belgrade
Faculty of Philology
Belgrade, Serbia
(milica.ikonic.nesic
|misko)@fil.bg.ac.rs

Abstract

Research on Serbian named entity recognition (NER) has evolved from early rule-based approaches to neural and transformer-based models, culminating in the TeslaNER+ dataset, which contains over 150,000 sentences annotated with eight entity types and linked to Wikidata. This paper explores the use of TeslaNER+ for relation extraction (RE), focusing on identifying and classifying semantic relations between entities in unstructured text. A predefined set of relations tailored to various domains, including literary texts is employed. After analyzing existing RE schemas, 19 relevant relations are selected. The methodology combines rule-based patterns with LLM-based prompting, leveraging lexically grounded patterns and prompts derived from key terms. Experiments were conducted on a subset of 23,000 sentences, and the paper reports initial results for selected relations. The FST approach achieves a precision of 0.77 and recall of 0.39, whereas the LLM model attains a higher recall (0.97) but lower precision (0.65). These results are approximate, as they are computed on a sampled subset rather than the full dataset. In terms of efficiency, the FST approach incurs no computational cost and enables fast extraction, while the LLM approach is significantly slower and involves additional expenses.

Keywords: Serbian NLP; Named Entity Recognition; Relation Extraction; Rule-based methods; Large Language Models; Semantic Relations.

1 Introduction

Work on Serbian named entity recognition (NER) has a long history, beginning with early rule-based systems (Krstev et al., 2014) relying on electronic dictionaries and the Unitex[1] corpus processing tools, continuing with convolutional neural network approaches, and more recently transformer-based models. These tools were essential in developing TeslaNER+, a dataset of over 150,000 sentences. It combines automated annotations (Nešić et al., 2024) with manual curation and features seven NER classes linked to Wikidata: PERS, LOC, ORG, DEMO, ROLE, WORK, EVENT. Recently, an additional class, PRODUCT, has been introduced, although it is not yet annotated throughout the entire corpus. TeslaNER+ integrates manually curated datasets from prior research and expands them with newly prepared texts and an additional linked-data layer.

The goal of this research is to use this dataset for the relation extraction (RE) task to detect and classify semantic links between annotated entities in unstructured text. Relation identification detects pairs of entities that are related to each other in the text, while relation classification assigns a predefined label to each identified entity pair, indicating the nature of their relationship. To the best of our knowledge, no prior research on relation extraction from Serbian texts has been published.

The main contributions of this study are: 1) adapting an existing relation schema to the characteristics of Serbian texts from various domains, including literary, 2) constructing a lexically grounded rule-based framework for relation detection using finite-state transducer (FST) patterns, and 3) exploring the complementarity of rule-based and large language model (LLM) approaches for relation extraction. The results contribute to the development of structured semantic resources derived from corpora and support further research in knowledge graph construction and information extraction for Serbian.

In this paper, we present the first preliminary results for the selected set of relations. Section 2 presents an overview of related work in Serbian

[1] https://unitexgramlab.org/

named entity recognition, types of relation extraction with relevant approaches in multilingual settings. Section 3 describes the dataset, the selection of relations, and the proposed methodology, including both rule-based and prompt-based approaches. Section 4 provides descriptive statistics of the dataset and the distribution of entity types and relations. Section 5 outlines the evaluation setup and reports the obtained results. Section 6 discusses the findings, challenges, and limitations of the approaches. Finally, Section 7 concludes the paper and outlines directions for future work.

2 Related Research

2.1 Overview of Serbian NER

Research on named entity recognition (NER) for Serbian spans more than two decades and reflects a gradual evolution from rule-based approaches to modern neural and LLM-based methods. Early work focused on the linguistic properties of proper names and their representation in lexical resources, as explored in the Prolex project (Vitas et al., 2009). This line of research was further extended through the use of electronic dictionaries and corpus processing tools, enabling the development of rule-based NER systems grounded in rich morphological descriptions (Vitas and Krstev, 2012).

Subsequent efforts addressed the recognition and normalization of named entities, emphasizing the integration of linguistic knowledge with corpus-based methods (Krstev et al., 2012). Rule-based approaches based on local grammars and finite-state technologies, particularly within the Unitex framework, proved effective for Serbian due to its morphological complexity (Krstev et al., 2014). At the same time, multilingual perspectives were introduced through resources such as NERosetta, which aimed to position Serbian within a broader cross-lingual NER space (Krstev et al., 2016).

More recent research has shifted toward machine learning and neural approaches, including the development of models for named entity linking, such as SrpCNNeL, which integrates entity recognition with linking to external knowledge bases (Nešić et al., 2024). Parallel corpora and linked data have also been explored as a means to enhance semantic interoperability and support cross-lingual NER and linking tasks (Stanković et al., 2024).

The latest developments highlight the complementary strengths of different paradigms: rule-based, neural, and LLM-based approaches, support an integrated, holistic approach to NER and NEL. Rather than replacing earlier methods, these paradigms contribute distinct advantages, combining linguistic precision with contextual flexibility and scalability. This progression is summarized in recent work tracing the development of Serbian NER from early rule-based systems to contemporary LLM-driven methods (Nešić et al., 2025). Overall, the evolution of Serbian NER and NEL demonstrates a convergence of methodologies, leading to more robust, adaptable, and semantically enriched systems.

2.2 Relation Types

The most common cases of RE focus on the extraction of binary relations within a sentence (sentence-level RE). Also, Diaz-Garcia and Lopez (2025) describe document-level RE (relationships between entities extend beyond individual sentences, spanning entire documents), multilingual RE (extracting relations from multilingual text), open RE or OpenRE (unveiling previously unknown relation types between entities within open-domain corpora), multimodal RE (integration of text and visual data to enhance understanding of relationships), Few-Shot and Low-Resource RE (developing models that generalize well even with minimal labeled examples), and distant RE (extracting relations between entities based on distant supervised data).

For example, in the sentence "Marko Ilić dismissed Ivan Rogić from his staff," an NEL system should recognize "Marko Ilić" and "Ivan Rogić" as entities and identify "dismissed" as the relationship linking them. Some RE systems rely on predefined types (e.g., `DismissedBy`, `TermEmployee` as short for "employment termination"), while some Open Information Extraction (OpenIE) systems extract structured, domain-independent relational triples: (subject; relation; object), directly from unstructured text without requiring a predefined schema, e.g., (Marko Ilić, dismissed, Ivan Rogić).

Watson NLP RE [2] encapsulates algorithms for extracting 32 relation types between two entity mentions, some of which are:

- `parentOf`: exists between a Person and their children or stepchildren.

[2] https://dataplatform.cloud.ibm.com/docs/content/wsj/analyze-data/watson-nlp-block-relation-extraction.html?context=dph

- `educatedAt`: exists between a Person and the Organization at which s/he is or was educated.
- `colleague`: exists between two Persons who are part of the same Organization.

RE supports the identification of relationships in text and plays a key role in building and populating knowledge graphs. It is essential for downstream applications such as question answering (Chen et al., 2019). Additionally, RE contributes to knowledge acquisition in the humanities, particularly in graph-based datasets, enabling more advanced information retrieval and analysis (Zhu et al., 2017).

2.3 Relation Extraction Techniques

Rule-Based Approaches characterized early relationship extraction systems, which used predefined patterns or syntactic rules to identify relationships. For example, patterns like "X dismissed Y" could help identify a relationship between entities "X" and "Y" (Diaz-Garcia and Lopez, 2025).

Supervised Machine Learning trains classifiers, like Support Vector Machines (SVMs) or decision trees, using labeled datasets to learn relationship patterns. The features of such labeled datasets can be part-of-speech tags, dependency parse trees, and word embeddings. The limitations of supervised learning models are the need for extensive labeled data; however, there is still no guarantee that they will generalize well to unseen relationships.

Deep Learning and Neural Network Approaches use neural network models which process text sequentially or in parallel. These models, e.g., recurrent neural networks (RNNs), convolutional neural networks (CNNs), and transformers, improve relationship extraction accuracy and adaptability by capturing complex patterns and long-range dependencies. Transformer-based models try to understand the surrounding context and use contextual embeddings to enhance relationship extraction.

Distant Supervision is used to align text with existing structured data (such as knowledge graphs), based on the assumption that co-occurring entities in a text, which are also present in the knowledge base with a known relationship, are likely related. This method is suitable for producing large quantities of training data, but it can also generate noise due to incorrect assumptions. Dependency parsing identifies the syntactic structure of sentences and thereby helps detect relationships between entities, while end-to-end neural models combine entity recognition and relation extraction in a single process to reduce error propagation and improve efficiency.

Our approach focuses on extracting binary relations within a single sentence (sentence-level RE), similar to most other RE systems. Since a large part of our corpus consists of literary texts (more than one third), our relation inventory was adapted to the characteristics of this genre. In fiction, relations such as `spouseOf`, `siblingOf`, `parentOf`, `residesIn`, `relativeOf`, and `partOfOrg`, occur frequently, while business-oriented relations (e.g., `shareholdersOf`, `competitor`, `clientOf`) are comparatively rare.

3 Methodology

From the 32 relations defined in Watson NLP RE, we selected and adapted 19 relations relevant for various domains represented in our dataset. The difference is that we distinguish `managerOf` between persons and `managerOfOrg` between person and organization, and we introduced new relations: `eventPers` and `eventLoc`. The proposed methodology combines two complementary approaches: a rule-based approach and an LLM-based approach. For each relation, a list of lexical triggers and phrases is defined and used both for constructing FST patterns and for generating prompts in the LLM-based method.

The set of relations is predefined, and possible subject–object combinations are specified:

- PERS–PERS: `colleagueOf`, `managerOf`, `parentOf`, `relativeOf`, `siblingOf`, `marriedTo`, `spouseOf`
- PERS–ORG: `educatedAt`, `managerOfOrg`, `affiliatedWith`
- PERS–LOC: `residesIn`, `bornAt`, `diedAt`
- PERS–WORK: `authorOf`
- ORG–ORG: `partOfOrg`
- ORG–LOC: `basedIn`
- LOC–LOC: `partOfLoc`
- EVENT–PERS: `eventPers`
- EVENT–LOC: `eventLoc`

3.1 Dataset

The approach was applied and evaluated on a subset of 23,000 sentences extracted from TeslaNER+. The subset contains the following distribution of annotated entities: LOC (27K), PERS (25K), ROLE (17K), ORG (10K), DEMO (6K), WORK (2K), EVENT (1K), and PRODUCT (1K).

The input to the system consists of text seg-

mented into sentences and enriched with named entity annotations, as illustrated with the sentence 'After all, Goethe married Christiane Vulpius, not Madame von Stein.':

```
<s file_id="burmag-zorgoj.tsv" num="1756">Uostalom,
i <ner class="PERS" qid="Q5879">Gete</ner> se oženio
<ner class="PERS" qid="Q66746">Kristijanom Vulpijus
</ner>, a ne <ner class="ROLE" qid="Q1378024">
gospođom</ner> Fon <ner class="PERS" qid="None">
Štajn</ner>."</s>
```

Each sentence is annotated with entity spans, their types (e.g., PERS, ROLE), and, where available, links to external knowledge base Wikidata (QIDs) (Vrandečić and Krötzsch, 2014). The relation extraction module processes such annotated sentences by identifying candidate pairs of entities and applying predefined patterns (rules) or prompts to determine whether a semantic relation holds between them.

In this example, the system identifies a relation of type `marriedTo` between two entities of type PERS: *Gete* and *Kristijanom Vulpijus*. The relation is triggered by the lexical pattern *se oženio* ("has married"), which serves as a strong indicator of a marital relationship in Serbian. Although additional entities are present in the sentence (e.g., *gospođom* and *Štajn*), the system correctly filters out irrelevant pairs based on entity types and contextual constraints.

The resulting output is represented in a structured format, where the detected relation is explicitly encoded as follows:

```
<nel type="marriedTo">
<ner class="PERS" qid="Q5879">Gete</ner>
<ner class="PERS" qid="Q66746">Kristijanom Vulpijus
</ner> </nel>
```

This example demonstrates how the combination of entity annotation, lexical triggers, and pattern-based or prompt-based processing enables the extraction of semantically meaningful relations from complex sentence structures.

3.2 Rule-based Approach

The FST-based approach to relation extraction relies on the Unitex corpus processing system and is grounded in morphological dictionaries and local grammars (Krstev, 2008; Vitas and Krstev, 2012). Texts are first processed using rich lexical resources that provide morphological and syntactic information, enabling precise recognition of lexical units and their inflected forms. RE is then performed through cascades of FSTs, implemented as local grammars, which encode patterns corresponding to specific semantic relations. These grammars capture lexical triggers, syntactic configurations, and contextual constraints to identify related entity pairs within sentences. The cascade architecture allows the sequential application of increasingly complex patterns and recognition of various relations in the same sentence, allowing an NE to participate in several relations. This rule-based approach is particularly effective in handling morphologically rich languages such as Serbian, although it requires substantial human knowledge and manual effort for grammar design and maintenance.

When designing FSTs for relation extraction, we experimented with two approaches applied to several relations: a relaxed one that relies only on annotated NEs and a set of triggers and a strict one that takes into account numerous syntactic variations (e.g. word order), and possible modifications of NEs and triggers. The obtained results showed that the first approach leads to low precision, and the second one to low recall. Therefore, we decided to produce FSTs that are relaxed with some additional constraints. The Figure 1 presents one such FST designed for `spouseOf` relation extraction. It basically relies on a set of triggers, e.g., *žena*, *supruga* ("wife", "spouse") with an additional constraint that two recognized PERS entities are of the opposite sex (in this case the first is female and the second is male), which corresponds to the traditional understanding of persons united in marriage. Intermediate nodes (`notPers` in presented FST) allow for optional tokens such as punctuation, intervening words, or NEs of other types, capturing variation in natural language expressions.

The Figure 2 presents a FST for the extraction of the `bornAt` relation that links a PERS entity and a LOC entity. This FST also relies on triggers which can be verbs (*poticati* 'come from', *doći na svet* 'lit. come to the world'), nouns (*rodno mesto* 'birthplace'), or noun phrases (*poreklom iz* 'originating from'), but it also takes into account different word order (PERS coming before LOC and vice versa). Intermediate nodes (`notPersLoc` in presented FST) allow for optional tokens such as punctuation, intervening words, or NEs of other types, while nodes `Date` allow the insertion of complex dates.

3.3 LLM-based Approach

Wadhwa et al. (2023) revisit the evaluation of generative relation extraction by using human assessment instead of exact matching. Their results show that few-shot prompting with GPT-3 achieves near

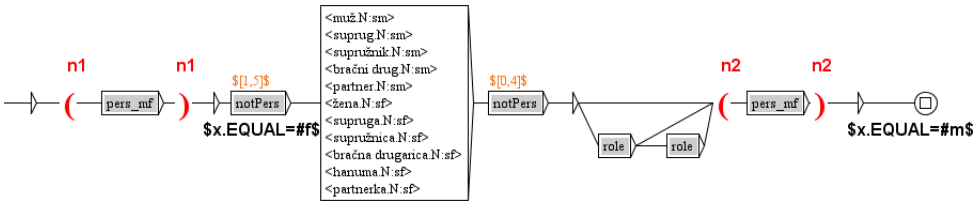


Figure 1: A simplified FST for the extraction of `spouseOf` relation (output omitted)

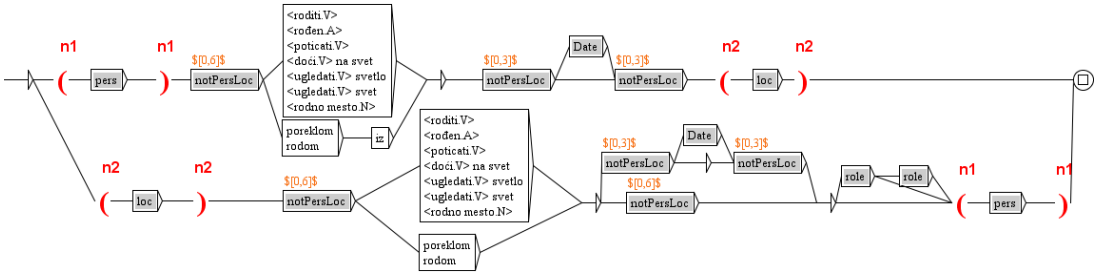


Figure 2: A simplified FST for the extraction of `bornAt` relation (output omitted)

state-of-the-art performance, comparable to fully supervised models, while Flan-T5 performs weaker in few-shot settings but reaches state-of-the-art results when fine-tuned with Chain-of-Thought explanations generated by GPT-3.

The LLM-based approach formulates relation extraction as an instruction-following task, where an LLM is prompted to identify and classify semantic relations between pairs of pre-annotated entities in text. Instead of relying on explicitly trained classifiers, the model leverages its pretrained knowledge and contextual understanding to infer relations directly from the input sentence or passage.

For each sentence, candidate pairs of entities are generated, and a structured prompt is constructed. The prompt includes: (i) the input passage, (ii) a list of recognized entities with their types, (iii) a predefined set of allowed relation labels, and (iv) task-specific instructions. The model has to decide whether a relation exists between a given entity pair and, if so, assign exactly one label from the predefined set.

To ensure consistency and facilitate automatic evaluation, the model is constrained to produce output in a fixed JSON format containing the head and tail entity identifiers, the predicted relation, a short textual evidence span extracted from the passage, and a confidence score. Lexical triggers associated with each relation are optionally included in the prompt to guide the model, although the final decision must rely on contextual evidence rather than surface patterns alone.

The prompt for the classification of the candidate pair is (it is translated in English for the sake of readability):

```
PAIR_CLASSIFICATION_PROMPT = """
You are a relation classification system for Serbian
literary text.

Task:
Decide whether there is a semantic relation
between the two target entities in the given
passage.
If yes, assign exactly one relation label
from the allowed set.
If no relation is clearly supported,
return "NO_RELATION".

Allowed relation labels:
{allowed_relations}

Typical Serbian lexical triggers
(hints only, not proof):
{trigger_block}

Rules:
- Use only the evidence from the sentence.
- Prefer the most specific relation.
- Use only the given entities.
- Return only valid JSON.
- If the relation is not clearly supported,
  return NO_RELATION.
- Evidence must be a short exact quote
  from the passage.
- Confidence must be a number between 0 and 1.

Return JSON in exactly this format:
{{
  "head": "{head_id}",
  "tail": "{tail_id}",
  "relation": "LABEL",
  "evidence": "short exact quote from the text
  or empty string",
  "confidence": 0.0
}}

Passage:
{passage}

Entities:
{entities_json}

Target pair:
- Head: {head_id} ({head_text}, {head_type})
- Tail: {tail_id} ({tail_text}, {tail_type})
"""
```

This approach allows flexible adaptation to different domains and relation schemas without additional model training. However, its performance

depends on prompt design, the clarity of the relation set, and the model's ability to handle ambiguity and complex sentence structures, which are particularly common in literary texts.

4 Statistics

Our rule-based system extracted a total of 1,676 relations from the dataset, covering a diverse set of semantic relation types, with an uneven distribution of relations (see Figure 3). Overall, the distribution reveals a dominance of location- and organization-related relations, while more fine-grained interpersonal and event-based relations occur less frequently. More precisely, the number of identified relations involving the entity ORG is 1,106 for five different relations, that is, 66.0% of relations retrieved. Similarly, the number of identified relations involving the entity LOC is 973 for six relations (58.1%). On the contrary, not many PERS–PERS relations were retrieved – 261 occurrences for seven relations (15.6%). The relatively low frequency of certain relations may indicate limitations in either lexical coverage or pattern detection.

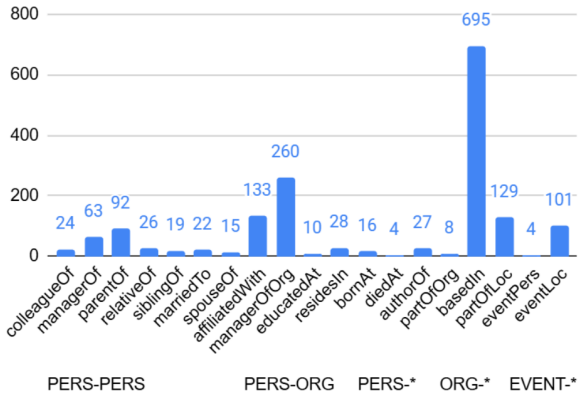


Figure 3: Results of RE by the rule-based system

We used the GPT-4.1-mini API for the prompt-based approach on a subset of 6,600 sentences, rather than the full dataset, due to slow performance. The total cost for this processing was $97. It extracted 24,607 relation candidates, of which 13,907 have a confidence score (CF, given by LLM) ≥ 0.9, and 14,139 have ≥ 0.85.

The productivity of relations was very different. Figure 4 show statistics, where for very productive (`colleagueOf`, `affiliatedWith`, `partOfOrg`, `basedIn`, `partOfLoc`), the threshold for CF $\geq$ 0.85. For evaluation we selected 2,629 relation candidates. The largest portion is dominated by LOC–LOC relations (`partOfLoc`, 8,642 instances CF ≥ 0.85, 12,616 total), followed by ORG–LOC (`basedIn`, 2,034 CF ≥ 0.85, 3,783 total) and ORG–ORG (`partOfOrg`, 1,022 CF ≥ 0.85, 1,541 total), indicating a strong emphasis on spatial and organizational hierarchies.

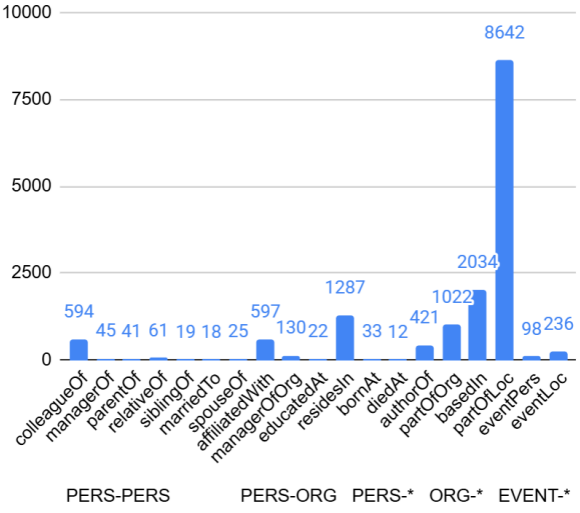


Figure 4: Results of RE by the LLM-based system

In contrast, other PERS–PERS relations are less frequent: `colleagueOf` with 594 candidates has CF ≥ 0.85(2,411 total), while more specific relations such as `marriedTo`, `siblingOf`, and `relativeOf` are sparsely represented. Similarly, PERS–ORG relations (722) are mainly driven by `affiliatedWith`, 597 candidates have CF ≥ 0.85 (1,532 total), whereas `educatedAt` is underrepresented.

PERS–LOC relations (1,332) are also skewed toward `residesIn` having 1,287 candidates, with `bornAt` and `diedAt` appearing only marginally. Among PERS–WORK relations, `authorOf` is also well represented (421), while EVENT-related relations remain relatively limited (98 + 236).

5 Evaluation

Sentences from the dataset presented in Section 3.1 and annotated using the rule-based system were transformed into so-called WebAnno TSV 3.2 File format.[3] Figure 5 presents examples of relation annotation in the INCEpTION environment (Klie et al., 2018) of several sentences from literary texts.

[3] https://webanno.github.io/webanno/releases/3.4.5/docs/user-guide.html#sect_webannotsv

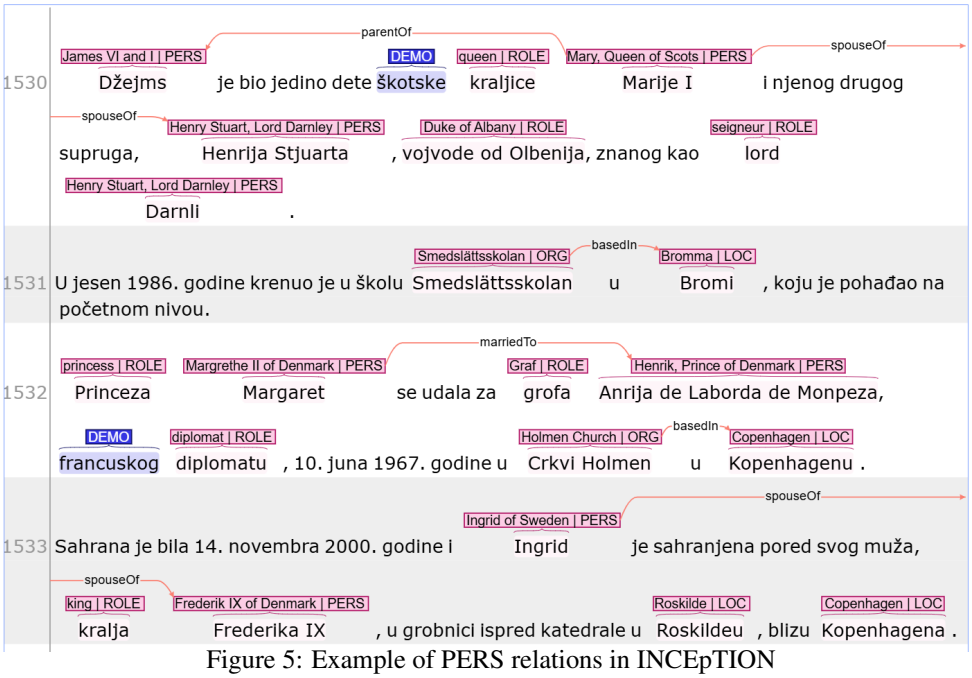


Figure 5: Example of PERS relations in INCEpTION

Named entities are annotated as PERS and LOC, and semantic relations are marked between relevant pairs. In the sentence 1530, the entities *Džejms* and *Marije I* are linked by the relation `parentOf`, triggered by the lexical cue *je bio dete* ("was child"), and the same PERS *Marija I* is linked to *Henrija Stjuarta* by the `spouseOf`. In the example 1531, ORG *Smedslättsskolan* is `basedIn` LOC *Bromma*. In the 1352 sentence, the entities *Margaret* and *Anri de Labord de Monpez* are linked by the relation `marriedTo`, indicated by the phrase *se udala* ("has married"). These examples illustrate how different types of interpersonal relations are identified through lexical triggers and contextual cues within narrative text.

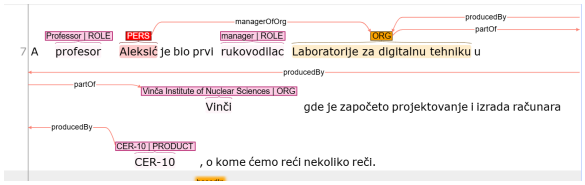


Figure 6: Example of ORG relations in INCEpTION

Illustrative examples of relations between entities of different types are presented in Figure 6. *Aleksić* is annotated as PERS, *profesor* "professor" and *rukovodilac* "manager" as ROLE, and *Laboratorije za digitalnu tehniku* "Lab for digital technology" as ORG. The relation `managerOfOrg` links the role *rukovodilac* to the organization, while additional relations such as `partOf` capture structural associations between organizations. The relation `producedBy` links an ORG entity with a PRODUCT entity.

Figure 7 presents the evaluation of extracted relations using FST-based approach, showing the high precision for most relations (blue). The evaluation results for relations extracted using LLM-based approach are presented in Figure 8. The LLM approach shows more deviations than the FST approach when extracting different relations. The results are approximate, as they are based on an evaluated sample of 2,629 candidates, rather than the full dataset.

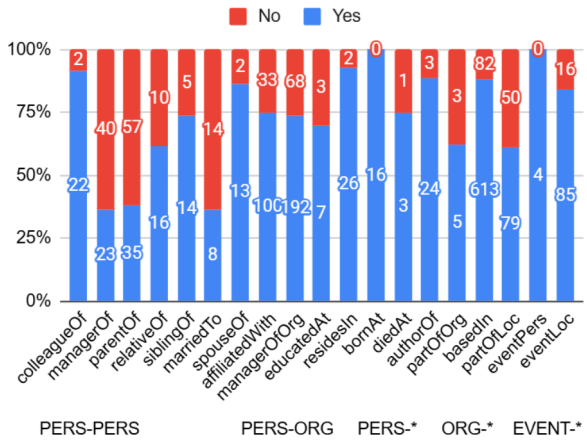


Figure 7: Evaluation of FST-based relation extraction

The precision of extracted relations across different relation types is presented in Figure 9. Over-

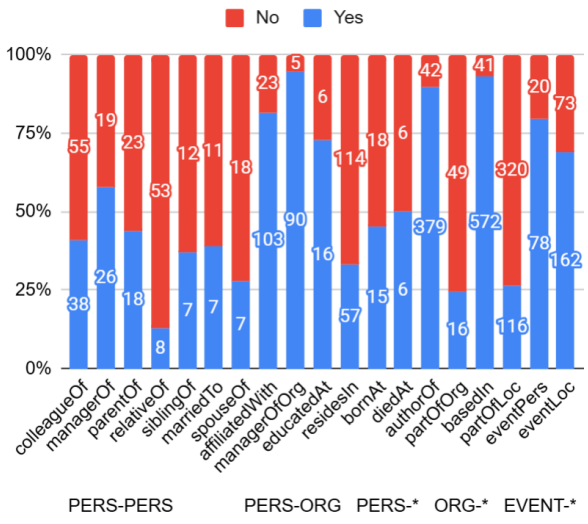


Figure 8: Evaluation of LLM-based relation extraction

all, performance is high, with overall precision $P = 0.77$. The best precision ($P = 1$) is observed for `eventPers` and `bornAt`, but not many occurrences were extracted. The most productive is `basedIn` (695) with $P = 0.88$. The best precision for a relation with a significant number of candidates was achieved for `residesIn` with $P = 0.93$. In contrast, `managerOf`, `parentOf`, and `marriedTo` show lower precision, indicating greater difficulty in capturing these from texts. The worst results were for `marriedTo` $P = 0.36$, where the problem might be the existence of a similar relation `spouseOf`, which was difficult to distinguish both for models and for evaluators.

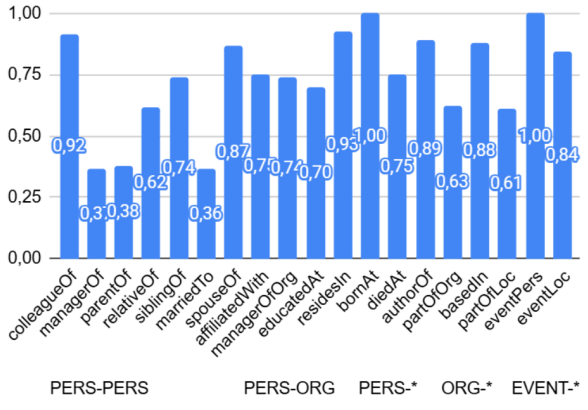


Figure 9: Precision of FST-based extraction

These results suggest that the FST model performs best on structurally explicit and less ambiguous relations, while more context-dependent social relations remain more challenging.

Figure 10 presents the precision of LLM-based relation extraction. Among PERS–PERS relations, moderate performance is observed, with `managerOf` and `parentOf` achieving the highest scores, while `relativeOf` remains notably lower. In the PERS–ORG group, `affiliatedWith` shows the highest score ($P = 0.95$), followed by `managerOfOrg` and `educatedAt`, indicating that institutional relations are more reliably detected.

For PERS–* relations, `authorOf` stands out with a high score, while relations such as `bornAt` and `residesIn` achieve moderate results. Within the ORG–* category, `basedIn` performs strongly, whereas `partOfOrg` shows lower values. Finally, EVENT–* relations, including `eventPers` and `eventLoc`, demonstrate relatively good performance. Overall, the results suggest that more structured and frequent relations are recognized more accurately, while less frequent or more complex relations are more prone to errors.

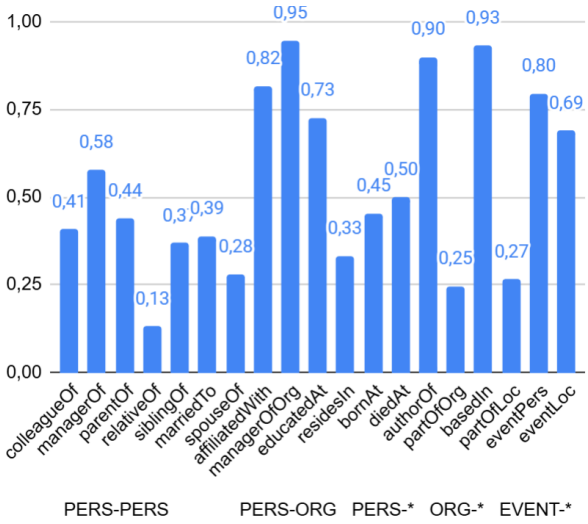


Figure 10: Precision of LLM-based extraction

Out of 2,638 evaluated instances, the LLM approach yielded 1,721 correct (Yes) and 917 incorrect (No) predictions. An additional 56 correct relations were identified by the FST approach which were not extracted by the LLM. In total, the FST system produced 1,285 correct (Yes) and 391 incorrect (No) predictions. For a fair comparison, recall was approximated on the same subset of sentences used for the LLM evaluation (6,600 sentences), where the FST system retrieved 796 candidates, of which 697 were correct and 99 incorrect. The results indicate that the FST approach achieves a precision of 0.77 and approximate recall

of 0.39 whereas LLM model yields much higher recall (0.97) at the cost of reduced precision (0.65).

6 Discussion

RE faces several well-known challenges, including ambiguity and polysemy, complex sentence structures, domain-specific language and limited generalization across domains, as well as the scarcity of labeled data. Additionally, texts often contain multiple overlapping relationships among several entities simultaneously, further increasing the difficulty of the task. We also observed these challenges in our dataset, particularly in literary texts, where nuanced language use and intricate narrative structures make relation identification and classification more demanding.

A key drawback of the LLM prompt-based approach is its relatively slow processing speed, coupled with a dependence on paid services, which may limit scalability and practical deployment.

LLM overgenerates, which can be illustrated with the sentence *Konferenciju je otvorio. . . , Sonja Liht ispred Fonda za otvoreno društvo, Meri Blek ispred Unicefa i Džim Stivens ispred Svetske banke.* ("The conference was opened by. . . , Sonja Licht on behalf of the Open Society Foundation, Mary Black on behalf of UNICEF, and Jim Stevens on behalf of the World Bank.") It generated 5 relations `affiliatedWith` instead of 3 (marked Yes):

- *Sonja Liht - Fonda za otvoreno društvo* (Yes)
- *Sonja Liht - Unicefa* (No)
- *Meri Blek - Unicefa* (Yes)
- *Meri Blek - Svetske banke* (No)
- *Džim Stivens - Svetske banke* (Yes)

The LLM exhibits difficulties with very long sentences, as processing times become excessively long and the model tends to generate nearly all possible combinations of entities. Training a specialized transformer-based model on the prepared dataset is expected to combine the strengths of both the rule-based and LLM-based approaches, leveraging the high precision of the former and the broader coverage and contextual flexibility of the latter. The dataset produced in this work enables such training. We expect that this integrated approach will lead to more robust and scalable relation extraction, particularly in complex and domain-specific texts such as literary corpora.

7 Conclusion

The presented results are promising and prove that use of FST and LLM-based approaches have their benefits and that the best performance may be achieved when used both. Future research will explore a range of applications of relation extraction, particularly its role in knowledge graph construction and enrichment. Additional directions include its integration into question answering systems, as well as its use in improving information retrieval and text summarization. Further potential lies in applying relation extraction to sentiment and opinion analysis, social network analysis, and domain-specific areas.

In future work, we plan to investigate how strong negative sampling, span filtering, and localized context representations can enable efficient search over all possible spans in an input sentence, as demonstrated by Eberts and Ulges (2020), thus making comprehensive relation extraction more feasible.

Acknowledgments

This research was supported by the Science Fund of the Republic of Serbia, #7276, Text Embeddings–Serbian Language Applications– TESLA and based upon work from COST Action CA23147 GOBLIN-Global Network on Large-Scale, Cross-domain and Multilingual Open Knowledge Graphs, supported by COST (European Cooperation in Science and Technology, https://www.cost.eu)

References

Zi-Yuan Chen, Chih-Hung Chang, Yi-Pei Chen, Jijnasa Nayak, and Lun-Wei Ku. 2019. UHop: An Unrestricted-Hop Relation Extraction Framework for Knowledge-Based Question Answering. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 345–356, Minneapolis, Minnesota. Association for Computational Linguistics.

Jose A Diaz-Garcia and Julio Amador Diaz Lopez. 2025. A Survey on Cutting-edge Relation Extraction Techniques Based on Language Models. *Artificial Intelligence Review*, 58(9):287.

Markus Eberts and Adrian Ulges. 2020. Span-Based Joint Entity and Relation Extraction with Transformer Pre-Training. In *ECAI 2020*, pages 2006–2013. IOS Press.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart De Castilho, and Iryna Gurevych.

2018. The INCEpTION Platform: Machine-assisted and Knowledge-oriented Interactive Annotation. In *Proceedings of the 27th international conference on computational linguistics: system demonstrations*, pages 5–9.

Cvetana Krstev. 2008. *Processing of Serbian – Automata, Texts and Electronic Dictionaries*. University of Belgrade, Faculty of Philology, Belgrade.

Cvetana Krstev, Jelena Jaćimović, and Duško Vitas. 2012. Recognition and Normalization of Some Classes of Named Entities in Serbian. In *Proceedings of the Fifth Balkan Conference in Informatics (BCI '12)*, pages 52–57, New York, NY, USA. ACM.

Cvetana Krstev, Ivan Obradović, Miloš Utvić, and Duško Vitas. 2014. A System for Named Entity Recognition Based on Local Grammars. *Journal of Logic and Computation*, 24(2):473–489.

Cvetana Krstev, Anđelka Zečević, Duško Vitas, and Tita Kyriacopoulou. 2016. NERosetta for the Named Entity Multi-lingual Space. In *Human Language Technology. Challenges for Computer Science and Linguistics: 6th Language and Technology Conference, LTC 2013, Poznań, Poland, December 7-9, 2013. Revised Selected Papers*, pages 327–340, Cham. Springer International Publishing.

Milica Ikonić Nešić, Saša Petalinkar, Ranka Stanković, and Ruslan Mitkov. 2025. From Zero to Hero: Building Serbian NER from Rules to LLMs. In *Proceedings of the First Workshop on Comparative Performance Evaluation: From Rules to Language Models*, pages 87–96.

Milica Ikonić Nešić, Saša Petalinkar, Ranka Stanković, Miloš Utvić, and Olivera Kitanović. 2024. SrpCNNeL: Serbian Model for Named Entity Linking. In *2024 19th Conference on Computer Science and Intelligence Systems (FedCSIS)*, pages 465–473. IEEE.

Ranka Stanković, Milica Ikonić Nešić, Olja Perisic, Mihailo Škorić, and Olivera Kitanović. 2024. Towards Semantic Interoperability: Parallel Corpora as Linked Data Incorporating Named Entity Linking. In *Proceedings of the 9th Workshop on Linked Data in Linguistics@ LREC-COLING 2024*, pages 115–125.

Duško Vitas and Cvetana Krstev. 2012. Processing of Corpora of Serbian Using Electronic Dictionaries. *Prace Filologiczne*, LXIII:279–292.

Duško Vitas, Cvetana Krstev, and Denis Maurel. 2009. A Note on the Semantic and Morphological Properties of Proper Names in the Prolex Project. In Satoshi Sekine and Elisabete Ranchhod, editors, *Named Entities - Recognition, Classification and Use*, volume 19 of *Benjamins Current Topics*, pages 117–136. John Benjamins Publishing Company.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a Free Collaborative Knowledgebase. *Communications of the ACM*, 57(10):78–85.

Somin Wadhwa, Silvio Amir, and Byron C. Wallace. 2023. Revisiting Relation Extraction in the era of Large Language Models. *Proceedings of the conference. Association for Computational Linguistics. Meeting*, 2023:15566–15589.

Yongjun Zhu, Erjia Yan, and Il-Yeol Song. 2017. A Natural Language Interface to a Graph-based Bibliographic Information Retrieval System. *Data & Knowledge Engineering*, 111:73–89.

A Comparative Study of Synonym Generation in Russian with Instruction-Tuned Large Language Models in a Zero-Shot Setting

Maria Khokhlova
St Petersburg State University
Russia
m.khokhlova@spbu.ru

Abstract

The article presents the results of applying modern large language models to the task of generating synonyms for Russian verbs. The aim of the study is to evaluate the quality and characteristics of the generated synonyms depending on the model type and decoding parameters. Two models were employed for synonym generation: the general-purpose multilingual model Mistral and the Russian-language model T-lite.

The experimental material consisted of 18 motion verbs, for each of which synonyms were generated using an instruction-based approach. Multiple runs were performed at different temperature settings. The linguistic analysis of the results was complemented by the computation of quantitative metrics, including the average number of generated synonyms, diversity and stability coefficients, as well as the average frequency of lexical item repetition across runs.

The findings indicate a clear dependence of generation outcomes on temperature. At low temperature values, the models exhibit high stability but limited diversity of the generated synonyms. In contrast, increasing the temperature leads to greater variability and a reduction in repetition. A comparative analysis shows that T-lite produces significantly less noise, with generated lexemes being semantically closer to the target verbs, and better captures the specific features of the Russian language. The results can be applied to tasks such as lexical substitution and the automatic expansion of lexical resources.

Keywords: synonym generation, large language models (LLMs), lexical substitution, Russian language

1 Introduction

Large language models (LLMs) are applied to a wide range of tasks and demonstrate strong performance, particularly in generating texts. At the same time, this raises the question of their applicability to more specialized tasks, such as automatic synonym generation. Synonymy is one of the key lexical relations connecting linguistic units within a language system. The results produced by such models may have broad applications in linguistics, for example in the creation of exercises for language learners, text simplification, style transfer, summarization, or paraphrase generation. The aim of this paper is to compare the performance of different types of LLMs on the task of synonym generation focusing on the quality of the output.

The task of synonym generation belongs to a relatively recent direction in natural language processing and involves automatic selection of words or multi-word expressions that are semantically close to a given item. This naturally raises the first question of whether such models can be effectively applied to synonym generation, thereby supporting lexicographic work. This issue is particularly relevant for morphologically rich languages such as Russian. In addition, the presence of extensive polysemy and homonymy may introduce further challenges and errors in automatic synonym generation. A second research question concerns whether different models produce different results, and if so, what the nature of these differences is.

The paper is organized as follows. The first section motivates the relevance of the study and formulates its objectives. The second section provides a brief review of related work. The third section discusses the obtained results. The final section summarizes the study and outlines directions for future research.

2 Related Work

A great deal of attention has been paid to synonymy in linguistic research, as its accurate description is of fundamental importance both theoretically and practically. Synonym generation is closely related to the historically earlier task of lexical substitution, i.e., finding an appropriate replacement for a word in a given context. Within the SemEval-2007 competition, the task formulation and dataset description were introduced in (McCarthy and Navigli, 2007; McCarthy and Navigli, 2009). Participants were asked to replace target ambiguous words in English contexts, performing three subtasks: (1) selecting the most appropriate substitution; (2) identifying the top 10 candidate substitutions; and (3) determining whether a word is part of a multiword expression. The results produced by systems were compared against annotations provided by human annotators. It was shown that human-generated substitutions are more accurate than those produced by automated systems.

The development of transformer-based language models substantially changed the lexical substitution task. The BERT model (Devlin et al., 2019) showed that bidirectional contextual representations capture semantic information much more precisely than distributional approaches, enabling better contect-based prediction of lexical alternatives. Zhou et al. (2019) proposed one of the first BERT-based methods for lexical substitution, showing that pretrained contextual language models can generate high-quality substitute candidates.

With the emergence of neural network technologies, new models have been proposed for substitution tasks in text, which necessitates the evaluation of more recent algorithms. In particular, a comparison of approaches based on context2vec, ELMo, BERT, and XLNet is presented in (Arefyev et al., 2020). The results show the promise of neural models, while the authors also analyze types of semantic relations between a target word and its proposed substitutions.

A further breakthrough came with the emergence of large autoregressive language models. Brown et al. (2020) demonstrated that LLMs are capable of solving a wide range of downstream tasks with zero-shot and few-shot settings using only prompts, without task-specific fine-tuning. This paradigm shift laid the foundation for subsequent research on instruction tuning and prompted the use of LLMs for various tasks, including lexical generation.

Further studies made it possible to directly compare instruction-tuned models without additional task-specific training (Shi et al., 2024). Wei et al. (2021) demonstrated that instruction tuning substantially improves model performance on previously unseen tasks in the zero-shot setting. At the same time, instruction-tuned models remain sensitive to the formulation of the prompt, and their performance may degrade considerably even when semantically equivalent instructions are used (Sun, Shaib and Wallace, 2023).

Recent surveys emphasize the growing role of LLMs in semantic processing and lexical-semantic tasks (Mao et al., 2024). At the same time, the evaluation of lexical abilities of LLMs remains an active research area, with the benchmarks proposed for assessing lexical instruction following and lexical control (Ren et al., 2026).

In general, synonym selection can be viewed as the task of identifying words that can replace a given lexical unit in a specific context. The primary requirement for candidate words is the preservation of the meaning of the “corrected” text. At the same time, a narrower formulation of the task may reduce it to finding synonyms for isolated words, i.e., without contextual information. In this case, synonym identification may rely on synonym sets or hyponymy relations. Particular attention should be paid to the evaluation of proposed candidate words, which can be carried out using lexicographic resources (e.g., dictionaries or the WordNet thesaurus), expert annotation, or distributional models (vector representations) based on semantic similarity.

3 Methodology

3.1 Data

As the dataset, 18 motion verbs (denoting movement in space) were selected from the Academic Grammar of the Russian Language (Shvedova, 1980). These verbs are also attested in dictionaries and exhibit a large number of related lexical units forming synonymic sets with them. The study considers verbs characterized by single occurrence and unidirectional motion. These include: «бежать» (*bezhat’*) ‘to run’, «брести»

(*bresti*) ‘to trudge/wade’, «везти» (*vezti*) ‘to transport (by vehicle)’, «вести» (*vesti*) ‘to lead’, «гнать» (*gnat’*) ‘to drive/chase’, «гнаться» (*gnat’sya*) ‘to chase’, «ехать» (*ekhat’*) ‘to go (by vehicle)’, «идти» (*idti*) ‘to go/walk’, «катить» (*katit’*) ‘to roll (transitive)’, «катиться» (*katit’sya*) ‘to roll (intransitive)’, «лезть» (*lez’t’*) ‘to climb/creep’, «летать» (*letat’*) ‘to fly (habitual)’, «нести» (*nesti*) ‘to carry’, «нестись» (*nestis’*) ‘to rush’, «плыть» (*plyt’*) ‘to swim/sail’, «ползти» (*polzti*) ‘to crawl’, «тащить» (*tashchit’*) ‘to drag/carry’, «тащиться» (*tashchit’sya*) ‘to trudge/go slowly’.

3.2 Models

In this study, two models were used. First, a general-purpose multilingual model, Mistral (version Mistral-7B-Instruct-v0.3[1]). Second, we used a modern instruction-tuned model, T-lite (version T-lite-instruct-0.1[2]), which is oriented on processing and generating texts in Russian. The choice was motivated by the fact that models from the T-lite family demonstrate strong performance across a range of tasks. They are based on the Qwen 2.5 architecture[3] and were further fine-tuned on large Russian-language datasets, and therefore take into account language-specific characteristics.

A comparison of these two types of models makes it possible to identify the potential impact of fine-tuning on the quality of synonym generation, as well as to evaluate differences between them in the task of identifying semantically similar words.

3.3 Prompting

In the experiments, a zero-shot approach with instruction prompting was used, in which the models were given the task of generating synonyms for verbs from the list without providing any examples, but with an explicit specification of the output format and constraints:

“Generate synonyms for the verbs from the list. The number of synonyms for each verb is 10. The synonyms must not be repeated; they must correspond to verbs and be provided in their base (infinitive) form, separated by commas, in descending order of semantic similarity. Output the result in the following format:

verb1: synonym1, synonym2, …

verb2: synonym1, synonym2, …”.

In total, 8,217 synonyms were obtained during the experiments.

3.4 Hyperparameters of Inference

In the experiments, the following inference hyperparameters were used: temperature, *top_p*, *max_new_tokens*, and *repetition_penalty*. The temperature parameter controls the “creativity” of generation and is sensitive to its value, which affects whether the outputs are more predictable or, conversely, more diverse in nature. The following values were used in this study: 0.0, 0.3, 0.5, 0.7, and 1.0.

According to the hypothesis, at values close to 0, the model tends to generate more probable lexical items that align with dictionary entries, whereas values close to 1 lead to the appearance of less expected (though still semantically appropriate) units and may introduce additional “noise”. The *top_p* parameter (nucleus sampling) controls randomness in token selection based on cumulative probability mass, which must exceed a predefined threshold (0.7 in our experiments). The *max_new_tokens* parameter defines the maximum number of new tokens the model is allowed to generate, i.e., the length of the output (set to 50 in the experiments). The *repetition_penalty* parameter is used to penalize repeated output units (set to 1.05).

Due to the stochastic nature of generation in LLMs (when temperature is greater than zero), multiple runs were performed for each verb. This is motivated by the fact that the model can produce different outputs for the same prompt; therefore, repeated generation allows for more robust and stable results. In the experiments, five runs were conducted for each configuration (model, temperature, and verb), after which the results were aggregated using frequency-based statistics.

3.5 Evaluation Metrics

To evaluate the results, descriptive statistics and derived metrics were used to assess the diversity of the generated outputs.

[1] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

[2] https://huggingface.co/AnatoliiPotapov/T-lite-instruct-0.1

[3] https://huggingface.co/collections/Qwen/qwen25

Diversity is defined as the ratio of the number of unique synonyms obtained by merging the results of all model runs (i.e., the cardinality of the union of sets) to the total number of generated synonyms:

$$Diversity = \frac{|U_{i=1}^{n} w_i|}{\sum_{i=1}^{n} |w_i|},$$

where n is the number of model runs; $|U_{i=1}^{n} W_i|$ is the number of unique synonyms across all runs; $|W_i|$ is the number of synonyms generated in the i-th run.

The metric demonstrates how many different words the model generates at a given temperature across multiple runs. A higher score indicates fewer repeated outputs across model runs performed under identical generation settings, reflecting a greater variety of generated synonyms.

Stability was evaluated as the mean value of the Jaccard coefficient computed over all pairs of runs. For each pair, the ratio of the cardinality of the intersection of the generated synonym sets to the cardinality of their union was calculated, after which the resulting values were averaged:

$$Stability = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} \frac{|w_i \cap w_j|}{|w_i \cup w_j|},$$

where n is the number of model runs; W_i is the set of synonyms generated in the i-th run; and W_j is the set of synonyms generated in the j-th run.

The coefficient corresponds to the overlap (intersection) of results across different generations. A higher coefficient score indicates greater output stability, meaning that the model produces more consistent (i.e., more similar or overlapping) results across repeated runs.

Diversity and stability were selected as the first step for evaluation and the primary criteria because they directly quantify variability and consistency of model outputs. In addition, the *average repeat frequency* of individual synonyms across runs was computed. For each verb, the number of runs in which a given generated lexical item appears was calculated, and these values were then averaged.

All generated items were included in the calculations regardless of their semantic correctness. Consequently, the reported metrics characterize the overall generation behaviour of the models rather than the quality of the generated synonyms. The synonym lists generated by the two models were further subjected to qualitative analysis, providing an expert linguistic evaluation of outputs.

4 Results

4.1 Mistral

The results produced by Mistral differ considerably from those generated by T-lite, both in lexical quality and in the interpretation of the synonym generation task.

Table 1 presents the values of the statistics for the Mistral model.

	0.0	**0.3**	**0.5**	**0.7**	**1.0**
average number of synonyms	9.89	8.62	9.10	8.67	9.28
diversity	0.18	0.42	0.51	0.59	0.61
stability	1.00	0.40	0.30	0.25	0.19
average repeat frequency	5.97	2.45	2.04	1.84	1.64

Table 1: Performance statistics (Mistral).

The model does not consistently generate exactly 10 lexical items despite the instruction explicitly specified in the prompt. The proportion of runs in which the required number of synonyms was produced is 19.8%. In many cases, the generated synonyms exceed a single word in length (e.g., «бегать быстро» [*begat' bystro*] 'to run fast', «бежать на большую дистанцию» [*bezhat' na bol'shuyu distantsiyu*] 'to run a long distance', «катиться по кругу» [*katit'sya po krugu*] 'to roll in a circle', «лежать неподвижно» [*lezhat' nepodvizhno*] 'to lie motionless').

In some cases, instead of a synonym, the model provides explanatory paraphrases: for instance, «ползти» [*polzti*] is glossed as 'to move using arms and legs' or 'to move on all fours', while «тащиться» [*tashchit'sya*] is rendered as 'to move along an inconvenient path' or 'to move over an uneven surface'. Occasionally, the proposed word is accompanied by clarification (e.g., «совершать» [*sovershat'*] with the sense of 'to do').

At the same time, the model does not demonstrate high lexical diversity: a large number of repetitions are observed, and the generated variants often differ only in their prepositional phrase components (e.g., «поддерживать на плаву» [*podderzhivat' na plavu*] 'to keep afloat', «поддерживать в состоянии» [*podderzhivat' v*

sostoyanii] ‘to maintain in a state’, «поддерживать в равновесии» [*podderzhivat’ v ravnovesii*] ‘to keep in balance’, «поддерживать в жизни» [*podderzhivat’ v zhizni*] ‘to sustain in life’, or «поддерживать в действии» [*podderzhivat’v deistvii*] ‘to keep in operation’).

Another characteristic feature of Mistral is the production of malformed lexical items. For example, such items as «агать» [*agat’*], «ваться» [*vat’sia*], «довать» [*dovat’*], «вать» [*vat’*] and «ься» [*’sia*] are not valid Russian lexical forms but parts of words. Such outputs most likely result from subword tokenization errors or incomplete decoding and illustrate one of the weaknesses of the model for lexical generation in Russian.

Another notable phenomenon is the generation of English examples despite the instruction requesting Russian synonyms. The output also contains invented lexical items, such as «спринтировать» (*sprintirovat’*, from English ‘to sprint’), as well as items that do not correspond to verbs (e.g., the noun «погонщик» (*pogonshchik*) ‘herder/driver’). It is noteworthy that for the verb «ползти» (*polzti*) ‘to crawl’, at temperature settings 0.00, 0.03 and 0.05, only English-language examples were produced (“crawl”, “slither”, “scuttle”, “wriggle”, “slink”, “scramble”, “scurry” and “skulk”), whereas Russian synonyms appeared only at temperature settings 0.07 and 1.00.

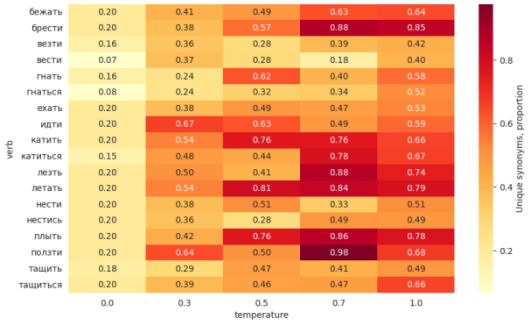


Figure 1: A heatmap of diversity by verbs (Mistral).

The heatmap (see Fig. 1) shows that, in general, the model demonstrates a tendency to generate more diverse synonyms as temperature increases. However, in some cases (e.g., for the verbs «лезть» [*lez’t’*] ‘to climb/creep’ and «летать» [*letat’*] ‘to fly’), maximum diversity is already achieved at t=0.7 and even decreases at t=1.0.

4.2 T-lite

The model generally succeeds in generating verbs belonging to the same lexical field. For example, for the verb «бежать» [*bezhat’*] ‘to run’, T-lite frequently produces the lexical items «бегать» [*begat’*] ‘to run habitually’, «мчаться» [*mchat’sya*] ‘to race’, «нестись» [*nestis’*] ‘to speed’, «убегать» [*ubegat’*] ‘to run away’ and «спешить» [*speshit’*] ‘to hurry’, that denote rapid movement. For the verb «брести» [*bresti*] ‘to trudge’, the model generates such lexical items as «плестись» [*plestis’*] ‘to plod’, «ковылять» [*kovyliat’*] ‘to hobble’, «тащиться» [*tashchit’sya*] ‘to drag oneself’ and «волочиться» [*volochit’sya*] ‘to trail along’, which share the sememes of slow or difficult movement. The results suggest that the model captures fine-grained semantic distinctions related to the manner of motion rather than relying solely on general semantic similarity.

In some cases, during test runs, the model generated additional verbs that were not specified in the original prompt, which required post-processing of the results.

Table 2 presents the values of the evaluation metrics for the T-lite model.

	0.0	**0.3**	**0.5**	**0.7**	**1.0**
average number of synonyms	10.00	12.96	11.00	10.93	11.75
diversity	0.17	0.52	0.61	0.74	0.88
stability	1.00	0.27	0.24	0.13	0.06
average repeat frequency	5.88	1.94	1.72	1.35	1.14

Table 2: Performance statistics (T-lite).

In contrast to Mistral, the model generates a more diverse set of synonyms, which in most cases do not exceed one word in length. However, in a number of cases, the model appears to generate descriptive expressions rather than lexical synonyms, effectively explicating the meaning of the source verb. Examples include «идти пешком» [*idti peshkom*] ‘to walk on foot’, «нести в руках» [*nesti v rukakh*] ‘to carry in one’s hands’, «вести за собой» [*vesti za soboy*] ‘to lead along’ and «плыть по течению» [*plyt’ po techeniiu*] ‘to drift with the current’. Although these phrases are semantically meaningful, they do not satisfy

the definition of single-word lexical substitution adopted in this study.

Another recurring phenomenon is the generation of text continuations rather than synonyms. For instance, the results include «вести к» [*vesti k*] ‘to lead to’, «вести через» [*vesti cherez*] ‘to lead through’, «взбираться на» [*vzbirat’sya na*] ‘to climb onto’ and «подняться в небо» [*podniat’sya v nebo*] ‘to rise into the sky’, suggesting that the model interprets the task as one of text continuation rather than lexical substitution.

T-lite occasionally produces semantically inappropriate items. For example, for the verb «везти» (*vezti*) ‘to transport; to carry by vehicle’, T-lite occasionally generates «говорить» (*govorit’*) ‘to speak’, which is unrelated to the semantics of motion. Another example is the colloquial verb «тащиться» (*tashchit’sya*) ‘to drag oneself; to trudge’ which produces «врубаться» (*vrubat’sia*) ‘to get into; to understand (slang)’, reflecting the secondary colloquial meaning of «тащиться» (‘to enjoy’ or ‘to be enthusiastic about’) rather than its literal meaning dealing with motion. This observation demonstrates that the model is sensitive to highly frequent colloquial senses and fails to correctly disambiguate polysemous verbs.

The model adheres more strictly to the prompt taking into account the number of items: the majority of runs (72.3%) produced exactly 10 synonyms; however, increasing temperature leads to an increase in the volume of generated outputs.

Fig. 2 presents a heatmap showing the results for the T-lite model according to the diversity metric. The model consistently produces more diverse outputs as temperature increases. It is important to note that, overall, the generated synonyms are semantically close to the corresponding verbs. For example, for «брести» (*bresti*) ‘to trudge’, the model produces: «ковылять» (*kovylyat’*) ‘to hobble’, волочить ноги (*volochit’ nogi*) ‘to drag one’s feet’, «ползти» (*polzti*) ‘to crawl’, «шаркать» (*sharkat’*) ‘to shuffle’, «плестись» (*plestis’*) ‘to plod’.

The model did not generate outputs at t=0.0 for the verbs «везти» (*vezti*) ‘to transport (by vehicle)’ and «нестись» (*nestis’*) ‘to rush’. One possible explanation is the use of greedy decoding, whereby the model deterministically selects the most probable token sequence. In some cases, this decoding strategy may fail to generate a complete response, resulting in empty outputs.

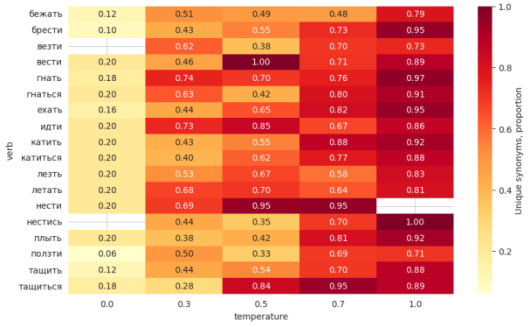


Figure 2: A heatmap of diversity by verbs (T-lite).

In addition, no substitutions were produced for the verb «нестись» (*nestis’*) at t=1.0, which may be explained by the stochastic nature of the model. At higher temperature values, this stochasticity can manifest as “noise” and generation errors.

4.3 Discussion

The analysis of the results reveals notable differences in the candidate synonyms generated by the two models. Overall, T-lite predominantly generates lexical synonyms, whereas Mistral frequently fails to adhere to the lexical substitution task, instead producing meaning descriptions, text continuations, parts of the original prompt or explanatory reasoning accompanying its responses. These findings suggest that the differences between the models are not limited to the number of correct responses but also involve distinct strategies for interpreting the instruction. While T-lite generally treats the prompt as a lexical substitution task, Mistral often interprets it as a more general text generation task.

For both models, the proportion of unique synonyms increases with rising temperature, i.e., the diversity of generated outputs becomes higher. However, this is also accompanied by an increase in “noise” and erroneous outputs, particularly in the case of Mistral. Conversely, the stability metric demonstrates the robustness of both models at t=0.0, where the results are nearly identical across runs, whereas at t=1.0 maximum variability is observed, reflecting increased stochasticity in generation.

The T-lite model exhibits a lower degree of overlap between runs as temperature increases

compared to Mistral; at the highest temperature, the intersection between runs is almost negligible. This also indicates greater generative variability and creativity of the model.

The average repeat frequency metric shows that at low temperature values, synonyms that are consistently reproduced across all runs dominate, whereas with increasing temperature, the proportion of unique lexical items that appear only once increases.

5 Conclusion

The analysis of the generated lexical items shows that instruction-tuned LLMs capture semantic relations between Russian motion verbs, but also exhibit some limitations associated with the complex nature of lexical semantics of Russian verbs, including aspectual characteristics, polysemy or stylistic variation.

The results confirm the feasibility of using LLMs for generating synonyms for a given lexical unit; however, the selected parameter settings significantly influence the obtained outputs. Varying the temperature parameter allowed control over the stochasticity and diversity of the generated items. At low temperature values, the systems exhibit a high rate of repetition among outputs, i.e., generation is largely deterministic, whereas higher temperature values lead to more variable outputs, which are nevertheless characterized by instability across runs. At intermediate values, a balance between stability and diversity of generation is achieved.

The T-lite model demonstrated a lower number of errors and irrelevant outputs compared to the multilingual Mistral model. The latter exhibits weaker adherence to the constraints (including lexical ones) imposed by the prompt. Our observations suggest that instruction-tuned LLMs perform lexical generation rather than dictionary-based synonym selection, showing both the flexibility and some limitations of zero-shot approach.

In future work, it is planned to include a larger number of models fine-tuned on Russian-language data for comparative evaluation. Another direction of research may involve assessing the extent to which modern Russian instruction-tuned LLMs are capable of reproducing synonymic relations documented in lexicographic resources.

Acknowledgments

The author acknowledges the financial support of Russian Science Foundation (project No. 22-18-00189-P “Structure and Functionality of Stable Multiword Units in Russian Everyday Speech”).

References

Diana McCarthy and Roberto Navigli. 2007. SemEval-2007 Task 10: English Lexical Substitution Task. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic, pages 48–53.

Diana McCarthy and Roberto Navigli. 2009. The English Lexical Substitution Task. In *Language Resources and Evaluation*, 43(2), Special Issue on Computational Semantic Analysis of Language: SemEval-2007 and Beyond, Agirre E., Màrquez L., Wicentowski R. (eds.), pages 139–159. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Wangchunshu Zhou, Tao Ge, Ke Xu, Furu Wei, and Ming Zhou. 2019. BERT-based Lexical Substitution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3368–3373, Florence, Italy. Association for Computational Linguistics.

Nikolay Arefyev, Boris Sheludko, Alexander Podolskiy, and Alexander Panchenko. 2020. A Comparative Study of Lexical Substitution Approaches Based on Neural Language Models. *arXiv preprint arXiv:2006.00031 [cs]*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss,

Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33: 1877–1901.

Ning Shi, Bradley Hauer, and Grzegorz Kondrak. 2024. Lexical Substitution as Causal Language Modeling. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 120–132, Mexico City, Mexico. Association for Computational Linguistics.

Jason Wei, Maarten Paul Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew Mingbo Dai, and Quoc V. Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jiuding Sun, Chantal Shaib, and Byron C. Wallace. 2023. Evaluating the Zero-shot Robustness of Instruction-tuned Language Models. *arXiv preprint arXiv:2306.11270*

Rui Mao, Kai He, Xulang Zhang, Guanyi Chen, Jinjie Ni, Zonglin Yang, and Erik Cambria. 2024. A survey on semantic processing techniques. *Information Fusion*, Vol. 101, Art. 101988.

Huimin Ren, Yan Liang, Baiqiao Su, Chaobo Sun, Hengtong Lu, Kaike Zhang, and Chen Wei. 2026. LexInstructEval: Lexical Instruction Following Evaluation for Large Language Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, *40*(30), pages 25117–25123.

Natalia Yu. Shvedova (ed.). 1980. *Russkaya grammatika (AG-80) [Russian Grammar (AG-80)]*, Vol. 1: Phonetics. Phonology. Stress. Intonation. Word Formation. Morphology. Moscow: Nauka, 789 p.

Cognitive and Corpus-Based Approaches to Language Processing

Assessing Reading Development through Finger-Voice Span: a Comparative Study of Bulgarian and Italian

Andrea Nadalini[1], Claudia Marzi[1], Marcello Ferro[1], Vito Pirrelli[1], Maria Todorova[2], Valentina Stefanova[2], Tsvetana Dimitrova[2], Hristina Kukova[2] and Svetla Koeva[2]

[1]Institute for Computational Linguistics "A. Zampolli", Italian National Research Council
{andrea.nadalini,claudia.marzi,marcello.ferro,vito.pirrelli}@cnr.it

[2]Institute for Bulgarian Language, Bulgarian Academy of Science
{maria,valentina,cvetana,hristina,svetla}@dcl.bas.bg

Abstract

This study investigates finger–voice span (FVS) as an index of reading development in primary school children, through a cross-linguistic comparison between Italian and Bulgarian. FVS measures the spatial distance, in characters, between the finger position in a finger-point oral reading session and the voice onset of the word being vocalized during the same session, providing insight into the coordination of visual, motor, and linguistic processes during reading. Data were collected from 231 children in grades 2-5, during oral reading tasks administered through the ReadLet application. Finger movements and voice articulation were synchronously recorded for subsequent alignment and analysis. To ensure reliable temporal alignment for Bulgarian, a subset of recordings was manually annotated at the token level to evaluate an Automatic Speech Recognition (ASR) pipeline adapted from the Italian version. The ASR system, based on WhisperX and enhanced with energy-based segmentation and dual alignment, provided reliable temporal alignment, supporting its use for large-scale analyses. FVS measures were extracted from more than 44,000 tokens and analysed using linear mixed models. Results revealed that both grade and language significantly influenced FVS. Italian children exhibited higher FVS values than their Bulgarian peers across all grade levels, while FVS increased with grade in both languages, reflecting shared developmental changes in reading coordination. These results provide evidence of shared developmental changes in FVS across languages, alongside language-specific differences. The study highlights the potential of the finger-tracking technique combined with ASR for a scalable, cross-linguistic assessment of reading proficiency in naturalistic settings.

Keywords: reading, finger-tracking, reading, finger-voice span (FVS), automatic speech recognition (ASR), Bulgarian, Italian.

1 Introduction

Oral reading of connected text relies on the tight coordination between ocular and articulatory movements. Eye movements are faster and more flexible than articulatory movements. They typically scan a text forward and backward using a variety of strategies, such as forward saccades, regressions, refixations, and word skipping. In contrast, the articulatory output of a skilled reader is characterised by smooth, continuous articulation, with few repetitions or disfluencies and only occasional pauses at linguistically complex boundaries (Inhoff et al., 2011; De Luca et al., 2013; Ziaka and Protopapas, 2025).

To coordinate these processes effectively, readers rely on a fine-tuned motor control strategy. This involves temporarily storing decoded units in a phonological buffer, integrating them into larger prosodic structures, and planning and executing articulation based on the buffered material (Laubrock and Kliegl, 2015).

This dynamic can be investigated by measuring the eye-voice span (EVS), defined as the temporal or spatial distance between the onset of the first fixation on a word and its articulation. Skilled readers can maintain more information in the phonological buffer and adjust their reading pace to larger linguistic units, such as chunks and phrases, reflecting efficient prediction, fluent decoding, and effective speech planning. Less proficient readers, by comparison, tend to show a smaller and less variable EVS, with eye movements more tightly aligned to oral output, reflecting slower word-by-word processing. By capturing the coordination between eye movements, lexical access, and speech planning, EVS provides an informative index of reading proficiency (Nadalini et al., 2023; Marzi et al., 2025).

It has recently been shown (Taxitari et al., 2021; Ferro et al., 2024) that finger-tracking data provide

a compelling alternative to eye-tracking evidence, enabling high–temporal resolution measurement of reading behaviour in naturalistic settings. Finger-tracking captures the timing and trajectory of the reader's index finger as it slides along the lines of a written text during finger-point reading, yielding fine-grained evidence of processing without requiring specialised laboratory equipment. In this study, we used the ReadLet infrastructure, a cloud-based platform delivering reading tasks via a tablet application, which enables simultaneous recording of finger movements and voice articulation in oral reading (Ferro et al., 2018).

Crucially, finger-tracking measures have been shown to strongly correlate with established eye-tracking indices, such as fixation durations and saccadic patterns, in both adult and developing readers (Crepaldi et al., 2022; Marzi et al., 2020). In addition, finger-voice span (FVS) captures a coordination pattern similar to the eye–voice span. In early readers, haptic movements are tightly coupled with gaze, indicating that finger pointing supports visual attention and sequential scanning of letters and syllables (Mesmer and Lake, 2010). As reading proficiency develops, however, the role of the finger shifts from supporting visual alignment to tracking articulation timing, with its movements becoming more synchronised with the speech stream (Nadalini et al., 2023; Marzi et al., 2025).

This convergence suggests that finger movements can reliably reflect underlying oculomotor control and lexical processing. In addition, the portability, scalability, ease of implementation, and relatively low cost of finger-tracking make it particularly well suited for large-scale studies in educational contexts and across diverse languages, such as Bulgarian and Italian (Lento et al., 2024; Marzi et al., 2026). These features support broader access to developmental reading data and facilitate cross-linguistic comparisons, with potential applications in educational practice and literacy policy (Nadalini et al., 2025).

Here, we present a comparative analysis of FVS in Italian and Bulgarian primary school pupils. To ensure reliable token-level alignment in Bulgarian data, we first evaluated an automatic speech recognition (ASR) system by comparing its output against manually time-stamped recordings, using the Italian-developed system as a reference (Srivastava et al., 2026). Thus, this study addresses two main research questions: (1) Does Finger-Voice Span increase with reading development in Bulgarian? (2) Does the developmental trajectory of FVS differ between Bulgarian and Italian?

2 Participants

The dataset consists of 231 primary school pupils from grades 2-5, recruited in Pisa (Italy, 109 recording sessions) and Sofia (Bulgaria, 122 recording sessions). Participation required parental consent and followed no selection criteria based on reading proficiency. All participants had normal or corrected-to-normal vision, no reported hearing impairments, and no diagnosed neurodevelopmental disorders, learning disabilities, or cognitive impairments. Participants were native or dominant speakers of the language tested at their respective location. Demographic characteristics of both Bulgarian and Italian participants are reported in Table 1.

Grade	N	Female	Male	Mean age (sd)
		BG		
2	20	7	13	8.64 (0.33)
3	38	16	22	9.57 (0.37)
4	32	16	16	10.50 (0.35)
5	32	24	8	11.50 (0.37)
		IT		
2	18	9	9	7.65 (0.35)
3	22	14	8	8.78 (0.20)
4	32	15	17	9.74 (0.33)
5	37	17	20	10.00 (2.50)

Table 1: Distribution of Bulgarian (BG) and Italian (IT) participants by grade level. Age is reported in decimal months.

3 Material and Methods

Data were collected using a 10.1-inch Android tablet. Speech were recorded with a bidirectional electret microphone with a 150-6800Hz frequency range and a sensitivity of -47dBV/Pa. The audio signal was further resampled at 48000Hz, 16bit, stereo and compressed to audio-WEBM format at 128kbps.

The reading materials consisted of narrative texts designed to be comparable across languages in terms of length and lexical, morphological, and syntactic properties. The same set of narratives was originally written in Italian and subsequently translated into Bulgarian following procedures aimed at preserving structural and linguistic equivalence (Pirrelli and Koeva, 2024). Each narrative was

structured into five episodes with gradually increasing length and complexity. Children read a grade-specific subset of each narrative, ranging from two episodes in grade 2 to the full five-episode set in grade 5. All texts were processed with language-specific annotation tools (Dell'Orletta et al. 2011, for Italian; Karagiozov et al. 2011, for Bulgarian). Lexical frequency measures were derived from Subtlex (Crepaldi et al., 2016) and the Bulgarian National Corpus (Koeva et al., 2012). Summary statistics for the reading materials are reported in Table 2.

	Ep. 1	Ep. 2	Ep. 3	Ep. 4	Ep. 5
			BG		
Word Count	119.75 (7.14)	133.25 (14.13)	140.75 (6.85)	151.00 (12.99)	155.50 (9.33)
Word Frequency	5.42 (1.32)	5.27 (1.46)	5.18 (1.51)	5.31 (1.47)	5.31 (1.47)
Word Length	4.57 (2.28)	4.78 (2.68)	4.88 (2.82)	4.80 (2.80)	4.82 (2.94)
Sentence Count	13.75 (0.50)	11.50 (0.58)	10.00 (0.00)	8.00 (0.82)	7.50 (0.58)
Sentence Length	8.70 (0.27)	11.59 (1.15)	14.08 (0.68)	18.92 (1.09)	20.82 (1.99)
Dependency Length	1.73 (1.42)	2.03 (2.09)	2.22 (2.22)	2.30 (2.45)	2.59 (2.94)
			IT		
Word Count	141.00 (0.82)	152.00 (0.82)	166.00 (4.83)	169.75 (3.20)	178.00 (7.96)
Word Frequency	5.76 (1.27)	5.45 (1.47)	5.46 (1.49)	5.41 (1.51)	5.38 (1.53)
Word Length	4.21 (2.17)	4.66 (2.46)	4.71 (2.53)	4.87 (2.76)	5.07 (3.01)
Sentence Count	14.00 (0.00)	11.50 (0.57)	10.00 (0.00)	8.00 (0.82)	7.50 (0.58)
Sentence Length	10.07 (0.00)	13.33 (0.67)	17.00 (0.00)	21.80 (2.25)	23.57 (1.81)
Dependency Length	1.57 (1.39)	1.78 (1.79)	1.93 (2.04)	2.06 (2.43)	2.33 (2.97)

Table 2: Descriptive statistics (mean and standard deviation) for Bulgarian (BG) and Italian (IT) reading texts by text episodes.

4 Manual Annotation of Bulgarian data

Manual annotation was carried out on a subset of Bulgarian audio recordings, corresponding to 40 text episodes out of a total of 387 recordings. The sample included ten recordings for each grade level (grades 2-5). Annotators were asked to mark the time onset and offset of each vocalized word in an audio recording, and type the transcript into the resulting time interval using Praat's TextGrid editor (Boersma and Weenink, 2024).

The subset of reading sessions was selected to ensure coverage of different text episodes and a balanced distribution across male and female participants. Each audio file was time-aligned with a tokenized version of its corresponding written text, with tokens being defined as individual orthographic words. Annotation included the creation of synchronised tiers for audio–text alignment, ensuring correspondence between the acoustic signal and the text. Alignment was supported by spectrogram and waveform visualisations, with appropriately set frequency ranges and temporal resolution. This enabled precise localisation of acoustic events and improved discrimination of temporally adjacent segments. Token boundaries were manually marked based on acoustic criteria, taking into account phoneme transitions, coarticulation effects, and reductions typical of spontaneous speech. In cases of ambiguous boundaries, a conservative strategy was adopted, with transitional segments assigned to the relevant token. This typically occurred in fast reading conditions, when prepositions or conjunctions merge with adjacent words, making exact boundary identification difficult.

Additional labels with annotators' comments were applied to mark occurrences of word repetitions, fragmented or interrupted words, hesitations, filled pauses, and instances of syllabification produced during reading. These phenomena can be further categorised according to established typologies of speech disfluencies (Zhang, 2020), including:

- Repetitions – immediate reiteration of a word, words, parts of words during reading.
- Self-interruptions / broken words – abrupt cessation of a word before completion.
- Restarts – abandonment of an initiated word or phrase followed by a new attempt at its pronunciation.
- Filled pauses – vocalized hesitation markers such as uh, um, or other non-lexical fillers.
- Unfilled pauses – silent intervals exceeding a predefined duration threshold.
- Prolongations – atypically extended articulation of a sound.
- Syllabification – deliberate or unintentional production of a word in separate syllables.
- Incorrect pronunciation – deviations from the expected phonological realisation of a word.

Description of the annotation principles, annotation labels and representative examples is provided in Appendix A.

5 The ASR algorithm

We used the ASR pipeline originally developed by Srivastava et al. (2026), built on the WhisperX framework (Bain et al., 2023). The first step of the pipeline is an energy-based pause detection algorithm that segments an input waveform into prosodically coherent units using energy (loudness) thresholds. Prosodic units are then transcribed using a pre-trained, language-specific Whisper Large V3 model, which produces segment-level transcriptions. These transcriptions are subsequently normalized for punctuation and diacritics before being passed to a dual alignment module. The alignment module features separate optimization components for token onset (Grosman, 2022) and offset (Conneau et al., 2020) time alignments. The final step of the algorithm is a token-level post-processing stage that merges these alignments, correcting systematic temporal shifts and ensuring consistency across neighbouring tokens. The final output consists of token boundaries with precise temporal alignment augmented with acoustic and prosodic cues, explicitly accounting for disfluencies in spontaneous speech.

6 ASR evaluation

ASR performance was evaluated using multiple complementary metrics. First, we measured temporal deviations (deltas) between automatically detected boundaries and manual annotations, separately for onset and offset timestamps. Next, we computed Intersection-over-Union (IoU[1]) across token segments to assess overall temporal alignment quality. Finally, as an aggregate measure of performance, we computed the proportion of tokens whose onset and offset boundaries deviated by less than 50 ms from the manual reference (i.e., within a phonetic-scale tolerance). We refer to this metric as the Collared Matching Rate (CLMR). Results are shown in Table 3[2]. A consistent pattern emerges across all evaluation metrics, with higher temporal alignment accuracy for Italian than for Bulgarian. This is reflected in both boundary-based and overlap-based measures. Mean and median onset and offset deviations were lower for Italian (36–38 ms mean, 19–23 ms median) than for Bulgarian (64–67 ms mean, 36–44 ms median), indicating more precise token boundary localisation. This advantage is further supported by overlap-based evaluation, with higher IoU values for Italian (80%) compared to Bulgarian (72–78%). Similarly, the Collared Matching Rate (CLMR), measuring the proportion of words aligned within a ±50 ms tolerance, was substantially higher for Italian ($\approx$ 70%) than for Bulgarian ($\approx$ 36%). Despite this performance gap, median timing errors remained below 50 ms for both languages, suggesting that Whisper maintains a reasonable level of temporal precision in both cases. The observed differences may be attributed to a combination of factors, including language-specific phonetic and prosodic characteristics, differences in training data coverage for Whisper models, and greater acoustic variability in children's oral production in Bulgarian.

Overall, the magnitude of the language difference is not uniform, across metrics. While IoU and boundary deviation measures show moderate though consistent gaps between Italian and Bulgarian, the CLMR exhibits a substantially larger divergence, suggesting higher sensitivity to strict temporal tolerance criteria. This pattern suggests that small systematic shifts in token boundaries have limited impact on overlap-based measures but a stronger effect on tolerance-based evaluation. Thus, the three metrics provide converging evidence of cross-linguistic differences, while capturing complementary aspects of alignment accuracy. As a result, these metrics suggest that while the ASR pipeline generalises across languages, its performance is modulated by language-specific properties that affect boundary detection accuracy.

7 FVS Preprocessing

For each word in the text, spatial finger-voice span (FVS) measures the spatial distance (in characters) between the finger position within a word and the onset word articulation, namely how far the finger is when the voice starts pronouncing it. Overall, FVS values were extracted for a total of 53,938 word tokens (27,357 from Bulgarian children, 26,581 from Italian children). Tokens were excluded if tracking time and speech duration fell below 35 ms or ex-

[1]For each time-stamped token, IoU is calculated by dividing the overlapping duration (intersection) between the model estimate and the ground truth, by the total combined duration of the two intervals. IoU provides a gross estimate of the overall accuracy of the model time stamps.

[2]ASR results for Italian are taken from Srivastava et al. (2026).

Metrics	Italian	Bulgarian
Mean Onset Delta (ms)	38	64
Median Onset Delta (ms)	23	44
Mean Offset Delta (ms)	36	67
Median Offset Delta (ms)	19	36
Mean IoU (%)	80.86	71.52
Median IoU (%)	86.04	77.54
Mean CLMR (%)	70.20	36.03
Median CLMR (%)	69.74	36.64

Table 3: ASR performance comparison in the two languages.

ceeded 3.5 seconds, as well as in cases where FVS exceeded 20 characters. This filtering resulted in a final dataset of 41,926 observations included in the analyses.

8 FVS Analyses

Statistical analyses were conducted in R. Separate linear mixed models were fitted for Bulgarian and Italian using the *lme4* package (Bates et al., 2015) to examine the effects of grade level grades 2-5) on spatial onset FVS. Each model included random intercepts for word token, reader and text, accounting for repeated measures and item- and participant-level variability. Significance of the fixed effects was assessed using Type III Wald chi-square tests implemented in the *car* package (Fox et al., 2007). Model results are reported in Table 4 and model predictions are shown in Figure 1.

Grade level significantly affected onset FVS in both Bulgarian (χ^2 = 11.07, df = 3, p = .011) and Italian children (χ^2 = 9.19, df = 3, p = .027), indicating developmental changes across the primary school years. The developmental trajectories, however, differed between the two languages. In Bulgarian, predicted FVS values increased progressively across grade levels. In Italian, FVS also increased from Grade 2 to Grade 4, but showed a slight decrease in Grade 5. Although the analyses were conducted separately for the two language groups, Figure 1 descriptively shows higher predicted FVS values for Italian readers across all grade levels. The different developmental trajectories observed in Bulgarian and Italian suggest that changes in FVS may not follow a uniform developmental pattern. Rather than reflecting a simple linear increase in reading proficiency, FVS is likely shaped by multiple interacting factors, including reading experience, language-specific characteristics, and the increasing complexity of the reading material across grade levels.

BULGARIAN DATA						
Fixed Effects						
		Est.	S.E.	t val.	d.f.	p
(Intercept)		1.56	0.44	3.56	48.84	0.001
gradeLevel3		1.07	0.51	2.10	125.12	0.038
gradeLevel4		1.22	0.52	2.34	122.54	0.021
gradeLevel5		1.72	0.52	3.31	121.31	0.001
Random Effects						
Group	Parameter		Std. Dev.			
token	(Intercept)		0.82			
idSession	(Intercept)		1.75			
idDoc	(Intercept)		0.28			
Residual			3.49			
ITALIAN DATA						
Fixed Effects						
		Est.	S.E.	t val.	d.f.	p
(Intercept)		3.18	0.58	5.47	38.11	0.001
gradeLevel3		0.65	0.72	0.91	114.86	0.366
gradeLevel4		1.85	0.67	2.78	115.32	0.006
gradeLevel5		1.35	0.65	2.08	115.54	0.040
Random Effects						
Group	Parameter		Std. Dev.			
token	(Intercept)		1.27			
idSession	(Intercept)		2.14			
idDoc	(Intercept)		0.42			
Residual			3.69			

Table 4: LMM results for spatial onset Finger-Voice Span in Bulgarian and Italian separately

9 Discussion

9.1 Developmental findings

In the present study, we examined developmental changes in finger–voice span (FVS) in Bulgarian and Italian primary school children. Separate analyses revealed a significant effect of grade level in both language groups, indicating developmental changes in FVS across the primary school years. Italian children consistently exhibited higher FVS values than their Bulgarian peers across all grade levels. The developmental trajectories slightly differed between the two groups: Bulgarian children showed a steady increase in FVS across grades, whereas Italian children displayed an increase up to grade 4 followed by a slight decrease in grade 5. A supplementary model including both language groups did not show a significant language x grade interaction, suggesting that the overall effect of grade was

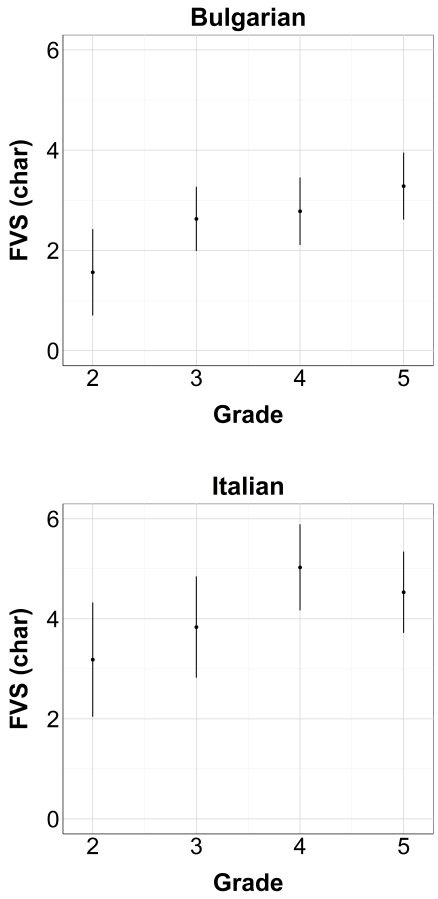


Figure 1: LMM estimates of average FVS values across grades for Bulgarian (top panel) and Italian (bottom panel) children. Error bars represent 95% prediction intervals.

broadly comparable across languages.

The tendency for Italian children to exhibit higher FVS values than Bulgarian children be due to several factors. First, it may reflect differences in lexical and morphosyntactic predictability between the Italian and Bulgarian texts. Additionally, it may be a consequence of the coarser time-stamps of Bulgarian automated transcriptions. An earlier estimated onset of word articulation (relative to the real onset) may, in fact, underestimate the distance of the finger from the voice. Finally, it may also be the outcome of different oral reading styles and teaching practices. However, the absence of a significant language x grade interaction in the combined model suggests that these cross-linguistic differences do not substantially modify the overall developmental trajectory of FVS. Despite qualitative differences in the observed patterns, both groups exhibit an overall increase in FVS with reading experience, pointing to broadly shared developmental constraints. During the early stages of reading acquisition, motor anchoring strategies such as finger pointing are thought to support visual attention and sequential decoding. As reading proficiency increases, however, this coupling gradually becomes less dominant, giving way to a more flexible coordination of visual, motor, and linguistic processes (Marzi et al., 2025).

9.2 Implications

The present findings suggest that tracking the development of finger-voice span (FVS) may provide a means of identifying both typical and atypical patterns in early reading learning during primary school years. Given that reading difficulties often emerge subtly in this period, FVS may offer a complementary measure for monitoring reading development.

From an applied perspective, the results suggest that finger-tracking can be implemented using standard tablet devices, enabling unobtrusive collection of multimodal reading data in classroom environments. Building on prior implementations in Italian schools, the present study shows that a similar protocol can be deployed in a different linguistic context with minimal adaptation. This suggests that the approach is suitable for scalable data collection and repeated assessment of reading development at both group and individual levels. Such fine-grained longitudinal data may, in fact, support learner-specific analyses that could inform personalised feedback on reading development.

9.3 Methodological considerations and limitations

An important methodological advantage of the present framework lies in the use of synchronised multimodal data streams, which allow for cross-channel constraints that can improve signal quality during processing. In particular, information from one modality can be used to mitigate noise or temporal uncertainty in another. For example, misalignment or drift in finger-tracking data can be partially corrected using temporally aligned speech signals in oral reading, jointly anchored to the text (Marzi et al., 2025). Conversely, speech data can provide coarse-grained information about the reader's position in the text. However, these benefits depend on the quality and robustness of each individual modality, and residual errors in ASR or segmentation may propagate across streams. This introduces a potential source of shared error that should be considered

when interpreting fine-grained temporal measures such as FVS.

9.4 Multimodal integration and future directions

Taken together, the present findings are consistent with a model in which FVS development reflects the interplay between general maturational factors and language-related characteristics that influence baseline levels, while developmental changes across grades remain broadly comparable across languages, in line with evidence on reading speed and automatisation (Marzi et al., 2026). Future research should further disentangle the relative contribution of linguistic, educational, and age-related factors to FVS development, ideally using fully age-matched samples and more tightly controlled instructional conditions. Extending this approach to a wider range of languages would further clarify the generalisability of the observed developmental patterns.

Acknowledgments

The present study is carried out within the project National Scientific Program "Development of Scientific Research and Innovation in the System of Bulgarian Preschool and School Education" Agreement SP-13/04.11.2025

References

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *arXiv preprint arXiv:2303.00747*.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of statistical software*, 67:1–48.

Paul Boersma and David Weenink. 2024. Praat: Doing phonetics by computer. https://www.praat.org/. Version 6.x.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition.

D Crepaldi, S Amenta, P Mandera, E Keuleers, and M Brysbaert. 2016. Frequency estimates from different registers explain different aspects of visual word recognition. In *International meeting of the Psychonomic society, Granada, Spain*, pages 5–8.

Davide Crepaldi, Marcello Ferro, Claudia Marzi, Andrea Nadalini, Vito Pirrelli, and Loukia Taxitari. 2022. *Finger Movements and Eye Movements During Adults' Silent and Oral Reading*, pages 443–471. Springer International Publishing, Cham.

Maria De Luca, Maria Pontillo, Silvia Primativo, Donatella Spinelli, and Pierluigi Zoccolotti. 2013. The eye-voice lead during oral reading in developmental dyslexia. *Frontiers in human neuroscience*, 7:696.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. 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*, pages 73–83.

Marcello Ferro, Claudia Cappa, Sara Giulivi, Claudia Marzi, Oaufae Nahli, Franco Alberto Cardillo, and Vito Pirrelli. 2018. Readlet: Reading for understanding. In *2018 IEEE 5th International Congress on Information Science and Technology (CiSt)*, pages 1–6. IEEE.

Marcello Ferro, Claudia Marzi, Andrea Nadalini, Loukia Taxitari, Alessandro Lento, and Vito Pirrelli. 2024. Readlet: a dataset for oral, visual and tactile text reading data of early and mature readers. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13595–13609.

John Fox, Gregor Gorjanc Friendly, Spencer Graves, Richard Heiberger, Georges Monette, Henric Nilsson, Brian Ripley, Sanford Weisberg, Maintainer John Fox, and MASS Suggests. 2007. The car package. *R Foundation for Statistical computing*, 1109(1431):84039–5.

Jonatas Grosman. 2022. Fine-tuned XLS-R 1B model for speech recognition in Italian. https://huggingface.co/jonatasgrosman/wav2vec2-xls-r-1b-italian.

Albrecht W Inhoff, Matthew Solomon, Ralph Radach, and Bradley A Seymour. 2011. Temporal dynamics of the eye–voice span and eye movement control during oral reading. *Journal of Cognitive Psychology*, 23(5):543–558.

Diman Karagiozov, Svetla Koeva, Maciej Ogrodniczuk, and Cristina Vertan. 2011. Atlas–a robust multilingual platform for the web. In *Proceedings of the German Society for Computational Linguistics and Language Technology Conference (GSCL 2011)*, pages 223–226.

Svetla Koeva, Ivelina Stoyanova, Svetlozara Leseva, Rositsa Dekova, Tsvetana Dimitrova, and Ekaterina Tarpomanova. 2012. The Bulgarian National Corpus: Theory and practice in corpus design. *Journal of Language Modelling*, 0(1):65–110.

Jochen Laubrock and Reinhold Kliegl. 2015. The eye-voice span during reading aloud. *Frontiers in psychology*, 6:1432.

Alessandro Lento, Andrea Nadalini, Marcello Ferro, Claudia Marzi, Vito Pirrelli, Tsvetana Dimitrova, Hristina Kukova, Valentina Stefanova, Maria Todorova, and Svetla Koeva. 2024. Assessing reading literacy of Bulgarian pupils with finger–tracking. In *Proceedings of the Sixth International Conference on Computational Linguistics in Bulgaria (CLIB 2024)*, pages 140–149.

Claudia Marzi, Marcello Ferro, Andrea Nadalini, Vito Pirrelli, Maria Todorova, Tsvetana Dimitrova, Valentina Stefanova, Hristina Kukova, and Svetla Koeva. 2026. Comparable Reading Development in Bulgarian and Italian: Cross-Linguistic Insights from a Finger-Tracking Study. *Languages*, 11(4):70.

Claudia Marzi, Andrea Nadalini, Alessandro Lento, Manu Srivastava, Alice Todesco, Vito Pirrelli, and Marcello Ferro. 2025. Oral text reading as a multisensory task. *Lingue e linguaggio*, 24(1):141–156.

Claudia Marzi, Anna Rodella, Andrea Nadalini, Loukia Taxitari, and Vito Pirrelli. 2020. Does finger-tracking point to child reading strategies? In *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020)*, pages 204–210.

Heidi Anne E Mesmer and Karen Lake. 2010. The role of syllable awareness and syllable-controlled text in the development of finger-point reading. *Reading Psychology*, 31(2):176–201.

Andrea Nadalini, Claudia Marzi, Marcello Ferro, Alessandra Cinini, Paola Cutugno, and Davide Chiarella. 2025. Inferential reading skills in high school: A study on comprehension profiles. *Education Sciences*, 15(6):654.

Andrea Nadalini, Claudia Marzi, Marcello Ferro, Loukia Taxitari, Alessandro Lento, Davide Crepaldi, and Vito Pirrelli. 2023. Eye-voice and finger-voice spans in adults' oral reading of connected texts: Implications for reading research and assessment. *The Mental Lexicon*, 18(3):366–400.

Vito Pirrelli and Svetla Koeva. 2024. Developing materials for assessing reading literacy and comprehension of early graders in Bulgaria and Italy. *Foreign Language Teaching*, 51(1).

Manu Srivastava, Marcello Ferro, Vito Pirrelli, and Gianpaolo Coro. 2026. Enhancing token boundary detection in disfluent speech. *Intelligent Systems with Applications*, 29:200614.

Loukia Taxitari, Claudia Cappa, Marcello Ferro, Claudia Marzi, Andrea Nadalini, and Vito Pirrelli. 2021. Using mobile technology for reading assessment. In *2020 6th IEEE congress on information science and technology (CiSt)*, pages 302–307. IEEE.

Hui Zhang. 2020. *The Distribution of Disfluencies in Spontaneous Speech: Empirical Observations and Theoretical Implications*. Doctoral dissertation, University of Pennsylvania.

Laoura Ziaka and Athanassios Protopapas. 2025. Dynamic adjustment of the eye–voice span and articulation duration in the course of multi-item naming tasks: Evidence for lockout scheduling. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Appendix A Acoustic Annotation of Transcribed Data

A.1 Annotation labels

All non-word audio segments as well as pauses and annotators‘ comments were marked in triangular brackets <>.
Ex.
избязло <pronunciation mistake – избягало ’run’>
по- <partial word>
в-няко-<unclear pronunciation>

A.2 Pause annotation

Annotations were grouped into functionally similar categories (e.g., restarts vs. repetitions). Minimum duration thresholds were also defined for coding speech pauses.

These annotations enable additional analyses of reading behaviour. Pauses were distinguished according to their position: Pauses were distinguished according to their position:

- constituent boundary (marking the end of a sentence, a clause or a phrase);
- internal constituent pauses (hesitations, repetitions, sighing).

A.3 Examples of pause annotation

Constituent boundary pauses:
Ex.
Тара.PROPN <pause1>, както се.REFL.CL <pause2> казваше.V <pause3> (3rd-grader)
Tara <pause1>, as (she was) REFL.CL <pause2> called, <pause3>

In the example, pause1 and pause3 mark a constituent boundary (of an embedded clause), while pause2 unexpectedly separates the reflexive clitic (се) from the verb (казваше ’named’), thereby resulting in the clitic being associated with the preceding wh-constituent.

Internal constituent pauses:
Ex.
благородна- <pause1> -та кралица на <pause2> паяците <pause3> в моята <pause4> градина <highly muted> <pause5> (3rd-grader)
noble- <pause1> -the queen of <pause2> spiders-the <pause3> in my-the <pause4> garden <highly muted> <pause5>

In this example, pause1 occurs before the definite article (-та), which may be analysed as a clitic-like element. Pause2 separates the preposition на (’of’) from its governed noun паяците (’the spiders’), while pause4 separates two noun phrases. Finally, pause3 and pause5 mark constituent boundaries.

A.4 Annotation observations

Observations made during the annotation may provide an additional perspective on the assessment of reading skills. They allow for the monitoring of a range of interrelated parameters, including reading speed and fluency, as well as comprehension of words in context and lexical breadth.

Experiment Design of a Balanced Eye-Tracking Study for Measuring the Cognitive Impact of the Stimulus Word "Disinformation"

Ruslana Margova
Institute GATE
5, James Bourchier Blvd
Sofia, Bulgaria
ruslana.margova@gate-ai.eu

Irina Temnikova
Institute GATE
5, James Bourchier Blvd
Sofia, Bulgaria
irina.temnikova@gmail.com

Abstract

This paper presents the design of a controlled eye-tracking experiment for investigating the cognitive impact of the stimulus word "disinformation" in Bulgarian news texts. Rather than reporting experimental findings, the paper describes the methodological framework developed for a balanced eye-tracking study that will examine whether the presence of this word influences readers' visual attention and gaze behaviour. The proposed design is intended for a Tobii Pro Spectrum eye tracker (600 Hz) and to compare matched pairs of authentic Bulgarian news headlines and leads in which the stimulus word is either preserved or removed while keeping the remaining textual context as similar as possible. The methodology controls for participant-related factors, including age, gender, education, prior knowledge and topic interest, as well as text-related factors such as length, readability and complexity. The experiment is designed to measure fixation duration, gaze duration and regression frequency in order to identify potential differences in cognitive processing associated with the presence of the stimulus word. By documenting the rationale behind the experimental design, the balancing strategy and the methodological challenges, this paper provides a reproducible framework for future eye-tracking studies on misinformation in Bulgarian, a low-resource language for which such experimental research is currently lacking. The resulting eye-tracking dataset and the accompanying corpus are intended to be released as open resources after completion of the study.

Keywords: eye-tracking, experimental design, disinformation, news, design.

1 Context and Motivation

Eye-tracking has become an established methodology for investigating how readers process truthful and misleading information. Previous studies have shown that gaze behaviour, including fixation duration and regressive eye movements, differs when readers encounter factually incorrect content, making eye-tracking a promising approach for studying misinformation processing. However, most eye-tracking studies on misinformation have been conducted for English, while evidence for other languages remains scarce. Extending such research to Bulgarian is scientifically important for several reasons. Bulgarian is a low-resource language in Natural Language Processing, and, to the best of our knowledge, there is currently no publicly available eye-tracking corpus dedicated to misinformation or news credibility in Bulgarian. Furthermore, Bulgarian differs from English in its use of the Cyrillic script, rich inflectional morphology, and language-specific lexical and syntactic characteristics. These linguistic differences may influence reading behaviour and visual attention, making it inappropriate to assume that findings obtained for English can be directly generalised to Bulgarian readers. In this study, we adopt the definition of "disinformation" proposed by the European Commission as "false or misleading content that is spread with the intention to deceive or secure economic or political gain, and which may cause public harm"[1]. We investigate whether the explicit presence of the stimulus word "disinformation" affects readers' visual attention during the processing of Bulgarian news headlines. Building on this rationale, the primary contribution of this paper is a reproducible methodological framework that controls for participant-related factors (e.g., age, gender, education, prior knowledge, and topic interest) and text-related factors (e.g., length, readability, and complexity), while isolating the effect of the stimulus word "disinformation". By documenting the methodological decisions and design challenges, we aim to establish a foundation for future eye-tracking studies on misinformation in Bulgarian.

[1] https://digital-strategy.ec.europa.eu/en/policies/online-disinformation

2 Related Work: Eye-Tracking for Misinformation

Eye-tracking has been widely used to investigate how readers process truthful and factually incorrect information. Previous studies have shown that eye-movement behaviour can distinguish between truthful and misleading news and can support predictive models based solely on gaze measurements. For instance, studies in English have demonstrated that false headlines receive significantly less visual attention than truthful ones (Hansen et al., 2020). Beyond visual attention, previous research has investigated the psychological aspects of news presentation (Ladeira et al., 2022) and the integration of peripheral physiology with machine learning for automated disinformation detection (Filippi et al., 2025). A further behavioural indicator is that readers tend to exhibit more regressive eye movements when processing factually incorrect content (Bozkir et al., 2022). The influence of different interface elements, such as breaking-news formats and additional links, has also been analysed to understand how news presentation affects visual attention (Chu et al., 2009; Kruikemeier et al., 2020).

Although eye-tracking research has expanded to several Slavic languages, it has primarily focused on linguistic phenomena rather than misinformation. Previous studies include language acquisition in Russian (Ermolina, 2022) and Czech (Juřík et al., 2025), readability assessment for Russian (Laurinavichyute et al., 2019), detection of dyslexic reading patterns in Serbian (Vajs et al., 2022), pronominal anaphora resolution in Polish (Wolna et al., 2022), and the processing of verbal aspect in Polish (Klimek-Jankowska et al., 2018). These studies demonstrate the applicability of eye-tracking methodologies to Slavic languages, while also highlighting the absence of research on misinformation processing.

To the best of our knowledge, at the time of writing this article, no previous eye-tracking study has investigated misinformation (and related concepts) processing in Bulgarian. This paper addresses this gap by proposing a balanced experimental design that isolates the cognitive effect of the stimulus word "disinformation" while controlling for participant-related and text-related variables. The proposed methodology is intended to provide a reproducible framework for future eye-tracking studies on misinformation (and similar concepts) in Bulgarian.

3 Research Question

The objective of this study is to investigate whether the presence of the stimulus word "disinformation" influences readers' visual attention while processing Bulgarian news headlines.

Specifically, we address the following research question:

RQ1. Does the presence of the stimulus word "disinformation" influence readers' eye-movement behaviour during the reading of Bulgarian news headlines?

To answer this question, we propose a balanced eye-tracking experiment in which participants read matched pairs of authentic Bulgarian news headlines that differ only in the presence or absence of the stimulus word. The experiment is designed to examine whether the presence of the stimulus word is associated with differences in fixation duration and regression frequency.

4 Challenges in Design

We encountered several challenges when designing the eye-tracking experiment, summarised in Table 1.

Challenge	Proposed solution
Defining factuality	Exclusion of political news and manual selection of authentic news items.
Building comparable text pairs	Creation of matched headline pairs from authentic news by minimally editing the original texts.
Prior knowledge	Assessment through pre- and post-reading questionnaires.
Prior exposure	Control question identifying previously encountered news items.
Text complexity	Readability metrics and complexity measures for Bulgarian texts.
Participant bias	Balanced participant recruitment with respect to age, gender, education, and background knowledge.

Table 1: Main methodological challenges and the corresponding mitigation strategies adopted in the experimental design.

- Political news was excluded to minimise con-

founding effects arising from participants' political attitudes and prior beliefs. In addition, factual verification of political claims is often inherently disputed because such claims may involve subjective interpretation, incomplete evidence, or evolving contexts. Fact-checking organisations cannot comprehensively cover the news cycle and may prioritise claims according to editorial criteria or public attention ((Fernández-Roldán et al., 2023), (Hameleers and Van der Meer, 2020). Restricting the corpus to non-political news therefore improves experimental control and reduces variability unrelated to the stimulus word. We exclude political news for several reasons, including the political bias of readers and the challenge of balancing it. Political statements are often subjective or contain 'nuanced truths,' making them difficult to categorize as strictly 'true' or 'false' for a controlled experiment. Another problem is the fact-checkers' assessment of politicians' claims (Fernández-Roldán et al., 2023). Further, fact-checkers cannot check the whole news flow, and they are most likely to select news when they confirm prior attitudes (Hameleers and Van der Meer, 2020). Some scholars share that they use factually correct and factually incorrect news, but without mentioning how they decide which news is trustworthy or not (Menicocci et al., 2024).

- Building indicators for factually correct and factually incorrect news is a big challenge. Some scholars made a set of features as clickbait or logical fallacies (Zhang et al., 2018), others define credible news as news which the reader wants to share (Torres et al., 2018), third group of scholars make a pair of news - true and fabricated, but the methodology of fabrication is not pretty clear (Simko et al., 2019).

- News media with different views may word the same news in different ways. The selection of texts turns out to be a major challenge in the current experiment - to define what is true and how exactly disinformation is fabricated. Our initial approach was to construct matched text pairs using articles produced by professional fact-checking organisations. However, this approach proved unsuitable for three reasons. First, fact-checking articles represent a distinct journalistic genre whose style differs substantially from conventional news reporting, making direct comparison problematic. Second, fact-checkers typically debunk broader narratives rather than individual news items, which prevented the construction of comparable one-to-one text pairs. Third, many of the original social media posts or news items referenced in fact-checking reports are no longer publicly accessible. Because the objective of this study is to compare authentic news texts under controlled conditions, we instead constructed our own corpus of Bulgarian news headlines containing the stimulus word "disinformation". The word "disinformation" was selected as a primary stimulus because it likely functions as an emotional trigger and a cognitive anchor. From a psycholinguistic perspective, such high-stakes terms can induce an 'Emotional Stroop effect', where the negative valence of the word captures attentional resources, potentially increasing the cognitive load required to process the remaining sentence. By isolating this word, we aim to measure whether its presence triggers a shift from heuristic to systematic processing, detectable through increased fixation durations and regression frequencies (Ben-Haim et al., 2016), (Singh and Goel, 2024). The experiment is designed to examine whether the presence of the stimulus word is associated with longer fixation durations and more frequent regressive eye movements.
Therefore, we took a different approach – specifically for the experiment, we created a dataset of actual news headlines from online news outlets that contain the word disinformation in a title or subtitle or lead of up to 250 characters. These headlines were collected manually over the past three years. To create pairs of almost identical comparable texts, we reworked the titles so that the word disinformation was missing from them, without replacing it with synonymous words. In some cases, only this word was removed, in others, an entire syntactic structure had to be removed to avoid its use. As a result, we have pairs of titles containing the word "disinformation" and those that do not contain it. Our goal is to check whether the word disinformation triggers attention, as opposed to its absence in the

same context.

One of the limitations of this approach is that when we have pairs of texts, given randomly, part of them will already be familiar to our participants. The text will be presented as plain without additional visual stimuli. Participants are not informed about the specific focus on 'disinformation' beforehand to prevent demand characteristics—a phenomenon where subjects subconsciously alter their reading behavior to fit the perceived goals of the study. This ensures the eye-tracking data reflects naturalistic processing. They will be notified afterwards.

- Several reader-related factors influence the reader when deciding whether a text contains factually correct information or factually incorrect information. We tried to take into account as many of them as possible.
  - Reader's previous knowledge about the text's topic. We took this factor into account by selecting readers with different education levels, and as equal as possible number of female and male participants. Additionally, we specifically looked for readers with and without previous knowledge in the text's topics. We also asked the readers to state the fields in which they had more knowledge before the experiment, and after reading each text to answer the question whether they had prior knowledge in the text's field.
  - Whether the text's topic is of interest to the reader. We took this detail into consideration by asking each reader whether they had personal interest in each text after reading it, and why.
  - The reading literacy and the media literacy of participants could be different and could influence the results. An additional post-reading questionnaire will help to measure the preparedness of the participants.
  - To account for reader fatigue, the number of texts is specifically calibrated. We measure tiredness through a self-report scale included in the post-reading questionnaire.
- Randomization of the order of texts is standard practice (Smith et al., 2014), (Walker, 2020).
- Text length, complexity and coherence. Text complexity affects eye movement metrics (such as fixation duration), which must be isolated to ensure we are measuring the impact of the stimulus word, not the difficulty of the sentence (Torres et al., 2021). The complexity in Bulgarian is not a much-analysed topic (Petrova and Bozhikova, 2022). We tried to take news headlines and leads of a similar length. The variables frequently used for measuring complexity are the average length of the words and their frequency in the language, both accounting for semantic difficulty, as well as sentence length, which is closely related to syntactic complexity (Torres et al., 2021). To measure the complexity of our text, we calculate the average sentence length, the average word length, and apply a Bulgarian language readability formula (Kazakov et al., 2025). Given the impossibility of having only texts with the same complexity, we aim to differentiate the results given their complexity. The coherence of a text, related to its meaningfulness, a notion associated with semantics rather than grammatical structure, we analyse for each stimulus, annotating as coherent or incoherent. Topic complexity is another problem. We do also consider that some topics may be too specialised for readers with a lower age or lower level of education.
- Given all the mentioned details under consideration, we wonder how to recruit the necessary types of participants. Participants will be recruited through advertisements posted in Facebook groups representing different interests and viewpoints.

5 Experiment Description

This section describes the settings of the planned experiment, whose actual execution is left for future work.

5.1 Eye-Tracker

The experiment will use the Tobii Pro Spectrum screen-based eye-tracker, designed for scientific research with a maximum sampling frequency of 600 Hz. The data collection will take place in a controlled indoor office environment. The Tobii Pro Lab software will be employed to collect, analyse, and visualize the gaze data.

5.2 Texts Used

The stimulus material (see Table 2 for a summary) consists of matched pairs of Bulgarian news headlines and leads, each limited to 250 characters. To ensure focus on the cognitive impact of the stimulus word, political topics are excluded. The design employs Areas of Interest (AOIs) to isolate the word "disinformation", enabling a precise comparison of fixation durations between the original and modified versions. For instance, a headline such as "CIA misled Iran with disinformation during the rescue of a US pilot" is compared against the version where the stimulus word is omitted.

Not all pairs permit such a clean single-word deletion: in some cases, removing "disinformation" without leaving an ungrammatical or semantically incoherent sentence required deleting an entire clause or syntactic structure rather than the word alone. Because such structural edits introduce changes in sentence length and syntactic complexity that could independently affect fixation duration and regression frequency, all 30 stimulus pairs will be classified prior to data collection by edit type: single-word deletion (structure otherwise unchanged) versus structural deletion (clause or phrase removed). [Report the resulting counts, e.g., "X pairs required only word deletion; Y pairs required structural deletion."] For pairs requiring structural deletion, we will additionally compute the character-length difference and token-level edit distance between the two versions of the pair; these measures will be entered as covariates in the statistical model, alongside the word-presence factor, so that the effect of the stimulus word itself can be disentangled from the effect of the accompanying structural change.

5.3 Participants

We plan to involve in this study participants from diverse age ranges (18–79), balanced by gender and education level. The experiment will begin with an initial pilot phase (see Table 2), followed by a larger sample to ensure statistical robustness. Potential selection biases will be mitigated through diverse recruitment channels.

5.4 Participant Recruitment and Eligibility

We plan to recruit participants for a paid eye-tracking experiment designed to study visual attention. To be eligible, participants must be at least 18 years old and have normal or corrected-to-normal vision. As in other eye-tracking experiments, participants should not be tired, should have had enough sleep, and should not have eye infections or medical conditions affecting the brain or cognitive processing. Potential selection biases will be mitigated through diverse recruitment channels.

Factor Category	Variables / Details
Within-subject factors	Presence of the stimulus word (with/without "disinformation")
Between-subject factors	Education; Gender; Prior knowledge
Pilot sample	10 participants
Stimulus material	30 text pairs

Table 2: Summary of the experimental design and sample requirements.

The time commitment will be of one hour and a half, including setup and data collection, and participants will be paid for their time.

5.5 Experimental Procedure

We plan to conduct the actual experiment in the following way: Upon arrival, participants will complete a short agreement with basic information and a formal agreement regarding data privacy. The eye-tracker will be calibrated for each participant, followed by a brief practice trial to ensure that the participant feels comfortable with the equipment. During the session, the participants will read the stimuli while their eye movements are recorded. We will focus on fixations, saccades, and gaze duration: fixations indicate where and for how long participants look, saccades represent eye movements, and gaze duration captures the time spent on relevant parts of the text. After the experiment, the participants will be debriefed about the experiment and its aims.

5.6 Guidelines

Before starting the task, participants will be instructed to read the presented headlines and leads naturally, at their own pace, as they would normally do when browsing news online. No specific emphasis will be placed on any word to ensure natural reading behavior. **To ensure data quality, participants will be asked to focus on a calibration fixation cross prior to each stimulus. After reading the text, they will manually advance the screen to**

complete the corresponding post-reading questionnaire.

5.7 Post-Reading Questionnaire

Immediately following the reading task, the participants will complete a follow-up survey designed to assess their interest and knowledge about the news items. These answers will later be compared with their gaze patterns. The post-reading questionnaire consists of the following items:

- **Factuality:** Participants must answer whether they perceive the information as factually correct, factually incorrect, or whether they cannot answer. Since it is possible that one part of a text is true while another is not, the participants will be asked to generalise and to classify the item as false if they have even the slightest doubt.
- **Prior Knowledge:** The participants must state whether they have previous knowledge of the topic.
- **Interest:** The participants must state whether the topic is interesting to them.
- **Prior Exposure:** To mitigate the potential bias of prior exposure to the news items, a control question—"Have you encountered this news story before?"—is included in the post-reading questionnaire, allowing for the exclusion of familiar texts from the final statistical analysis to ensure data purity.

6 Ethical Questions

We aim to follow all best practices and legal and ethical requirements in conducting this experiment. Specifically, we:

- are filing an ethical permission application to the Ethical Committee of our host institution;
- plan to provide each participant with a document disclosing how their personal data will be processed and ensuring that their identities and beliefs remain confidential from other participants;
- will include in the document a warning that infrared sensors are used, to prevent participation by users who might have issues with such sensors;
- will include only participants who are in the accepted working age, i.e., at least 18 years old;
- will pay participants for their time, reflecting our country's basic salaries.

7 Limitations

A more fundamental limitation concerns the interpretability of the presence/absence manipulation itself: it does not fully isolate the emotional valence of "disinformation" from its lexical properties (length, frequency, morphological complexity), all of which are independently associated with fixation duration in reading research. Because the compared conditions differ in whether the word is present at all, rather than swapping it for a matched neutral word, observed differences in fixation duration cannot be unambiguously attributed to emotional processing as opposed to word length or frequency effects. Future iterations of this design could include a third condition substituting a length- and frequency-matched neutral word (e.g., "информация" ["information", which shares morphological structure with "дезинформация" while lacking its negative valence) to more precisely isolate the emotional-Stroop-like effect from general lexical processing costs. Although the proposed design addresses several methodological challenges, certain further limitations remain. These include the limited availability of suitable Bulgarian news texts, the exclusion of highly political content to minimise bias, the relatively small corpus of manually collected headlines, and the possibility that some participants may already be familiar with individual news items. Despite controlling for participant- and text-related variables, unobserved individual differences may still influence reading behaviour.

8 Conclusions and Further Work

This paper has detailed a methodological framework for a balanced eye-tracking experiment investigating the cognitive impact of the stimulus word "disinformation" in Bulgarian news texts, addressing key challenges related to text selection, participant characteristics, ethical considerations, and experimental control.

A central contribution of the proposed design is the isolation of the stimulus word "disinformation", allowing its potential influence on visual attention to be investigated under controlled conditions. The

use of matched text pairs aims to maximise internal validity while preserving the ecological validity of authentic news texts. We also acknowledge the limitations of the proposed design, including the relatively limited availability of suitable Bulgarian news texts and the possibility that some participants may already be familiar with individual news items.

The proposed methodology lays the foundation for a large-scale eye-tracking study of misinformation processing in Bulgarian. Future work will implement the experiment, analyse eye-movement measures such as fixation duration, saccadic behaviour, and gaze duration, and evaluate whether the presence of the stimulus word is associated with systematic differences in reading behaviour. We also plan to extend the study to additional news formats, including audio and video, and to integrate complementary physiological measures (e.g., cognitive load indicators) (Zagermann et al., 2016). The resulting corpus and anonymised eye-tracking dataset are intended for open-access release to support reproducibility, comparative studies, and further research.

Acknowledgments

This research is part of the GATE project funded by the Horizon 2020 WIDESPREAD-2018-2020, TEAMING Phase 2, the Programme under grant agreement no. 857155, the Programme "Research, Innovation and Digitalization for Smart Transformation" 2021-2027 (PRIDST) under grant agreement no. BG16RFPR002-1.014-0010-C01 and, the project BROD (Bulgarian-Romanian Observatory of Digital Media), funded by the Digital Europe Programme of the European Union under contract number 101226153.

References

Moshe Shay Ben-Haim, Paul Williams, Zachary Howard, Yaniv Mama, Ami Eidels, and Daniel Algom. 2016. The emotional Stroop task: assessing cognitive performance under exposure to emotional content. *Journal of visualized experiments: JoVE*, (112):53720.

Efe Bozkir, Gjergji Kasneci, Sonja Utz, and Enkelejda Kasneci. 2022. Regressive saccadic eye movements on fake news. In *2022 Symposium on Eye Tracking Research and Applications*, pages 1–7.

Sauman Chu, Nora Paul, and Laura Ruel. 2009. Using eye tracking technology to examine the effectiveness of design elements on news websites. *Information Design Journal (IDJ)*, 17(1).

Iuliia Ermolina. 2022. Acquisition of verbal aspect in L2 English by advanced learners with L1 Russian and L1 Norwegian: A web-based eye tracking study. Master's thesis, UiT Norges arktiske universitet.

Alejandro Fernández-Roldán, Carlos Elías, Carlos Santiago-Caballero, and David Teira. 2023. Can we detect bias in political fact-checking? Evidence from a Spanish case study. *Journalism Practice*, pages 1–19.

Eleonora De Filippi, Arijit Nandi, and Alexandre Pereda Baños. 2025. Misinformation detection through multimodal machine learning analysis of eye-tracking and physiological responses: A proof-of-concept study. In *Pattern Recognition. ICPR 2024 International Workshops and Challenges*, pages 18–33, Cham. Springer Nature Switzerland.

Michael Hameleers and Toni GLA Van der Meer. 2020. Misinformation and polarization in a high-choice media environment: How effective are political fact-checkers? *Communication research*, 47(2):227–250.

Christian Hansen, Casper Hansen, Jakob Grue Simonsen, Birger Larsen, Stephen Alstrup, and Christina Lioma. 2020. Factuality checking in news headlines with eye tracking. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 2013–2016.

Vojtěch Juřík, Libor Juhaňák, Alexandra Ružičková, Nicol Dostálová, and Zuzana Juříková. 2025. Experimental dataset on eye-tracking activity during self-regulated learning. *Scientific Data*, 12(1):967.

Dimitar Kazakov, Stefan Minkov, Ruslana Margova, Irina Temnikova, and Ivo Emanuilov. 2025. Towards creating a Bulgarian readability index. In *Proceedings of the First Workshop on Advancing NLP for Low-Resource Languages*, pages 192–200.

Dorota Klimek-Jankowska, Anna Czypionka, Wojciech Witkowski, and Joanna Błaszczak. 2018. The time course of processing perfective and imperfective aspect in Polish: Evidence from self-paced reading and eye-tracking experiments. *Acta Linguistica Academica*, 65(2-3):293–351.

Sanne Kruikemeier, Sophie Lecheler, and Ming M Boyer. 2020. Learning from news on different media platforms: An eye-tracking experiment. In *Studying Politics Across Media*, pages 75–96. Routledge.

Wagner Junior Ladeira, Marlon Dalmoro, Fernando de Oliveira Santini, and William Carvalho Jardim. 2022. Visual cognition of fake news: the effects of consumer brand engagement. *Journal of Marketing Communications*, 28(6):681–701.

Anna K Laurinavichyute, Irina A Sekerina, Svetlana Alexeeva, Kristine Bagdasaryan, and Reinhold Kliegl. 2019. Russian Sentence Corpus: Benchmark measures of eye movements in reading in Russian. *Behavior research methods*, 51(3):1161–1178.

Stefano Menicocci, Viviana Lupo, Silvia Ferrara, Andrea Giorgi, Eleonora Serra, Fabio Babiloni, and Gianluca Borghini. 2024. Fake-news attitude evaluation in terms of visual attention and personality traits: A preliminary study for mitigating the cognitive warfare. *Behavioral Sciences*, 14(11):1026.

Daniela Petrova and Violeta Bozhikova. 2022. Sentiment and complexity analysis on two databases in Bulgarian language–final estimation. In *2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, pages 01–04. IEEE.

Jakub Simko, Martina Hanakova, Patrik Racsko, Matus Tomlein, Robert Moro, and Maria Bielikova. 2019. Fake news reading on social media: an eye-tracking study. In *Proceedings of the 30th ACM conference on hypertext and social media*, pages 221–230.

Sakshi Singh and Sanjay Goel. 2024. The impact of cognitive load on belief in fake news and real news. In *AMCIS 2024 Proceedings*.

Nicholas D Smith, Fiona C Glen, Vera M Mönter, and David P Crabb. 2014. Using eye tracking to assess reading performance in patients with glaucoma: A within-person study. *Journal of ophthalmology*, 2014(1):120528.

Débora Torres, Wagner R Sena, Humberto A Carmona, André A Moreira, Hernán A Makse, and José S Andrade Jr. 2021. Eye-tracking as a proxy for coherence and complexity of texts. *PloS one*, 16(12):e0260236.

Russell Torres, Natalie Gerhart, and Arash Negahban. 2018. Epistemology in the era of fake news: An exploration of information verification behaviors among social networking site users. *ACM SIGMIS Database: The DATABASE for Advances in Information Systems*, 49(3):78–97.

Ivan Vajs, Vanja Ković, Tamara Papić, Andrej M Savić, and Milica M Janković. 2022. Spatiotemporal eye-tracking feature set for improved recognition of dyslexic reading patterns in children. *Sensors*, 22(13):4900.

Callum Walker. 2020. Eye-tracking the reader experience. In *An Eye-Tracking Study of Equivalent Effect in Translation: The Reader Experience of Literary Style*, pages 215–289. Springer.

Agata Wolna, Joanna Durlik, and Zofia Wodniecka. 2022. Pronominal anaphora resolution in Polish: Investigating online sentence interpretation using eye-tracking. *Plos one*, 17(1):e0262459.

Johannes Zagermann, Ulrike Pfeil, and Harald Reiterer. 2016. Measuring cognitive load using eye tracking technology in visual computing. In *Proceedings of the sixth workshop on beyond time and errors on novel evaluation methods for visualization*, pages 78–85.

Amy X Zhang, Aditya Ranganathan, Sarah Emlen Metz, Scott Appling, Connie Moon Sehat, Norman Gilmore, Nick B Adams, Emmanuel Vincent, Jennifer Lee, Martin Robbins, et al. 2018. A structured response to misinformation: Defining and annotating credibility indicators in news articles. In *Companion Proceedings of the The Web Conference 2018*, pages 603–612.

Everyday Speech of Russian Students: From Field Recordings to the ESC Speech Corpus (Data Processing and Representation)

Tatiana Sherstinova
HSE University-Saint Petersburg, Russia
tsherstinova@hse.ru

Irina Petrova
HSE University-Saint Petersburg, Russia
irina.petrova1011@gmail.com

Aleksei Melnik
HSE University-Saint Petersburg, Russia
melnik-a@russian-short-stories.ru

Karina Azarevich
HSE University-Saint Petersburg, Russia
kiazarevich@edu.hse.ru

Sofiia Chepovetskaia
HSE University-Saint Petersburg, Russia
soncheroz1503@gmail.com

Viktoria Deborina
HSE University-Saint Petersburg, Russia
vddeborina@edu.hse.ru

Abstract

This paper presents the current state of the ESC corpus, which is the largest collection (more than 2,000 hours of recordings) of spontaneous recordings of everyday Russian speech obtained in natural communicative environments. The data are collected using a long-duration speech recording methodology: volunteers record their everyday communication over extended periods of time, typically during a full "speech day", using professional audio recorders or high-quality mobile devices. The corpus data document the current state of youth everyday speech, including the active use of neologisms, emerging colloquial vocabulary, and elements of youth slang. The paper provides updated statistics on the current stage of corpus development and focuses on the methodological and technological aspects of corpus construction. In particular, the processing pipeline relies on diarization and automatic speech recognition systems for generating initial transcripts of the recordings. Another important stage of speech data preparation is the anonymization of personal information, as a substantial portion of everyday conversational recordings contains private and sensitive content. The ESC corpus can serve as a testbed for evaluating NLP tools for Russian. The resulting resource is expected to be valuable both for fundamental research on contemporary Russian spoken discourse and youth communicative practices, and for applied tasks such as the development and training of automatic speech recognition systems, language models, and other natural language processing technologies.

Keywords: corpus linguistics, spoken Russian, everyday speech, youth speech, speech corpora, speech technologies, speech data anonymization, oral discourse, speech processing.

1 Introduction

Everyday spoken speech is a truly unique type of linguistic material. It is the form of language we use in most communicative situations, helping us deal with everyday tasks both in private life and in professional settings. Its production does not require paper, writing tools, or any digital devices. This mode of communication has existed at all times, linking us with our distant ancestors and, most likely, remaining an essential means of communication for future generations as well.

Paradoxically, however, we know much less about spoken speech than about written language. Most grammars, dictionaries, and linguistic statistics are based on written texts, while only a relatively small number of studies dealing with spoken data rely on so-called "laboratory" speech — that is, language material recorded under

specific communicative conditions (in recording studios, over the phone, or in interview settings). The properties of such data differ considerably from those of natural spontaneous speech. As a result, everyday spoken language still lacks a sufficiently detailed theoretical and statistical description.

Despite the fact that in recent decades more spoken corpora have been developed (see, for example, (CLARIN 2026; Du Bois 2000; BNC 2007)), both the overall number of such resources and the volume of audio data they contain remain clearly insufficient for most of the world's languages. A notable exception is Japanese, where everyday language use has been the focus of linguistic research since the 1940s (Shibamoto 1987; Shibata 1983; Campbell 2004).

Until recently, the largest source of authentic spoken data for Russian has been the "One Day of Speech" corpus (ORD), recorded in 2007–2016 in Saint Petersburg using a continuous speech recording methodology. The corpus reflects everyday communication of 128 informants representing different social groups of the urban population—across gender, age, professional, and social status categories—which has made it possible to carry out a wide range of phonetic, lexical, grammatical, pragmatic, and sociolinguistic studies based on these data (Asinovsky et al. 2009; Bogdanova-Beglarian et al. 2016). However, language, as a living system, is constantly evolving, and material recorded ten years ago can no longer fully reflect its current state. This is especially true of youth speech.

This paper presents a new audio resource of everyday Russian speech, focused specifically on the speech of young people and students. It provides updated statistics on the current stage of corpus development and highlights key methodological and technological aspects of its construction.

2 ESC Speech Corpus: Methodology and Stages of Development

The ESC corpus of youth and student speech is a spoken corpus designed to document and study everyday Russian based on authentic spontaneous speech used in both private and professional communication (Sherstinova and Petrova 2024; Sherstinova et al. 2025). The development of the ESC corpus began in March 2023. It is based on the continuous speech recording methodology originally developed for the "One Day of Speech" (ORD) corpus (see above). Its key feature, however, is the focus on youth speech: the corpus primarily includes recordings of speakers aged 18–27. This age group — socially active and sensitive to current trends — makes it possible to capture the most up-to-date state of the language at a given point in time.

At the initial stage of the project, the main goal was to establish a stable workflow for continuous data collection. Over the past three years, a recording process has been set up that allows the corpus to be steadily expanded with new audio data. Informants participating in the project complete a set of documents, including: (1) informed consent to participate in the study; (2) consent to the processing of personal data; (3) a sociological questionnaire; (4) a speech day diary; (5) a psychological test; and (6) a passive vocabulary test.

For recording, informants are provided with professional audio devices — typically Zoom H1n or Tascam DR-05X recorders (output format: WAV, stereo, 16-bit, 44,100 Hz), which they carry with them throughout their regular day.

At present, the ESC corpus contains over 2,000 hours of original recordings collected from 165 participants, including 122 women and 43 men. The statistics as of early April 2026 are presented in Table 1. Thus, in terms of overall volume, the ESC corpus has already surpassed its predecessor, the ORD corpus, which contains approximately 1,500 hours of audio recordings.

Gender	**Informants (N, %)**		**Recording Hours (N, %)**	
Female	122	73.94%	1574	77.81%
Male	43	26.06%	449	22.19%

Table 1: Gender Distribution of ESC Recordings.

A noticeable gender imbalance can be observed, largely due to the conditions under which the ESC corpus has been developed. The project was initiated at the Faculty of Philology of HSE University, where the majority of students are women. The university setting accounts for the strong predominance of recordings from student informants aged 18–24 (85.45%). The average age of participants is 21.7 years.

For the vast majority of informants, the current place of residence is Saint Petersburg and the

Leningrad region (89.09%), where the recordings are collected. At the same time, geographic diversity is ensured by the fact that more than half of the informants (54.55%) indicated other regions as their place of birth. In addition, about 5% of participants are citizens of other countries (Kazakhstan, Belarus, Kyrgyzstan, Moldova, Iran, and Uzbekistan). Russian is the native language for almost all informants (97.58%). The corpus also includes bilingual student informants whose native languages are Belarusian, Moldovan, and Persian.

In addition to using professional audio recorders, participants were also allowed to use their smartphones as recording devices. This option was chosen by 32 informants (19.39%), including 8 who made recordings both with their own devices and with the recorders provided for all informants. Table 2 presents the distribution of recording devices.

Recording Method	Informants (N)	Average Recording Duration
Audio recorder	133	10:37:25
Smartphone	24	27:35:58
Both devices	8	16:27:13
Total	165	12:15:33

Table 2. Recording methods used by informants

On average, about 12 hours[1] of recordings were obtained per informant. The use of a smartphone as a recording device tends to increase the total duration of collected audio: up to 27.5 hours when recording only on a phone and up to 16.5 hours when combining both recording options. This difference in average values is mainly explained by the limited time for which audio recorders are provided to participants.

Informants may take part in the recording more than once. This option was used by 8 informants (half of whom used both types of devices). These participants account for 18.81% of the total recording time currently available in the ESC corpus. The largest amount of data obtained from a single informant is 134 hours of recordings, which corresponds to approximately 11 average "speech days"[2].

The second stage after collecting the audio data is its processing. At this stage, long audio files, with an average duration of about four hours, are reviewed by experts and manually segmented into macro episodes, which are "large segments unified by the place of communication, its conditions, and participants" (Sherstinova 2015) typically lasting 15–40 minutes. These macro episodes become the main unit of storage and description in the corpus.

During segmentation, extended fragments that do not contain speech are removed. In addition, since the primary focus is on "live" unprepared speech, fragments containing "secondary" (recorded or mediated) speech — such as background radio or film audio, or public transport announcements — may also be excluded if they are not involved in the ongoing interaction, that is, if they do not elicit any response or commentary from the informant and do not influence his/her speech behavior.

Thus, a typical 12-hour "speech day" of an informant, originally consisting of three four-hour audio files, is represented in the corpus as a set of *n* episodes, each accompanied by a brief description of the recorded situation. This description includes: (1) the type of communication; (2) the social roles and relationships of the speakers; (3) the location; and (4) a brief outline of each episode.

The episodes are then automatically transcribed using a pipeline that includes Whisper v3 for automatic speech recognition (ASR) and *pyannote* for diarization, i.e., the identification and separation of different speakers' utterances. The use of automated speech processing tools significantly speeds up the workflow. However, the quality of automatic transcription for recordings made in natural (often quite noisy) conditions remains imperfect, so the ASR output is subsequently reviewed and corrected by experts, being the most labor-intensive stage in corpus construction.

Before being published on the corpus website, the transcripts undergo an additional stage of data anonymization: all personal and sensitive information is replaced and encoded.

In this paper, we focus in particular on several aspects of corpus construction, including the use

[1] This figure reflects the total duration of audio recordings obtained from informants. Naturally, the amount of "actual speech" is significantly smaller and varies across informants.

[2] A speech day is defined as a continuous recording lasting between 8 and 16 hours that captures all of the informant's speech activity over the course of a day.

of speech recognition systems for generating initial transcripts, issues related to diarization, as well as the need for personal data anonymization and new approaches to morphological annotation.

3 Speech Diarization and ASR

To evaluate the quality of speech recognition for recordings made in real field settings, characterized by high levels of noise and overlapping speech, a series of experiments on speech diarization and ASR for spontaneous spoken language was conducted. These experiments assessed the performance of automatic diarization tools, i.e., methods for identifying and separating speakers in a speech stream. The selected models and associated tools were subsequently integrated into the overall audio data processing pipeline of the corpus. The analysis involved comparing the recognized text, automatically generated timestamps, and speaker labels with corpus transcripts produced manually by experts.

Since the ESC corpus does not yet contain a sufficiently large amount of expert-annotated speech transcripts, it was decided to conduct the initial experiments on data from the ORD corpus, which serves as a prototype and for which a substantial amount of expert-annotated transcripts is available. The reference dataset included 195 episodes of everyday speech recordings collected using the "speech day" methodology.

The expert annotation of the reference dataset was organized as a table in which each utterance is associated with a unique speaker identifier, a brief description of the episode, the name of the audio file, the start time, and the duration of the utterance. The transcripts include markers of syntagmatic and phrasal segmentation, as well as indications of pauses and paralinguistic phenomena (Bogdanova-Beglarian et al. 2016). Overlapping speech is represented within a single utterance, with multiple speakers marked using the symbols "@" or "#" (Asinovsky et al. 2009).

In contrast, the automatically generated data include speaker labels, along with start and end timestamps for each segment. Differences between the expert and automatic annotations concern speaker identification, timestamp format, and the representation of overlapping speech: in diarization output tables, speaker identifiers are reused within each audio file, timestamps indicate the beginning and end of each segment, and overlapping speech is represented as separate utterances with overlapping time intervals.

For evaluation purposes, the annotation formats were unified in order to eliminate differences that do not affect semantic content but may distort the metrics when comparing hypotheses and references. Thus, all punctuation was removed, all words were converted to lowercase, the letter "ё" was replaced with "е", and numerals (both in digits and written out) were replaced with the marker "NUM". At the same time, during the conversion of speaker labels, there was a risk of errors that could lead to incorrect calculation of evaluation metrics. Overlapping speech in the expert annotation was not transformed, as it is not possible in such cases to extract separate timestamps for individual speakers. In addition, occasional spelling errors are present in the reference annotation. These risks were minimized through additional testing of the processing algorithms and manual verification of aligned word pairs from the expert and automatic annotations.

The audio data processing pipeline for evaluation consisted of four stages: 1) normalization, 2) diarization and segmentation of audio into fragments, 3) speech recognition, and 4) metric calculation.

First, the audio files were converted to a unified format and resampled to the frequency required for ASR (16 kHz), and their volume levels were normalized. Next, diarization was performed to identify speakers and segment the audio accordingly. The resulting segments were saved as separate files and passed to the speech recognition model. Finally, the obtained transcripts were used to compute evaluation metrics for both diarization and recognition quality.

Standard metrics for evaluating diarization and speech recognition quality were used (Karpov, Kipyatkova 2012): DER (Diarization Error Rate), DWER (Diarization Word Error Rate), SAWER (Speech Attributed Word Error Rate), and WER (Word Error Rate) (Ali and Renals 2018; Munteanu et al. 2006). The WER, DWER, and SAWER metrics are based on counting substitutions, deletions, and insertions of words, whereas DER evaluates diarization quality at the level of time segments.

For speech recognition, the Whisper Large v3 model was used. Its performance and comparison with acoustic model developed by NTR in 2023

(NTR) were previously reported in (Sherstinova, Mikhaylovskiy, Kolpashchikova 2024). For diarization, we used *pyannote* (Bredin et al. 2020), an open-source framework whose models achieve DER values of 34.4 and 24 on the DIHARD and ETAPE datasets, respectively.

The results of DER calculations for the reference and test datasets show considerable variation in diarization quality. The minimum DER value was 10.3, while the maximum values approach 100. At the same time, for 10% of the recordings, the diarization error does not exceed 27.9. The mean DER across the dataset is also relatively high (68.3), with a median of 52.6. As expected, the best and most stable results are obtained for recordings with simpler conditions: minimal background noise, no more than one or two speakers, and clear, sufficiently loud speech.

The mean values of WER, DWER, and SAWER range from 46.1 to 64, with median values between 46.2 and 62.9. For less than 10% of the recordings, WER is 27.4, DWER is 24.0, and SAWER is 30.6. The 90th percentile reaches 62.6 for WER, 81.1 for DWER, and 96.8 for SAWER, with maximum values approaching 100. Although the minimum values (ranging from 8.4 to 13.3) indicate that some high-quality recognition results are achieved, the system — similarly to diarization — shows substantial variability in performance.

At present, 869 macro-episodes of the ESC corpus have been processed using diarization and ASR models, and their expert review is ongoing.

4 Anonymization of Personal Data in Transcribed Speech

Since recordings of everyday speech typically contain a large amount of personal information about informants and their interlocutors, the next important stage in the development of the ESC corpus involved establishing principles for anonymizing textual data prior to their online publication, in accordance with international data protection standards and Russian legislation. As in studies devoted to the ORD corpus (Sherstinova 2019), anonymization of ESC transcripts addresses two main tasks: removing or masking personal information that can be used to identify a speaker, and encoding obscene language.

For the ESC corpus, seven categories of data subject to anonymization were identified, and abbreviations were introduced for each of them:

- PROPN — given names,
- SURN — surnames,
- PATR — patronymics,
- ORG — organization names,
- STD — names of educational institutions,
- PLACE — toponyms,
- OBSC — obscene language.

Additionally, a separate category, NON, was introduced to mark elements subject to anonymization that appeared in automatically generated transcripts but were later identified as recognition errors or model hallucinations (Holtzman et al. 2020) and were therefore removed from the final edited version of the transcript.

The anonymization procedure was carried out using three different approaches, depending on the category of the word or phrase to which it was applied:

4.1 Proper Names

The anonymization of proper names, including surnames and patronymics, requires special attention, as names and surnames are not only the most frequent but also the most sensitive type of personal information that can be used to identify speakers. When anonymizing real given names, surnames, and patronymics, substitutes were selected so that the rhythmic structure of the utterance remained unchanged (i.e., preserving the number of syllables and the position of stress within the word). Each anonymized element was marked with the symbol %:

Marina Leonidovna → *Karina*% *Eduardovna*%
Vladimir Il'ich → *Arkadiy*% *Kuz'mich*%
Gordeev → *Avdeev*%.

For each episode, a separate list was created containing anonymized given names, surnames, and patronymics, along with their original counterparts. This was necessary to preserve the recognizability of individuals mentioned in the conversation: for example, if a person named *Marina* is mentioned in an episode and anonymized as *Karina*%, the same substitute is used consistently throughout that episode. Ideally, this approach should also be extended across all episodes within a single "speech day."

4.2 Names of Organizations, Places, and Educational Institutions

Names of organizations, locations, and educational institutions were replaced with a general category label enclosed in angle brackets:

Vy vse iz Griboedova[3]? → *Vy vse iz* <*university*>?

In addition to the categories <*university*> and <*organization*>, other frequent replacements include <*street*>, <*address*>, and <*toponym*>, the latter covering various types of geographical entities, as well as <*group*> or <*name*> (e.g., student associations or film clubs).

4.3 Obscene Language

In words belonging to obscene language, the entire root or part of the root was replaced with asterisks (*), corresponding to the number of letters, for example: *bl*t'*, *s**a*, etc. This approach masks offensive language while preserving the possibility of recognizing these words.

4.4 Statistics of Anonymization Results

At present, 4,304 utterances from 38 episodes have been anonymized and verified. Among them, 364 unique utterances (8.6%) contain at least one sensitive element. This means that sensitive data occur in approximately every 11–12 utterances. In total, 619 units were anonymized. Table 3 presents the absolute frequencies of these units in the corpus, along with their percentage relative to the total number of such units (%).

Category	Frequency	%	ipm
PROPN	335	54.12	3938
OBSC	154	24.88	1810
PLACE	50	8.08	588
PATR	29	4.69	341
SURN	25	4.04	294
STD	15	2.42	176
ORG	6	0.97	71
NON	5	0.81	–

Table 3: Distribution of lexical items subject to anonymization.

Thus, the vast majority of elements requiring anonymization belong to the category of proper names. Moreover, if given names, surnames, and patronymics are considered as a single category, their combined share reaches 63%. It should also be noted that elements related to obscene language account for almost 25%, making this the second most frequent category. This is followed by the category of toponyms, which accounts for about 8% of all anonymized elements.

These data were obtained from a sample of 85,071 tokens, which allows us to estimate the *ipm* (*instances per million*) of these elements in the text and thereby assess the workload involved in expert verification of the anonymization process. The corresponding values are presented in the right column of Table 3.

Further work in this direction includes developing approaches to automate the anonymization and masking of sensitive information with minimal involvement of experts. In particular, these procedures should allow for the automatic generation of substitute given names, surnames, and patronymics that preserve the syllabic structure of the original utterance.

5 Corpus Website Functionality

The ESC corpus website is currently operating in test mode at https://esc-corpus.ru/. To develop and test its functionality, a set of 38 communicative macro-episodes has been selected, containing 12.85 hours of speech data produced by 59 speakers. This collection reflects the diversity of everyday life of students in Saint Petersburg, including personal, professional, and social interaction. The corresponding transcripts contain 59,162 word tokens across 4,601 utterances (Sherstinova et al. 2025).

At present, corpus search is implemented at two levels: (1) the macro-episode level, which allows filtering speech data by communication conditions, participants' social characteristics, and type of interaction, and (2) the transcript level. The macro-episode level makes it possible to retrieve conversations based on these parameters and to view their full transcripts from beginning to end (rather than only immediate context windows, as is common in corpus linguistics).

[3] *Griboedova* is an informal name for one of the buildings of HSE University in Saint Petersburg.

Standard search is currently implemented by substring and keyword. In the near future, search by lemmas and morphological features is planned. The audio recordings themselves are not publicly available, since spoken data contain a large amount of personal information, including biometric data.

6 Morphological annotation of spoken speech

Morphological annotation is an essential component of any linguistic corpus, as it enables the tagging of parts of speech and grammatical features. However, when working with spontaneous spoken speech, the use of parsers designed for written texts often leads to inaccurate results. Models trained on written data are oriented toward normative language use, whereas spoken speech widely includes discourse markers and functionally reinterpreted forms that do not conform to these expectations, which leads to reduced accuracy in morphological analysis. The situation is further complicated by interruptions (incomplete words), hesitations, and other features of spontaneous speech.

Existing parsers (e.g., *MyStem*, *pymorphy*, *Stanza,* etc.) have been trained primarily on written, standardized texts, such as literature, blogs, news, and official documents, which further reinforces the mismatch described above. As a result, the distinctive features of spoken speech are underrepresented in the training data. Consequently, when analyzing spoken material with standard morphological parsers, the error rate increases: word boundaries may be identified incorrectly, and errors occur in part-of-speech tagging and grammatical feature assignment.

Therefore, one of the key practical tasks addressed using ESC data is the adaptation of existing automatic morphological annotation tools to the specifics of spoken speech. The ESC corpus, particularly its transcripts of everyday communication, provides a solid basis for developing such solutions. In particular, work is underway to adapt automatic morphological annotation for spoken data based on a manually annotated subcorpus.

A comparative evaluation of several widely used parsers for Russian (*MyStem*, *pymorphy3*, *Natasha*, *spaCy*, *Stanza*) was conducted on the corpus material. To ensure comparability, a unified tagging scheme was adopted, after which the automatically assigned tags were compared with a manually annotated gold standard. The evaluation was carried out using standard metrics, including accuracy, precision, recall, and F1-score, as well as a qualitative analysis of typical errors.

The most stable results were demonstrated by *MyStem* and *Stanza*; however, both systems show difficulties in tagging interjections, particles, and discourse markers, which play an important pragmatic role in conversational speech. Overall, the results confirm that models trained on written corpora tend to interpret spoken forms in line with the norms of standardized language, which leads to formally correct but functionally and pragmatically inaccurate analyses. This highlights the need to develop a morphological parser for Russian specifically designed for processing spontaneous spoken speech.

7 Conclusions

This paper presents the current state of the ESC corpus of everyday student speech and describes key practical aspects of its development, including the use of diarization models and speech recognition systems, which make it possible to speed up the addition of new speech transcripts to the corpus. Issues of personal data anonymization are also addressed. The ESC project is considered a long-term initiative, and the collection of audio data is planned to continue in the coming years.

Since the creation of the corpus is a multi-stage and labor-intensive process, its expansion proceeds unevenly across different stages. The fastest stage is the collection of new audio recordings, as it requires the least effort. In contrast, segmentation into macro-episodes and their annotation, which are necessary for preparing audio files for automatic transcription, are currently performed manually, which significantly slows down the growth of the final transcripts. Another bottleneck is the expert correction of automatically generated speech transcripts. Because both of these stages require expert involvement, the growth of text data lags far behind the accumulation of new recordings.

As shown in Section 6, a manually annotated subcorpus is used to adapt automatic morphological annotation to spoken speech. The data obtained make it possible to test alternative analysis strategies and, in the future, to develop a tool specifically designed for spontaneous discourse. This will make it possible to integrate

the corresponding modules into the corpus and its demonstration website.

In addition to textual transcripts of everyday speech, the original audio recordings represent a particularly valuable component of the corpus, as they make it possible to "hear" authentic student communication in Russian. These materials can be highly useful both for didactic purposes in teaching Russian as a foreign language and for speech technologies — for training speech systems and voice assistants to communicate in a natural way, taking into account the lexical, grammatical, phonetic-prosodic, and pragmatic features of real speech. However, since the speech signal constitutes a biometric characteristic that can often be used to identify a speaker, the publication of audio recordings in open access is only possible under conditions of substantial acoustic modification of the signal, which represents a specific form of anonymization.

The ESC corpus can also serve as a testbed for evaluating NLP tools for Russian. The resulting resource is expected to be valuable both for fundamental research on contemporary Russian spoken discourse and youth communicative practices, and for applied tasks such as the development and training of automatic speech recognition systems, language models, and other natural language processing technologies.

Acknowledgments

This work is an output of a research project HSE-BR-2025-032 implemented as part of the Basic Research Project at HSE University. The implementation of the project has been made possible thanks to the active involvement of HSE University philology students, who contribute to the corpus on a volunteer basis at all stages of its development. We also express our gratitude to all anonymous informants and their interlocutors who provided recordings of their everyday communication for the ESC corpus.

References

Tatiana Sherstinova and Irina Petrova. 2024. ESC Corpus of Spoken Russian: Everyday Student Conversations Captured through Continuous Speech Recording in Natural Communicative Environments. In Alexey Karpov and Vlado Delić, editors, Speech and Computer, 26th International Conference, SPECOM 2024, Belgrade, Serbia, November 25–28, 2024. Lecture Notes in Artificial Intelligence, volume 15299, pages 151–162.

Tatiana Sherstinova, Aleksey Melnik, Irina Petrova, Karina Azarevich, Valeria Melkozerova, and Sofia Chepovetskaya. 2025. "Okie dokie, here's the no-cap truth!": Everyday Russian Youth Speech in Corpus Representation (Structure and Application of the ESC Sound Corpus). Komp'juternaja Lingvistika i Intellektual'nye Tehnologii / Annual International Conference on Computational Linguistics and Intellectual Technologies, Dialogue 2025, volume 23, pages 331–344.

Tatiana Sherstinova, Nikolay Mikhaylovskiy, Evgeniya Kolpashchikova, and V. Kruglikova. 2024. Bridging Gaps in Russian Language Processing: AI and Everyday Conversations. In 35th Conference of Open Innovations Association (FRUCT), Tampere, Finland, April 24–26, 2024, pages 253–258.

Tatiana Sherstinova, Rostislav Kolobov, and Nikolay Mikhaylovskiy. 2023. Everyday Conversations: a Comparative Study of Expert Transcriptions and ASR Outputs at a Lexical Level. In Proceedings of SPECOM 2023 / LNCS 14338/14339. Springer, pages 43–56.

Janet S. Shibamoto. 1987. Japanese Sociolinguistics. Annual Review of Anthropology, volume 16, pages 261–278.

Takeshi Shibata. 1983. Study of Language Behavior over 24 Hours [Issledovanie yazykovogo sushchestvovaniya v techenie 24 chasov]. Translated by V. M. Alpatov. In V. M. Alpatov, editor, Linguistics in Japan [Yazykoznanie v Yaponii]. Moscow: Raduga, pages 134–141.

Nick Campbell. 2004. Speech & Expression; the Value of a Longitudinal Corpus. In M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, and R. Silva, editors, Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04), Lisbon, pages 183–186.

Reference Guide for the British National Corpus. http://www.natcorp.ox.ac.uk/docs/URG.xml. Last accessed: 10/01/2025.

Ari Holtzman et al. 2020. The Curious Case of Neural Text Degeneration. In Proceedings of

the International Conference on Learning Representations.

Natalia Bogdanova-Beglarian, Tatiana Sherstinova, Olga Blinova, Olga Ermolova, Ekaterina Baeva, Gregory Martynenko, and Anastasia Ryko. 2016. Sociolinguistic Extension of the ORD Corpus of Russian Everyday Speech. In A. Ronzhin et al., editors, Speech and Computer, Lecture Notes in Artificial Intelligence, volume 9811. Springer.

Aleksey A. Karpov and Irina S. Kipyatkova. 2012. Metodologiya otsenivaniya raboty sistem avtomaticheskogo raspoznavaniya rechi. Izvestiya vysshikh uchebnykh zavedenii. Priborostroenie, volume 55, number 11, pages 38–43.

Matthew Maciejewski, Dominik Klement, Ruizhe Huang, Matthew Wiesner, and Sanjeev Khudanpur. 2024. Evaluating the Santa Barbara Corpus: Challenges of the Breadth of Conversational Spoken Language. In Interspeech 2024, pages 2155–2159.

Whisper Model Homepage. https://openai.com/index/whisper/.

NTR Homepage. https://ntr.ai. Last accessed: 2026/04/10.

John W. Du Bois et al. 2000–2005. Santa Barbara Corpus of Spoken American English, parts 1–4. Philadelphia: Linguistic Data Consortium.

The Corpus of Spontaneous Japanese. https://clrd.ninjal.ac.jp/csj/misc/preliminary/index_e.html. Last accessed: 28/03/2025.

Awais Ali and Steve Renals. 2018. Word Error Rate Estimation for Speech Recognition: e-WER. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, volume 2, pages 20–24.

Cosmin Munteanu, Graeme Penn, Ronald Baecker, Elaine Toms, and D. James. 2006. Measuring the Acceptable Word Error Rate of Machine-Generated Webcast Transcripts. In Proceedings of the 9th International Conference on Spoken Language Processing.

Alexander Asinovsky, Natalia Bogdanova, Marina Rusakova, Svetlana Stepanova, Anastasia Ryko, and Tatiana Sherstinova. 2009. The ORD Speech Corpus of Russian Everyday Communication “One Speaker’s Day”: Creation Principles and Annotation. In Text, Speech and Dialogue, Lecture Notes in Computer Science, volume 5729.

Tatiana Sherstinova. 2015. Macro Episodes of Russian Everyday Oral Communication: towards Pragmatic Annotation of the ORD Speech Corpus. In A. Ronzhin et al., editors, Speech and Computer, Lecture Notes in Artificial Intelligence, volume 9319, pages 268–276.

One Speech Day Corpus Online. https://ord.spbu.ru/.

CLARIN Spoken Corpora. https://www.clarin.eu/resource-families/spoken-corpora.

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. 2020. pyannote.audio: Neural Building Blocks for Speaker Diarization. In Proceedings of ICASSP 2020 — IEEE International Conference on Acoustics, Speech and Signal Processing, pages 7124–7128.

Multi-word Expressions in the Electronic Dictionary *Tēzaurs*: Integrating Data from Different Sources

Gunta Nešpore-Bērzkalne, Mikus Grasmanis, Ilze Lokmane
Institute of Mathematics and Computer Science, University of Latvia
Raiņa bulvāris 29, Riga, Latvia
gunta.nespore@lumii.lv, mikus.grasmanis@lumii.lv, ilze.lokmane@lu.lv

Abstract

In addition to single word entries, the electronic Latvian dictionary *Tēzaurs* contains more than 76,000 multi-word expressions (MWEs), which are represented in separate entries with the same structure as other entries. A group of MWEs (ca. 5000 entries) was effectively left out of some processes as they could not be automatically interpreted: MWE variants shown in ambiguous parentheses, functional words indicating metadata not separated from lexeme text, etc. These entries mostly originated from printed dictionaries, which has led to repetitive and inconsistent data and required significant manual revision. Efforts are underway to integrate these MWEs into the *Tēzaurs* data network; the problematic entries have been restructured and the principles for data consolidation have been defined.

Keywords: multi-word expressions, lexical variation, condensed headwords, dictionary data integration.

1 Introduction

Tēzaurs[1] is the largest Latvian electronic explanatory dictionary with more than 410,000 entries, which was started as a compilation of nearly 300 pre-existing dictionaries and other sources but has since developed into a multifunctional lexical resource consisting of granular interlinked entities enriched with various kinds of metadata (Grasmanis et al., 2023). Besides entries for single words, *Tēzaurs* also contains ca. 76,000 entries for multi-word expressions (MWEs; dictionary entries that contain more than one orthographic word (Bauer, 2021: 5)). Thus, it is currently the most comprehensive resource of MWEs in Latvian lexicography.

Tēzaurs has evolved over many years as a compilation of numerous valuable but sometimes heterogeneous sources (both printed dictionaries and collections of terms, place names, taxonomies, idioms, etc.) with incomplete and inconsistent guidelines for consolidation and standardization or without any at all. Initially, the inclusion of each new source was purely additive, extending *Tēzaurs* with an additional set of entries mostly only loosely connected with entries coming from other sources. During this time, the *Tēzaurs* platform has also been improved providing new means for data normalization, standardization, and enrichment with additional or meta information, thus enabling a consolidation of information, a portion of which is presented in this paper.

In *Tēzaurs*, most MWEs have been assigned category tags according to their function and level of figurativeness (Rituma et al., 2024): see Table 1. This paper focuses on idiomatic MWEs, the principal sources of which are The Dictionary of Standard Latvian (Ceplītis, 1971–1996) and The Dictionary of Latvian Phraseology (Laua et al., 1996), the latter containing idiomatic MWEs only.

Although the format of MWEs derived from different sources was initially highly heterogeneous, most of the data was unified and integrated into *Tēzaurs* by automatic or semi-automatic methods. However, a group of MWEs (ca. 5000 entries; most of them – phraseological units) required significant manual revision. These entries contained data from printed dictionaries that could not be automatically interpreted: some headwords contained several MWE variants, and a part of them was metadata, not a lexeme (headwords representing multiple MWEs are hereinafter referred to as condensed headwords). The information was repetitive and inconsistent, and therefore difficult to process, as the same MWE in different entries may be represented differently. As a result, a headword could

[1] Available at https://tezaurs.lv or as data from https://repository.clarin.lv/repository/xmlui/handle/20.500.12574/160

appear the following way: ***dūc*** *(arī* ***rūc****)* ***kā*** *(****bišu****)* ***stropā*** *(arī* ***strops****,* ***pa stropu****), retāk* ***dūc*** *(arī* ***rūc****)* ***kā*** *(****bišu****)* ***spietā*** *(arī* ***spiets****)* (lit. '**buzz** (also **hum**) **as in a** (**bee**) **hive** (also **a hive**, **around a hive**), less often **buzz** (also **hum**) **as in a** (**bee**) **swarm** (also **as a swarm**') 'the noise made by many people'.

Currently, efforts are underway to fully integrate these MWEs into *Tēzaurs*, i. e., to rewrite condensed headwords in a consistent and machine-interpretable form by separating each MWE variant into its own database record, as well as deduplicate and consolidate the information about them. This process requires substantial manual effort but constitutes a necessary step to ensure that all MWEs are fully accessible for searching, processing, and research. Moreover, the manual review of the data helps to pinpoint and resolve additional issues related to MWE description and representation.

The next chapters outline the structure of MWE entries in *Tēzaurs* (2), the expansion of the initial headwords (3), and how these entries are consolidated and integrated into *Tēzaurs* (4). The last chapter gives conclusions and an insight into the future work for MWE representation.

2 MWE Entry Structure in *Tēzaurs*

The MWEs in *Tēzaurs* have separate entries that can contain the same elements as single-word entries: MWE lexical variants; tags of style and usage context; sense definitions; a section for adding usage examples from the Latvian National Corpora Collection (Saulīte et al., 2022). Additionally, MWE entries contain MWE category tags (see Table 1). As structurally independent units, MWE entries can be linked with other entries in *Tēzaurs*, including entries of their component-words (this is how MWEs are displayed in single-word entries: the entry of the respective word shows all MWEs that the word is linked with), as well as entries involved in the Latvian WordNet showing semantic relations between MWEs and other lexical units (Klints et al., 2024). They can also be linked with the Princeton English WordNet (Fellbaum, 1998). However, MWEs in *Tēzaurs* lack explicit information on their morphological and syntactic structure. An experiment for creating semiautomatic syntactic and structural annotation has been commenced, but the results are not publicly available yet.

Name of Category	No of MWEs
complex terms	23,777
taxonomic group names	18,200
multi-word proper nouns	16,993
phraseological units	8,953
collocations	3,065
foreign language MWEs	1,151
unclassified MWEs	4,348
Total number of MWEs	**76,467**

Table 1: The MWEs categories in *Tēzaurs* and the number of MWEs in them.

3 The Expansion of Condensed MWE Headwords

A list of condensed headwords was extracted, namely those containing (1) parentheses and (2) specific functional words. This list comprised ca. 5,000 entries. The following paragraphs outline the use of parentheses and functional words in the condensed headwords.

Ambiguous parentheses: Parentheses may have several functions that are difficult to interpret automatically: (1) the text in parentheses is optional, e.g., *dūc kā* ***(bišu)*** *stropā* (lit. 'buzzes as in a (**bee**) hive'); (2) the text in parentheses (without the functional word) or one of the words listed in parentheses may replace the text preceding the parentheses, e.g., *dūc (arī* ***rūc****) kā stropā* (lit. 'buzzes (also **hum**) as in a hive'); (3) the pronoun in parentheses is not a part of the lexeme but indicates the free arguments or modifiers of the MWE (as described in (Odijk and Kroon, 2024)), e.g., *izliet* ***(kāda)*** *asinis* (lit. 'to shed (**someone's**) blood') 'to kill'; *atvērt durvis* ***(kādam, kaut kam)*** (lit. 'to open the door (for **someone, something**') 'to create favorable conditions'; any word from a certain lexical group can take the place of the pronoun.

Functional words: Include *arī* 'also', *biežāk* 'more commonly', *retāk* 'less commonly'. E.g., *ņemt acis pirkstos* ***(retāk*** *rokā),* ***arī*** *turēt acis pirkstos* (lit. 'to take eyes in one's fingers (**less often** in hand), **also** to hold eyes in one's fingers') 'to look closely'.

The text of a lexeme is not always clearly discernible from functional words and it is not always possible to determine which portion of the MWE the functional word or the text in parentheses pertains to.

To transform condensed headwords into a set of single-lemma lexemes (variants of MWE), a format for describing lexeme variants needed to be selected. The lexeme variants already present in *Tēzaurs* (in both single-word and MWE entries) are represented as lists: each variant is stored in separate database records. However, previously the number of variants per lexeme has typically been small (up to three), whereas the condensed headwords can give a substantially larger number of variants. Presenting these as simple lists may not be user-friendly. An alternative approach was also considered, where MWE variants were represented by other, unambiguously interpretable condensed headwords (formulae). However, implementing such a solution would require substantial changes to the entire *Tēzaurs* system to apply the new format to the existing entries. Therefore, it was decided to start with explicitly expanding the condensed headwords into a list of lexeme variants for each entry. This approach also enables a more thorough analysis of existing data and better decisions regarding future MWE representation.

The condensed headwords were expanded as follows:

- Each MWE variant was moved to a separate record, with most parentheses removed; the only remaining (now unambiguous) parentheses indicate the free arguments or modifiers of the MWE;
- Functional words were converted into tags indicating frequency of use (except *arī* 'also', which gives no additional information).

Thus, the previously mentioned example ***ņemt acis pirkstos*** *(retāk* ***rokā****), arī* ***turēt acis pirkstos*** provides three variants: (1) *ņemt acis pirkstos*, (2) *ņemt acis rokā*, and (3) *turēt acis pirkstos*; furthermore, a tag is added to the second variant to indicate its less frequent use.

The structure of the other previously mentioned example ***dūc*** *(arī* ***rūc****)* ***kā*** *(****bišu****)* ***stropā*** *(arī* ***strops****,* ***pa stropu****), retāk* ***dūc*** *(arī* ***rūc****)* ***kā*** *(****bišu****)* ***spietā*** *(arī* ***spiets****)* is more complex; it is depicted in Figure 1. The formula starts with a black arrow on the left, and ends with another arrow on the right. Each possible path in this formula describes one variant of the MWE. The convergence of all possible paths provides the set of possible MWE variants. The *arī* 'also' clauses are depicted as branches, the *retāk* 'less often' clause is coloured in gray. The optional clause is depicted as a path skipping the optional word.

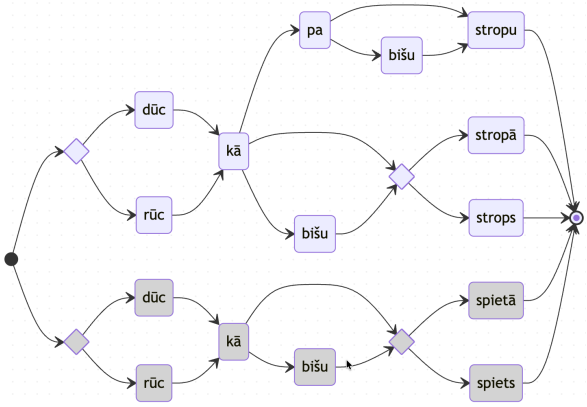


Figure 1: Structure of the MWE *dūc kā bišu stropā* (lit. 'buzzes as in a bee hive) 'the noise made by many people' with variants

4 Integration of the MWE Entries into the *Tēzaurs* Data

To integrate an MWE entry into the *Tēzaurs* data, an essential step is to ensure that the entry is linked to all its component entries. This enables dictionary users to navigate from the MWE entry to the corresponding single-word entries and back, as well as to find the MWE in the entries of its components.

After expanding the condensed headwords into several simple lexemes, it becomes apparent that many lexemes repeatedly occur in multiple entries. Our goal is to describe each lexeme only once; thus, we developed a consolidation methodology and process, as depicted in Figure 2. In brief, the process can be explained as *expand-deduplicate-combine-compose*. If, in addition to the lemma, all other parts of the lexemes also match, the consolidation occurs immediately. However, for many lexemes with the same lemma, deviations are noticeable in both the set of given senses and the additional information. At first glance, the deviations are partly significant and partly sporadic. In the following subsections, we describe which types of information deviate more frequently and how such cases were handled.

4.1 Different Sets of MWE Variants

Various entries that show the same lexeme may contain different sets of lexeme variants. As seen in Figure 2, the lexeme *sist nost* (lit. 'to beat down') is mentioned in all initial entries (data from different entries shown in different colours), unlike other

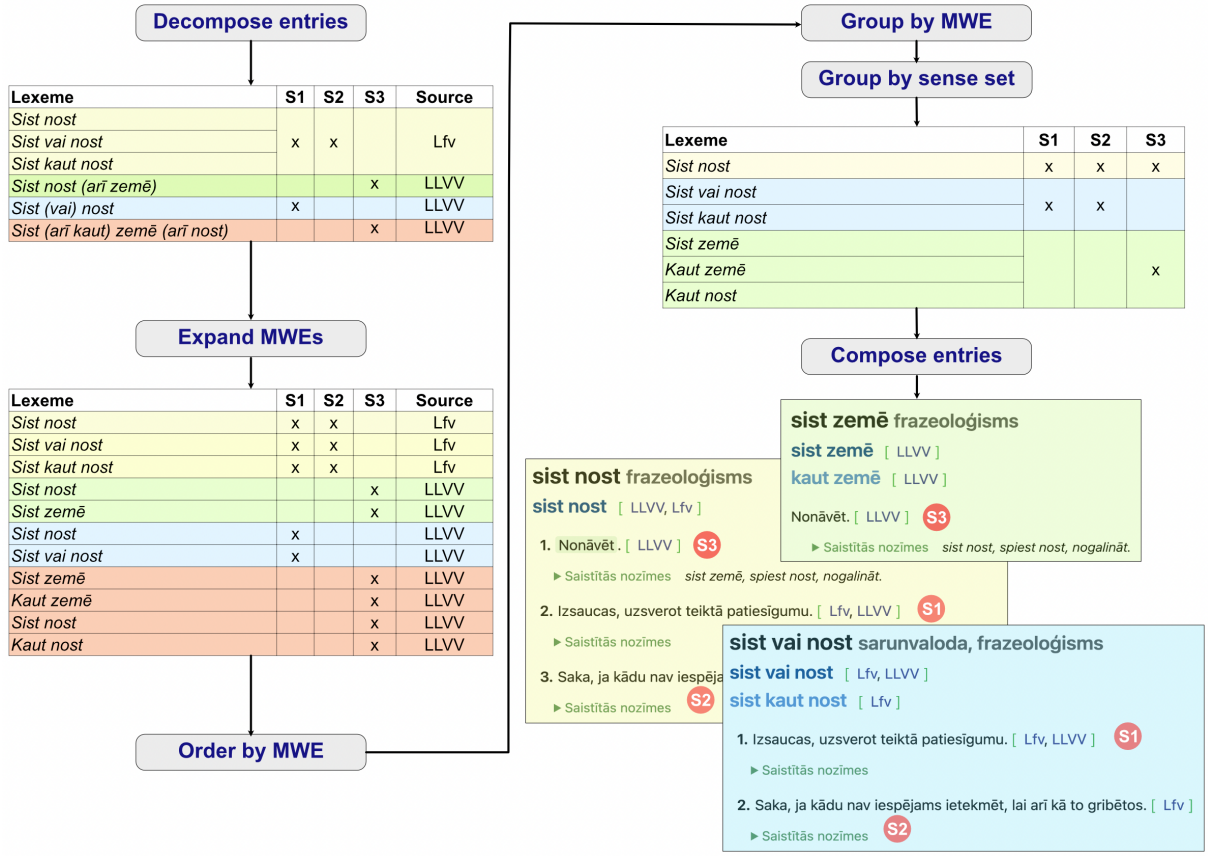


Figure 2: Processing of MWEs with overlapping data

variants. If these entries have an identical set of senses, in the final entry we include all lexeme variants with sources indicated for each of them. If the sets of senses differ between entries, the course of action described in the following section is taken.

4.2 Different Set of Senses

Entries from different sources may have varying levels of sense granularity and different sets of senses. This includes whether or not the literal meaning of an MWE is indicated (in the example in Figure 2, the literal sense is 'to kill' (S3), which is not given in the entry from (Laua et al., 1996)).

We group MWE variants into entries according to the set of senses (usually retaining the finest granularity) with the source indicated for each. Synonymous senses of different entries are linked with the synonymy link, which links the MWE with other synset elements of the corresponding sense, if there are any. As seen in Figure 2, after sorting MWE lexemes by sets of senses, the four initial entries were recompiled into three, where each MWE lexeme is only described once, providing all information corresponding to this variation.

4.3 How Many Entries?

Considering the previously described problems and solutions, a crucial question arises: whether MWEs with similar wording and identical meanings should be represented in a single entry or across multiple entries? What is the optimal solution between two extreme approaches: (1) considering each MWE with an even slightly different form to be independent and thus represented in a separate entry, or 2) grouping all MWEs that share the same senses in a single entry, where other metadata, e.g. register, could be specified separately for each variant? For example, one may consider the MWEs *kā putns kokā* (lit. 'like a bird in a tree'), and *kā putns zara galā* (lit. 'like a bird at the end of a branch'): they have the same syntactic structure and meaning 'to be somewhere for a very short time', and even express similar imagery; thus it can be questioned whether they should be in one entry or two. Furthermore, how many entries should there be for the MWEs *apsildīt kājas* (lit. 'to warm up the feet'), *apsildīt degunu* (lit. 'to warm up the nose'), and *ap-*

sildīt nagus (lit. 'to warm up the nails'), which all share the same structure and meaning, i.e., 'to be somewhere for a very short time', but use different imagery?

Ermakova et al. (2022) proposes representing MWE variants in one or two entries, or a specific entry group depending on the number and variety of the MWEs (Hub-node model of MWE entries), whereas (Gantar and Krek, 2022) distinguish MWE variants from independent MWEs primarily on the basis of their inventory of senses and number and sequence of components. In the PARSEME (Savary et al., 2026) annotation guidelines[2], an MWE is regarded as a sequence of lexicalised components; if one of the components is replaced, the resulting expression is treated as a different MWE (it should be noted, however, that the objective of PARSEME is the annotation of MWEs in corpora rather than their lexicographic description).

In connection with various lexical resources, the representation and encoding of MWE variants, as well as the best version of the lemma/canonical form have been widely discussed e.g., (Villavicencio et al., 2004; Odijk and Kroon, 2024; Skoumalová et al., 2024; Al-Haj et al., 2014; Vrbinc and Vrbinc, 2016). However, these studies rarely provide explicit criteria for distinguishing MWE variants from independent MWEs. Implicitly, it could be assumed that expressions falling outside the scope of described variation should be treated as independent MWEs.

It is clear that in the future, *Tēzaurs* will need more refined guidelines for distinguishing MWE variants from independent MWEs. Developing and applying such guidelines, however, constitutes a separate step of MWE processing, as they should be applied to the entire *Tēzaurs* data rather than only to the dataset discussed in this paper. At the present stage, our objective is to ensure that the newly integrated MWEs are incorporated consistently into the existing structure of *Tēzaurs*.

Therefore, whenever possible, we currently retain the division of MWEs into entries as inherited from the sources; however, decisions can not be postponed in cases where MWE division into entries differs across sources.

In these cases, we try to follow the Latvian lexicographic tradition, which is reflected by the principal MWE sources (Laua et al., 1996; Ceplītis, 1971–1996) as well. Regarding idiomatic MWEs, the common practice (Laua, 1992) in Latvian lexicography is to group MWEs with an identical meaning (set of senses), similar form and imagery into a single entry and view them as MWE variants. MWEs with identical or close meaning, similar form, but different imagery, on the other hand, are split into separate entries and viewed as synonyms.

In practice, the distinction between variants and synonyms is not clear-cut, as the closeness of imagery is difficult to assess. In borderline cases we maintain a more fine-grained entry division and create synonym links between the identical MWE senses. This enables the dictionary users to still access MWEs with identical meanings all at once – either in a single entry or through synonymy links.

5 Conclusions and Future Work

Extensive work has been carried out to integrate ca. 5,000 idiomatic MWEs into the dictionary and make them more accessible to users. Principles have been developed for consolidating data from diverse sources. Manual review of entries has improved the overall quality of the data.

Future work includes identifying more user-friendly methods of representing MWEs with numerous variants and refining the criteria for division of similar MWEs into entries. We also plan to introduce syntactic annotation for MWEs, with first experiments already underway.

Acknowledgments

This work was supported by State Research Programme Letonika – Fostering a Latvian and European Society project "Digital Resources and AI Technologies for the Sustainability of the Latvian Language (DigiLATE)" (VPP-IZM-LETONIKA-2025/1-0004) in synergy with the Latvian Council of Science project "Advancing Latvian computational lexical resources for natural language understanding and generation" (LZP2022/1-0443). We also thank our collegues Madara Stāde and Baiba Valkovska for valuable advice.

References

Hassan Al-Haj, Alon Itai, and Shuly Winter. 2014. Lexical representation of multiword expressions in morphologically-complex languages. *International Journal of Lexicography*, 27(2):130–170.

Laurie Bauer. 2021. *An Introduction to English Lexicology*. Edinburgh University Press, Edinburgh.

[2] https://parsemefr.lis-lab.fr/parseme-st-guidelines/2.0/?page=home

Laimdots Ceplītis, editor. 1971–1996. *Latviešu literārās valodas vārdnīca*, volume 1–8. Zinātne, Rīga.

Maria Ermakova, Alexander Geyken, Lothar Lemnitzer, and Bernhard Roll. 2022. Integration of Multi-Word Expressions into the Digital Dictionary of German Language (DWDS). Towards a Lexicographic Representation of Phraseological Variation. In *Dictionaries and Society. Proceedings of the XX EURALEX International Congress*, pages 851–860, Mannheim. IDS-Verlag.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech and Communication. Mit Press.

Polona Gantar and Simon Krek. 2022. Creating the Lexicon of Multi-Word Expressions for Slovene. Methodology and Structure. In *Dictionaries and Society. Proceedings of the XX EURALEX International Congress*, pages 549–562, Mannheim. IDS-Verlag.

Mikus Grasmanis, Pēteris Paikens, Lauma Pretkalniņa, Laura Rituma, Laine Strankale, Artūrs Znotiņš, and Normunds Grūzītis. 2023. Tēzaurs.lv – the experience of building a multifunctional lexical resource. In *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference*, pages 400–418. Lexical Computing CZ s.r.o.

Agute Klints, Mikus Grasmanis, Gunta Nešpore-Bērzkalne, Lauma Pretkalniņa, Madara Stāde, Normunds Grūzītis, Ilze Lokmane, Pēteris Paikens, Laura Rituma, and Andrejs Spektors. 2024. Tēzaurs as a digital multifunctional lexical resource. *Baltic Journal of Modern Computing*, 12(4):513–525.

Alīse Laua. 1992. *Latviešu valodas frazeoloģija*. Zvaigzne, Rīga.

Alīse Laua, Aija Ezeriņa, and Silvija Veinberga. 1996. *Latviešu frazeoloģijas vārdnīca*, volume 1–2. Avots, Rīga.

Jan Odijk and Martin Kroon. 2024. A canonical form for flexible multiword expressions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 91–101, Torino, Italia. ELRA and ICCL.

Laura Rituma, Gunta Nešpore-Bērzkalne, Agute Klints, Ilze Lokmane, Madara Stāde, and Pēteris Paikens. 2024. Classifying multi-word expressions in the latvian monolingual electronic dictionary tēzaurs.lv. In *6th International Conference Computational Linguistics in Bulgaria (CLIB)*, pages 113–118.

Baiba Saulīte, Roberts Darģis, Normunds Grūzītis, Ilze Auziņa, Kristīne Levāne-Petrova, Lauma Pretkalniņa, Laura Rituma, Pēteris Paikens, Artūrs Znotiņš, Lauma Strankale, Kristīne Pokratniece, Ilmārs Poikāns, Guntis Bārzdiņš, Inguna Skadiņa, Anda Baklāne, Valdis Saulespurēns, and Jānis Ziediņš. 2022. Latvian national corpora collection – korpuss.lv. In *13th Language Resources and Evaluation Conference (LREC)*, pages 5123–5129.

Agata Savary, Manon Scholivet, Carlos Ramisch, Takuya Nakamura, Eric Bilinski, Sara Stymne, Voula Giouli, Stella Markantonatou, Vasile Păiș, Maria Mitrofan, Louis Estève, Bruno Guillaume, Verginica Barbu Mititelu, Jaka Čibej, Roberto A. Díaz Hernández, Victoria Fendel, Polona Gantar, Olha Kanishcheva, Cvetana Krstev, Chaya Liebeskind, Irina Lobzhanidze, Aleksandra Marković, Gunta Nešpore-Bērzkalne, Adriana Pagano, Mehrnoush Shamsfard, Ranka Stanković, Vahide Tajalli, Carole Tiberius, and Aakanksha Padhye. 2026. Parseme 2.0 multilingual corpus of multiword expressions. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 4819–4834.

Hana Skoumalová, Marie Kopřivová, Vladimír Petkevič, Tomáš Jelínek, Alexandr Rosen, Pavel Vondřička, and Milena Hnátková. 2024. Lemur: A lexicon of czech multiword expressions. In Voula Giouli and Verginica Barbu Mititelu, editors, *Multiword Expressions in Lexical Resources: Linguistic, Lexicographic, and Computational Perspectives*, pages 1–37. Language Science Press, Berlin.

Aline Villavicencio, Ann Copestake, Benjamin Waldron, and Fabre Lambeau. 2004. Lexical encoding of MWEs. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, pages 80–87, Barcelona, Spain. Association for Computational Linguistics.

Alenka Vrbinc and Marjeta Vrbinc. 2016. Canonical forms of idioms in online dictionaries. *GEMA Online Journal of Language Studies*, 16:1–15.

Optimizing LLM-based Discourse Analysis in Low-Resource Languages: A Case Study of Bulgarian Euro Adoption Narratives

Dobromir Tsolyov
GATE Institute / Sofia, Bulgaria
dobromir.tsolyov@gate-ai.eu

Keith Kiely
GATE Institute / Sofia, Bulgaria
keith.kiely@gate-ai.eu

Georgi Goranov
GATE Institute / Sofia, Bulgaria
georgi.goranov@gate-ai.eu

Iglika Ivanova
GATE Institute / Sofia, Bulgaria
iglika.ivanova@gate-ai.eu

Stanimir Dimitrov
GATE Institute / Sofia, Bulgaria
stanimir.dimitrov@gate-ai.eu

Abstract

This study explores the "complexity ceiling" of LLMs in analyzing culturally specific Euro adoption narratives in Bulgarian. Navigating this high-subtext content requires pipelines that balance abstract reasoning with rigorous stability. Using the IWH and FABLE methodologies, which comprise a total of 11 dimensions scored $\{0, 1, 2\}$, we evaluated model performance on a dataset sampled for diversity via vector embeddings.

We operationalize a complexity ceiling as the highest-complexity configuration within a bounded search space that satisfies a predefined stability criterion. The ceiling was identified via "descent to feasibility": beginning from the highest complexity parameters within the predefined search space, and iteratively reducing internal and external parameters until the stability criterion was met. Stability was measured using the "80/80 rule," requiring at least 80% of samples to maintain modal consistency in 8 out of 10 independent iterations, across all dimensions. Results indicate that pipelines capable of both stability and higher-order reasoning are achievable through calibrated trade-offs between internal and external complexity. The leading configuration met the 80/80 criteria for 10 out of 11 dimensions, suggesting that stable multi-dimensional discourse analysis is achievable under carefully calibrated complexity conditions, even in low-resource settings.

Keywords: LLMs, low-resource languages, critical discourse analysis, political speech

1 Introduction

Automated discourse analysis of politically charged content in low-resource languages presents a distinctive class of challenges that general LLM benchmarks are ill-equipped to anticipate. Bulgarian political narratives surrounding the Euro adoption debate are dense with cultural subtext, historical analogy, and idiomatic framing that resist straightforward analytical decomposition. Recent evaluations in other less resourced European languages such as Latvian and Polish similarly report degraded performance and heightened vulnerability in multilingual LLMs, supporting the view that low-resource politico-cultural contexts remain a structurally hard regime for current models (Dargis et al., 2024). When an LLM is tasked with scoring such content across multiple abstract dimensions simultaneously, the interaction between linguistic complexity and inferential load creates conditions under which output consistency degrades — not uniformly across dimensions, but in ways that are specific to the combination of task structure, prompt design, and model configuration. Standard performance benchmarks, calibrated on cleaner, higher-resource tasks, offer no reliable basis for predicting where this degradation begins or how severe it will be (Gürses and Ceylan, 2026).

This unpredictability is the core problem this study addresses. Applying the IWH and FABLE frameworks (Kiely, 2025; Patel et al., 2025; Sehat et al., 2024) to Bulgarian Facebook posts requires a pipeline that is not merely capable of higher-order

reasoning, but demonstrably stable under repeated inference to be trusted as an analytical instrument. Without this stability, scores cannot be meaningfully interpreted as evidence of discourse properties, regardless of how sophisticated the underlying model is. Inversely, stability without sufficient reasoning capacity would lead to consistently inaccurate results. The research objective is therefore to identify high-performing pipeline configurations that jointly balance scoring stability and reasoning depth within politically and culturally complex content.

Unique contributions: We contribute (1) an operational definition and search procedure for an LLM complexity ceiling on high-subtext political discourse in a low-resource linguistic environment, (2) a stability benchmark (80/80) tailored to multi-dimensional discourse scoring, and (3) empirical evidence about how internal and external complexity interact in a low-resource, real-world setting.

2 Theoretical Framework

The theoretical framework for this study integrates geopolitical power cycle theory with multi-stage misinformation harm assessments to analyze automated discourse in low-resource linguistic environments. By combining the **Indicators of Weak Hegemony (IWH)** with the **FABLE** framework, this study evaluates the complexity ceiling of Large Language Models (LLMs) in navigating high-subtext sociopolitical narratives.

2.1 Literature Review: Social Science Approaches to Hegemony and Media Deception

Hegemonic power is commonly understood as cyclical, moving through phases of rise, consolidation, and decline under pressures such as military overreach, institutional erosion, and domestic polarization (Patel et al., 2025: 10). In a Gramscian tradition, hegemony refers not only to coercive capacity but to the ability of dominant groups to manufacture consent through culture, political institutions, and everyday social relations (Kiely, 2025: 2). Within this cycle, a weak hegemony is marked by the fragmentation of once-dominant narratives and the erosion of trust in democratic institutions, as competing projects struggle to redefine what counts as “common sense” (Kiely, 2025: 1). Recent work on LLMs’ political worldviews shows that model outputs on ideological questions can vary systematically across prompts, and that reliability and cross-prompt consistency are themselves important objects of study in political communication research (Ceron et al., 2024; Alali et al., 2026) The Indicators of Weak Hegemony (IWH) translate this abstract trajectory into observable dimensions: Fragmented Consensus captures discursive breakdowns of liberal or democratic consensus, Counter-Hegemonic Forces track the visibility of actors and narratives that challenge the established order, and Crisis of Authority reflects contestation of institutional legitimacy and moral-intellectual leadership (Patel et al., 2025: pp. 8–9; Kiely, 2025: p. 3). These and related indicators, such as increased reliance on coercion over consent, are encoded as separate analytical dimensions in our annotation scheme.

Resistance to hegemonic decline is mirrored in the information ecosystem, where media deception exploits cognitive heuristics—rules of thumb—to shape public perception with minimal argumentative substance (Olariu, 2022: 30). Research in political psychology and media studies identifies analytical thinking as a key dynamic shield against such manipulation (Olariu, 2022: 29). In this context, implicit claims function as assertions conveyed through rhetorical strategies or contextual gaps rather than explicit factual statements, steering interpretation without direct evidential support (Nair, 2025: 27). These mechanisms of deception and inference are central to the FABLE dimensions discussed below and motivate our focus on how reliably an LLM can score such high-subtext content across multiple socio-political indicators.

2.2 Indicators of Weak Hegemony (IWH)

The IWH framework operationalizes structural markers of decline to monitor socio-political stability Patel et al. (2025: 8). It defines three primary dimensions:

Fragmented Consensus: This measures the breakdown of a unified public or political agreement. Historically, internal division and the death of a "liberal consensus" signal an irreversible phase of institutional decay (Patel et al., 2025: 9).

Counter-Hegemonic Forces: This assesses the presence of narratives or groups that actively challenge the established status quo. (Such forces often include "organic intellectuals" emerging from marginalized groups to advance alternative worldviews (Kiely, 2025: 3).

Crisis of Authority: This evaluates the perceived loss of legitimacy in established institutions and leadership. It is often exacerbated when leadership styles contribute to the erosion of moral authority, appearing to place figures above the law (Patel et al., 2025: 70).

In the broader IWH scheme, these sit alongside additional indicators such as reliance on coercion rather than consent and deficits of moral-intellectual leadership, which we encode as separate analytical dimensions in our annotation scheme.

2.3 The FABLE Framework: Measuring Misinformation Harm

To prioritize fact-checking efforts, this study utilizes the FABLE framework, which assesses the potential urgency and magnitude of misinformation as a harm (Sehat et al., 2024: 310). Unlike binary truth-matching, FABLE evaluates content across five variable dimensions (Sehat et al., 2024: 373):

Fragmentation (Social): The degree to which information breaks down social cohesion or questions the trust and functioning of public institutions (Sehat et al., 2024: 376).

Actionability: The extent to which content encourages the audience to take specific, often disruptive, actions, with an emphasis on potential physical harm (Sehat et al., 2024: 377).

Believability: The effectiveness of a message to a specific community, often measured by the absence of high-quality, publicly accessible refutations (Sehat et al., 2024: 380).

Likelihood of Spread: The potential for virality, determined by the "influencer" status of the spreader or the content's format, such as sensationalist text or video (Sehat et al., 2024: 381–382).

Exploitativeness: The degree to which the narrative preys on human fears, emotional triggers, or the lack of resources in vulnerable groups like children or the elderly (Sehat et al., 2024: 383–384).

2.4 Operationalizing Believability: The Role of Logical Fallacies

The dimension of **Believability** (FABLE-3) is operationalized through the identification of **logical fallacies**, which serve as rhetorical "fingerprints" used to counterfeit argumentation (Olariu, 2022: 33). These fallacies act as markers that artificially increase content credibility by bending cognitive heuristics (Olariu, 2022: 30). Key fallacies include:

Anecdotal Evidence (The Anecdote): Using a particular situation to generalize and ascribe general significance or representation power (Olariu, 2022: 33). This generalizes personal stories to broader worldviews that shape understanding and behavior Nair (2025: 17).

Slippery Slope: A fallacy used as a "prophylactic schemata" to avoid perceived emotional costs by arguing that a small first step will lead to catastrophic events (Olariu, 2022: 31).

In practical application, identifying these tactics allows LLMs to decompose content into rhetorical strategies (Jeong et al., 2025: 6934). However, Believability has proven to be a "problematic dimension" for automated analysis, frequently failing to meet the 80/80 standard.

3 Methodology

3.1 Data Sampling and Scoring

3.1.1 Data Sampling

The data used in these experiments contains Facebook posts from 02.05.2025 to 22.06.2025, centered around the debate around former president Rumen Radev's proposal for a referendum on joining the Eurozone in 2026. It was obtained through the Meta Content Library, filtered by the keywords "euro" and "referendum". It initially consisted of 445 Facebook posts from various public profiles. For this study, we analyzed only the textual data and ignored any other media. Many exact duplicates were removed as well, with most of them being the same post, shared from multiple profiles. The cleaned dataset we took our samples from consists of 302 posts.

For the optimal pipeline search, the objective was to identify a subset that was both sufficiently representative of the full dataset's diversity and small enough to make 10-run stability testing across all candidate configurations computationally feasible within the available budget. Therefore, 7% of the data was sampled in a way to ensure maximum diversity. First, all of the posts were separated into three groups, based on their previously annotated euro adoption stance ("pro", "anti" and "neutral"). For each group, the data was vectorized using the BAAI/bge-m3 model from HuggingFace Chen et al. (2024) for embedding, and separated into clusters using KMeans. The sample closest to each centroid

was chosen for a test sample. To avoid bias, we selected an equal number of representatives from each group.

3.1.2 Scoring

Political scientists have mapped the theoretical foundations to the specifics of the analysis by translating them into a total of 62 questions in English with possible answers: “Yes”, “No”, and “Unclear” (Appendix E).

The 62 questions were derived by operationalizing abstract theoretical dimensions, based primarily on the work of Gramsci (1971), referred to as the IWH and FABLE frameworks into observable binary indicators specifically tailored to the Bulgarian Eurozone debate. Political scientists on our team mapped each of the 11 dimensions — six from IWH (e.g., Fragmented Consensus) and five from FABLE (e.g., Believability) — to a set of 4–15 questions designed to detect specific rhetorical strategies or sociopolitical markers. For example, under IWH-1 (Fragmented Consensus), Q2 ("Does it contrast specific demographic groups?") was constructed to capture the "two societies" narrative central to Bulgarian polarization. Similarly, FABLE-3 (Believability) is operationalized through questions like Q37 ("Are personal anecdotes used as representative?"), which identify logical fallacies that serve as "rhetorical fingerprints" to artificially increase content credibility. Each question carries a weight reflecting its theoretical importance and empirical power to discriminate between narrative types in the Bulgarian context.

Question weights were established based on the core principle of theoretical importance and empirical discriminating power within the Bulgarian socio-political context. We assigned weights between -1 and 1 to each question to capture both confirmatory and disconfirmatory evidence. For example, in IWH-1 (Fragmented Consensus), Q2 regarding "Us/Them contrasts" is assigned a higher weight (0.6) than Q5 (0.2) because it is considered a more central indicator of the specific "two societies" narrative in Bulgaria.

The system features a bidirectional weighting approach, most notably in the FABLE-3 (Believability) dimension: Positive weights (+0.2 to +0.4): Applied to "high-susceptibility" fallacies like anecdotal evidence or slippery slopes, which empirical research (Marin et al., 2024) suggests are harder for audiences to detect and thus increase a message’s believability. Negative weights (-0.1 to -0.3): Applied to "low-susceptibility" fallacies like ad hominem attacks or extreme polarization, which audiences can often recognize as flawed, thereby decreasing the content’s perceived credibility. Ultimately, these weighted "Yes" responses are summed to produce a raw dimension score, which is then mapped to a final ordinal score (0, 1, 2) through dimension-specific thresholds designed to attenuate noise and ensure analytical stability.

Let the full question set be $Q = q_1, q_2, ..., q_{62}$, where each question q_k belongs to exactly one dimension $d \in D$ and one methodology $m \in \{IWH, FABLE\}$, while every dimension is associated with between 4 and 15 questions. Each question q_k has an associated weight $w_k \in [-1, 1]$, capturing both confirmatory and disconfirmatory evidence. Each question receives a response $a_{i,k} \in \{Yes, No, Unclear\}$ for text i. Only "Yes" contributes to the score:

$$y_{i,k} = 1; \quad (\text{if } a_{i,k} = \text{Yes})$$
$$y_{i,k} = 0; \quad (\text{if } a_{i,k} = \text{No, Unclear})$$

Let $Q_d \subseteq Q$ denote the subset of questions belonging to dimension d. The raw score for text i on dimension d is: $r_{i,d} = \sum_{k \in Q_d} w_k \cdot y_{i,k}$ Each dimension d has a set of dimension-specific thresholds $\theta_{d,1} < \theta_{d,2}$ that map the raw score to the final ordinal score $f_{i,d} \in \{0, 1, 2\}$:

$$f_{i,d} = \begin{cases} 0, & \text{if } r_{i,d} < \theta_{d,1} \\ 1, & \text{if } \theta_{d,1} \leq r_{i,d} < \theta_{d,2} \\ 2, & \text{if } r_{i,d} \geq \theta_{d,2} \end{cases}$$

This elaborate score system thus plays the role of a first-order stability measure, as it attenuates output variance from higher-order reasoning by introducing a discrete buffer against noise and minor fluctuations.

The final output of the LLM pipeline consists of three components: the answer, certainty score (1-5), and a justification in English. The calculations happening beyond that are deterministic.

3.2 Complexity Ceiling

3.2.1 Definition

The concept of a state of complexity beyond which LLMs fail to produce reliable outcomes is already present in the recent scientific literature. Shojaee et al. (2025) identify a complexity cliff in deterministic puzzle environments: Tower of Hanoi, River Crossing, and similar tasks. They found out that

when complexity is gradually increased, the performance of reasoning models improves steadily up until a point at which it "falls off a cliff". Susnjak et al. (2024) define a complexity threshold in a more general sense, in a simulated setting. They also come to the conclusion that there is a point beyond which excess complexity becomes a systemic barrier to reliable model performance. Both studies are concerned with the point of breaking, and both measure proximity to that point through accuracy against a fixed correct answer, either real or simulated. Our case differs on both of these points. First, we are interested in finding the spot right before excess complexity breaks performance. Second, the output is a set of 11 ordinal scores across politically charged text, where correctness cannot be verified by comparison to a fixed reference. In our case, we need to find such a state of balance between internal and external complexity that:

$$c^* = max\{c \in C | S(c) \geq \tau\}$$

Where:

c^* – the complexity ceiling (the optimal configuration)
c – a configuration in the complexity space C, defined by the vector of internal parameters (hyperparameters) and external parameters (context load)
C – the full complexity space being searched
$S(c)$ – the stability function evaluated at configuration c
τ – the stability threshold

In practice, C is a discretized and partially explored subset of the full configuration space, and c^* should be interpreted as a local optimum within this bounded search.

3.2.2 Complexity Types

External Complexity External complexity is defined by context. It is important to note that, although the texts being analysed are a very important part of the context themselves, we did not manipulate them at all. Regardless of the complexity configurations, we always analyse one full text per LLM call. In terms of complexity manipulation, three context engineering strategies were applied.

Batch Prompting Batch prompting (Lin et al., 2024) makes LLM prompting more efficient by combining instructions in fewer, bigger prompts. The authors acknowledge that from a certain point on, the more tasks (complexity) are added to a prompt, the harder it gets for the model to perform well.

The 62 questions were clustered together in two meaningful ways: based on methodology (IWH, FABLE), and based on methodological dimension.

- **Methodology-based:** every text was analysed with a total of 2 prompts, with each of them containing 32 (IWH) and 30 (FABLE) questions.
- **Dimension-focused:** every text was analysed by a total of 11 prompts, with each of them containing between 4 and 15 questions. The Believability dimension of FABLE contains 15 questions, as it is the only dimension where questions carry negative weights.

In-context Learning We employed an in-context learning approach (Wang et al., 2024), in which the input prompt is conditionally parameterized with dimension- or methodology-specific orienting guidelines prior to each inference batch. The guidelines were taken from general guideline documents for human critical discourse analysis (CDA) coders. This was planned as a measure to ease complexity and provide guidance. All configurations in this study are zero-shot — no labelled examples are provided to the model. The orienting guidelines therefore serve as the primary scaffolding for task framing in the absence of demonstrations.

System-level Instructional Conditioning We also put the same guidelines in the system prompt, again conditionally invoked, based on the specific methodology or dimension instructions the input prompt contains (Cheng and Mastropaolo, 2026).

Internal Complexity

Model We worked entirely with Gemini 3 Flash. At the time of selection, it was the most affordable frontier reasoning model reaching above 90% on both Multilingual Q&A (MMMLU) and Commonsense reasoning across 100 Languages and Cultures (Global PIQA) scores (Google DeepMind, 2025). While the approach is model-agnostic in design, its empirical behavior may vary across reasoning LLM architectures and requires validation beyond the single model used in this study.

Thinking level Gemini 3 has four thinking level options - "high", "medium", "low" and "minimal". These levels form a discrete, ordered internal complexity axis, which is just the comfortably

controllable parameter our search for a complexity ceiling requires. The lower the thinking level, the lower the internal complexity load.

Temperature Temperature governs the entropy of the model's output distribution: higher values flatten the probability distribution over possible tokens, increasing lexical and inferential diversity at the cost of determinism. Inversely, lower values concentrate probability mass, producing more focused but less exploratory outputs.

Together, the internal and external parameters defined above constitute the bounded configuration space C for this study. The leading configuration identified as described in 3.3 is therefore a **local optimum within this space**. As it is constrained by the available hyperparameter structure of the selected model and the computational budget, it should not be interpreted as a statement about the global capability limits of the model or of reasoning LLMs in general.

3.3 Descent to Feasibility

3.3.1 Definition

In contrast with the complexity cliff method (Shojaee et al., 2025), where complexity is gradually increased until reaching a point-break, we search for the ceiling in the opposite way. We set the complexity parameters high and gradually decrease them until the stability criteria are met:

$$c_{t+1} = c_t - \Delta c \text{ until } S(c_{t+1}) \geq \tau$$

Where Δc is the reduction step across the complexity dimensions. The algorithm terminates at $c^* = c_{t+1}$, the first point where the feasibility condition is satisfied.

The non-monotonic nature of LLM reasoning under increasing complexity loads (Shojaee et al., 2025) renders bottom-up search strategies suboptimal for identifying the global complexity ceiling. In a non-monotonic stability landscape, incremental ascent risks terminating at the first locally stable configuration encountered - a low-complexity point that satisfies the stability criterion but falls short of a genuinely demanding pipeline. By contrast, descending from a reasonably high point of complexity and stopping at the first stable configuration increases the likelihood that the identified configuration reflects a higher-complexity reasoning regime compared to bottom-up search strategies.

3.3.2 Process

After we define the highest starting point in terms of complexity within the constrained search space, the gradual descent of complexity parameters starts in the following order:

Thinking level The first variable to be reduced is the model's thinking budget — the parameters governing the depth and allocation of abstract reasoning prior to output generation. This ordering is motivated by the interaction between extended reasoning and output instability: high reasoning token allocations, when combined with complex instruction sets and elevated temperature, have been shown to amplify output variance and reduce cross-run modal consistency. Accordingly, thinking parameters are treated as the highest-leverage source of instability and are addressed first, with all other parameters held constant.

Context If the stability criterion remains unmet following reduction of thinking parameters, the process moves on to the context. We don't go with the smaller batch prompts immediately - before that we try to ease complexity by adding general guidelines first to the input prompt, then to the system prompt. Only after this doesn't work out, dimension-specific prompts are employed. This decomposition significantly reduces the simultaneous inferential load imposed on the model without altering the analytical scope of the task. With the smaller prompts, the cycle repeats - starting from the highest thinking-level point, and going down. Bigger batch prompts are more time and energy efficient - they need fewer calls per analysis, therefore our search strategy starts with lowering thinking parameters rather than instruction load.

Temperature Temperature is treated as a fixed anchor throughout stages Thinking level and Context and reduced last. This design serves two purposes. First, it isolates the effect of each descending step: holding temperature constant ensures that stability gains are attributable to reductions in thinking level or context complexity specifically, rather than to a more constrained output distribution (Atil et al., 2025). Second, elevated temperature keeps both complexity dimensions genuinely active simultaneously — reducing it prematurely would suppress the reasoning process rather than calibrate it, producing a spurious stability that reflects reduced expressiveness rather than genuine pipeline robustness. Temperature reduction is therefore reserved

for cases where the full descent at higher entropy fails to yield a shortlisted configuration.

3.4 Stability Benchmark

Here we start from the assumption of Atil et al. (2025), which states that stability depends on the specific combination of task type, output format, and prompt structure. In our case, this means that instability in Bulgarian high-subtext scoring is not predictable from general benchmarks. We needed a systematic way to measure it directly.

3.4.1 Definition - 80/80 rule

At least 80% of all texts sampled for the optimal pipeline search should produce the same final score at least 8 times after 10 independent runs across all 11 dimensions, provided that modal consistency for all texts across all dimensions averages above 80%. For every dimension $d \in D$, we calculate the stability score as follows. Given:

N – total number of texts in the pilot sample
$R = 10$ – number of independent runs per text
$s_{i,d,r} \in \{0, 1, 2\}$ – the score assigned to text i on dimension d in run r
$\hat{s}_{i,d}$ – the modal score of text i on dimension d across all R runs:

$$\hat{s}_{i,d} = \text{argmax}_{v \in \{0,1,2\}} \sum_{r=1}^{R} \mathbf{1}\{s_{i,d,r} = v\}$$

A text i is modally consistent on dimension d if its modal score appears in at least 8 out of 10 runs:

$$M_{i,d} = 1\Big\{ \sum_{r=1}^{R} 1\{s_{i,d,r} = \hat{s}_{i,d}\} \geq 0.8R \Big\}$$

The dimension-level stability score is then:

$$S_d = \frac{1}{N} \sum_{i=1}^{N} M_{i,d}$$

3.4.2 Rationale

Why 10 Runs According to Atil et al. (2025), and Wang and Wang (2025), 5 runs are enough to get most information about LLM output stability, with the latter study showing that every subsequent run beyond that has considerably flattening returns.

Why Modal Consistency Although consistency and accuracy are clearly two distinct terms, Wang et al. (2023) demonstrated empirically that the most common outcome from different chain-of-thought reasoning pipelines has much higher probability to be the correct one, linking modal consistency across different complex thinking processes to higher accuracy. In our case, higher temperature and thinking levels ensure that every run produces new, different reasoning patterns. However, as ground truth is not directly observable, we treat stability under a heavy complexity load strictly as a **proxy** rather than a measure of correctness.

Why 80/80 80% is a standard reliability benchmark across content analysis and (Neuendorf, 2002; McHugh, 2012; Lombard et al., 2002). In light of the intrinsic ambiguity of politically charged Facebook posts in Bulgarian, it's unrealistic to expect that all texts can be clearly defined according to the standards of all 11 dimensions. Therefore, we decided to loosen-up the stability criteria in that regard to 80% of the texts being required to pass this threshold. To further address concerns about statistical precision at small sample size, we add the requirement of at least 80% of average modal consistency for all texts, across all dimensions. When the central tendency of the distribution is itself above the criterion value, the conclusion is robust to sample size variation - large effects are detectable even at small N (Cohen, 1988). Furthermore, our 80/80 rule has an analogical precedent in a famous approach for the stability assessment of another area where higher-order reasoning is concerned - the educational process. Block and Anderson (1975) have assumed that a learning mastery process can be considered stable if 80% of the students pass 80% of the criteria on every single subject.

4 Results

4.1 Search Space and Algorithm

We have defined a search space for this study, based on the complexity parameters described above, which can be seen in Table 1.

Due to resource restrictions, configurations under temperature of 0.8 were not fully tested. A total of 14 configurations were tested following Appendix B: Algorithm 1, with 2 other configurations tested outside of this scope for referential purposes. The results for all tested configurations are listed in Appendix A: Table 2.

Eventually, three of the configurations with temperature set at 1 came very close to meeting the 80/80 criterion (Figure 1, Appendix C: Figure 2).

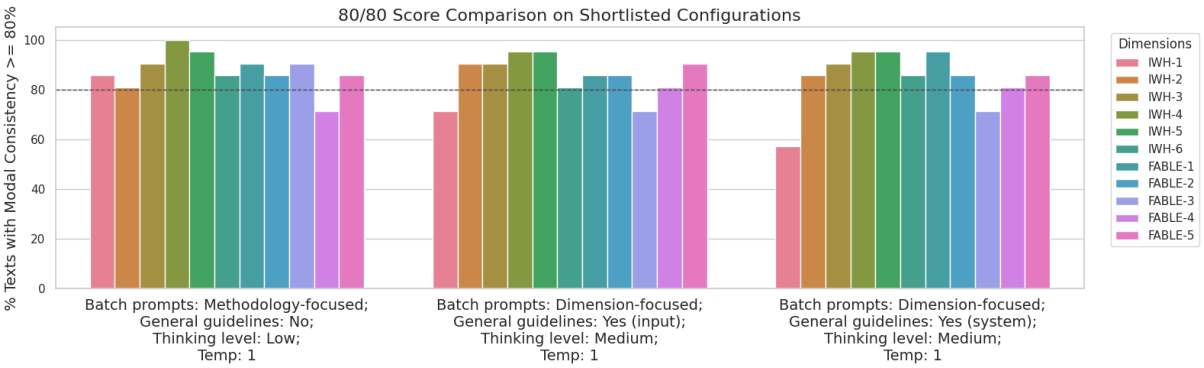


Figure 1: Comparison between the shortlisted configurations, returned by Algorithm 1.

Complexity Parameter	Values
Batch Prompting	Methodology-focused (big), Dimension-focused (small)
General Guidelines	No, Yes (input prompt), Yes (system prompt)
Thinking Level	Medium, Low
Temperature	1.0, 0.8

Table 1: Search space for the tested complexity configurations.

4.2 Problematic Dimensions

The shortlist threshold of $k = 9$ reflects two empirically grounded observations: first, IWH-1 exhibited instability across the majority of tested configurations, suggesting genuine interpretive ambiguity in this dimension; second, FABLE-3's instability under dimension-focused prompts is attributable to its disproportionate question load (15 questions vs. 4-6 for all other dimensions), making it a structural rather than semantic source of variance. Requiring 11/11 stability would therefore conflate pipeline performance with dimension-specific factors.

4.3 Shortlisted Configurations

Three configurations were shortlisted - one with methodology-focused batch prompts, and another two with dimension-focused ones.

4.3.1 Methodology-focused Batch Prompts

The complexity configuration with highest stability had methodology-specific batches, "low" thinking level and temperature of 1, when the batches were put into the system prompt (Figure 1). Exactly the same configuration produced clearly less stable results, when the batch prompts were put into the input prompt part (Appendix D: Figure 3).

4.3.2 Dimension-focused Batch Prompts

Two nearly identical configurations came close to fully meeting the 80/80 criteria as well (Figure 1), differing only in whether the guidelines were placed in the input prompt or the system prompt. Both had 9 out of 11 dimensions pass our stability threshold, suggesting that placement of general guidelines is not a decisive stability factor at this complexity level. Here, the annotation guidelines came on top of a pipeline that failed on 3 out of 11 dimensions and directly improved stability on IWH-2 in both cases.

More dimension-specific configurations satisfied the 80/80 standard with the same success rate (9 out of 11) but they were not shortlisted due to their lower levels of thinking complexity.

5 Discussion

Balance Between Internal and External Complexity The results point to two structurally distinct near-ceiling configurations, each reflecting a different balance between internal and external complexity. The first concentrates complexity in the context through methodology-focused batch prompts, while keeping the thinking level low. The second spreads context load across smaller, dimension-focused prompts and compensates with a higher reasoning budget at medium thinking level.

These configurations do not merely differ in parameter values; they appear to excel at different analytical tasks. The methodology-focused, low-thinking configuration achieves stable results not only on dimensions with relatively direct textual evidence, but also on IWH-1 and FABLE-3, which proved the hardest to stabilize overall. A plausible explanation is that a broader contextual frame combined with constrained inferential depth reduces the model's tendency to over-interpret inherently

ambiguous dimensions.

The dimension-focused configurations, by contrast, appear better suited to dimensions whose indicators are not explicitly stated in the text but require higher-order inference. FABLE-3 (Believability) remains a consistent instability point for this configuration type, but we attribute this primarily to its disproportionate question load: 15 questions compared to 4 to 6 for every other dimension. This interpretation is supported by the strong stability performance of all other dimension-specific prompts under the same configuration (except for IWH-1), suggesting that medium-thinking pipelines can achieve stable outputs under constrained prompt complexity.

The Role of General Guidelines When we put general-level guidelines, the prompt-heavy configurations' stability declined consistently. The same measure had, also sustainably, the opposite effect when applied to the smaller dimension-specific prompt configurations. We can hypothesise that a possible explanation here could be that when complexity is predominant in the context, adding steering context adds even more complexity, thus causing even more instability. This observation needs further research for a precise explanation, but it demonstrates the subtle balance between external and internal complexity.

6 Conclusion

While the approach is not inherently language-specific, its applicability to other low-resource environments remains an open empirical question. Politically charged and culturally specific narratives will always require high reasoning capacity and stability. Not a trade-off, or a compromise - both of them. This work demonstrates that guiding a reasoning LLM with strong multilingual capacity from a high-complexity state toward a stable configuration is a viable strategy for achieving both analytical depth and scoring consistency.

However, we should also acknowledge this study's **limitations**. This study has used only one reasoning model, a constrained search space, and a single political event in one low-resource language. Broader empirical validation across LLMs, larger search spaces, political contexts, and languages is needed before the approach can be considered fully operational. Importantly, this study evaluates stability independently of ground-truth correctness.

Future work includes optimizing the search for complexity ceiling with a larger search space, adapting the methodology to non-political use cases and other low-resource languages. We also plan to pair the 80/80 criterion with human-annotated validity checks and other sources or approximations of ground truth. Last but not least, we are already working on applying this methodology to multimodal media content.

References

Salah Feras Alali, Mohammad Nashat Maasfeh, Mucahid Kutlu, and Saban Kardas. 2026. Measuring political stance and consistency in large language models.

Berk Atil, Sarp Aykent, Alexa Chittams, Lisheng Fu, Rebecca J. Passonneau, Evan Radcliffe, Guru Rajan Rajagopal, Adam Sloan, Tomasz Tudrej, Ferhan Ture, Zhe Wu, Lixinyu Xu, and Breck Baldwin. 2025. Non-determinism of "deterministic" LLM settings.

James H. Block and Lorin W. Anderson. 1975. *Mastery Learning in Classroom Instruction*. Macmillan, New York.

Tanise Ceron, Neele Falk, Ana Barić, Dmitry Nikolaev, and Sebastian Padó. 2024. Beyond prompt brittleness: Evaluating the reliability and consistency of political worldviews in LLMs. *Transactions of the Association for Computational Linguistics*, 12:1378–1400.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Zaiyu Cheng and Antonio Mastropaolo. 2026. An empirical study on the effects of system prompts in instruction-tuned models for code generation.

Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2 edition. Lawrence Erlbaum Associates.

Roberts Darģis, Guntis Bārzdiņš, Inguna Skadiņa, Normunds Grūzītis, and Baiba Saulīte. 2024. Evaluating open-source LLMs in low-resource languages: Insights from Latvian high school exams. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 289–293, Miami, USA. Association for Computational Linguistics.

Google DeepMind. 2025. Gemini 3 flash: Model card and evaluation. Accessed: 2026-04-15.

Antonio Gramsci. 1971. *Selections from the Prison Notebooks of Antonio Gramsci*. International Publishers, New York.

Ömer Alperen Gürses and İsmail Ceylan. 2026. Consistency over accuracy: run-to-run stability of contemporary large language models on turkish curriculum-aligned theoretical anatomy multiple-choice questions. *BMC Medical Education*, 26:285.

Jiwon Jeong, Hyeju Jang, and Hogun Park. 2025. Large language models are better logical fallacy reasoners with counterargument, explanation, and goal-aware prompt formulation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6933–6952, Albuquerque, New Mexico. Association for Computational Linguistics.

Keith Peter Kiely. 2025. Wearing gramsci's glasses: transition, conspiracy and disinformation in bulgaria through the lens of hegemony. *Journal of Political Ideologies*, 0(0):1–20.

Jianzhe Lin, Maurice Diesendruck, Liang Du, and Robin Abraham. 2024. Batchprompt: Accomplish more with less.

Matthew Lombard, Jennifer Snyder-Duch, and Cheryl Campanella Bracken. 2002. Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research*, 28(4):587–604.

Pinja M. Marin, Marjaana Lindeman, and Annika M. Svedholm-Häkkinen. 2024. Susceptibility to poor arguments: The interplay of cognitive sophistication and attitudes. *Memory & Cognition*, 52(7):1579–1596.

Mary L. McHugh. 2012. Interrater reliability: The kappa statistic. *Biochemia Medica*, 22(3):276–282.

Anushka Manchanda Nair. 2025. Multi-stage LLM reasoning for automated detection and classification of high-impact misinformation. Master's thesis, Massachusetts Institute of Technology.

Kimberly A. Neuendorf. 2002. *The Content Analysis Guidebook*. Sage, Thousand Oaks, CA.

Oana Olariu. 2022. Critical thinking as dynamic shield against media deception: Exploring connections between the analytical mind and detecting disinformation techniques and logical fallacies in journalistic production. *Logos Universality Mentality Education Novelty: Social Sciences*, 11(1).

Ketan Patel, Christian Hansmeyer, Nandan Desai, and Aditya Ajit. 2025. American hegemony at a critical juncture, lessons from history's great powers. *Frontiers in Political Science*, Volume 7 - 2025.

Connie Moon Sehat, Ryan Li, Peipei Nie, Tarunima Prabhakar, and Amy X. Zhang. 2024. Misinformation as a harm: Structured approaches for fact-checking prioritization. *arXiv preprint arXiv:2312.11678*.

Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. 2025. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. *arXiv preprint arXiv:2506.06941*.

Teo Susnjak, Timothy R. McIntosh, Andre L. C. Barczak, Napoleon H. Reyes, Tong Liu, Paul Watters, and Malka N. Halgamuge. 2024. Over the edge of chaos? excess complexity as a roadblock to artificial general intelligence.

Julian Junyan Wang and Victor Xiaoqi Wang. 2025. Assessing consistency and reproducibility in the outputs of large language models: Evidence across diverse finance and accounting tasks.

Liang Wang, Nan Yang, and Furu Wei. 2024. Learning to retrieve in-context examples for large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1752–1767, St. Julian's, Malta. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models.

Appendix A Table 2: 80/80 scores for each configuration tested

#	Batch type	General Guidelines	Thinking Level	Temperature	Dimensions Passing
1	Methodology-Focused	No	Medium	1.0	7/11
2	Methodology-Focused	Yes (Input)	Medium	1.0	5/11
3	Methodology-Focused	Yes (System)	Medium	1.0	6/11
4	Methodology-Focused	Yes (System)	Medium	0.8	6/11
5	Methodology-Focused	No	Low	1.0	10/11
6	Methodology-Focused	Yes (Input)	Low	1.0	7/11
7	Methodology-Focused	Yes (System)	Low	1.0	7/11
8	Methodology-Focused	Yes (System)	Low	0.8	8/11
9	Dimension-Focused	No	Medium	1.0	8/11
10	Dimension-Focused	Yes (Input)	Medium	1.0	9/11
11	Dimension-Focused	Yes (System)	Medium	1.0	9/11
12	Dimension-Focused	No	Low	1.0	9/11
13	Dimension-Focused	Yes (Input)	Low	1.0	9/11
14	Dimension-Focused	Yes (System)	Low	1.0	7/11

Table 2: 80/80 results for each configuration tested

Appendix B Algorithm 1

Algorithm 1 Descent to Feasibility

Input: $\tau = 0.80$, $R = 10$, acceptance threshold $k = 9$, D: set of dimensions ($|D| = 11$), $S_d(c, R)$: stability score of dimension d under configuration c across R independent runs

Output: *shortlist*: shortlist (all c with passed $\geq k$); c^* if full ceiling found

shortlist $\leftarrow \varnothing$
$c^* \leftarrow \varnothing$
for $temp \in \{1.0, 0.8\}$ **do**
  for $batch \in \{\text{Methodology-Focused}, \text{Dimension-Focused}\}$ **do**
    for $guidelines \in \{Input\_Prompt, System\_Prompt\}$ **do**
      for $thinking \in \{Medium, Low\}$ **do**
        $c \leftarrow (batch, guidelines, thinking, temp)$
        $passed\_count \leftarrow |\{d \in D \mid S_d(c, R) \geq \tau\}|$
        if $passed\_count = |D|$ **then**
          $shortlist \leftarrow shortlist \cup c$
          $c^* \leftarrow c$
          return $shortlist, c^*$
        else if $passed\_count \geq k$ **then**
          $shortlist \leftarrow shortlist \cup c$
return $shortlist, c^*$

Appendix C Figure 2: Average Modal Consistency

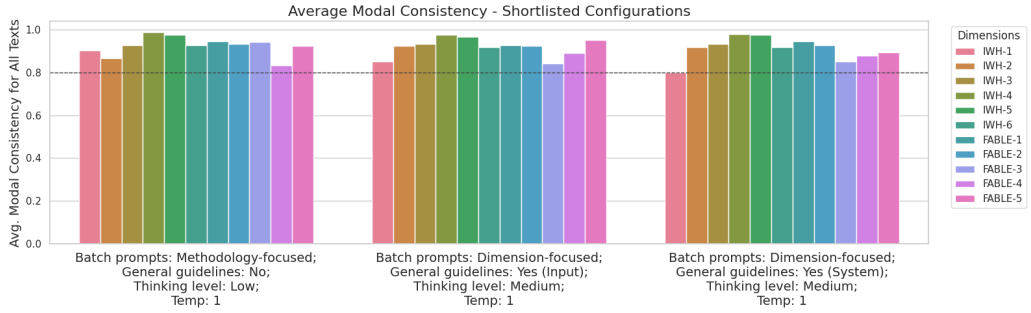


Figure 2: Average modal consistency comparison between the three shortlisted configurations.

Appendix D Figure 3: Leading Config with Batch Prompts in System Prompt vs Batch Prompts in User Prompt

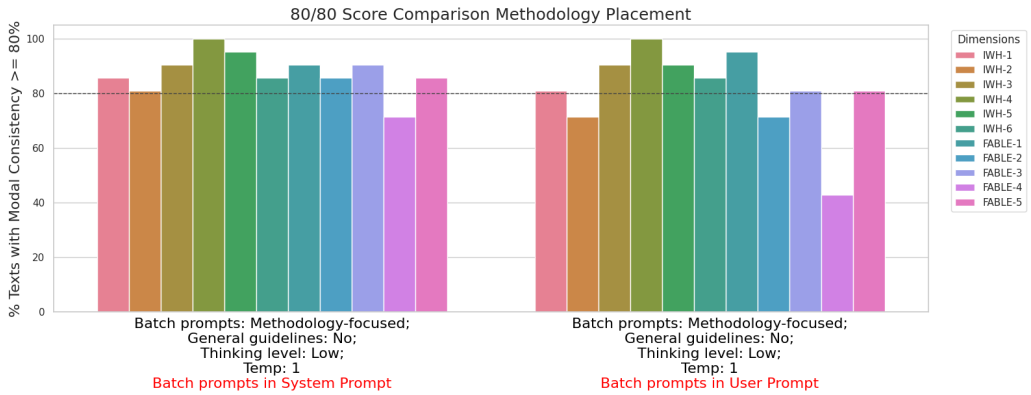


Figure 3: Comparison between placing the methodology in the system prompt, as opposed to the user prompt. The other configurations remain the same.

Appendix E Coding Questionnaire

Below is the complete list of questions used in the evaluation phase of the study:

IWH-1: Fragmented Consensus (5 items):
Q1. Does the text highlight polls or anecdotes that Bulgarians are deeply divided on the euro? (Yes/No/Unclear)
Q2. Does it contrast specific demographic groups (urban vs rural, educational levels, wealthy vs poor, young vs old, elite vs people)? (Yes/No/Unclear)
Q3. Does it cite recent electoral volatility linked to the currency issue? (Yes/No/Unclear)
Q4. Are coalition rifts over the euro portrayed as irreconcilable? (Yes/No/Unclear)
Q5. Is the societal split itself framed as a central crisis that must be addressed? (Yes/No/Unclear)

IWH-2: Counter-Hegemonic Forces (5 items):
Q6. Are organized opposition actors (petition committees, civic groups) spotlighted? (Yes/No/Unclear)
Q7: Does the content call for or depict a specific, organized action (e.g., a named protest, a petition with a link)? (Yes/No/Unclear)
Q8: Does the content use strong rhetorical calls for general resistance or civil disobedience (e.g., "resistance is a duty")?
(Yes/No/Unclear) Q9. Are alternative media sources or Telegram channels presented as information hubs? (Yes/No/Unclear)
Q10. Is the counter-elite narrative presented as gaining momentum or influence? (Yes/No/Unclear)

IWH-3: Crisis of Authority (4 items):
Q11. Does it reference repeated elections, cabinet collapses, or leadership resignations? (Yes/No/Unclear)
Q12. Are court battles (e.g., Constitutional Court decisions) cited as blocking policy implementation or a major democratic process (such as a referendum or popular initiative)? (Yes/No/Unclear)
Q13. Is low public trust in parliament, BNB, or government ministries emphasized? (Yes/No/Unclear)
Q14. Does it characterize current leaders as "illegitimate" or "without mandate"? (Yes/No/Unclear)

IWH-4: Reliance on Coercion (3 items):
Q15. Are state-aligned outlets accused of censoring opposition voices or presidential speeches? (Yes/No/Unclear)
Q16. Is there mention of legal threats, fines, or systematic smear campaigns against dissenters? (Yes/No/Unclear)
Q17. Does the content claim that organized political opposition (e.g., parties, formal movements) are being systematically silenced or censored? (Yes/No/Unclear)

IWH-5: Ideological Contradictions (4 items):
Q18. Does it accuse anti-euro elites of maintaining euro bank accounts abroad while opposing adoption? (Yes/No/Unclear)
Q19. Does it highlight MPs pricing real estate in euros while rejecting public euro debate? (Yes/No/Unclear)
Q20. Are governing parties portrayed as pro-EU publicly but corrupt or hypocritical privately? (Yes/No/Unclear)
Q21. Are there examples of sudden policy U-turns ("a referendum was unnecessary, now it is acceptable")? (Yes/No/Unclear)

IWH-6: Moral/Intellectual Void (5 items):
Q22. Does it deride economists or academics as "Soros puppets" or "on the payroll"? (Yes/No/Unclear)
Q23. Is expert analysis replaced by conspiracy language or populist slogans? (Yes/No/Unclear)
Q24. Are complex fiscal issues reduced to emotive metaphors (e.g., euro slavery)? (Yes/No/Unclear)
Q25. Does it lament the absence of authentic leaders" with principles? (Yes/No/Unclear)
Q26. Are appeals made to raw emotion over evidence ("voice of people, voice of God")? (Yes/No/Unclear)

FABLE-1: Fragmentation (6 items):
Q27. Does the post use a rhetorical frame that divides society into two opposing, irreconcilable groups (e.g., "the people vs. the elite," "patriots vs. traitors")? (Yes/No/Unclear)
Q28. Does it portray Bulgarian state institutions as fundamentally untrustworthy? (Yes/No/Unclear)
Q29. Does it attack mainstream Bulgarian media credibility systematically? (Yes/No/Unclear)
Q30. Does it question the honesty or patriotism of "ordinary Bulgarians" who disagree? (Yes/No/Unclear)
Q31: Does the post claim that the outcome of an election/referendum/constitutional court decision was (or will be) fraudulent or rigged? (Yes/No/Unclear)
Q32: Does the post claim that the process for initiating or conducting an election/referendum was illegally blocked or sabotaged? (Yes/No/Unclear)

FABLE-2: Actionability (6 items):
Q33. Is there an explicit call to sign petitions, share content, protest, or block legislation? (Yes/No/Unclear)
Q34. Are specific dates, places for offline action, or livestream details provided? (Yes/No/Unclear)
Q35. Does it urge immediate action with urgency language ("now", "before it's too late")? (Yes/No/Unclear)
Q36. Does it single out specific individuals or offices for pressure or harassment? (Yes/No/Unclear)

Q37. Does it link euro grievances to vulnerable groups (children, pensioners) to mobilize protection instincts? (Yes/No/Unclear)
Q38. Is the message part of a coordinated campaign or series (organized hashtags)? (Yes/No/Unclear)

FABLE-3: Believability (15 items):
High-Risk Believability Factors (Increase Score):
Q39. Does it rely primarily on personal anecdotes or testimonials as evidence? (Yes/No/Unclear)
Q40. Does it use "everyone knows" or appeal to common sense? Yes/No/Unclear)
Q41. Does it present slippery slope arguments ("euro will lead to complete sovereignty loss")? (Yes/No/Unclear)
Q42. Does it exploit knowledge gaps ("experts won't tell you this")? (Yes/No/Unclear)
Q43. Does it make causal claims without proof ("euro caused Greece crisis")? (Yes/No/Unclear)
Q44. Is the primary format narrative/storytelling rather than factual? (Yes/No/Unclear)
Q45. Does it target audiences with lower formal education/media literacy? (Yes/No/Unclear)
Q46. Is sharing motivated by identity expression or persuasion? (Yes/No/Unclear)

Neutral Believability Factors:
Q47. Does it mix factual information with selective evidence? (Yes/No/Unclear)
Q48. Is it published by outlet with unclear editorial standards? (Yes/No/Unclear)

Low-Risk Believability Factors (Decrease Score):
Q49. Does it rely heavily on ad hominem attacks? (Yes/No/Unclear)
Q50. Does it contain obvious logical contradictions? (Yes/No/Unclear)
Q51.Does it use extreme polarization that may trigger skepticism? (Yes/No/Unclear)
Q52. Does it lack any supporting evidence whatsoever? (Yes/No/Unclear)
Q53. Is the tone so extreme it appears fabricated? (Yes/No/Unclear)

FABLE-4: Likelihood of Spread (5 items):
Q54. Is the author an influencer, major party page, or national TV/media site with significant reach? (Yes/No/Unclear) Note: Code based on the immediate publisher of the content being analyzed. If a personal blog quotes a national TV station, the author is the blogger, not the station.
Q55. Is the content published on high-traffic, public platforms (Facebook, TikTok, YouTube)? (Yes/No/Unclear)
Q56. Are viral elements present (hashtags, mentions, paid promotion, or boosting evidence)? (Yes/No/Unclear)
Q57. Does it contain arresting visuals, caps-lock headlines, emojis, or eye-catching formatting? (Yes/No/Unclear)
Q58. Is it tied to active news cycles (e.g., Radev's speeches, parliamentary debates, EU announcements)? (Yes/No/Unclear)

FABLE-5: Exploitativeness (4 items):
Q59. Does it warn of imminent threats (poverty, sovereignty loss, economic "robbery") to generate fear? (Yes/No/Unclear)
Q60. Does it emphasize powerlessness with specific examples ("160 MPs rob the people")? (Yes/No/Unclear)
Q61. Does it specifically target groups with fewer resources or higher vulnerability (rural populations, elderly)? (Yes/No/Unclear)
Q62. Is the content exclusively in Bulgarian, targeting monolingual audiences with limited cross-referencing ability? (Yes/No/Unclear)

Corpus-Based Approaches to Style, Register, and Meaning

Quantifying Literary Style: A Corpus-Based Study

Iglika Nikolova-Stoupak, Gaël Lejeune, Eva Schaeffer-Lacroix

Sens Texte Informatique Histoire, Sorbonne Université, Paris, France

iglika.nikolova-stoupak@etu.sorbonne-universite.fr,
{gael.lejeune, eva.lacroix}@sorbonne-universite.fr

Abstract

We propose a highly explainable method for the identification of a literary work's style based on a large number of quantitative characteristics. Drawing inspiration from Biber's method of classifying text (notably, as written versus oral) through multi-dimensional analysis (Biber, 1988), we argue that with the help of contemporary computational methods and an extension of the number of metrics to encompass previously unmeasurable textual aspects such as semantics, a single work's style may be efficiently defined. The analysis explores metrics defined by Biber as well as established readability features and high-level semantic ones (e.g. the use of metaphors). This short paper showcases the process of corpus compilation and related statistical analysis that leads to the definition of several texts' distinctive features as those that demonstrate the highest deviation from the established baseline for the corpus. We also determine the most overrepresented vocabulary items for the selected works. In the framework of our associated wider project, the determined features will then be used to derive adapted versions of the selected texts that seek to retain their essence. The artefacts will be evaluated through qualitative human-based surveys of reader experience. We release the associated corpus and the code used for statistical analysis.

Keywords: literary style, corpus analysis, readability, stylometry, lexical frequency

1 Introduction

The study of a text's style is especially relevant in the current technological era, where it is used, for instance, in the identification of AI-generated text. State-of-the-art classifiers of style typically make use of advanced neural methods. As a result, their function lacks interpretability. What we are proposing herein is a method for style quantification that makes use of a large number of linguistic characteristics, ranging from shallow, length-based ones to ones that make use of neural models and semantics (e.g. metaphor identification). A number of these features have been established as relevant within computer-assisted stylistics. We define a literary work's style as its key deviations in these features against a pre-calculated baseline. The validity of this definition is to be established within the underlying project's next steps, which include human-based evaluation through a survey of reader experience.

The attained conclusions about quantification of the style of a work of literature may be utilised by human authors and Natural Language Processing (NLP) tools to derive literary adaptations that retain the original work's style, thereby addressing the common criticisms that authenticity is lost through the process of adaptation, and that the reading experience associated with an adapted text is strikingly different (Hutcheon, 2006; Thiel, 2013). Our methods' main contributions are: explainability, replicability, inclusion of metatextual characteristics (such as the time and place of writing) as elements of style, and the focus on a single literary text as a stylistic entity.

The associated corpus and the code used for statistical analysis are accessible in the following GitHub repository. We encourage collaboration in the direction of expanding the linguistic feature set and improving its robustness.

2 Background and Related Work

The study of stylistics dates from the 19th century, and it met significant development around the middle of the 20th century. Spitzer (1948) saw a literary author's style as the innovation that distinguishes their texts from other texts associated with the same temporal and geographical framework. He conducted microscopic studies of vocabulary use and

was met by criticism concerning his method's subjectivity and reliance on intuition (Spitzer, 1948: p. 26). Wellek and Warren (1949) challenged the prevalent theory that literary language is merely a reflection of a given language's development, claiming that different genres are associated with different language use and that literature itself often exerts influence on language as a whole. They defined style as "deviations and distortions from normal usage". Wellek and Warren (1949) also suggested that style encompasses characteristics beyond lexical items, such as syntax, sound and rhythm. Hatzfeld (1953) included in the notion of style a variety of linguistic elements, all the way to a text's semantics. Contrary to the common focus on the author, he also viewed an isolated literary work as a main stylistic unit. Toolan (2014)'s claims about the stylistics of literature include that language correlates with meaning, that every encountered linguistic characteristic is meaningful, and that an individual work can never be analysed exhaustively.

Stylometry, also known as "computational stylistics", is the discipline concerned with the quantitative study of style. Early work, such as that of Mendenhall (1901), sought recurring linguistic patterns in corpora through manual analysis, while modern stylometry uses computational methods, thereby achieving greater objectivity and scope. Traditionally, textual features such as word frequencies, parts of speech (POS) and vocabulary richness were discussed, and additional processing typically included Principal Component Analysis (PCA). Eder et al. (2015) define stylometry as the investigation of the relationship between writing style and underlying metadata, such as authorship, textual genre, and date of composition, the main focus being on authorship. The method of defining stylistic specificity through comparison with a reference corpus, utilised in the present project, is well established. For instance, Mahlberg (2013) identifies idiosyncrasy in literary dialogue based on a corpus of non-literary language (Mahlberg, 2013).

Later stylometric classification methods include machine learning (such as logistic regression and SVM models), pre-trained models to learn authorial representations (notably, BERT) and, ultimately, Large Language Models (LLMs), which require no fine-tuning and can identify authorship based on increasingly small amounts of text (Huang et al., 2024). Common contemporary applications of stylometry include the binary classification of human versus machine-produced text and the attribution of a text to a specific language model. Comparing the performance of different LLMs at authorship identification in Early Modern English drama, Hicke and Mimno (2023) argue that the models' overall high performance is also associated with limitations, such as confident wrong attributions, which may be influenced by an author's overall popularity and prevalence in training data.

The traditional linguistic features utilised in stylometry highly coincide with those associated with readability, the study of texts' difficulty for the purpose of matching them to an appropriate audience (DuBay, 2004). Established formulas of readability, such as Flesch Reading Ease and Dale-Chall, rely on length-based measures (e.g. the number of words in a sentence) and/or the mapping of vocabulary items against frequency lists. A dichotomy of lexical and syntactic features is often utilised, and additional categories, such as "discourse", may be added. Like stylometry, readability progressed from methods based on feature engineering to black box deep learning methods. Neural classifiers are used to map text to gold labels, such as CEFR proficiency levels. Interestingly, there is evidence that state-of-the-art models closely make use of traditional readability features, proving their continuous relevance (Deutsch et al., 2020).

3 Corpus

We compiled a corpus of 100 highly acclaimed books in English. For the purpose, we made use of the ranking provided by TheGreatestBooks.org[1], which in turn is based on multiple online book rankings. To facilitate intertextual comparison, we excluded works such as plays, poetry and essays. Our final list contains mostly novels as well as additional texts that contain long narrative (such as *The Odyssey* and *One Thousand and One Nights*). 65% of the works were originally written in English, and the others include translations from various modern European languages, as well as Ancient Greek, Latin and Arabic. Thirteen authors are represented by more than one work, Dostoevsky and Hemingway coming on top of the list with four works each. The corpus includes texts composed between 700 BCE (*The Odyssey*) and 1987 (*Beloved*). All texts are available online free of charge, and the majority are included in Project Gutenberg. No additional

[1] https://thegreatestbooks.org/

OCR process was required. For the full list of texts, please refer to Appendix A.

It is important to note that broad variables pertaining to the corpus's texts, such as year of writing, original language and literary genre will undoubtedly exert influence in the statistical analysis. We consider this a problem inasmuch as the variables are not equally represented within the corpus, thereby leading to possible biases in the calculated baseline, and we therefore encourage and plan the application of the defined analysis to additional and larger corpora. However, the influence of these variables within a definition of the distinctive style of an individual literary work does not undermine our primary objective. We argue that they contribute to an isolated work's style along with more narrow (e.g. author-specific) features, and that an adapted version of the text that seeks to retain its essence is to keep track of both. Please consult our robustness analysis of the corpus pertaining to original language and year of writing in Appendix B.

With planned microscopic human-based evaluation in mind, we singled out five of the texts in the corpus for the purpose of defining their distinctive characteristics and in extension, within our proposed framework, their style: *Moby Dick*, *The Picture of Dorian Gray*, *Anna Karenina*, *Les Misérables*, and *The Odyssey*. Seeking variety in terms of represented languages and countries as well as years of composition, we selected works that are canonical in the framework of literary adaptation, as the next steps of our larger project involve the application of the attained conclusions to derive adapted versions of the texts.

4 Methods

4.1 Preprocessing

All texts in the corpus underwent a pipeline of initial preprocessing based on hand-crafted rules. All metatextual information, including prefaces (authorial or not), glossaries, notes, page numbers, and text between square brackets, was removed. Titles, including those of volumes, parts, and chapters, were also removed. Text resembling URLs as well as special symbols, typically resulting from a previously conducted OCR process, were removed. Extra spaces and empty lines were removed, and lines separated by line-break hyphens were merged.

4.2 Linguistic Features

We quantified 274 different features within the compiled corpus. Please see Table 1 for representative feature examples. For the full set of features utilised in the statistical analysis, please refer to Appendix C.

Broad feature category	Examples of included features
Tense, aspect and verbal morphology	Past tense; Present tense; Perfect aspect; Present participles; Infinitives
Reference and pronominal use	First-person pronouns; Third-person pronouns; Demonstrative pronouns; Demonstrative determiners
Syntax and clause structure	By-passives; Agentless passives; That-complements after verbs; Causal subordinators; Clausal coordination
Modification and phrase structure	Attributive adjectives; Predicative adjectives; Total adverbs; Total prepositional phrases
Lexical and grammatical distribution	Verbs; Nouns; Proper nouns; Punctuation; Content-to-function-word ratio
Lexical richness, rarity and readability	Type-token ratio; Hapax legomena; Out-of-vocabulary words; Mean word length; Words per sentence
Stance, modality and discourse	Hedges; Amplifiers; Conjuncts; Possibility modals; Private cognition verbs; Public reporting verbs
Narrative, semantic and higher-level features	Dialogue; Indirect speech; Death-related words; Mean concreteness; Metaphor use; Humor probability

Table 1: Examples of linguistic features, included in the statistical analysis.

A large number of the utilised features are traditionally used within the fields of readability and stylometry and cover the general categories "length-based", "vocabulary-related", "syntax-related", and "discourse-related" (DuBay, 2004; Collins-Thompson, 2014; Deutsch et al., 2020; Eder et al., 2015). Another large group of features is inspired by Biber's work on multidimensional text classification and includes categories such as "tense and aspects markers", "place and time adverbials", "questions", "passives" and "subordination" (Biber, 1988). The remaining features come from general literary theory, whilst special attention is accorded to elements pertaining to children's literature; the reason being that literary adaptation, which we are planning to apply our conclusions to, typically addresses a younger audience. Beauvais (2023)'s book

on children's literature has been especially useful in the process of feature engineering. Features based on literary theory include, for instance, the length of a text's first sentence as well as the presence of humour and figurative language. Due to their strong semantic implication, the last two features were estimated using transformer-based neural models, fine-tuned for binary classification of the phenomena in question: humor-no-humor[2] and Metaphor-Detection-XLMR(Wachowiak et al., 2022). For our benchmarking of the utilised neural models, please refer to Appendix D.

The assembled list of linguistic features has no claims of being comprehensive; rather, it offers a multi-faceted and consistent quantitative description of literary texts.

4.3 Statistical Analysis

The final set of features was selected not only on theoretical grounds, but also based on the feasibility of their automatic extraction. We privileged the use of automatic annotation tools as readily implementable in Python, such as the spaCy library for lemmatisation, POS tagging and morphological analysis.[3] In some cases, such as for the identification of time and space adverbials, we resorted to hard-coded lexical lists.

Wherever possible, we based our calculations on 100 tokens of text, following Biber's guidelines. Instead of single values per feature, we calculated the following set of aggregators: average value, minimal value, maximal value, and standard deviation within the text. This differentiation has been shown to provide additional insight into readability-based characteristics (Wilkens et al., 2022) and reflects the importance of considering not only feature frequencies but also their position and distribution within a text (Hoey, 1983).

We opted for the inclusion of the five texts preselected for qualitative analysis and human-based evaluation in the corpus-wise calculation of all features, as they take part in the same population of acclaimed narrative-based literary works. To define these texts' distinguishing style, we determined their top five deviating features with regard to the corpus baseline. Deviation was measured using z-scores[4]. To ensure that the identified features represent distinct stylistic dimensions, we filtered out highly correlated features using pairwise Pearson correlation with a threshold of 0.70.

In addition to the described quantitative features, we explored the most overrepresented lexical items within the five investigated texts as compared against a 10k-word frequency list based on Project Gutenberg and provided by Wiktionary[5]. To calculate deviation in word frequency, we used the following formula:

$$\text{Deviation} = \left(\frac{f_{\text{book}}(w)}{f_{\text{ref}}(w)} - 1 \right) \times 100 \qquad (1)$$

where:

- $f_{\text{book}}(w)$ = relative frequency of the word w in the text
- $f_{\text{ref}}(w)$ = relative frequency of the word w in the corpus

For instance, a value of 100% means that a word is used two times more frequently than expected. We disregarded proper nouns as well as word fragments (e.g. "don", "ll").

5 Results

316 out of the 37,401 existing feature pairs were determined to be strongly correlated. As expected, the majority of strong correlations are associated with different aggregators of the same variable as well as other logically related features, such as word-based and lemma-based type-to-token ratio. Less obvious high dependencies include these between "number of verbs (maximal)" and "number of nouns (average)" and between "number of verbs (average)" and "punctuation (maximal)", both with a value of 0.96.

Table 2 presents the top three deviating features per investigated work. Demonstrative pronouns are the most highly deviating feature for two of the texts (*The Picture of Dorian Gray* and *Les Misérables*); a related feature, demonstrative determiners, makes the list for *Moby Dick*. Most of the remaining features are related to POS and grammatical categories (verbs, pronouns, adjectives, subordination). A shallow length-based feature (sentence

[2] https://huggingface.co/mohameddhiab/

[3] We used the `en_core_web_sm` model, which reports approximately 97% POS tagging accuracy on the OntoNotes 5.0 benchmark (AI, 2025). POS tagging forms the basis for subsequent lemmatisation and morphological analysis.

[4] Defined as: $z = \frac{x-\mu}{\sigma}$, where x is the book value, μ the corpus mean, and σ the corpus standard deviation.

[5] https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists/PG/2006/04/1-10000

Book	Top deviating features
The Picture of Dorian Gray	Demonstrative pronouns (sd) ↑; Past tense (sd) ↑; Predictive modals (max) ↑
Moby Dick	Seem/appear (max) ↑; Demonstrative determiners (avg) ↑; By-passives (max) ↑
Anna Karenina	Private cognition verbs (max) ↑; Metaphor use (max) ↑; Present participles (avg) ↑
Les Misérables	Demonstrative pronouns (max) ↑; Nouns (avg) ↑; Time adverbials (max) ↑
The Odyssey	Words per sentence (min) ↑; Conjuncts (avg) ↑; Other subordination (avg) ↑

Table 2: Top three most significantly deviating features for the five investigated texts. The arrows denote the direction of deviation.

length) comes first in the case of *The Odyssey*. All top-three deviations are upward compared to the corpus baseline. For a more detailed table of results that includes five features per text as well as the associated values within the text in question and the whole corpus, please refer to Appendix E.

See Appendix F for the top 10 most overrepresented vocabulary items in the five investigated texts. The words associated with *Moby Dick*, a novel whose narrative unfolds within the self-contained microcosm of a whaling ship, demonstrate the highest values of overuse. The word "whale", most overused globally, is narrowly linked to the text's topic. The majority of all identified items are content words (notably, nouns) that are intuitive given the respective work's content (e.g. "painter" for *The Picture of Dorian Gray*, "countess" for *Anna Karenina*). Occasionally, different forms based on the same lemma are present in the lists (e.g. "peasants" and "peasant" for *Anna Karenina*). For the purpose of this study, we decided to keep such words as separate entries, as is the case in the utilised frequency list. In *Les Misérables*, the original French language's influence is clearly observable through the words "sous", "reverie" and "sombre", all of French origin. It is important to note that this specificity pertains to the specific utilised English translation of *Les Misérables* rather than the original work.

6 Conclusion and Future Directions

We compiled a corpus of 100 acclaimed literary works with strong narrative elements that span across genres, time periods and countries of origin. We went on to define a large number of quantitative textual characteristics and to use them in the establishment of a baseline pertaining to the corpus. Subsequently, we proposed a method for establishing specific texts' distinctive features as those exhibiting the highest deviations from the corpus baseline. Planned future work will focus on the following steps: 1) use of the identified stylistic features to derive (through hybrid manual and automatic methods) literary adaptations that seek to retain the original works' distinctive characteristics 2) evaluation of the resulting adaptations against neutral baselines via LLM-based as well as human evaluation. The former will consider the number of tokens necessary for recognition of the texts, and the latter will be based on a survey consisting of comprehension questions and questions describing the overall reading experience.

7 Limitations

Firstly, the current study is limited to English text and its cross-lingual relevance remains to be determined. Also, as explained, broad factors such as a work's original language of writing, genre and period of composition are all considered to play a role in its style. An alternative study seeking to control these variables would require the crafting of a more uniform and/or larger corpus. We also note that, while traditional stylistics tends to pursue a qualitative analysis of literary features, including their intended or observed effects on the reader, our project follows the tradition of stylometry in that it does not attempt a meaning-based analysis.

Although OCR-related errors or typographical inconsistencies are minimised through the processes of corpus composition and preprocessing, they have not been eliminated completely and, therefore, may influence the achieved statistics. Also, the multiple utilised linguistic features are not perfectly crafted. In particular, the ones that include hard-coded lists (such as the list of non-child-friendly words) may not be exhaustive, and the ones that make use of automatic tools (such as POS tagging) are not error-free. In particular, while the neural models used to annotate humour and metaphoric language offer a proxy of the phenomena that is consistent between texts, their annotations do not necessarily match human opinion.

References

Explosion AI. 2025. spacy: Facts & figures. https://spacy.io/usage/facts-figures. Accessed July 2026.

Clémentine Beauvais. 2023. *Écrire comme une abeille : la littérature jeunesse, de la lecture à l'écriture*. Gallimard Jeunesse, Paris.

Douglas Biber. 1988. *Variation Across Speech and Writing*. Cambridge University Press, Cambridge.

Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research. *International Journal of Applied Linguistics*, 165(2):97–135.

Tovly Deutsch, Masoud Jasbi, and Stuart M. Shieber. 2020. Linguistic features for readability assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics.

William H. DuBay. 2004. *The Principles of Readability*. Impact Information, Costa Mesa, CA.

Maciej Eder, Jan Rybicki, and Mike Kestemont. 2015. Stylometry with r: A package for computational text analysis. *The R Journal*, 8:107–121. Https://doi.org/10.32614/RJ-2016-007.

Helmut Anthony Hatzfeld. 1953. *A Critical Bibliography of the New Stylistics Applied to the Romance Literatures, 1900–1952*. Number 5 in University of North Carolina Studies in Comparative Literature. University of North Carolina Press, Chapel Hill.

Rebecca M. M. Hicke and David Mimno. 2023. T5 meets tybalt: Author attribution in early modern english drama using large language models.

Michael Hoey. 1983. *On the Surface of Discourse*. George Allen & Unwin, London.

Baixiang Huang et al. 2024. Authorship attribution in the era of large language models. In *Proceedings of the ACM Conference*.

Linda Hutcheon. 2006. *A Theory of Adaptation*. Routledge, New York.

Michaela Mahlberg. 2013. *Corpus Stylistics and Dickens's Fiction*. Routledge, London.

Thomas Corwin Mendenhall. 1901. A mechanical solution of a literary problem. *Popular Science Monthly*, 60:97–105. Accessed January 2019.

Leo Spitzer. 1948. *Linguistics and Literary History: Essays in Stylistics*. Princeton University Press, Princeton.

Elizabeth Thiel. 2013. Downsizing dickens: Adaptations of oliver twist for the child reader. In Anja Müller, editor, *Adapting Canonical Texts in Children's Literature*, pages 143–162. Bloomsbury Academic, London.

Michael Toolan. 2014. The theory and philosophy of stylistics. In Peter Stockwell and Sara Whiteley, editors, *The Cambridge Handbook of Stylistics*, pages 13–31. Cambridge University Press, Cambridge.

Lennart Wachowiak, Enrica Troiano, and Roman Klinger. 2022. Drum up support: Systematic analysis of image-schematic conceptual metaphors. In *Proceedings of the 3rd Workshop on Figurative Language Processing*, Abu Dhabi. Association for Computational Linguistics.

René Wellek and Austin Warren. 1949. *Theory of Literature*. Harcourt, Brace and Company, New York.

Rodrigo Wilkens, Thomas François, Anaïs Tack, and Cédrick Fairon. 2022. Fabra: French aggregator-based readability assessment toolkit. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1217–1233.

Appendix A Corpus "100 Novels"

#	Title	Author	Year	Original language	Translator
1	Ulysses	James Joyce	1922	English	–
2	In Search of Lost Time	Marcel Proust	1913	French	Charles Kenneth Scott-Moncrieff
3	The Great Gatsby	F. Scott Fitzgerald	1925	English	–
4	The Catcher in the Rye	J. D. Salinger	1951	English	–
5	One Hundred Years of Solitude	Gabriel García Márquez	1967	Spanish	Gregory Rabassa
6	Nineteen Eighty Four	George Orwell	1949	English	–
7	Moby Dick	Herman Melville	1851	English	–
8	Don Quixote	Miguel de Cervantes	1605	Spanish	John Ormsby
9	The Sound and the Fury	William Faulkner	1929	English	–
10	Anna Karenina	Leo Tolstoy	1877	Russian	Constance Garnett
11	Crime and Punishment	Fyodor Dostoevsky	1866	Russian	Constance Garnett
12	Pride and Prejudice	Jane Austen	1813	English	–
13	Lolita	Vladimir Nabokov	1955	English	–
14	War and Peace	Leo Tolstoy	1869	Russian	Aylmer & Louise Maude
15	Wuthering Heights	Emily Brontë	1847	English	–
16	To Kill a Mockingbird	Harper Lee	1960	English	–
17	The Lord Of The Rings	J. R. R. Tolkien	1959	English	–
18	The Brothers Karamazov	Fyodor Dostoevsky	1880	Russian	Constance Garnett
19	The Trial	Franz Kafka	1925	Czech	David Wyllie
20	The Adventures of Huckleberry Finn	Mark Twain	1884	English	–
21	Madame Bovary	Gustave Flaubert	1857	French	Eleanor Marx Aveling
22	The Stranger	Albert Camus	1942	French	Stuart Gilbert
23	The Odyssey	Homer	$\sim$ -700	Ancient Greek	Samuel Butler
24	The Grapes of Wrath	John Steinbeck	1939	English	–
25	The Divine Comedy	Dante Alighieri	1308	Italian	Henry Francis Cary
26	The Magic Mountain	Thomas Mann	1924	German	H. T. Lowe-Porter
27	To the Lighthouse	Virginia Woolf	1927	English	–
28	Jane Eyre	Charlotte Brontë	1847	English	–
29	Middlemarch	George Eliot	1871	English	–
30	Heart of Darkness	Joseph Conrad	1899	English	–
31	Alice's Adventures in Wonderland	Lewis Carroll	1865	English	–
32	Mrs. Dalloway	Virginia Woolf	1925	English	–
33	Catch-22	Joseph Heller	1961	English	–
34	The Master and Margarita	Mikhail Bulgakov	1967	Russian	Michael Glenny
35	Invisible Man	Ralph Ellison	1952	English	–
36	The Iliad	Homer	$\sim$ -700	Ancient Greek	Alexander Pope
37	Les Misérables	Victor Hugo	1862	French	Isabel Florence Hapgood
38	Great Expectations	Charles Dickens	1860	English	–
39	The Red and the Black	Stendhal	1830	French	Horace Barnett Samuel
40	Frankenstein	Mary Shelley	1818	English	–
41	On the Road	Jack Kerouac	1957	English	–
42	The Little Prince	Antoine de Saint-Exupéry	1943	French	Katherine Woods
43	Absalom, Absalom!	William Faulkner	1936	English	–
44	One Thousand and One Nights	unknown	$\sim$ 800	Arabic	Edward William Lane
45	Journey to the End of The Night	Louis-Ferdinand Céline	1932	French	Ralph Manheim
46	Lord of the Flies	William Golding	1954	English	–
47	Brave New World	Aldous Huxley	1932	English	–

#	Title	Author	Year	Original language	Translator
48	David Copperfield	Charles Dickens	1849	English	–
49	The Old Man and the Sea	Ernest Hemingway	1952	English	–
50	Beloved	Toni Morrison	1987	English	–
51	Gone With the Wind	Margaret Mitchell	1936	English	–
52	Animal Farm	George Orwell	1945	English	–
53	The Life And Opinions Of Tristram Shandy	Laurence Sterne	1759	English	–
54	The Idiot	Fyodor Dostoevsky	1869	Russian	Eva Martin
55	Dracula	Bram Stoker	1897	English	–
56	Under the Volcano	Malcolm Lowry	1947	English	–
57	The Leopard	Giuseppe Tomasi di Lampedusa	1958	Italian	Archibald Colquhoun
58	The Sun Also Rises	Ernest Hemingway	1926	English	–
59	The Golden Notebook	Doris May Lessing	1962	English	–
60	Things Fall Apart	Chinua Achebe	1958	English	–
61	Gulliver's Travels	Jonathan Swift	1726	English	–
62	Rebecca	Daphne du Maurier	1938	English	–
63	The Aeneid	Virgil	∼ -19	Latin	John Dryden
64	Robinson Crusoe	Daniel Defoe	1719	English	–
65	Midnight's Children	Salman Rushdie	1981	English	–
66	The Scarlet Letter	Nathaniel Hawthorne	1850	English	–
67	Their Eyes Were Watching God	Zora Neale Hurston	1937	English	–
68	A Farewell to Arms	Ernest Hemingway	1929	English	–
69	The Metamorphosis	Franz Kafka	1915	Czech	David Wyllie
70	The Castle	Franz Kafka	1926	Czech	Anthea Bell
71	A Passage to India	E. M. Forster	1924	English	–
72	The Plague	Albert Camus	1947	French	Robin Buss
73	The Portrait of a Lady	Henry James	1881	English	–
74	Candide	Voltaire	1759	French	William F. Fleming
75	Slaughterhouse-Five	Kurt Vonnegut	1969	English	–
76	As I Lay Dying	William Faulkner	1930	English	–
77	A Portrait of the Artist as a Young Man	James Joyce	1916	English	–
78	All Quiet on the Western Front	Erich Maria Remarque	1928	German	A. W. Wheen
79	The Count of Monte Cristo	Alexandre Dumas	1844	French	unknown
80	The Man Without Qualities	Robert Musil	1930	German	Burton Pike
81	Little Women	Louisa May Alcott	1868	English	–
82	The Good Soldier	Ford Madox Ford	1915	English	–
83	The Picture of Dorian Gray	Oscar Wilde	1890	English	–
84	Emma	Jane Austen	1815	English	–
85	Buddenbrooks	Thomas Mann	1901	German	H. T. Lowe-Porter
86	For Whom the Bell Tolls	Ernest Hemingway	1940	English	–
87	Demons	Fyodor Dostoevsky	1872	Russian	Richard Pevear & Larissa Volokhonsky
88	The Age of Innocence	Edith Wharton	1920	English	–
89	Orlando	Virginia Woolf	1928	English	–
90	Dune	Frank Herbert	1965	English	–
91	Native Son	Richard Wright	1940	English	–
92	The Unbearable Lightness of Being	Milan Kundera	1984	Czech	
93	Dead Souls	Nikolai Gogol	1842	Russian	D. J. Hogarth
94	Doctor Zhivago	Boris Pasternak	1957	Russian	Max Hayward & Manya Harari
95	The Bell Jar	Sylvia Plath	1963	English	–
96	Charlotte's Web	E. B. White	1952	English	–
97	In Cold Blood	Truman Capote	1966	English	–
98	Treasure Island	Robert Louis Stevenson	1883	English	–

#	Title	Author	Year	Original language	Translator
99	Vanity Fair	William Make-peace Thackeray	1847	English	–
100	The Heart Is A Lonely Hunter	Carson McCullers	1940	English	–

Appendix B Corpus Robustness Analysis

We conducted a robustness analysis in order to measure the influence of the variables of time and original language of writing in relation to our corpus. For this purpose, we removed the following from the original corpus: all works whose original language of writing was not English (35 in number), and all works composed outside the years 1860–1970 (14 in number; the time span was selected based on the general timeline of literary modernism while ensuring that the resulting corpus remained approximately half the size of the original corpus). The resulting corpus consists of 49 works. For each linguistic characteristic used in the project, we compared its mean value in the full and restricted corpora. We report the following information: the percentage change between the two means and the Hedges' g (a standardised measure of the magnitude of the difference).

Statistic	Absolute % change	Absolute Hedges' g
Mean	6.4576	0.1196
Median	3.9748	0.0656
Standard deviation	8.4231	0.1427
Minimum	0.0000	0.0000
Maximum	65.9864	0.4450

Table 4: Summary of the absolute changes in feature-wise statistics between the full and restricted corpora.

Overall, the restricted corpus shows only limited value differences compared to its full counterpart. The mean and median absolute Hedges' g values (0.120 and 0.066, respectively) are both below the conventional threshold for a small effect ($g = 0.20$). Even the largest observed effect ($g = 0.445$) is below the 'medium' threshold. The largest differences are associated with the following features: humour, adverbials of place and time and length of first sentence.

Appendix C Quantitative Linguistic Features

Verb tense and aspect
Past tense
Present tense
Perfect aspect
Present participles
Past participles

Adverbials and questions
Place adverbials
Time adverbials
Direct wh-questions

Pronouns and reference
Pronouns, total
First-person pronouns
Second-person pronouns
Third-person pronouns
Demonstrative pronouns
Indefinite pronouns
Demonstrative determiners

Nominal and verbal form
Nominal suffixes
Gerundive nominals in *-ing*
Words differing from lemma
Infinitives

Voice and predication
By-passives
Agentless passives
Copular constructions
Existential constructions

Complementation and relatives
That-complements after verbs
That-complements after adjectives
Wh-free relatives
Sentence relatives

Subordination
Causal subordinators
Concessive subordinators
Conditional subordinators
Other subordinators

Modification and phrase structure
Total prepositional phrases
Attributive adjectives
Predicative adjectives
Total adverbs
Phrasal coordination
Clausal coordination

Lexical and grammatical categories
Punctuation
Verbs
Nouns
Proper nouns
Content-to-function-word ratio

Lexical richness and rarity
Type-token ratio (words)
Type-token ratio (lemmas)
Hapax legomena (lemmas)
Out-of-vocabulary words

Readability and length
Mean word length
Words per sentence
Syllables per sentence
Letters per word

Syllables per word
Number of tokens*
Number of sentences*
Words in first sentence*
Characters in first sentence*

Stance and discourse markers
Conjuncts
Downtoners
Hedges
Amplifiers
Emphatics
Discourse particles

Modality and reporting
Possibility modals
Necessity modals
Predictive modals
Public reporting verbs
Private cognition verbs
Suasive/directive verbs
Seem/appear

Negation
Analytic negation
Synthetic negation

Narrative and discourse structure
First-person narration
Dialogue
Indirect speech
Fragments without verbs
Non-finite fragments

Semantic and lexical content
Death-related words
Sex-related words
Total flagged words
Concreteness coverage
Mean concreteness
Concreteness sum

Other higher-level features
Metaphor use (%)
Mean metaphor probability
Metaphorical words total
Humor probability
Non-humor probability
Humor label

Figure 1: The quantitative linguistic features measured in the context of our corpus-based analysis. All features except those marked with * were measured per 100 textual tokens and are represented by average, minimum, maximum and standard deviation values within the given text.

Appendix D Neural Models Benchmarking

We performed manual benchmarking in relation to metaphor and humour detection as neural-based features that are of particularly subjective nature and therefore challenging to quantify. Our main goals were to test the models' consistency and their performance on literary text. For the metaphor detection model, we manually annotated a short (380-word) extract and compared it with the model's output. Our gold standard consisted of 14 metaphors, whereas the model identified 23, 6 of which coincided with our annotations. We determined that the model's threshold is rather low in that it marks a large number of commonly used fixed expressions as metaphors (e.g. "plenty of time", "to make out", "to have nothing to do with"). For the humour detection model, we manually annotated 20 random 100-token extracts. We identified 7 humorous extracts, while the model identified 13, with 7 false positives and 1 false negative. In its clear bias toward the positive label, it tended to mark dialogue, casual and adult language as humorous. Ultimately, we concluded that despite the models' biases, they provide valuable linguistic information for our project's purposes due to their highly consistent identification tendencies.

Appendix E Representative Characteristics

Book	Feature	Value	Mean	SD	Z
The Picture of Dorian Gray	Demonstrative pronouns (sd)	1.16	0.85	0.13	2.38
	Past tense (sd)	4.91	3.88	0.46	2.26
	Predictive modals (max)	8.00	5.24	1.29	2.14
	Copular constructions (max)	11.00	7.77	1.61	2.00
	Present tense (sd)	4.23	3.24	0.51	1.96
Moby Dick	Seem/appear (max)	10.00	4.51	1.27	4.33
	Demonstrative determiners (avg)	1.11	0.57	0.19	2.87
	By-passives (max)	4.00	2.35	0.63	2.64
	Attributive adjectives (avg)	5.98	3.83	0.90	2.40
	Non-finite fragments (max)	26.00	10.55	7.05	2.19
Anna Karenina	Private cognition verbs (max)	12.00	6.43	1.74	3.19
	Metaphor use (%) (max)	36.84	27.49	4.07	2.30
	Present participles (avg)	0.91	0.52	0.18	2.18
	Verbs (avg)	148.10	59.10	42.42	2.10
	Causal subordinators (max)	5.00	3.77	0.65	1.90
Les Misérables	Demonstrative pronouns (max)	20.00	8.51	1.97	5.82
	Nouns (avg)	305.94	70.93	52.48	4.48
	Time adverbials (max)	12.00	5.64	1.54	4.13
	Amplifiers (max)	12.00	5.84	1.55	3.97
	Out-of-vocabulary words (max)	84.00	35.20	12.94	3.77
The Odyssey	Words per sentence (min)	12.50	5.75	1.72	3.92
	Conjuncts (avg)	0.41	0.14	0.10	2.73
	Other subordinators (avg)	0.95	0.68	0.12	2.20
	Phrasal coordination (avg)	4.16	2.84	0.61	2.18
	Downtoners (max)	1.00	2.59	0.75	-2.13

Table 5: Most strongly deviating features for the five investigated texts. The columns denote the features' values for the text in question, their mean values and standard deviation for the corpus and the calculated z-values. Numbers are rounded to two decimal points.

Appendix F Lexical Analysis

Work	#	Word	Overuse (%)
Moby Dick	1	whale	71434
	2	aye	8214
	3	mast	6744
	4	sir	4438
	5	mates	3287
	6	rigging	2848
	7	carpenter	2743
	8	jaw	2274
	9	aloft	2150
	10	aft	2113
The Picture of Dorian Gray	1	duchess	7025
	2	sir	3971
	3	painter	2278
	4	portrait	2148
	5	coloured	2113
	6	gilt	1981
	7	studio	1980
	8	hideous	1844
	9	monstrous	1724
	10	valet	1681
Anna Karenina	1	countess	4268
	2	princess	2737
	3	marsh	2323
	4	divorce	2300
	5	everyone	2164
	6	peasants	1883
	7	yes	1665
	8	coachman	1487
	9	peasant	1458
	10	onto	1428
Les Mis-érables	1	sous	3841
	2	convict	3031
	3	mayor	2423
	4	bishop	2034
	5	grating	1854
	6	sir	1753

Work	#	Word	Overuse (%)
	7	a	1686
	8	reverie	1495
	9	indescribable	1323
	10	sombre	1321
The Odyssey	1	raft	3862
	2	meats	3754
	3	maids	3152
	4	groaning	2769
	5	mast	2680
	6	goddess	2651
	7	armour	2546
	8	offerings	2446
	9	gods	2213
	10	barley	2105

Table 6: Top 10 most overused words in the five investigated works as compared against the Project Gutenberg 10k frequency list. All values are rounded to whole numbers.

Imploring you to investigate more languages: A corpus-based contrastive study of Ukrainian and English causative constructions

Iryna Karamysheva
Lviv Polytechnic National University
iryna.d.karamysheva@lpnu.ua

Jakob Horsch
Slovak Academy of Sciences
jakob.horsch@korpus.sk

Abstract

This study provides a corpus-based contrastive analysis of causative constructions in English and Ukrainian of the type S1 FORCES/CAUSES P1 [S2 P2/V_{inf}]], which are simultaneously secondary predication constructions. Defining 'causative construction' as a form-function pairing where a primary subject (S1) exerts a causal force on a secondary subject (S2) to perform an action (P2), the research moves beyond raw frequency counts to explore the interaction between lexical fillers and constructional slots. Our corpus study is based on 247,737 tokens from COCA and 221,132 tokens from GRAC. Using Simple Collexeme Analysis and employing log-odds ratio as an association measure, the study addresses two questions: (1) identifying the most prototypical Complement Taking Predicates (P1) in each language, and (2) determining how the analytic character of English vs the synthetic character of Ukrainian shapes a 'causative continuum' via a network of such constructions in contrasted languages.

The findings reveal that English exhibits decentralized prototypicality, i.e. functional leaders coexist with semantically specialized, high-association verbs. Ukrainian, however, shows unified prototypicality with specific lexemes dominating frequency and association strength. Additionally, we find significant cross-linguistic variation: while English uses an objective case for the secondary subject (S2), reflecting the Latin term 'Accusativus cum Infinitivo', Ukrainian remains sensitive to case. The study finds that Ukrainian uses case marking (Accusative vs. Dative) as a structural diagnostic for 'directness', where high-force coercive verbs govern the Accusative, while persuasive verbs tend toward the Dative, underscoring that a language's morpho-logical character can influence its semantic net-work of causation.

Keywords: causative construction, English, Ukrainian, collexeme analysis, corpus study

1 Introduction

In this corpus-based contrastive analysis, we investigate non-finite complement constructions with a causative meaning. Widely studied in English, they have been described variously, including 'Accusativus cum Infinitivo', 'object control constructions', 'equi-type structures', 'small clauses', and 'periphrastic/analytical causatives'. Causative constructions encode one event as the argument of a second event, hence they have also been called 'secondary predication constructions'. Consider the examples in English (1) and Ukrainian (2), which illustrate the syntactic roles involved:

(1) *When we*$_{S1}$ *forced*$_{P1}$ [*him*$_{S2}$ *to retire*$_{P2}$] […]
(Family Guy, 2011)

(2) *Вони знають, що ніхто*$_{S1}$ *не змусить*$_{P1}$ [*їх*$_{S2}$ *повернути*$_{P2}$_*борги*$_{complement}$]
'They know that nobody will force them to repay their debts.' (https://lb.ua/, 2020)

As the examples (1-2) show, there is a primary predicate (P1) and a secondary predicate (P2) as well as a primary subject (S1), and a secondary subject (S2). This results in two predications. The secondary predication is expressed by an objective-case pronoun or common-case noun in English and an accusative/dative-case pronoun/ noun (Ukrainian) – S2 – in combination with an infinitive (P2). This predication becomes the argument of P1 as a non-finite complement.

In terms of semantics, S1 expresses a causal force/agent while S2 represents a patient that is impelled by force of reason to perform an act P2/V_{inf}. The overall meaning of the construction can thus be summarized as follows: [S1 FORCES (P1) S2 P2 /V_{inf}]. It is noteworthy that the causee S2 simultaneously performs two roles. First, it is a patient under the causal force from agent S1; second, it acts as an agent of the action expressed by P2/V_{inf}. Thus, S2, which can be regarded an object of P1, is simultaneously the logical subject of P2. In the literature, the combination of S2+P2

has been treated as one argument, expressed by the non-finite complement construction (Noonan, 2007: 53; Croft, 2022: 551-558). Hence the predicates (P1) are referred to as 'Complement Taking Predicates' (CTPs).

There is a large body of research on the English causative construction, ranging from studies against the background of mainstream generative grammar (e.g., Comrie, 1976) to non-transformational approaches (e.g., Kim and Michaelis, 2020: 168 ff.). Recently, the causative construction has also attracted attention from constructionist research, which traditionally relies on corpus data (e.g., Egan, 2008; Gilquin, 2010). These studies have investigated periphrastic causative constructions using sophisticated statistical methods that are collectively referred to as collostructional analysis. Developed by Stefanowitsch and Gries (2003; 2008), this method is used to measure the degree of interaction between words and constructions (Gilquin and Gries, 2026). For example, Gilquin (2006) employed multiple distinctive collexeme analyses to investigate the non-finite verb slots of ten English periphrastic causative constructions such as [X CAUSE Y $V_{to\text{-}inf}$] (Gilquin, 2010, 2024). Demonstrating its feasibility for contrastive studies, Gilquin also applied collexeme analysis by comparing the infinitive verb slots of the English causative construction [X MAKE Y V_{inf}] to its French equivalent [X FAIRE V_{inf} Y] (2015, 2024).

In the following, we present a contrastive corpus study of the English causative construction and its Ukrainian counterpart, relying on collostructional analysis to uncover formal differences and similarities. While the Ukrainian causative construction remains largely unexplored, some recent corpus studies have provided a first glimpse (Karamysheva, 2024, 2025). These studies, while limited in scope, have revealed that Ukrainian differs from English in some respects. For example, in English the non-finite V slot P2 can be filled not only by infinitives with and without to, but also with the Present Participle and Past Participle (Gilquin, 2006; Karamysheva, 2025). In Ukrainian, however, the V slot can only be filled by an infinitive.

In the following, we restrict ourselves to the P1/CTP slot, presenting the first-ever contrastive corpus-based investigation of English and Ukrainian causative constructions. Specifically, we will explore so-called causative verbs that fill the CTP slot. Our research questions are, thus, as follows:

1. Which CTPs are most prototypically used in the English and Ukrainian causative constructions?
2. Are there any differences between the two languages, and how can they be explained?

Following a "the more data, the better" approach (Hoffmann et al., 2019), we rely on large datasets extracted from the Corpus of Contemporary American English (COCA) (Davies, 2008) and the General Regionally Annotated Corpus of Ukrainian (GRAC) (Shvedova et al., 2017), following the consideration that both are large and representative. Assuming a constructionist approach that treats syntactic patterns that encode specific non-compositional meanings as form-meaning pairings in their own right (cf., e.g., Horsch and Boas, 2025: 16), we employ the term 'causative construction' in the following.

2 Methodology

2.1 Data preparation

First, we used synonymic expansion to compile lists of verbs that can be used as CTPs in the English and Ukrainian causative constructions. For English, we additionally relied on a variety of sources (e.g., Quirk et al., 1985; Noonan, 2007; Egan, 2008). The list of Ukrainian causative CTPs was compiled using the 'synonym function' of the online resource slovnyk.me[1], which is based on authoritative dictionaries of the Ukrainian language. We had to rely on this procedure because to our knowledge, there are no ready-made lists of CTPs that trigger this type of non-finite complement construction in Ukrainian. For both languages, we cross-checked the corpus data whether the selected verbs fit to the syntactic pattern N_1 V N_2 non-f V (N_3).

Our choice of corpora follows from the fact that Ukrainian, unlike English, is a low-resource language, with a very limited choice of corpora (as noted by Shvedova and Lukashevskyi, 2024: 14). Thus, while we acknowledge that a balanced, bidirectional corpus would be ideal for providing a *tertium comparationis* (TC) (as suggested by Hasselgård 2020), we settled on COCA and

[1] https://slovnyk.me/dict/synonyms (last accessed 13.03.2026)

GRAC, following Karamysheva, who has noted that "two separate corpora […] can be used to harvest data for contrastive analysis, provided that there is a correct choice of [TC], capturing a certain pattern, and reflected in the query built" (2024: 1). Henceforth we shall, therefore, use the constructionist notion of 'construction' (see above) to serve as a TC, allowing us to compare English and Ukrainian regarding CTPs.

Our queries differed slightly between the corpora: For COCA, we used the features offered by the online interface at english-corpora.org; for GRAC we used CQL. The search of causative constructions with specific CTPs presupposed two queries for each CTP, since the non-finite complement contains a causee/patient (S2) expressed either by a pronoun in the objective case or a noun in the common case (in English) or pronoun/noun either in the accusative or dative case, depending on the CTP (in Ukrainian). Consider the queries and the examples of constructions obtained (3-6):

(3) CAUSE PRON TO _v?i
Every talk has caused me to think about some topic a new way.
(BLOG http://lfb.org/today/the-center-of-the-conspiracy/, 2012)

(4) ENCOURAGE NOUN TO _v?i
School librarians encourage students to consider diverse and global perspectives
(ACAD: Knowledge Quest, 2015)

(5) [lemma="наказувати"] [tag=".*pron.*"] [tag=".*inf.*"]
[…] *поліція наказує їм залишатися на місці.*
'[…] the police order them to stay in place.'
(https://www.eurointegration.com.ua/, 2019)

(6) [lemma="радити"] [tag=".*noun.*" & tag!=".*pron.*"] [tag=".*inf.*"][2]
Через це експерти з кібербезпеки радять людям переходити на інші застосунки.
'Because of this, cybersecurity experts are advising people to switch to other apps.'
(https://nv.ua/, 2021)

In total, we extracted 27 English and 23 Ukrainian CTPs with semantics typically associated with the causative construction. In each case, the token frequencies of both queries were summed up to provide the general frequency of the construction per corpus (this value is required for collostructional analysis, see section 2.3 below). Moreover, to investigate whether the 'causative continuum' differs between the two languages (cf. Talmy, 1988; Comrie, 1989; Shibatani and Pardeshi, 2002), to the effect that the 'force' of causation can be more direct or less direct, as reflected by the semantic spectrum of causative verbs, we subdivided our lists of causatives into two groups: 1) 'coercive' with a more direct causation impact and 2) 'persuasive' with a milder form of causation, expressed by requests, advice or persuasion.

The list of English coercive verbs includes the following 11 CTPs, in order of descending token frequency: *make* (88,541), *force* (21,382), *cause* (14,652), *get* (9,379), *have* (5,370), *bring* (3,422), *compel* (1,915), *drive* (1,580), *command* (749), *press* (428), *oblige* (403). The English persuasive type includes the following 16 CTPs: *ask* (48,523), *encourage* (15,936), *lead* (10,993), *urge* (8,137), *beg* (3,047), *prompt* (2,910), *advise* (2,881), *persuade* (2,874), *challenge* (2,162), *dare* (1,137), *implore* (445), *influence* (347), *recommend* (187), *stimulate* (187), *beseech* (137), *plead* (013).

The list of Ukrainian coercive verbs includes 7 CTPs: *примусити*|*змусити* 'make, force' (52,473); *казати* 'force, oblige' (2,830); *велiти* 'command, order' (2,580); *наказувати* 'order' (2,145); *заставляти* 'force' (678); *силувати* 'force' (239); *приневолити* 'force' (173). The list of Ukrainian persuasive verbs is longer and includes 16 CTPs: *попросити*|*просити* 'ask' (86,354); *пропонувати*|*запропонувати* 'advise' (38,889); *порекомендувати*|*рекомендувати* 'recommend' (9,315); *радити*|*порадити* 'advise' (8,005); *спонукати* 'urge' (7,931); *умовляти*|*вмоляти* 'persuade' (2,052); *заохочувати* 'encourage' (1,753); *стимулювати* 'encourage' (1,670); *благати* 'beg, beseech' (1,597); *переконувати* 'persuade' (1,474); *підштовхувати* 'urge, encourage' (331); *підбурювати* 'persuade, coerce' (212); *схиляти* 'persuade, incline' (186); *підбивати* 'persuade, coerce' (130); *заклинати* 'beg, beseech' (58); *раяти*|*нараяти* 'advise, recommend' (57).

Note that some Ukrainian CTPs are treated as pairs, following two considerations: First, some CTPs (e.g.

[2] Note that the query for Ukrainian constructions with pronouns as S2 (Example (5), [lemma="наказувати"][tag=".*pron.*"] [tag=".*inf.*"]) yielded clean results, whereas the query with the noun tag, e.g. [lemma="наказувати"] [tag=".*noun.*"] [tag=".*inf.*"] returned sentences intermixed with pronouns, which called for the manual sorting out of such cases. Therefore, a more specialized query, cf. example (6), was required.

попросити|*просити* 'ask') are aspectual pairs of the same lexeme; second, some are allomorphic variants (e.g. *примусити*| *змусити* 'make, force'), which may be additionally treated as interchangeable synonyms.

2.2 Data extraction

To address the research questions, we annotated data from two corpora, COCA 2019[3] and GRAC v. 19a[4]. Table 1 provides an overview of the corpora and how many tokens were retrieved.

Corpus	Corpus size (tokens)	No. of hits (tokens)
COCA	1,001,610,938	247,737
GRAC	2,023,921,236	221,132

Table 1: Overview of corpora and tokens

The extracted data were saved in spreadsheets and annotated in *Libre Office Calc* (The Document Foundation, 2025) for the following variables:

- WORD (levels: e.g. *make, have…*)
- FREQUENCY WORD IN CORPUS
- FREQUENCY WORD IN CONSTRUCTION

These variables are required as input for collexeme analysis, which we discuss in the next section.

2.3 Data analysis: Collexeme analysis

We use statistical tests to explore semantic tendencies in the CTP slot of the English and Ukrainian causative constructions. Specifically, we employ simple collexeme analysis (Stefanowitsch and Gries, 2003; Stefanowitsch and Gries, 2008; Stefanowitsch 2013), which "identifies association patterns [...] between lexical items and syntactic constructions" (Hartmann and Ungerer, 2023: 608). Put succinctly, collexeme analysis "identifies lexemes that occur with above-chance frequency in a construction"; these are referred to as "attracted collexemes" (Hartmann and Ungerer, 2023: 608). Simple collexeme analysis works by calculating expected frequencies on the basis of contingency tables and applying "inferential statistical tests [to determine] the statistical significance of the observed distribution" (Stefanowitsch, 2014: 221).

Simple collexeme analysis outputs a list that shows "which verbs are typical for a given construction" (Stefanowitsch, 2014: 222). Following the assumption that "the verbs most typically occurring with a given construction reflect its meaning", this allows us to draw conclusions about the construction itself (Stefanowitsch, 2014: 222), allowing us to address the research questions. CTPs that turn out to be 'attracted' to the causative construction are referred to as 'collexemes'.

To carry out collexeme analysis, we employed Gries' (2024) *Coll.analysis_4.1* script for *R* (R Core Team, 2020). As association measure (AM)[5], we rely on the log odds (henceforth referred to as 'log-odds'), which has been described as the "gold standard […] for association" (Gries, 2022: 27). The main advantage of log-odds is that, unlike other commonly used measures (e.g. G^2 or t), it "behaves exactly like a measure that was designed to control frequency" (Gries, 2022: 26). In fact, apart from a results table in which collexemes are ranked by log-odds, *Coll.analysis_4.1* conveniently also outputs a graph plotting log-odds (the AM) against the y-axis and frequency against the x-axis (as logged co-occurrence frequencies).

The importance of distinguishing between AM and frequency will become more apparent when we turn to the results in section 3. For now, suffice it to say that corpus data can and do show all four possible combinations (as noted by Gries, 2022: 30):

- high frequency and high association
- high frequency and low association
- low frequency and high association
- low frequency and low association

Obviously, the fact that all combinations are possible has implications for interpreting corpus data, because it demonstrates that frequency or AMs by themselves do not necessarily correlate and therefore cannot be used for drawing valid conclusions about the relationship between a word and the construction in whose context it is examined (Gries, 2022). Rather, AMs and frequency should complement each other, to the effect that the latter is used to contextualize the former (Gries, 2022: 13).

[3] https://www.english-corpora.org/coca/
[4] https://uacorpus.org/
[5] We acknowledge that recently, there has been a trend towards "tupleization" (Gries, 2026), i.e., an "emerging consensus among users of collocation-based methods to ideally draw on a combination of different measures" (Hartmann and Ungerer, 2023: 609). Due to the space limitations of this publication, we have decided to restrict ourselves to log-odds.

3 Results

3.1 English

First, we present the results for English. Table 2 shows the results of simple collexeme analysis, including frequency counts. We see that of all CTPs, *implore* and *encourage* (log-odds: 7.09) are semantically most strongly attracted to the English causative construction, followed by *urge* (log-odds: 7.04). At the bottom of Table 1, least strongly attracted, we find *have* (log-odds: 0.58), *plead* (log-odds: 1.06) and *press* (log-odds: 2.31).

Word	Freq. in cxn	Freq. not in cxn	Rel.	Log-odds
implore	445	1497	attr.	7.09
encourage	15936	57316	attr.	7.09
urge	8137	29903	attr.	7.04
persuade	2874	14355	attr.	6.71
compel	1915	12227	attr.	6.46
beseech	137	876	attr.	6.45
beg	3047	22070	attr.	6.34
prompt	2910	21962	attr.	6.30
advise	2881	24699	attr.	6.17
ask	48523	628216	attr.	5.96
force	21382	296333	attr.	5.77
oblige	403	6261	attr.	5.56
make	88541	2207658	attr.	5.53
cause	14652	289988	attr.	5.38
dare	1137	25743	attr.	5.19
lead	10993	334528	attr.	4.93
stimulate	187	10726	attr.	4.26
challenge	2162	145204	attr.	4.11
command	749	50026	attr.	4.11
bring	3422	436031	attr.	3.47
drive	1580	252442	attr.	3.24
influence	347	104479	attr.	2.60
recommend	187	56751	attr.	2.59
get	9379	3341902	attr.	2.46
press	428	171233	attr.	2.31
plead	13	18116	attr.	1.06
have	5370	12310549	attr.	0.58

Table 2: Collexeme analysis results, English

Figure 1 shows that AM and frequency do not necessarily correlate, and that a complementary approach is therefore warranted (see discussion in section 2.2). For example, with 445 tokens, *implore* has a considerably lower frequency than *have* (5,370 tokens). Nevertheless, with a log-odds of 7.09, *implore* is much more strongly associated with the causative construction than *have* (log-odds: 0.58). Returning to the plot in Figure 1, this notion becomes clear. Recall that log-odds is plotted against the *y*-axis and logged co-occurrence frequency against the *x*-axis. This means that generally speaking, the more to the top right of the semantic space we find a CTP, the more prototypical it can be considered for the construction. Conversely, the more to the bottom left of the plot a word is found, the less prototypical it is for the construction. Thus, while *implore* has the highest AM (log odds: 7.09), it actually has a considerably lower frequency (445 tokens), making it less prototypical than e.g. *ask* (48,523 tokens, log-odds: 5.96) and *make* (88,541 tokens, log-odds: 5.53).

More prototypical CTPs include *encourage* (15,936 tokens, log-odds: 7.09), *force* (21,382 tokens, log-odds: 5.77) and *cause* (14,652 tokens, log-odds: 4.38). Conversely, *plead*, with a low frequency (13 tokens) and a low AM (log-odds: 1.06) may be considered the least prototypical. Finally, as discussed in section 2, there is also the possibility of high frequency combined with low association, such as *have* (5,370 tokens, log odds: 0.58). The plot also shows what might be termed clusters[6] of CTPs, for example *persuade, beg, prompt, advise*.

Finally, when considering the 'big picture', the plot reveals a rather heterogenous distribution, with a considerable spread across the *y*- and *x*-axes. Thus, there is quite a large difference between the lowest (*have*, 0.58) and highest (*implore,* 7.09) log-odds, making for a range of 6.51. This stands in contrast with the Ukrainian results, a point to which we return in section 3.2 below. While our range of causative verbs differs from that used in previous research, making comparisons difficult, we would like to note Gilquin's (2010) corpus study of English periphrastic causative constructions (with e.g. *make, cause* and *have*), which employed a related statistical method, multiple distinctive collexeme analysis. Gilquin found that the pattern [X *make* Y V_{inf}] turns out to be the most flexible, attracting the highest number of distinctive features in comparison to other verbs.

[6] By 'clusters' we mean CTPs in close proximity to each other on the plot, i.e., with similar frequencies (*x* axis) and log-odds values (*y* axis).

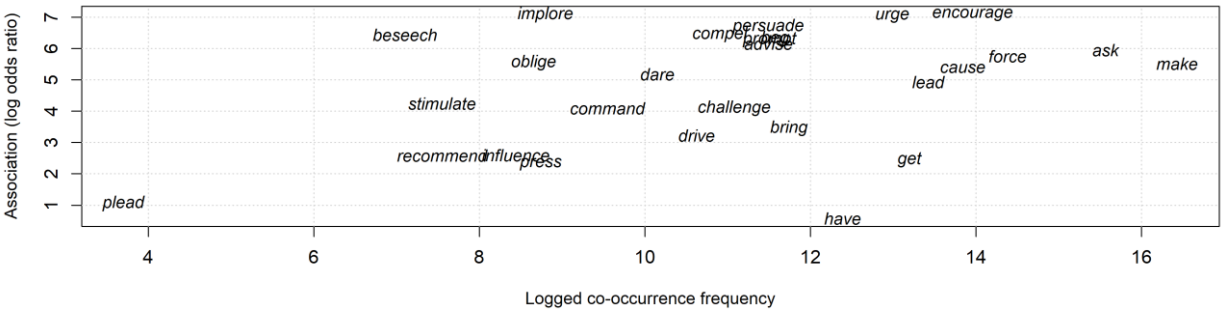


Figure 1: Plot of CTPs in the English causative construction, COCA

She speculates that "people associate [X *make* Y V_{inf}] [as] the most frequent and flexible periphrastic causative construction, with inducive causation because it is a meaning that is somehow more salient in people's minds than the other types of causation" (Gilquin, 2010: 142). In this context, it is intriguing that *make* is, by far, the most frequent verb in our data, with 88,541 tokens (cf. Table 2). However, it does not exhibit the highest AM (log-odds: 5.53), occupying more of a middle position, an observation we believe can be explained with its polysemous nature.

3.2 Ukrainian

Turning to the Ukrainian results (Table 3 and Figure 2), we see some striking differences to the English results from section 3.1. When considering the plot, we see a more homogenous distribution, with more spread along the *x*-axis but not as much over the *y*-axis. This is also evident from the fact that in the Ukrainian data, the log-odds values (plotted against the *y*-axis) range from 3.17 (*казати* 'say' used in the meaning 'to order') to 9.10 (*примусити*|*змусити* 'force, compel'), a difference of 5.93.

However, the lower bound in the Ukrainian data is an outlier (*казати*), without which the range would be just 3.68. What this suggests is that despite a spread in terms of frequency that is comparable to English, the attraction of CTPs to the causative construction in Ukrainian is more uniform, indicating less variability. That is, frequency does not appear to interact much with the AM, which could be interpreted as individual CTPs being rather equal in terms of how they are used in the Ukrainian causative construction. Turning to the individual CTPs, we see that with a high frequency and a high AM, *примусити*|*змусити* 'force, compel' (52,473 tokens, log-odds: 9.10) and *попросити*|*просити* 'ask, request' (86,354 tokens, log-odds: 7.50) are most prototypical for the Ukrainian causative construction.

Word	Freq. in cxn	Freq. not in cxn	Rel.	Log-odds
примусити| *змусити*	52473	70105	attr.	9.10
приневолити	173	524	attr.	8.01
спонукати	7931	27360	attr.	7.92
умовляти| *вмовляти*	2052	9146	attr.	7.64
заставляти	678	3028	attr.	7.63
попросити| *просити*	86354	720449	attr.	7.50
веліти	2580	13606	attr.	7.47
наказувати	2145	15668	attr.	7.14
силувати	239	1770	attr.	7.12
заохочувати	1753	13453	attr.	7.09
порекомендувати |*екомендувати*	9315	82904	attr.	6.98
пропонувати| *запропонувати*	38889	559252	attr.	6.65
благати	1597	20404	attr.	6.58
раяти|*нараяти*	57	726	attr.	6.58
порадити| *радити*	8005	107603	attr.	6.56
підбурювати	212	3138	attr.	6.43
стимулювати	1670	27316	attr.	6.33
підштовхувати	331	7022	attr.	6.07
заклинати	58	1307	attr.	6.01
переконувати	1474	39078	attr.	5.85
підбивати	130	4263	attr.	5.63
схиляти	186	7522	attr.	5.42
казати	2830	1105235	attr.	3.17

Table 3: Collexeme analysis results, Ukrainian

With low frequencies and (relatively) low AMs, the least prototypical CTPs are *раяти*|*нараяти* 'advise, recommend' (57 tokens, log-odds: 6.58) and *заклинати* 'beg, beseech' (58 tokens, log-odds: 6.01). These results may be due to the fact that *раяти*|*нараяти* 'advise, recommend' is considered rather dialectal and antiquated, whereas *заклинати* 'beg, beseech' is more characteristic of literary style than everyday conversation. Again, note that these CTPs are found in the middle of the table, making them rather inconspicuous. Only

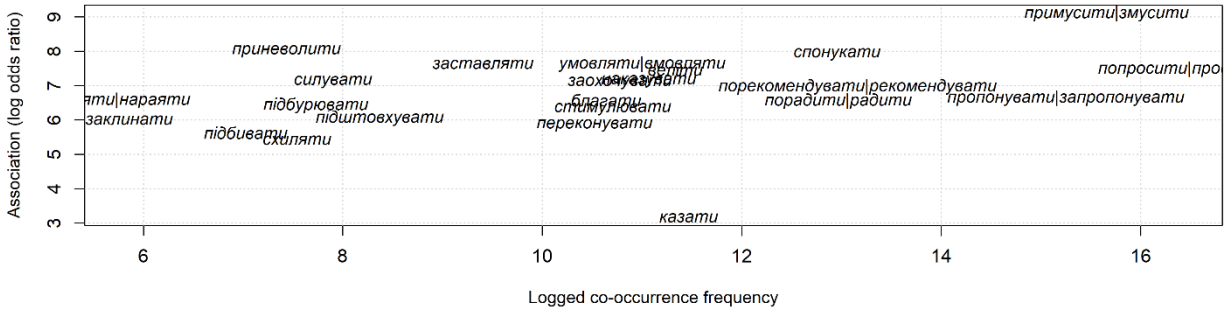


Figure 2: Plot of CTPs in the Ukrainian causative construction, GRAC

when taking into account their low frequencies can we appreciate their comparatively 'weak' prototypicality, underscoring the importance of the complementary approach, i.e., contextualizing association measures with frequencies (see section 2.2).

4 Discussion

The results of our simple collexeme analyses reveal similarities and differences between English and Ukrainian, with the main difference being that causative verbs prioritize different types of CTPs as prototypes. Thus, regarding research question 1, we conclude that in English the verbs *ask* (log-odds: 5.96) and *make* (log-odds: 5.53) are the most prototypical, with high corpus frequencies and high association measures (AM). By both measures, they clearly distinguish themselves from the other verbs. Interestingly, English features many verbs (e.g. *encourage* and *force*) with a vastly higher AM than the more generic *have*. In fact, the low AM of *have* (log-odds: 0.58) indicates what we call a 'polysemous drifter' – with a high overall frequency in the language and not specifically dedicated to the causative construction. In Ukrainian, in contrast, the prototypical 'core' is more compact. *Примусити/змусити* 'force, compel' and *попросити/просити* 'ask, request' emerge as the verbs most attracted to the construction. The high AM (e.g. *примусити*|*змусити* 'force, compel'; log odds: 9.10) suggests that when a Ukrainian speaker produces this construction, s/he is statistically much more likely to rely on a 'highly coercive' causative verb than an English speaker.

Our results also suggest that the 'causative continuum' – the scale from direct coercive manipulation to indirect persuasion – grades differently in English and Ukrainian. English displays heterogeneity and a high degree of lexical specialization; the wide range of log-odds (6.51) and the 'semantic clusters' (e.g. around *persuade* and *beg*, cf. Figure 1) indicate a rather fragmented continuum. Ukrainian, on the other hand, presents a more homogenous distribution. Excluding the outlier *казати* 'say', the AM range is significantly more restricted than in English. Moreover, the consistently high AM values in Ukrainian suggest that the construction 'attracts' its verbs/CTPs more strongly, leading to a more uniform attraction across the investigated lexicon. Regarding the second research question, we can, thus, conclude that differences outweigh similarities in the distributional behavior, even if the semantic intent remains the same with 16 CTPs in each language of the persuasive subtype.

On a final note, our Ukrainian data shows that some CTPs use dative case on the secondary subject S2, warranting the label Dativus cum Infinitivo (DcI). This concerns 3 out of 7 coercive CTPs (*казати* 'order', *велiти* 'command/order', *наказувати* 'order') and 4 out of 16 persuasive CTPs (*пропонувати*| *запропонувати* 'offer', *порекомендувати*| *рекомендувати* 'recommend', *радити* 'advise', *раяти*|*нараяти* 'advise, recommend'). It, thus, appears that in Ukrainian, in contrast with English, morphology/case is sensitive to verb semantics ('force dynamics' in Talmy's (1988) sense.). This is in line with previous research in Slavonic languages on object control constructions that prefer dative case (e.g., Dziwirek, 1998). In fact, there is a substantial body of research devoted to causation in different languages, impossible to cite in its entirety without going beyond the scope of this paper. These include studies with a focus on the typology of causatives, including their form, syntax and meaning, e.g. Dixon, who remarks that "languages differ a great deal in the syntax of their causatives and in the specific meanings attached to them" (2000: 33). Shibatani and Pardeshi note that "when a language allows alternative marking of the causee nominal, the accusative version generally conveys sociative causation, whereas the dative or

other oblique marking signals indirect causation" (2002: 99-100). Since it is the Ukrainian language that is understudied regarding typology, we consider the results obtained in this study of great value.

In this context, it can be stated that in Ukrainian causative constructions the accusative case marks direct impact, while dative marks interpersonal communication. Both retain 'complex transitivity' because they both follow this pattern, but they use case as a precision tool to define the relationship between the primary S1 and secondary S2 subjects. Ukrainian CTPs of the 'accusative cluster' (e.g. *примусити/змусити* 'make, force', *заставляти* 'force' and *силувати* 'compel') express high coercion/ direct manipulation. They are 'true' causatives in the sense that the causer acts directly on the secondary subject. The high token count for *змусити* 'force, compel' suggests this is the prototypical periphrastic causative in Ukrainian, much like *make* in English. The 'dative cluster' of coercive verbs can be regarded as less directive/more social (i.e., indirect manipulation). CTPs like *казати* (here in the sense of 'tell, command'), *веліти* 'command' and *наказувати* 'order', encode by means of the dative case that the secondary subject is not just an object being pushed, but an addressee of a speech act who is expected to carry out the secondary predication action. In contrast, the most prototypical persuasive CTP *попросити/просити* 'ask' takes the accusative, which suggests that 'asking' is still viewed as a direct social impact on the person, while the dative persuasives (e.g. *пропонувати* 'offer/advise', *порекомендувати* 'recommend' and *радити* 'advise' indicate a weaker causal link where the secondary subject is a volitional addressee rather than a manipulated object.

5 Conclusion

In this study of the causative construction, we used simple collexeme analysis to analyze 27 causative CTPs in English and 23 causative CTPs in Ukrainian. The results show interesting similarities and differences between the two languages. Most prominently, English appears more heterogeneous in how verbs are used in the construction, whereas Ukrainian turns out to be quite homogeneous.

In more visual terms, English is best described as exhibiting lexical dispersion, with a loose agglomeration of various verbs with weak-to-moderate attraction. Ukrainian, in contrast, features a more uniform and specialized set of CTPs that are more fundamentally attracted to the causative construction. Even mid-to-low frequency verbs in Ukrainian show attraction levels that rival the very highest peaks in English, where we also observe a long tail of verbs with very low association (log-odds <3.0). In contrast, in Ukrainian even the less frequent verbs maintain association strengths that would place them near the very top of the English list. Regarding our differentiation between 'coercive' and 'persuasive' causative CTPs, we observe that in Ukrainian, case marking acts as a structural diagnostic for different levels of 'directness' and 'interpersonal manipulation.' From a methodological perspective, our study underscores the importance of using an association measure such as the log-odds in addition to raw corpus frequencies (cf. section 2.3 and Gries, 2022: 30).

Being a pilot study, our findings are to be considered preliminary, with more extensive research with a focus on Ukrainian certainly being warranted. For example, registers will have to be taken into account, given the results for some verbs that are e.g. dialectal/ antiquated (see section 3.2). On a final note, due to limitations regarding time and resources, we were unable to discuss our results in the context of the stylistic/genre structure of the corpora[7]. However, we are convinced that taking into account the distribution of the studied constructions across different genres of the two corpora constitutes an interesting topic for further research.

Acknowledgments

We would like to thank two anonymous reviewers for their valuable suggestions on how to improve our manuscript. All remaining errors are our own.

Part of the work reported here was carried out in the context of the research project "Weight Effects in English and Slovak" (2024–26), generously funded by the European Union's *NextGenerationEU* program through the Recovery and Resilience Plan for Slovakia (project no. 09I03-03-V04-00713), in the context of which one of the authors was employed.

[7] We would like to express our thanks to an anonymous reviewer for pointing this out.

References

Comrie, Bernard. 1976. The syntax of causative constructions: cross-language similarities and divergences. In M. Shibatani (ed.) *Syntax and Semantics, Vol. 6: The Grammar of Causative Constructions*. New-York, San Francisco, London: Academic Press, pages 261–312.

Comrie, Bernard. 1989. *Language Universals and Linguistic Typology: Syntax and Morphology*. University of Chicago Press.

Croft, William. 2022. *Morphosyntax: Constructions of the World's Languages*. Cambridge: Cambridge University Press.

Davies, Mark. 2008. *The Corpus of Contemporary American English* (COCA): 520 million words, 1990-present. Available at: https://www.english-corpora.org/coca/

Dixon, Robert M.W. (2000). A typology of causatives: form, syntax and meaning. In R. M. W. Dixon and Alexandra Y. Aikhenvald (eds.) *Changing valency: case studies in transitivity*. Cambridge University Press, 2000, pages 30–83.

Dziwirek, Katarzyna. 1998. Reduced Construction in Universal Grammar: Evidence from the Polish Object Control Construction, *Natural Language and Linguistic Theory,* 16: 53–59.

Egan, Thomas. 2008. *Non-finite Complementation. A Usage-based Study of Infinitive and -ing Clauses in English*. Amsterdam-New York.

Gilquin, Gaëtanelle. 2015. Contrastive collostructional analysis: Causative constructions in English and French. *Zeitschrift für Anglistik und Amerikanistik*, 63(3): 253–272.

Gilquin, Gaëtanelle. 2010. *Corpus, Cognition and Causative Constructions*. Studies in Corpus Linguistics, V. 39, John Benjamins.

Gilquin, Gaëtanelle. 2006. The verb slot in causative constructions. Finding the best fit', *Constructions*, SV1. Available at: https://constructions.journals.hhu.de/article/view/445

Gilquin, Gaëtanelle. 2024. Transfer of collostructions: the case of causative constructions, *Corpus Linguistics and Linguistic Theory*, pages 1–21. Available at: https://doi.org/10.1515/cllt-2024-0023

Gilquin, Gaëtanelle and Gries, Stefan. Th. 2026. Collostructional analysis. In H. Nesi and P. Milin (eds.) *International Encyclopedia of Language and Linguistics*, 3rd Edition. Elsevier.

Gries, Stefan Th. 2022. What do (some of) our association measures measure (most)? Association? *Journal of Second Language Studies*, 5(1):1–33.

Gries, Stefan Th. 2024. Coll.analysis 4.1. A script for R to compute perform collostructional analyses.

Gries, Stefan Th. 2026. Tupleization. in H. Nesi and P. Milin (eds.) *International Encyclopedia of Language and Linguistics.* 3rd ed. Amsterdam: Elsevier.

Hartmann, Stefan. and Ungerer, Tobias. (2023) "Attack of the snowclones: A corpus-based analysis of extravagant formulaic patterns," *Journal of Linguistics*, 60(3): 599–634.

Hasselgård, Hilde. 2020. Corpus-based contrastive studies: Beginnings, developments and directions, Languages in Contrast, 20(2): 184–208.

Hoffman, Thomas, Horsch Jakob, and Brunner Thomas. 2019. The More Data, The Better: A Usage-based Account of the English Comparative Correlative Construction. *Cognitive Linguistics*, 30(1): 1–36.

Horsch, Jakob and Boas, Hans C. (2025) Construction Grammar-Based Research on Slovak, in Slovakia: A Case Study and a Road Map. *Slovak Speech*, 90(2): 11–50. Available at: https://doi.org/10.31577/slovrec.2025.2.2

Karamysheva, Iryna. 2024. Challenges of corpus-based contrastive studies: a case study of the English "make"-construction and its Ukrainian counterpart. *IEEE 19th International Conference on Computer Science and Information Technologies (CSIT)*, Lviv, Ukraine, pages 1–4.

Karamysheva, Iryna. 2025. The choice of corpora for contrastive analysis: a case study of the English causative construction with "encourage" and its Ukrainian counterpart.

Іноземна філологія [*Foreign Philology*]. Vol. 138, pages 133–147).

Kim, Jong-Bok and Michaelis, Laura A. 2020. *Syntactic Constructions in English*. Cambridge: Cambridge University Press.

Noonan, Michael. 2007. Complementation. in T. Shopen (ed.), *Language Typology and Syntactic Description*, vol. II: Complex constructions, 2nd edition. Cambridge: Cambridge University Press, pages 52–150.

Quirk, Randolph, Greenbaum, Sidney, Leech, Geoffrey, and Svartvik, Jan. 1985. *A comprehensive grammar of the English language*. Indexed by D. Crystal. London: Longman.

R Core Team. 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.R-project.org/

Shibatani, Masayoshi, and Pardeshi, Prashant. 2002. The causative continuum. In M. Shibatani (ed.). *The Grammar of Causation and Interpersonal Manipulation*. John Benjamins Publishing, pages 85–126.

Shvedova, Maria and Lukashevskyi, Arsenii. 2024. Creating parallel corpora for Ukrainian: A German-Ukrainian parallel corpus (ParaRook||DE-UK). *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING*. Torino, Italia. ELRA and ICCL, pages 14–22.

Shvedova, Maria *et al.* 2017. GRAC: General Regionally Annotated Corpus of Ukrainian. Kyiv, Lviv, Jena. Available at: Available at uacorpus.org.

Stefanowitsch, Anatol and Gries, StefanTh. 2003. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, 8(2): 209–243. Available at: https://doi.org/10.1075/ijcl.8.2.03ste

Stefanowitsch, Anatol and Gries, Stefan Th. 2008. Channel and constructional meaning: A collostructional case study. In G. Christiansen and R. Dirven (eds.) *Cognitive sociolinguistics: language variation, cultural models, social systems*. Berlin: Mouton de Gruyter (Cognitive linguistics research, 39), pages 129–152.

Stefanowitsch, Anatol. 2013. Collostructional analysis. In T. Hoffmann and G. Trousdale (eds.) The Oxford Handbook of Construction Grammar. Oxford: Oxford UP, pages 290-306. Available at: https://doi.org/10.1093/oxfordhb/9780195396683.013.0016

Stefanowitsch, Anatol. 2014. Collostructional analysis: A case study of the English into-causative. In T. Herbst, H.-J. Schmid, and S. Faulhaber (eds.) *Constructions Collocations Patterns*. Berlin, München, Boston: De Gruyter Mouton, pages 217–238.

Talmy, Leonard. 1988. Force Dynamics in Language and Cognition. *Cognitive Science*, 12(1): 49–100.

The Document Foundation. 2025. LibreOfficeCalc 25.8.4.2. Available at: https://www.libreoffice.org/discover/calc/

Balkan focus particles in Bulgarian: *bash* and *taman*

Ekaterina Tarpomanova

Sofia University Saint Kliment Ohridski
Sofia, Bulgaria

katya@slav.uni-sofia.bg

Abstract

This study investigates the Bulgarian focus particles *bash* and *taman*, two lexical items of the shared Balkan lexicon, with regard to their role in information structure. Using a corpus-based approach, the paper analyzes their function as particularizers that identify a focused constituent within a set of alternatives. The analysis is based on data from the Bulgarian web corpus CLASSLA-web.bg 2.0. The findings indicate that both particles are usually positioned adjacent to the focused constituent, most often in preposed position, though variation is attested. Overall, the study highlights both the shared and divergent properties of *bash* and *taman*, showing how their etymological origins and semantic development influence their present-day behavior as focus particles in Bulgarian.

Keywords: information structure, focus particles, corpus lingustics

1 Introduction

The present study examines two Bulgarian lexical items that belong to the shared Balkan lexicon, with particular emphasis on their potential function as focus particles. Although morpho-syntactic similarities are considered key features of the Balkan Sprachbund as they prove the deep convergence process that took place between the Balkan languages, the words they share are not with lesser importance in revealing the cultural and spiritual proximity of the Balkan peoples. The lexical items that are object of this study are not just shared loanwords in the Balkan languages as they play an important communicative role as focus particles interacting with the sentence focus and thus contributing for the correct interpretation of the information structure in discourse.

The study adopts a corpus-based approach to the analysis of the two focus particles in Bulgarian, with the aim of establishing their role in the information structure of the sentence. The theoretical account of focus particles follows the framework proposed by Ekkehard König (1991), while the analysis of Bulgarian focus particles draws on the model developed by Ruselina Nitsolova (2001), as well as two studies by Yovka Tisheva and Marina Dzhonova (2003, 2006) on the Bulgarian focus particle *samo* 'only'. The corpus analysis has a twofold objective: on the one hand, to provide quantitative data on the co-occurrence of focus particles with different word classes; on the other hand, to analyze – on the basis of authentic corpus examples – the types of sentential constituents that may fall within their scope. To this end, the CLASSLA-web.bg 2.0 is employed – a corpus of Bulgarian web, comprising 5,986,954,232 tokens from a wide range of genres and topics and enriched with linguistic annotation (Pungeršek et al., 2026).

2 Information structure

Information structure (Halliday, 1967) or information packaging (Chafe, 1976) is a central concept in contemporary linguistics, concerned with how speakers organize and present information within an utterance in relation to the communicative context and the common ground of the interlocutors. Rather than focusing on grammatical correctness or semantic content, information structure examines how meaning is shaped by the distribution of given and new information, the management of discourse, and the speaker's assumptions about the hearer's knowledge (Halliday, 1967; Lambrecht, 1994; Krifka, 2008). In recent decades, information structure has become increasingly relevant not only in theoretical linguistics and discourse analysis but also in related fields such as computational linguistics. Understanding how information is structured contributes to models of language processing and discourse representation and may

improve natural language generation systems by enhancing the analysis of spoken and written discourse.

Information structure has been analysed whithin different theoretical frameworks, but at its core, it addresses the ways in which speakers highlight certain elements of a sentence while backgrounding others. Concepts such as *topic* (or *topic – comment*) and *focus* (or *ground – focus*) play a crucial role in structuring discourse. For instance, the difference between *John bought the book* and *John bought the book* lies not in propositional content but in how information is packaged and interpreted.

Cross-linguistically, languages employ diverse strategies to encode information structure. While languages like English rely on intonation and stress, others, such as Japanese, use dedicated particles to mark topic and focus explicitly (Nakamura, 2006). Balkan languages have been classified as discourse-prominent languages as they code the communicative units not only by prosody, but also using syntactic devices such as word order, clitic doubling, subject position, etc. (Krapova, 2008: 181). Many languages use lexical tools to manage information structure, among which focus particles serve as a key device for marking sentence focus.

3 Focus and focus particles

Focus is an important notion in theories of information structure, referring to the part of an utterance that conveys the most informative, non-presupposed, or contrastive content. Broadly speaking, focus identifies what the speaker presents as new, informative, dominant, or contrary-to-expectations in relation to the part of the utterance considered as ground (Vallduví & Engdahl, 1996: 3). It plays a central role in how speakers manage the flow of information and guide the listener's attention to what is communicatively important. Focus has been defined as "that part of the utterence that answers the question under discussion" (Clopper & Tonhauser, 2011:1). Following the strategy of "question-answer congruence test", questions are used as an important tool to identify the focused constituent (Kroeger, 2017:4). Focus is not a uniform phenomenon but encompasses several subtypes, including information focus and contrastive (identificational) focus. Information focus introduces new information into the discourse, typically answering implicit or explicit questions (e.g., *What did John buy? – A car*), whereas, contrastive focus highlights alternatives and often implies a correction or opposition (e.g., *John bought a car, not a bike*) (Kiss, 1998; Krifka, 2008).

Focus particles are operators on the semantic structure of a sentence that mark the contrastive focus – an alternative among a set of alternatives that can be implemented in the case. The speaker only has a presupposition for these alternatives before he/she has received the information about the actual event that contrasts with his/her presupposition (Nitsolova, 2008: 479). Focus particles can be classified into several types based on their semantic contribution and pragmatic function. Two main types are to be outlined: inclusive (additive) which include some of the alternatives as possible values of the variable of their scope (*also*, *too*), and exclusive (restrictive), which imply that none of the alternatives satisfies the respective open sentence (*only*) (König, 1991: 33). König (1991: 122) notices a third type – particles of emphatic assessment of identity (*exactly*, *precisely*). Following Quirk's classification (Quirk et al., 1972), Nitsolova adopts the term particularizers for the same subclass of focus particles in Bulgarian – *tochno* 'exactly, precisely', *tăkmo* 'exactly, precisely', *imenno* 'namely', defining them as cases in which the predicate connects with some part of the alternatives (Nitsolova, 2008: 479).

4 *Bash* and *taman* as lexical items of the common Balkan lexicon

The lexical items that are object of analysis, *bash* and *taman*, both meaning 'exactly, precisely', belong to the shared Balkan lexicon which contains words that occur in the languages of the Balkan Sprachbund regardless of their origin (Asenova and Detrez, 2021). Similarly to other words of the Balkan lexicon, *bash* and *taman* are also found in Bosnian-Croatian-Montenegrin-Serbian (BCMS).

Among languages of the Balkans, *bash* occurs in Turkish, BCMS, Balkan Slavic (Bulgarian and Macedonian), Slovene and Albanian; it has regional use in Daco-Romanian. In some of the Balkan languages, including Bulgarian, *bash* is described in lexicographic resources as two separate lexical items: as an adjective (and substantive) with the meaning 'main' and as an adverb or particle with the meaning 'exactly,

precisely'. The adjective has been loaned from Turkish *baş* 'head, leader; beginning, basis', but the origin of the adverb/particle is disputable. In some sources, mainly Bulgarian and Serbian, it is stated that the adverb/particle has derived from the adjective, i.e. they are etymologically related and both have Turkish origin (Asenova and Detrez 2021: 304; BER, 1971: 37; ERSJ, 2006: 276), whereas Croatian dictionaries claim Slavic origin of the adverb/particle – *baš* has derived from Proto-Slavic **bъšь* 'entirely, quite', a petrified form of the noun *bъšь* 'hit, blow', cognate with the verbs *buhnuti, bušiti* 'to hit' (ERHJ 2016: 45-46).

In contemporary Bulgarian *bash* occurs as an uninflected adjective with the meaning 'main, the best, top' (1a) and as focus particle with the meaning 'exactly, precisely, just' (1b).

(1a) *Izbraha* ***bash*** *maystorite na vinata v Parvomay i Dalbok izvor.* 'The **top** master winemakers were selected in Parvomay and Dalbok Izvor.'

(1b) – *Krastosa edva li e pravilnata duma, otche.* – ***Bash*** *pravilnata si e.* '– Crossed is hardly the right word, Father. – It's **exactly** the right one.'[1]

For Serbian, Mišković defines two functions of the focus particle *baš*: specificatory when the particle specifies the focus and has a temporal, locative or deictic expression in its scope (2a), and emphatic when it focuses on certain process, state or quality (2b):

(2a) *Knjiga je* ***baš*** *na stolu.* 'The book is **right** on the table.'

(2b) ***Baš*** *peva.* 'He/she is really singing (beautifully).' (Miskovic, 2001)

As compared to Serbian (and BCMS in general), *bash* in Bulgarian has most often a specifying function, whereas the emphatic function is rather rare. The difference between BCMS and Bulgarian can be illustrated with the sentences below which are translations of *The Ugly Duckling* by Hans Christian Andersen:

(3a) Serb. *„**Baš** si ružan!", rekle su mu patke. „Ali nama to ne smeta, samo da se ne ženiš nekom od nas!"* (translation by Stanislav Vinaver)

(3b) Bulg. – *Ti si* ***mnogo*** *grozno! – kazaha divite patitsi. – Ala tuy ne ni vliza v rabotata, stiga samo da ne ti hrumne da se ozhenish za nyakoya oot nasheto semeystvo.* (translation by Svetoslav Minkov)

(3c) Eng. *"You are* ***terribly*** *ugly," they told him, "but that's nothing to us so long as you don't marry into our family."* (translation by Jean Hersholt)

Both Bulgarian and English have chosen an adverb to intensify the quality expressed by the adjective – *mnogo* 'very' and *terribly*, respectively, whereas for Serbian *bash* seems to be a natural choice in this context.

When used independently, *bash* expresses confirmation or agreement with a previous statement (4).

(4) – *Kaza, che tezi neshta sa na Sarina?* – ***Bash***. 'You said these are Sarina's things. – **Right**.'

The origin of *taman* is clear – in the Balkan languages, it is loaned from Ottoman Turkish (*tamam* 'whole, complete, accurate; just, exactly, precisely'), where it penetrated from Arabic (*tamām* 'full, entire, complete; completeness, wholeness, perfection') (Asenova and Detrez 2021: 313). It occurs (with different frequency) in almost all contemporary languages and dialects of the Balkans: Turkish, Balkan Slavic (Bulgarian and Macedonian), Balkan Romance (Daco-Romanian, Aromanian and Megleno-Romanian), Albanian, Greek, BCMS. In modern Arabic and Turkish, it has undergone strong pragmaticisation and its initial lexical meaning related to the notion of completeness has developed into a pragmatic one (Ruhi 2013; Al-Hanaktah and Hamdan 2024), with closest translation equivalents in English discourse markers such as *ok*, *right*, *fine*. In the Balkan languages, pragmatic development is not observed. The word *taman* is defined there as an adverb meaning 'exactly'.

In contemporary Bulgarian, *taman* is a colloquial word with several main uses: as an adverb of manner, indicating that something (clothes, shoes) is the right size (5a) or otherwise suitable (5b) in the speaker's view; as an adverb of time, indicating that the action is carried out immediately before or at the moment of speaking (5c), or in sentences in which the situation in the first clause, introduced by *taman*, takes place

[1] All Bulgarian examples are drawn from the CLASSLA-web.bg 2.0 corpus and are transliterated into Latin script. The lexical items under analysis are presented in bold both in the original and in the English translation, where possible, and the constituent falling within their scope is underlined. As noted by Tisheva and Dzhonova (2003: 58), the scope of focus particles and sentence focus usually coincide, but not always.

immediately before the situation in the second one (5d); as an adverb expressing logical relations (according to the classification in Nitsolova 2008: 450), functioning as a link between two utterances, confirming the preceding statement and adding a new argument in its support or a positive development of the situation (5e); and as a focus particle (5f).

(5a) *Roklyata ti e **taman**.* 'Your dress **fits perfectly**.'

(5b) *Na men lichno nastroyka 23 na klimatik si mi e **taman**…* 'Personally, setting 23 on the air conditioner suits me **just fine**...'

(5c) ***Taman** vlizahme v Balgariya*. 'We were **just** entering Bulgaria.'

(5d) ***Taman** reshih da zatvorya „krazhbite" i ti gi otvori nanovo.* 'I had **just** decided to stop the "thefts," and you opened them up again.'

(5e) *Idi bre, **taman** shte razberesh ot parva raka shto smartnostta e niska.* 'Come on, go there, **then** you'll understand firsthand why the mortality rate is low.'

(5f) … *osvobozhdenieto ot sorosoidnata sekta tryabva da zapochne **taman** ot Polsha!* '... liberation from the Sorosoid sect must begin **precisely** with Poland!'

Taman is rarely used independently; in such cases it expresses the speaker's approval that the circumstances in a given situation are good and meet expectations.

(6) *– Arzhentina i Iran poiskaha da se prisaedinyat kam BRIKS. – Da, da, **taman**.* '– Argentina and Iran have requested to join BRICS. – Yeah, yeah, **perfect**.'

The review of the etymology and semantic development of *bash* and *taman* shows that, although they have different origins and functions, their usage in contemporary Bulgarian is similar when they function as focus particles.

5 *Bash* and *taman* as particularizers

Both Balkan lexical items in Bulgarian are used as focus particles of the type of particularizers, i.e., they associate with a given alternative within a set of alternatives. Both items are characteristic of informal register and, in contrast to their primary Bulgarian synonyms *tochno* and *tăkmo* 'exactly, precisely', they are expressive and stylistically marked. *Bash* and *taman* occur significantly less frequently compared to their stylistically neutral synonyms: *tochno* has 2,825,380 occurrences, *tăkmo* – 217,070, *bash* – 66,576, and *taman* (in its three phonetic variants – *taman*, *tamăn*, and *tamam*) – 21,798. As has been shown, the two particles have different origins and inherit the semantic properties of their source lexemes: *bash* is linked to the concept of centrality or primacy, whereas *taman* is associated with the idea of completeness or perfection.

The present study examines the way in which they participate in the information structure of the utterance by identifying the focus constituent of the sentence, and, additionally, addresses the possible differences in their semantics as focus particles. This perspective directs attention primarily to their syntactic properties – namely, their position within the sentence and their scope, and secondly to their interpretation depending on the context. It is well known that the syntactic behavior of focus particles varies across languages; moreover, within a single language, different focus particles may exhibit distinct syntactic patterns. Focus particles can generally combine with all major constituent types – NP, VP, AP, AdvP, PP, and CP (Herrmann 2014: 185).

In Bulgarian, *bash* and *taman* occur in sentence-initial, sentence-final, and medial positions. They are typically preposed and in contact position with respect to the constituent they associate with (7a); more rarely, they are postposed (7b) or occur at a distance from that constituent (7c).

(7a) *Boyko Borisov **taman** tebe chaka, da mu nosish haber ot Varna…* 'Boyko Borisov is waiting **exactly** for you to bring him news from Varna.'

(7b) *Na tebe **bash** da ti obyasnyavam…* 'To you, of all people (lit. exactly), I should explain...'

(7c) *Tiho sya, **taman** Steven Pinker se opitva da podkovava britantsite, che nyamalo takova neshto kato chist kralski algliyski.* 'Quiet now, Steven Pinker is **just** trying to convince the British that there is no such thing as pure Queen's English.'

Therefore, most often, though not always, *bash* and *taman* identify the first right constituent as the focus. Table 1 presents the co-occurrence of the two focus particles and the first right constituent, which may belong to different word classes: noun (N), pronoun (Pron), adjective (A), adverb (Adv), verb (V), numeral (Num), conjunction (Conj), or preposition (P). The results can to some extent guide us in determining the scope of the focus particles, but further analysis shows that they can be misleading.

	bash	***taman***
N	17,108	539
Pron	10,574	2,608
A (A+N)	4,613 (2,140)	150 (79)
Num	506	1,053
Adv	2,224	423
V	1,231	3,364
P	12,688	1,575
Conj	534	445

Table 1: Co-occurrence of *bash* and *taman* with different word classes.

5.1 *Bash* and *taman* + NP

The first option to be checked is whether the two focus particles can have in their scope a noun phrase with different structure – a common noun, a proper noun, a pronoun, and a noun modified by an adjective. In the combination of the particles with a noun, the overwhelming predominance of occurrences of *bash* is immediately noticeable. In fact, in a large number of cases *bash* is used not as a focus particle but as an adjective (e.g. *bash opozitsionerite* 'the real oppositionists'). The expression *bash maystor* – the title of a well-known Bulgarian film that has come to be used ironically to refer to a person who is incompetent at their job – alone occurs 1,128 times. However, *bash* can have a scope over an NP with different type of structure – rather rarely if the NP consists of a noun (8a) and much more frequently if it consists of a proper noun (8b), of a pronoun (8c) or an adjective and a noun (8d).

(8a) *Sanchez si e* ***bash*** *krilo. I v Barselona si igraeshe krilo.* 'Sanchez is **exactly** a winger. He also played as a winger at Barcelona.'

(8b) *Taka che* ***bash*** *Hari e vinoven…* 'So Harry is the one to blame...'

(8c) *...****bash*** *toy ni nabuta da poluchavame samo ruski gaz!* '…it was **exactly** him who got us stuck with receiving only Russian gas!'

(8d) *...no faktite sochat, che* ***bash*** *desnite demokrati sa se kultivirali v kachestvoto si na obsluzhvasht personal okolo Todor Zhivkov.* '…but the facts suggest that it was **precisely** the right-wing democrats who were cultivated in their role as a kind of service personnel around Todor Zhivkov.'

The review of the results for *taman* shows that, as a focus particle, it much more rarely takes an NP within its scope, and this is usually a pronoun (9a). In most cases, *taman* is not used as a particularizer, but as an adjective (9b) or adverb (9c). In a number of sentences, although the first right constituent is an NP, as an adverb it modifies a VP or a clause (9d).

(9a) *Makar che ako beshe za byagane prez Dunava,* ***taman*** *toy beshe za tova...* 'Although, if it had been about escaping across the Danube, he was **just** the right person for it (lit. it was exactly him for it)…'

(9b) *Neshto ne e* ***taman*** *rabotata.* 'Things aren't quite **right**.'

(9c) *Ako ne iskate, ne glasuvayte,* ***taman*** *BSP shte izchezne okonchatelno.* 'If you don't want to, don't vote – **that way** BSP will disappear for good.'

(9d) *...da mahnat pone vecheryata,* ***taman*** *izlishni kilograma shte svalyat.* '…they should at least drop dinner – **that way** they'll lose some extra weight.'

For both *bash* and *taman*, the high frequencies in combination with pronouns are due to the annotation scheme of the corpus: pronominal adverbs (*bash togava* 'right then', *bash kogato* 'exactly when'), as well as reflexive pronouns functioning as derivational or grammatical particles (*taman si trăgvam* 'I am just leaving') or as ethic dative (*bash si e legenda* 'he is a legend indeed'), are annotated as pronouns. In addition, pronominal clitics cannot serve as the focus of a sentence (*taman go nauchih* 'I just teached him').

5.2 *Bash* and *tamam* + NumP

When analyzing the possible scopes of *bash* and *taman*, special attention should be paid to numeral phrases (for an analysis of NumP as part of DP, see Cinque 2005), due to the semantic compatibility of the particularizers with numerals. The particlarizers take in their scope NumPs with both ordinals (10a, 10b) and cardinals (10c), as well as numeral derivatives (10d). The co-occurrence of *bash* and *taman* with numerals is much more frequent as compared with their use with other quantifiers such as *mnogo* 'very, many', *malko* 'a little'.

(10a) *taka che si smeni iztochnitsite, a Svetlyo si vzimat* ***bash*** *5 bona.* 'so change your sources, and Svetlyo, they take **exactly** 5 grand.'

(10b) *Mozhem da si predstavim kak shte reagira, ako* ***bash*** *dvete partii-simvoli na statukvoto na prehoda izlyazat na vlast*. 'We can imagine how he

will react if **exactly** the two parties that symbolize the status quo of the transition come to power'

(10c) *...**taman** tretiyat den ot tridnevnata voyna.* '…**exactly** the third day of the three-day war.'

(10d) *...ima **taman** shestkratno uvelichenie na dobivite ot 1981 do 1989 g.* 'There is **exactly** a sixfold increase in output from 1981 to 1989.'

In combination with numerals, *taman* occurs more frequently than *bash*, despite the overall frequency of *taman* being approximately three times lower. A plausible explanation lies in the modal nuance associated with *taman*, which is absent in *bash*. When *taman* functions as a particularizer with a NumP in its scope, it additionally conveys the speaker's subjective stance, typically positive – expressing approval or the fulfillment of prior expectations, and in some cases irony. A distinction between modal and epistemic uses of scalar particles such as *at least* has been discussed in a number of studies (Grosz 2011; Biezma 2013; for the South Slavic languages Stefanov 2026, among others). In the case of the Bulgarian particularizers *bash* and *taman*, the latter exhibits stronger modal properties, likely inherited from its adverbial uses.

5.3 *Bash* and *taman* + AP

In nearly half of the cases, when the first right-hand constituent is an adjective, it forms part of an NP consisting of an adjective and a noun (in Table 1 – A+N). Beyond these instances, the two focus particles interact with adjectival phrases under conditions comparable to those observed with NPs: both particles may take an AP within their scope, although this occurs relatively infrequently and is more common with *bash* than with *taman*.

(11a) *Raykov izglezhda **bash** idealen za sluchaya.* 'Raykov seems **just** perfect for the case.'

(11c) *„Pampers" peleni ne kupuvam, no „Hagis" v U. sa si ama **taman** sashtite kato drugade.* 'I don't buy "Pampers" diapers, but "Huggies" in U. are just **exactly** the same as elsewhere."

5.4 *Bash* and *taman* + AdvP

Both focus particles may take adverbial phrases within their scope, although *bash* does so more frequently and combines with a wider range of adverb types, including temporal (12a), (12b), locative (12c), and manner adverbs (12d). As observed by Miskovic (2001) for Serbian, when temporal and locative adverbs fall within its scope, *bash* typically has a specifying function, whereas with manner adverbs it conveys emphasis; the latter use is relatively infrequent in Bulgarian. By contrast, *taman*, as in its interaction with NPs, is more often used as an adverb (12f) than as a particularizer when combined with adverbs. When functioning as a particularizer, it most commonly occurs with temporal adverbs (12e).

(12a) *...mi **bash** dnes vashite hora kazaha che ne e agent kato prezidenta…* 'well, it was **exactly** today that your people said he is not an agent like the president…'

(12b) *...taka se omazaha **bash** predizborno.* 'so they messed up **right** before the elections.'

(12c) *Vav videoto mnogo dobre kazvat koy otkade idva i snimkata na kolata na myasto si e **bash** napreki na idvashtoto semeystvo.* 'In the video they clearly say who comes from where, and the picture of the car is positioned **exactly** across the approaching family.'

(12d) *I ti ne mozh da mi kazhesh, che ti e **bash** ubavo kato kogato beshe po-mlad.* 'And you can't tell me it feels **exactly** as good as when you were younger.'

(12e) *I **taman** sega shte ti go trusnat v glavata s novi zakoni na ES.* 'And **just** now they will dump it on your head with new EU laws.'

(12f) *Prodalzhavay, **taman** po-skoro shte ni otarvesh ot prisastvieto si.* 'Go on, **that way** you'll get rid of your presence sooner.'

5.5 *Bash* and *taman* + VP

With verb phrases, *taman* displays a considerably higher frequency; however, it is almost invariably used as an adverb (13a) rather than as a particularizer. In certain cases, *taman* allows for ambiguity between an adverbial/adjectival and a focus-particle interpretation, with disambiguation relying on prosody in spoken language or broader contextual cues in written discourse (13b). *Bash*, in turn, only rarely takes VPs within its scope. Many apparent instances result from the annotation of participles as verbal forms, although these function as adjectives within NPs (e.g. *bash protestirashtite partii* 'exactly the protesting parties'). Most commonly, *bash* scopes over the copula, and more rarely over other verbs (13c). A characteristic construction involves the copula combined with a reflexive pronoun in ethical dative, which introduces a subjective stance on the part of the speaker. The predicative complement in such

constructions may be realized by various phrase types, including NP (13d), (13e), AP (13f), NumP (13g), and PP (13h). With both the copula and full verbs, when used as particularizers, *bash* and *taman* select as focused constituent the complement within the VP. In copular sentences, the particularizers move to a sentence-initial position due to prosodic factors in Bulgarian.

(13a) *...vecherniyat prelet **taman** zapochva, chovek tryabva da si otvarya ochite.* '…the evening migration is **just** beginning; one has to keep their eyes open.'

(13b) ***Taman** si e za Kitay.* 'It's **perfect** for China.' / ***Taman** si e za Kitay.* 'It's **exactly** for China.'

(13c) *A dneshnite arabi **bash** iskat titli, slava, da badat zabelyazani, a ne pari.* 'And what today's Arabs **really** want is titles, fame, to be noticed, not money.'

(13d) *Uff... păk makedontsite bili s debeli glavi – **bash** sa si balgari!* 'Ugh… and they say Macedonians have thick heads – they are **exactly** Bulgarians!'

(13e) ***Bash** si e prerazhdane v dvizhenie.* 'It **really** is reincarnation in motion.'

(13f) ***Bash** si e falshiva.* 'It **really** is fake.'

(13g) ***Bash** si e 1000MB.* 'It is **exactly** 1000MB.'

(13h) ***Bash** si e za tam.* 'It **really** belongs there. (lit. for there)'

5.6 *Bash* and *taman* + PP

Both particles may take prepositional phrases within their scope, with this tendency being particularly pronounced for *bash*. The PPs involved may be headed by a range of prepositions. In the case of *bash*, the highest number of occurrences is found with the highly frequent Bulgarian preposition *na* 'of, to', whereas *taman* most commonly combines with *za* 'for'. This distribution can be attributed to the adjectival semantics of *taman*, which conveys meanings such as 'suitable (for)' or 'perfect (for)'.

(14a) *Prostitutka nalyetya **bash** na shefa na politsiyata.* 'A prostitute went **straight at the chief of police**.'

(14b) *...ama **bash** v toz sektor nashiyat „aviatsionen" spetsialist ne otgovarya.* '…but in **exactly** this sector our "aviation" specialist is not responsible.'

(14c) *za da mozhe **bash** za uikenda da ni se stovaryat tezhki gramotevitsi.* 'so that heavy thunderstorms hit us **exactly** on the weekend.'

(14d) *…nisko letyasht voennomorski hidroplan se spuska **taman** kam nego.* 'a low-flying naval seaplane is descending **right** toward him.'

(14e) *Nyama osnovaniya Ivo da se pritesnyava ot sada i to **taman** v Sofiya.* 'There are no grounds for Ivo to worry about the court – **especially** in Sofia.'

(14f) *...**taman** kato sherifa ot diviya zapad!* '…**just** like the sheriff from the Wild West!'

5.7 *Bash* and *taman* + CP

Both *bash* and *taman* may interact with complementizer phrases. However, the results presented in Table 1 should be interpreted with caution, as corpus data involving conjunctions frequently contain cases of misannotated grammatical ambiguity, for instance between conjunctions and prepositions (*bash kato druzhkata ti* 'exactly as your friend') or between conjunctions and particles (*bash pa tam* 'exactly there'). At the same time, *k*-words (i.e. interrogative and relative pronouns and pronominal adverbs) are included in the results under the category of pronouns.

A comparison between the two particularizers shows that *bash* more frequently takes CPs within its scope than *taman*. Both particles combine with the modal complementizer *da*, which in Bulgarian functions as a substitute for the old Slavic infinitive and has grammaticalized into a preverbal morpheme (Asenova 2002: 151). The sequence *bash da* occurs 237 times, and *taman da* 932 times; however, in these cases *bash* predominantly functions as a particularizer (15a), whereas *taman* is used almost exclusively as an adverb – temporal or expressing logical relation (15b). CPs introduced by the factive complementizer *che* ('that') only rarely fall within the scope of *bash* (15c), and there are no such instances with *taman* in the corpus. Both particularizers can take within their scope CPs introduced by *ako* 'if' (15d, 15e), but *taman* retains its adverbial meaning in most cases.

(15a) *...**bash** da sa tupali Bayern redovno e malko presileno! Biha gi samo 2 pati.* '**Exactly** that they have beaten Bayern regularly is a bit exaggerated! They beat them only twice.'

(15b) *I **taman** da si traghnat, Bush kazal: …* 'And **just** as they were about to leave, Bush said: …'

(15c) *Ti ne znaesh… hahaha… i samo smenyash nikuviti… hahaha… i taka… **Bash** che ne smenyam nikoga "nikuviti", a drugi go pravyat ot rusofilska zloba poradi tozi fakt, pisheiki s moya!* 'You don't know… haha… and you only change nicknames… haha… and so on… **Exactly** that I never change "nicknames", while others do it out of pro-Russian spite because of this fact, writing with mine!'

(15d) *Ti ne razbirash – **bash** ako beshe pobedil fashizmat, sega Makedoniya, polovin Gartsiya i tryala Dobrudzha […] shteshe da e nasha!* 'You don't understand – exactly if fascism had won, now Macedonia, half of Greece, and all of Dobruja […] would have been ours!'

(15e) *Zdravey, kolega, i az ako mozhe da se kandidatiram za stoykata, ponezhe mnogo iskam da si vzema edna, no to tuk mi idvat mnogo skapo i sega si porachah zumeri ot Rusiya, i **taman** ako mi ya dadesh, shte e prefektno.* 'Hello, colleague, if possible, I'd also like to apply for the stand, since I really want to get one, but from here they come out very expensive, and now I've ordered buzzers from Russia, and **just** if you give it to me, it will be perfect.'

6 Concusion

The analysis of the corpus data show that both *bash* and *taman* can take scope over a wide range of syntactic constituents, including NP, NumP, AP, AdvP, VP, PP, and CP, though with notable differences in frequency and distribution. *Bash* more frequently functions as a focus particle, and does so with all types of phrases except VPs, typically exhibiting a specifying function. Still, when used with NPs disambiguation is always necessary. In contrast, *taman* is more often used as an adverb, which is why it shows a higher frequency with VPs. When functioning as a particularizer, most often with NumPs and PPs, *taman* tends to convey an additional modal nuance, expressing the speaker's (positive) attitude.

References

Isra'a Al-Hanaktah and Jihad M. Hamdan. 2024. On the Pragmatic Functions of the Idiomatic Expression 'Tamam' in Jordanian Arabic. *Journal of Language Teaching and Research*, Vol. 15, No. 3.

Petya Asenova. 2002. *Balkansko ezikoznanie*. Veliko Tarnovo: Faber.

Petya Asenova and Raymond Detrez. 2021. *Rechnik na obshtobalkanskata leksika v knizhovnite ezitsi*. Sofia, Universitetsko izdatelstvo "Sv. Kliment Ohridski".

BER 1971: Vladimir Georgiev, Ivan. Galabov, Yordan Zaimov, Stefan Ilchev. *Balgarski etimologichen rechnik*. Sofia, Izdatelstvo na BAN.

ERHJ 2016: Ranko Matasović, Tijmen Pronk, Dubravka Ivšić, Dunja Brozović Rončević. *Etimološki rječnik hrvatskoga jezika*. *1. svezak. A – Nj*. Zagreb, Institut za hrvatski jezik i jezikoslovlje.

ERSJ 2006: Aleksandar Loma (urednik), Marta Bjelitić, Jasna Vlajić-Popović, Marija Vučković, Maja Đokić, Aleksandar Loma, Snežana Petrović. *Etimološki rečnik srpskog jezika, Sveska 2: BA–BD*. Beograd, Institut za srpski jezik SANU.

María Biezma. 2013. Only one 'at least': Refining the role of discourse in building alternatives. In K. Shwayder, editor, *Proceedings of Penn Linguistics Colloquium 36*, pages 11–19.

Wallace Chafe. 1976. Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In Charles N. Li, editor. *Subject and Topic*, pages 25–55. New York: Academic Press.

Guglielmo Cinque. 2005. Deriving Greenberg's Universal 20 and Its Exceptions. *Linguistic Inquiry*, 36(3), pages 315-332.

Cynthia G. Clopper and Judith Tonhauser. 2011. On the Prosodic Coding of Focus in Paraguayan Guaraní. In Mary Byram Washburn, Katherine McKinney-Bock, Erika Varis, Ann Sawyer, and Barbara Tomaszewicz, editors. *Proceedings of the 28th West Coast Conference on Formal Linguistics*, pages 249-257.

P. G. Grosz. (2011). A uniform analysis for concessive at least and optative at least. *Proceedings of SALT 21*, pages 572-591.

M. A. K. Halliday. 1967. Notes on Transitivity and Theme in English. *Journal of Linguistics*, Vol. 3, No. 2, pages 199-244.

Annika Herrmann. 2014. *Modal and Focus Particles in Sign Languages: A Cross-Linguistic Study*. Berlin, Boston, De Gruyter Mouton.

Katalin É. Kiss. 1998. Identificational Focus and Information Focus. *Language*, 74(2), pages 245-273.

Ekkehard König. 1991. *The Meaning of Focus Particles: A Comparative Perspective*. London and New York: Routledge.

Iliyana Krapova. 2008. Obshti cherti na sintaktichniya stroezh na taka narechenata lyava periferiya na izrechenieto v balkanskite ezitsi. *Imeto v pametta na ezika. Sbornik v pamet na profesor Boris Simeonov*, pages 181-198. Sofia, Figura.

Manfred Krifka. 2008. Basic Notions of Information Structure. *Acta Linguistica Hungarica*. Vol. 55, No. 3-4, pages 243-276.

Paul Kroeger. 2017. Basic Concepts in Information Structure: Topic, Focus, and Contrast. *GIALens 2017*, Vol. 11, No. 1.

Knud Lambrecht. 1994. *Information Structure and Sentence Form: Topic, Focus, and the Mental Representations of Discourse Referents*. Cambridge University Press.

Mirjana Mišković. 2001. The particle *baš* in contemporary Serbian. *Pragmatics,* 11 (1), pages17–30.

Hiroaki Nakamura. 2006. Topic-Comment Articulation in Japanese: A Categorial Approach. *Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation*, pages 198-205. Tsinghua University Press.

Ruselina Nitsolova. 2001. Za kontrastnija fokus s chastici v balgarskiya ezik. *Savremenni sintaktichni teorii*, pages 76-85. Plovdiv.

Ruselina Nitsolova. 2008. *Balgarska gramatika. Morfologiya*. Sofia, Universitetsko izdatelstvo "Sv. Kliment Ohridski".

Kuzman Pungeršek, Taja and Peter Rupnik, Vít Suchomel, and Nikola Ljubešić. 2026. The Growing Gains and Pains of Iterative Web Corpora Crawling: Insights from South Slavic CLASSLA-web 2.0 Corpora. In submission.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1972. *A grammar of contemporary English.* Beccles and London, Longman.

Şükriye Ruhi. 2013. The interactional functions of tamam in spoken Turkish. *Dil ve Edebiyat Dergisi / Journal of Linguistics and Literature*, 10:2, pages 9-32.

Martin Stefanov. (2026). Restrictive particles as discourse markers in South Slavic languages (contrastive study on the *at least* particles in Bulgarian, Serbian, Croatian, and Slovene). *New Directions in Slavic Studies: Language, Literature, and Political Discourse. Proceedings of the 6th International Conference GranaSlavic2025*, University of Granada, 2 – 4 June 2025, pages 645-654.

Yovka Tisheva and Marina Dzhonova. 2003. Lexical Markers on the Information Structure Level. In Anissava Miltenova, David Birnbaum, Sarah Slevinski, editors. *Computational Approaches to the Study of Early and Modern Slavic Languages and Texts. Proceedings from the International Conference on Electronic Description and Edition of Slavic Sources*, pages 53-76. Sofia, Boyan Penev Publishing Centre.

Yovka Tisheva and Marina Dzhonova. 2006. Lexical Tools for Focus Marking – The Case of *samo* in Bulgarian. *Formal Approaches to South Slavic and Balkan Languages – 5*, pages 147-150.

Enric Vallduví & Elisabet Engdahl. 1996. The Linguistic Realisation of Information Packaging. *Linguistics*, 34(343), pages 459-519.

When Bulgarian ‘Still’ Meets Ukrainian ‘Already’

Ivan Derzhanski, Olena Siruk
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
{iad58g,olebosi}@gmail.com

Abstract

Languages vary in whether and how they use words for ‘still’, ‘already’, or both, for expressing the meaning ‘as early/soon as’ (‘no later than’). If both can in principle be used, the eligibility of one or the other may depend on the syntactic category and lexical choice of the modified constituent, the position of the event on the time axis and the speaker’s view of the situation. Some questions arise: What does the latter point include? How is this meaning of a ‘still’ or ‘already’ word related to its other uses? If the same word can also express the contrary meaning ‘as late/recently as’ (‘no sooner than’), is the ambiguity tolerated, or how is it resolved? Finally, in a translation, can the word choice in the original play a part too?

We look at pairs of sentences in which Bulgarian *още* ‘still’ corresponds to Ukrainian *вже* ‘already’ from a parallel Bulgarian–Ukrainian corpus, containing both texts written in one of these languages and translated into the other and texts translated into both from six other languages. We compare them to similar sentences where Ukrainian also uses its ‘still’ word, *ще*. We observe that the correspondence *още* : *вже* is more frequent in translations from languages which actively use their ‘already’ word for ‘as early/soon as’ than from those which do not. But a strong dependence on the semantics of the modified constituent can also be established. We discuss the details of the distribution and its presumable reasons.

Keywords: Bulgarian language, lexical semantics, parallel corpus, particles, temporal adverbials, translation, Ukrainian language.

1 Introduction

Languages vary in whether and how they use words for ‘still’, ‘already’, or both, as modifiers of time adverbs, temporal subordinate clauses and other expressions with temporal semantics in the meaning ‘as early/soon as’ (‘no later than’). As Galton (1976: 383) puts it, ‘we should not be surprised if what is objectively the same date, at the same measurable distance from the present, can be depicted in contrasting ways in different languages’. He comments on the comparison of the Russian sentence (1a) with its German counterpart (1b):[1]

(1) a. *Впрочем ещё Августин указывал …*
b. *So wies schon der hl. Augustin hin.*
‘Already Augustine pointed out …’

‘In Russian, the saint falls still within his period of antiquity, known to be far removed from the present, whilst it was still ongoing; this distance need not even be explicitly stated. In the German rendering, on the other hand, it is noted as remarkable that such a statement was already made by Augustine much earlier than one might perhaps expect.’

But it is not that one language exclusively uses a ‘still’ word for ‘as early as’ and another an ‘already’ one. Whilst in German only *schon* is possible with a subject noun phrase (Мустайоки, 1988: 110), other constituents also allow *noch*. Both schemes may even cooccur within the same sentence; here is an example in German as well as one in Italian:

(2) *Du wußtest damals im voraus, daß ich dir gehorsam sein würde, ich hatte es schon vor der Beichte, und noch eh ich dich kannte, versprochen.* ‘You knew in advance that I would be obedient to you; I had promised it even before the confession, and even before I knew you.’ (H. Hesse, *The Glass Bead Game*)

(3) *Pare che già fanciullo fosse toccato dalla grazia divina e laggiù in Castiglia leggesse i*

[1] To save glossing space, words meaning ‘still’ and ‘already’ will be marked by single and double underlining, respectively. Appendix A presents the ‘still’ and ‘already’ words of all languages mentioned in the paper, including (for completeness, in brackets) words that do not themselves appear in the text.

libri degli arabi e dei dottori greci ancora impubere. 'It seems that even as a child he was already blessed by divine grace, and in his native Castile he read the books of the Arabs and the Greek doctors while still in his teens.' (U. Eco, *The Name of the Rose*)

'Facing such parallels, a linguist cannot help but wonder whether the meanings differ' (Мустайоки, 1988: 111).

Regarding German, Löbner (1989: 202) notes that (4a) 'is hard to distinguish from' (4b):

(4) a. *Sie kommt noch heute.*
b. *Sie kommt schon heute.*
'She is coming this very day.'

Yet in general the two are not equivalent: unlike *schon*, *noch* is restricted to time specifications which include the reference time, so that replacing *heute* 'today' with *übermorgen* 'the day after tomorrow' in (4a) is problematic 'unless it is clear that for independent reasons reference is implicitly to the day after tomorrow'.

In Russian, too, both schemes are in use:

(5) a. *Уже Щерба обращал внимание на необходимость изучения языковых ошибок.*
b. *Ещё Щерба обращал внимание на необходимость изучения языковых ошибок.*
'Already Shcherba drew attention to the need to study language errors.'

(6) a. *Уже в Твери я понял, что еду зря.*
b. *Ещё в Твери я понял, что еду зря.*
(on a journey from Moscow to St Petersburg) 'Already in Tver [i. e., about one quarter of the way] I realised I was travelling in vain.'

The difference between them has been formulated in a variety of ways. Апресян (1979: 512–3) appeals to the concept of speaker's time: in (5a) the linguist Lev Scherba is thought of as the speaker's contemporary, and in (5b) as someone from another period. Богуславский (1996: 240–1) highlights the different reference points: in (6a) Tver is presented as being *near* the start of the journey, and in (6b) as *far* from the destination. Thus *уже* marks the event as having happened *early*, and *ещё* as having happened *long ago* (Мустайоки, 1988: 113–4).

That said, some languages use one of these words for 'no later than', but not the other. One such is Bulgarian, where only *още* has this function. This may be a regional feature. Reiter (1987) shows that the very limited use of the 'already' word, compared to German *schon*, is characteristic of the Balkan languages, especially the more eastern ones. Here is an example that he gives as an illustration of Bulgarian *още* translating German *schon* (p. 117–118):

(7) *Нейният мъж ще получи още сега, приживе на папа, известно обезщетение, само четвъртинка от покупната цена на къщата… ← Ihr Mann bekommt schon jetzt, bei Papas Lebzeiten, eine Abstandssumme, ein Viertel bloß von der Hauskaufsumme …* 'her husband gets compensation already, in Papa's life-time — a quarter of the purchase price of the house.' (T. Mann, *Buddenbrooks*)

Note that *вече* is unable to be focussed, in particular it can't be in the focus of interrogation and carry the interrogative clitic (9), unlike *още* (10) and in contrast to Finnish *jo* (8):

(8) a. *Onko kirje jo valmis?* 'Is the letter ready yet?'
b. (I) *Onko kirje jó valmis?*
(II) *Joko kirje on valmis?*
'Is the letter ready so soon?' (Välikangas, 1982: 375)

(9) a. *Готово ли е вече писмото?*
b. **Вече ли е готово писмото?*

(10) a. *Пише ли се още писмото?* '*Is* the letter still being written?'
b. *Още ли се пише писмото?* 'Is the letter *still* being written?'

The motivation of the use of 'still' and/or 'already' words for 'no later than' (in any language) can be assumed to be such as hinted in (Galton, 1976: 383) and elaborated in (Мустайоки, 1988: 134): beginning with 'whilst it is still time t, already event e happens', either 'still t' replaces 'already' or 'still' is lost and 'already' adopts (so to say) t as an argument (Figure 1).

The diagram in Figure 2 illustrates the first paraphrase in the style of (van der Auwera, 1993): the real situation in which e already happens at t can be thought of in contrast with an anterior or imaginary (potential or counterfactual) one in which it does

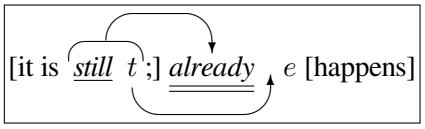


Figure 1: Reconstruction of the motivation for using ‘still’ and ‘already’ for ‘as early as’

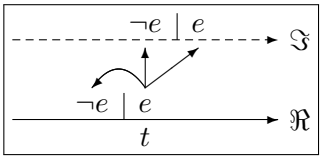


Figure 2: ‘at t, already e happens’

not (yet) do so, as well as a posterior imaginary one in which it does. And as per the second paraphrase it is still t when e happens, and one might say that the contrast is with a later situation in which it is no longer t, as well as a potential or counterfactual one where t strikes before e. But since t is a time, it makes sense to have it trade places with e and to align the times of the real and the imaginary axis, as shown in Figure 3. (Whether e can be said to happen after t depends on its aspectual characteristics, hence the question mark and the dashed border at the possible change of state.) In this way the diagram in Figure 3(b) becomes identical to the one in Figure 2 in the relation of the situation to the imaginary axis, where e only happens after t; the difference is that on the real axis the former ‘looks’ backwards and the latter forwards. This essentially confirms Богуславский (1996)’s description, although it seems not to tell the whole story, even for Russian:

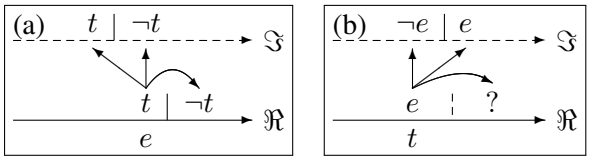


Figure 3: ‘it is still t when e happens’

(11) *Зевс знал, как много горя и зла принесет с собой в мир сын златой Афродиты. Он хотел, чтобы умертвили его ещё при рождении.* ‘Zeus knew how much grief and evil the son of golden Aphrodite would bring into the world. He wanted him to be killed at birth.’ (Nikolay Kun, *Greek Myths and Legends*)

What is highlighted here is not that this was *long before* a future reference point; Zeus’s idea must have been to annihilate Eros before he has managed to work any mischief, that is, *as soon as possible* after his appearance in the world. Clearly more factors are at work, even more so in view of the cross-linguistic variation.

Lexicographers are not always attentive to these uses of ‘still’ and ‘already’ words. The multi-volume *Dictionary of the Bulgarian Language* lists the sense ‘as early as’ under the word ***още*** (Павлова and Гунова, 2002) as its 4th meaning, with nine illustrations of use of *още* with temporal or circumstantial adjuncts (e. g., *Още като ученик се интересувал от археология* ‘Still a school student, he was interested in archeology’). No examples of use with a subject are given. There is one – taken from an older dictionary (Аргиров et al., 1927: 277–278) – in the entry for ***вече*** (Чоролеева, 1979), under its 9th meaning (*Вече Априлов замисляше за български университет* ‘Already Aprilov was thinking of a Bulgarian university’); this use, however, is marked as obsolete.

The *Dictionary of the Ukrainian Language* does not list ‘no later than’ as a separate meaning of ***вже*** (Радченко, 1979); one example of such use of the word is given (*Уже сьогодні їх заручають, А через тиждень і повінчають* ‘Already today they are being engaged, and in a week they shall be married’), but with no mention of its being different from the others in the entry, although it is, witness the fact that it is the only one in which *още* and not *вече* would be employed if translated into Bulgarian (preserving the general structure). In the entry for ***ще*** (Горобець, 1980) the sense discussed here is listed as the 3rd meaning, with ten examples of use, though one of them (*Ще треті півні не співали* ‘The third cocks had not crowed yet’) illustrates a different sense from the rest. In this dictionary, too, there are no examples where the modified phrase is a subject.

We will now look at these Bulgarian and Ukrainian words more closely, and especially at the correspondences of Bulgarian *още* to Ukrainian *вже*.

2 On the Parallel Corpus

The working Bulgarian–Ukrainian parallel corpus is composed entirely of fiction (mostly novels, but also short stories), including both original Bulgarian and Ukrainian texts and translations from other languages. The material has been obtained from electronic libraries or from paper editions through scanning, optical character recognition and error correc-

tion by *ad hoc* software tools and by hand. The current version is made of ten sectors, each composed of texts with the same original language and counting approximately 15.2 mln words on the Bulgarian and 13.4 mln words on the Ukrainian side. This amounts to a total size close to 28.6 million words. Two sectors contain translations from English and two from Russian (because of the larger amount of text available); the remaining original languages are Bulgarian, French, German, Italian, Polish and Ukrainian. All texts are aligned at sentence level.[2]

3 Results

In the corpus we found **652** pairs of sentences in which Bulgarian *още* corresponds to Ukrainian *вже*, with both expressing the meaning 'as early as'.

3.1 Distribution by sector

The distribution across the sectors of the corpus (that is, by original language) is very uneven: very many examples are found in translations from German, not quite so many from Polish, few from Russian, Italian and French, and very few from English (Table 1). The original text may feature a 'still' word (though this happens very seldom except in Bulgarian), neither 'still' nor 'already', an 'already' word used as a sentential modifier (this is rare, too, except in German), or one with a 'local' use, that is, applied to a constituent as in Ukrainian (the most common scenario):

(12) *Но го познаха още отдалече.* ‖ *Та впізнали його вже здалеку.* ← ru *Но узнали его ещё издали.* 'But they recognised it from some distance away.' (K. Bulychev, *Those Who Survive*)

(13) *Ако дохождаш например в четири часа следобед, аз още от три часа ще почна да се чувствувам щастлива.* ‖ *Якщо ти прийдеш, скажімо, о четвертій дня, то я вже з третьої години відчуватиму себе щасливим.* ← fr *Si tu viens, par exemple, à quatre heures de l'après-midi, dès trois heures je commencerai d'être heureux.* 'If, for example, you come at four o'clock in the afternoon, then at three o'clock I shall begin to be happy.' (A. de Saint-Exupéry, *The Little Prince*)

(14) *Всичко започна още тогава.* ‖ *Це вже тоді у нього почалося.* ← de *Damals fing es schon mit ihm an.* 'It all began back then already.' (A. Seghers, *The Seventh Cross*)

(15) *Но още първия ден се убеждавам, че съм само сянка на някогашните египетски владетели!* ‖ *Але вже в перший день переконуюсь, що я лише тінь давніх володарів Єгипту!* ← pl *Lecz już w pierwszym dniu przekonywam się, że jestem tylko cieniem dawnych władców Egiptu!* 'But on the very first day I realise that I am merely a shadow of the former rulers of Egypt!' (B. Prus, *Pharaoh*)

The zeroes in the English and French cells of the last row of Table 1 echo (Välikangas, 1982: 392): "In Spanish and Italian, constructions such as *ya entonces*, *già da bambino* are quite common. French stands apart. It is not that *déjà* is not freely used before an attribute or an epithet (*cet homme (est) déjà vieux* 'this man (is) already old'), or that it is not possible (alongside *dès*, which is more common) before a circumstantial phrase (*déjà au Moyen Age* 'already in the Middle Ages', *déjà hier* 'just yesterday'): yet the texts give the impression – confirmed for the spoken language by my informants – that the combination *déjà*+modified phrase is relatively rare. *Déjà* frequently appears after the modified phrase (*L'an dernier déjà* 'last year already', *depuis longtemps déjà* 'for a long time already'), and even more often it is detached from the rest of the sentence." Among the *още* : *вже* 'as early as' examples in the corpus there is one whose French original illustrates this detachment:

(16) *Още когато защищавах сирачетата, побързах да се избавя от тях, отстранявайки ги от полесражението.* ‖ *Уже в ту пору, коли я був оборонцем сиріт, я, прагнучи здихатися їх, велів їм ховатися.* ← fr *Déjà, du temps que je protégeais les orphelines, je commençais par me débarrasser d'elles en les envoyant se cacher.* 'During the period in which I was protecting female orphans, I had already begun to get rid of them by sending them into hiding.' (J.-P. Sartre, *The Words*)

[2] See (Siruk and Derzhanski, 2013) for more details on the general make-up of the corpus, as well as (Derzhanski and Siruk, 2024) and (Держански and Сирук, 2025) for recent examples of using the corpus in comparative research of selected grammatical categories and function words in Bulgarian and Ukrainian.

	Bg	De	E1	E2	Fr	It	Pl	R1	R2	Uk	total
‘still’ (local)	73	4	0	0	0	0	6	0	1	—	**84**
neither ‘still’ nor ‘already’	—	24	12	20	33	21	11	0	9	—	**133**
‘already’ (sentential)	—	17	1	2	2	2	2	1	0	—	**26**
‘already’ (local)	—	116	0	0	0	25	95	31	44	100	**409**
total	**73**	**161**	**13**	**22**	**35**	**48**	**114**	**32**	**54**	**100**	652

Table 1: Lexical choices in the original texts corresponding to Bulgarian *още* : Ukrainian *вже* ‘as early as’

Välikangas goes on to report that in English, too, such usage is discouraged by some style guides, e. g., (Beesley and Hart, 1963): “[Already] can only modify a verb. It cannot modify the whole sentence, another adverb or adjective. [For translating Finnish *jo*, one is advised to use] some emphatic adverb or adverbial phrase, such as ‘even’, ‘as early as’, etc.”. Again, the relevant part of the corpus provides one example in which *already* might be said to be preposed to an adverbial, but it is detached from it by a comma in the same way as *déjà* is in (16):

(17) *Съсредоточи вниманието си върху Сермак, същият Сермак, който още през първия си мандат в Градския съвет бе обръщал неведнъж с главата надолу този улегнал орган на властта […]* ‖ *Саме на Сермаку слід зосередити свою увагу, це він уже під час свого першого терміну в міській раді перевернув цей статечний орган влади догори дриґом […]* ← en *It was on Sermak that he concentrated, the Sermak who had already, in his first term in the City Council, turned that sedate body topsy-turvy more than once […]* (I. Asimov, *Foundation*)

Thus the zeroes in the bottom cells of the E1, E2 and Fr columns in Table 1 are actually what should be expected.

3.2 Yesterday. Today. Tomorrow

For an idea of how the position of the event on the time axis affects the expression in both languages, we shall look at the use of *още/ще* and *вече/вже* as modifiers of the frequent time adverbs *вчера/вчора* ‘yesterday’, *днес/сьогодні* ‘today’ and *утре/завтра* ‘tomorrow’.

Table 2 reports how many times the word meaning ‘still’ or ‘already’ occurs adjacent to one of these time adverbs in the corpus, regardless of the syntactic structure. It is evident that in both languages the word for ‘today’, *днес/сьогодні*, is the most frequent one in these contexts; but they are followed by Bulgarian *утре* ‘tomorrow’ and Ukrainian *вчора* ‘yesterday’ (whose totals are similar), with Bulgarian *вчера* ‘yesterday’ and Ukrainian *завтра* ‘tomorrow’ (whose totals are even closer) coming last. Let us look at these expressions more closely.

Neither language features more than a few occurrences of the word for ‘already’ next to ‘yesterday’. Both employ a postposed ‘still’, which yields essentially the same meaning analysed as modifying the sentence (‘as of yesterday, still …’) or the adverb (‘as late as yesterday’), to describe a situation which obtained on the day before in contrast with the day of reference, although there is only one occasion on which they do so together (18). In Ukrainian a preposed *ще* can have the same effect (19), nearly always with an imperfective verb, but exceptionally with a perfective one as well (20). Bulgarian differs, according to the evidence of the corpus. The meaning ‘as late as yesterday’ is often (on 40 of 75 occasions where Ukrainian has *ще вчора* and Bulgarian neither *още* nor *вече*) expressed by *до вчера* ‘until yesterday’ (21), but by a preposed *още* only once, in a translation from Ukrainian (22):

(18) *Този пеньоар вчера още го нямаше.* ‖ *Цього халата вчора ще не було.* ← de *Den Schlafrock hatte sie gestern noch nicht.* ‘She didn’t have that dressing gown yesterday.’ (E. M. Remarque, *The Black Obelisk*)

(19) *Вчера още ѝ казвам: Данушке, ще останеш ли вярна на Збишко?* ‖ *Ще вчора я її питаю: «Данусько, будеш вірна Збишкові?»* ← pl *Wczoraj jeszcze mówię jej: Danuśka, a dotrzymasz-li ty Zbyszkowi?* ‘Only yesterday I said to her: “Danusia, will you keep faith with Zbyszko?”’ (H. Senkiewicz, *The Teutonic Knights*)

(20) *Вчера още, заплашен от заминаването на киевския воевода, князът беше решил да спре да воюва, докато не събере още сили […] — но днес отново стоеше начело на близо дванайсетхилядна армия […]*

	~ *още*	*още* ~	~ *вече*	*вече* ~	total
вчера	12	43	2	2	**59**
днес	60	203	68	1	**332**
утре	2	153	20	1	**176**
total	**74**	**399**	**90**	**4**	567

	~ *ще*	*ще* ~	~ *вже*	*вже* ~	total
вчора	14	130	6	4	**154**
сьогодні	33	82	53	58	**226**
завтра	10	6	19	32	**67**
total	**57**	**218**	**78**	**94**	447

Table 2: Numbers of occurrences of 'still' and 'already' words next to 'yesterday', 'today' and 'tomorrow' on the Bulgarian and the Ukrainian side of the corpus

‖ *Ще вчора князь, засмучений тим, що від нього йде київський воєвода, вирішив був уже до появи нових підкріплень війну припинити […], сьогодні ж він стояв на чолі майже дванадцятитисячної армії […]* ← pl *Wczoraj jeszcze książę, zagrożony odejściem wojewody kijowskiego, już był postanowił wojnę aż do chwili przybytku sił zawiesić […] — dziś stał znów na czele blisko dwunastotysięcznej armii […]* 'Only yesterday, through the disaffection of the Voyevoda of Kyiv, the prince had determined to defer the war until troops should come to his assistance; […] — and, today, he stood at the head of twelve thousand men […]' (H. Senkiewicz, *With Fire and Sword*)

(21) *Тъкмо вземаше хляб от полицата, която е над полицата с чиниите, до вчера там имаше седем самуна, а днес са вече пет.* ← *Саме діставав хліб з полиці, що над мисником, ще вчора там було сім буханців, а сьогодні вже п'ять.* 'He was just taking some bread from the shelf above the dish rack; yesterday there had been seven loaves, but today there were only five.' (V. Zemlyak, *Green Mills*)

(22) *Още вчера са пасли спокойно някъде по рътлините, а сега вървят уморени, падат и стават, препъват се в неравния паваж […].* ← *Ще вчора спокійно паслися десь по балках, а зараз ідуть виморені, поперепадавшись, спотикаючись по бакаюватому бруку […].* 'Yesterday they had still been peacefully grazing in gullies, and now they were moving along exhausted, suffering, tripping on the uneven pavement […].' (O. Honchar, *The Cathedral*)

In (23) the preposition *до* connects graphically with the adverb *вчера*, yielding *довчера* 'until yesterday', again with teh interpretation 'as late as yesterday'.

(23) *Роксолани — една дума, която за Европа довчера не означаваше нищо.* ← *Роксолани. Слово, яке для Європи ще вчора нічого не значило.* 'Roxolani. A word that meant nothing to Europe just yesterday.' (P. Zahrebelnyi, *Roksolana*)

Elsewhere Bulgarian *още вчера* stands for 'as early as yesterday' (24), which in Ukrainian can also be expressed as *вже вчора*, but this correspondence, too, only occurs once, in a translation from Russian, and influence from the original seems likely (25).

(24) *[…] аз още вчера си помислих, че ти не можеш да я приготвиш като моя готвач.* ‖ *Я ще вчора знав, що ти не зумієш зробити цієї страви так, як мій кухар.* ← de *[…] das habe ich mir schon gestern gedacht, daß du diese Speise nicht machen kannst wie mein Koch.* 'I was thinking all yesterday that you could not make this pie like my cook.' (W. Hauff, *Nosey, the Dwarf*)

(25) *Сега ми е ясно — казах аз — защо още вчера вашите съотечественици знаеха и за Алиса, и за долитането на нашия кораб.* ‖ *Тепер для мене зрозуміло, — сказав я, — чому вже вчора ваші співвітчизники знали і про Алісу, і про приліт нашого корабля.* ← ru *Теперь мне понятно, — сказал я, — почему уже вчера ваши соотечественники знали и об Алисе, и о прилёте нашего корабля.* '"Now I understand," I said, "why yesterday your compatriots already knew about Alice and the arrival of our ship."' (K. Bulychev, *Alisa's Voyage*)

With 'today', too, a postposed 'still' is used for 'as late as today' or, with negation, 'not yet today' (26).

(26) *Днес още не мога — отговори Юлка.* ‖ *Мені сьогодні ще не можна, — відповіла Юлька.* ← ru *Мне сегодня ещё нельзя, — сказала Юлька.* '"I can't [play volleyball] today yet [*sc.*

so soon after my appendectomy],” said Yulka.’ (K. Bulychev, *One Hundred Years Ahead*)

Bulgarian is consistent in using *днес още* for ‘as late as today’ and *още днес* for ‘as early as today’, bar the occasional exception, whereas Ukrainian employs *сьогодні ще* for the former meaning, *ще сьогодні* for both (27) and *вже сьогодні* for the latter (28–29). But also often (on 76 occasions out of 137 where Bulgarian has *още днес* and Ukrainian neither *ще* nor *вже* the time adverb is followed by the emphatic particle *ж* (30).

(27) *След малко се върна с поздрав от Лигия и с обещание, че още днес ще му донесе отговора ѝ.* ‖ *Невдовзі він повернувся з вітанням од Лігії й пообіцяв, що ще сьогодні принесе відповідь.* ← pl *Po chwili wrócił z pozdrowieniem od Ligii i z obietnicą, że dziś jeszcze odniesie jej odpowiedź.* ‘After a while he returned with a greeting from Lygia, and promised to deliver her answer that day.’ (H. Senkiewicz, *Quo Vadis*)

(28) *А щом жреците още днес изпращат делегация, трябва да са слаби.* ‖ *А жерці вже сьогодні висилають делегацію, отже, вони почувають своє безсилля.* ← pl *A jeżeli kapłani już dziś wysyłają delegację, muszą być słabi.* ‘And if the priests are already sending a delegation today, they must be weak.’ (B. Prus, *Pharaoh*)

(29) *Важното е, че тя ще бъде негова, и то още днес.* ‖ *Передовсім вона буде йому належати, й це вже сьогодні.* ← pl *Przede wszystkim będzie jego i to dziś jeszcze.* ‘First and foremost, she will be his, this very day.’ (ibid.)

(30) *[…] кълна се в духовете на предците, какво ли не бих дал да го сторя още днес!* ‖ *[…] ларами присягаюся — багато віддав би я, щоб вчинити так сьогодні ж!* ← ru *[…] ларами клянусь — я отдал бы многое, чтобы сделать это сегодня.* ‘[…] by the lares I swear — I’d give a lot to be able to do so today!’ (M. Bulgakov, *The Master and Margarita*)

The meaning ‘tomorrow still = as late as tomorrow’ (31–32) is evidently in very low demand.

(31) *Във всеки случай и утре все още ще съм ранен, без пари, без ангажименти и без Мари — значи не е належащо.* ‖ *А втім, я і завтра буду сидіти з розбитим коліном, без грошей, без роботи і без Марі — так що нічого спішного тут немає.* ← de *Allerdings werde ich morgen auch noch verletzt, ohne Geld, ohne Engagement und ohne Marie sein — es ist also nicht dringend.* ‘To tell the truth, tomorrow I’ll still be injured, with no money, no bookings and no Marie — so it’s not really urgent.’ (H. Böll, *The Clown*)

(32) *Сбогувам се с фараона и моля боговете да ми позволят и утре да го поздравя като фараон…* ‖ *Прощаюся з фараоном і благаю богів, щоб вони дали мені привітати його завтра ще як фараона…* ← pl *Żegnam faraona i proszę bogów, ażeby pozwoliły mi jutro powitać cię jeszcze jako faraona…* ‘I take my leave from the pharaoh and pray to the gods that they may allow me to greet you still as pharaoh tomorrow…’ (B. Prus, *Pharaoh*)

For ‘as early as tomorrow’, Bulgarian uses *още утре*, which corresponds to Ukrainian once (33); sometimes we find *вже завтра* (34), but more often (on 61 occasions out of 134) the particle *ж* (35).

(33) *Едната ще отиде още утре заран в тайника волво.* → *Одна ще завтра вранці буде в тайнику.* ‘One will go into the Volvo’s secret compartment first thing tomorrow morning.’ (B. Raynov, *Typhoons with Gentle Names*)

(34) *Може би ще се окаже още утре, че трябва да се махна.* ‖ *Можливо, мені вже завтра доведеться тікати.* ← de *Vielleicht heißt's schon morgen: fort.* ‘Maybe I’ll have to leave tomorrow already.’ (A. Seghers, *The Seventh Cross*)

(35) *Трябва да пиша на татко още утре.* ‖ *Треба завтра ж татові написати.* ← fr *Il faut que j'écrive à papa dès demain.* ‘I need to write to Dad tomorrow.’ (G. de Maupassant, *Bel-Ami*)

In the exceptional example (33) one notes the adjacency of *завтра* to *вранці* ‘in the morning’, which when used alone usually combines with *ще* (and refers to the morning of the same day (36); when it refers to the following or any later morning, the choice is *вже* (37), but there are 21 occurrences of the former and 5 of the latter).

(36) *Ами онези, Доренце, аз още сутринта ги проверих.* ← *A ті, Дорочко, я ще вранці перевірив.* 'As for those, Dora, I already checked them in the morning.' (V. Drozd, *Yrii*)

(37) *И сам Сюлейман я забрави още на другата сутрин.* ← *І сам Сулейман забув про неї уже вранці.* 'And Suleiman himself had already forgotten about her on the [next] morning.' (P. Zahrebelnyi, *Roksolana*)

So we see that the disbalance in Table 2 results from several factors: Bulgarian normally expresses 'as late as' by ~ *още* and 'as early as' by *още* ~, Ukrainian uses ~ *ще* for the former, *ще* ~ for both and *вже* ~ for the latter; among the combinations of these two meanings with the three time adverbs, the more frequent ones are 'as late as yesterday' (often marked in Bulgarian by the preposition *до* 'until' rather than *още* or *вече*) and 'as early as today/tomorrow' (often marked in Ukrainian by the particle *ж* rather than *ще* or *вже*).

An 'already' word postposed to a time adverb (which was not discussed above) can be interpreted compositionally, as a modifier of the sentence. But it can serve as a translation of an 'as early as' construction, thanks to the proximity of the meaning:

(38) *Вчера вече започнаха на Козелшчина.* 'Yesterday, they already started in Kozelshchyna.' ← *Уже вчора на Козельщині й почали.* (Y. Hutsalo, *The frenzied, mad Kirik*)

(39) *А утре вече ще се натикаме в устата на звера.* 'And tomorrow we will already push our way into the mouth of the beast.' ← *Ось і ми вже завтра поліземо в пащеку смерті.* (M. Stelmakh, *The Four Fords*)

Note that in the following (outwardly parallel) example the meaning is different: it is not about today's breaking of the speaker's fast but about the breaking of today's fast (although this is only revealed by the context):

(40) *Аз пък днеска вече се облажих.* ← *А я вже сьогодні скоромину їв.* 'Well, I have broken my fast already today.' (M. Stelmakh, *The Geese and Swans Are Flying*)

Let us now turn to the distribution of Ukrainian *ще* and *вже* as bearers of the meaning 'as early as' expressed in Bulgarian by *още*. Table 3 summarises the numbers of their occurrences with the three time adverbs we looked at, dividing the data for *сьогодні* 'today' by tense and modality (we mostly see it with future or imperative verb forms and expressions of possibility and necessity).

	ще	*вже*
вчора 'yesterday'	45 (24)	1 (25)
сьогодні 'today' (past, present)	0	6 (28)
сьогодні 'today' (irrealis)	50 (27)	20 (29)
завтра 'tomorrow'	1 (33)	14 (34)

Table 3: Number of uses of Ukrainian *ще* and *вже* for 'as early as' with 'yesterday', 'today' and 'tomorrow'

The very strong association of *вчора* 'yesterday' with *ще* and of *завтра* with *вже* may be supported by the frequent opposition of something still being the case at an earlier time, such as yesterday or today, to something else already happening at a later time, such as today or tomorrow (21).

Apart from *вчора* 'yesterday', *ще* is preferred by, e. g., *напередодні* 'on the day before', *того/тої (самого/-ої)* t 'on that (same) t' (where t is a time noun), *до* … 'before …', *звечора* 'since the evening', *вдосвіта* 'at dawn'. Apart from *завтра* 'tomorrow', *вже* is preferred by, e. g., *наступнього/-ої* t 'on the following t', *другого дня* 'on the next/second day', *на третій день* 'on the third day', *зараз* 'now, right away', *тепер* 'now, presently', *через* T 'T later' (where T is a length of time).

In brief, *ще* tends to be chosen if the time is located in the past relative to the reference point, and *вже* if it is in the future; looking at Fig. 2 and Fig. 3(b) again, in both cases the contrast on the real axis is with the reference point. We see this in (41), where the two markers cooccur:

(41) *Аз приключих всичко това толкова бързо, че още на следващата сутрин напуснах острова и бях в Биен преди обед.* ‖ *Я так поспішав, що вже назавтра вранці поїхав з острова і прибув до Б'єна ще до полудня.* ← fr *Je fis une telle diligence, que dès le lendemain matin, je partis de l'île et me rendis à Bienne encore avant midi.* 'I used so much diligence that the next morning I left the island and arrived at Bienne before noon.' (J.-J. Rousseau, 'The Confessions')

Here the first event (departure from the island) is placed in the morning *after* the time thereto discussed, then the next reference point is set at noon and the second event (arrival at Bienne) is *before* it.

With time adverbs such as *сьогодні* 'today', which are often encountered with both particles, the difference can be elusive. But it can be said that the meaning of *вже сьогодні* is closer to 'right now', and of *ще сьогодні*, to 'before tomorrow'; ***вже*** *сьогодні* is perceived as a qualified 'already', and *ще* ***сьогодні*** as a qualified 'today'. This intuition is fully in line with the motivation for the two constructions (Figure 1) and the reconstruction of the former as 'today is the time for event e to already happen' and of the latter as 'it will still be today when event e happens'.

4 Conclusions

Ukrainian is one of the languages in which a word for 'already' can be used, alongside a word for 'still', as a modifier of an expression with temporal semantics to express the meaning 'as early/soon as' ('no later than'). These languages are numerous, but the details differ (as shown by the tangible influence of the source language in the data drawn from our corpus: the examples are by far most numerous in translations from German, where the phenomenon is most pervasive, followed by a near-tie between Ukrainian originals and translations from Polish). This adds to the complexity and theoretical interest of the issue. Notably, however, the closely related Bulgarian is not such a language: it only uses 'still' (*още*), never 'already' (*вече*), in this function.

We have shown how a study of texts in parallel Bulgarian and Ukrainian versions can serve to efficiently detect occurrences of the Ukrainian 'already' word (*вже*) so used as a prelude to comparing them to uses of 'still' (*ще*) in similar contexts and analysing the correlation with the semantics of the temporal expression – in particular, the asymmetry between past and future (exemplified by 'yesterday' and 'tomorrow') – and the effect of the choice of one of the two modifiers on the meaning of the sentence. We have also seen how in certain cases the two languages resort to alternative means, such as the preposition *до* in Bulgarian or the particle *ж* in Ukrainian, for expressing a similar sense.

Such findings are of value for language learners, translators, writers of reference grammars and compilers of mono- and bilingual dictionaries; and within NLP, for sentence- and word-level alignment in parallel corpora, for the design of bilingual lexical resources and for machine translation.

References

J. van der Auwera. 1993. 'Already' and 'Still': Beyond Duality. *Linguistics and Philosophy*, 16(6):613–653.

A. R. Beesley and K. Hart. 1963. *A Guide to Finnish Mistakes in English: Grammatical Notes, Points to Avoid and Over 400 Practice Sentences*. Kielipalvelu ARB (2. ed.).

I. Derzhanski and O. Siruk. 2024. The verbal category of conditionality in Bulgarian and its Ukrainian correspondences. In *Proceedings of the Sixth International Conference on Computational Linguistics in Bulgaria (CLIB 2024)*, pages 187–195, Sofia. Department of Computational Linguistics, Institute for Bulgarian Language, Bulgarian Academy of Sciences.

H. Galton. 1976. "Noch" und "schon" im Deutschen und Russischen. *Folia Linguistica*, X(3/4):377–384.

S. Löbner. 1989. German *schon–erst–noch*: An integrated analysis. *Linguistics and Philosophy*, 12:167–212.

N. Reiter. 1987. "Schon"-lose Balkansprachen. *Cahiers balkaniques*, 12:102–133.

O. Siruk and I. Derzhanski. 2013. Linguistic corpora as international cultural heritage: The corpus of Bulgarian and Ukrainian parallel texts. *Digital Presentation and Preservation of Cultural and Scientific Heritage*, 3:91–98.

O. Välikangas. 1982. La notion de 'déjà' et les mots qui servent à la rendre dans quelques langues européennes. *Neuphilologische Mitteilungen*, 83(4):371–404.

Ю. Д. Апресян. 1979. Английские синонимы и синонимический словарь (postamble). In А. И. Розенман and Ю. Д. Апресян, editors, *Англо-русский синонимический словарь*, pages 500–543. Москва: Русский язык. [Ju. D. Apresjan. 1979. Anglijskie sinonimy i sinonimičeskij slovar' (postamble). In A. I. Rozenman and Ju. D. Apresjan, editors, *Anglo-russkij sinonimičeskij slovar'*, pages 500–543. Moskva: Russkij jazyk].

Ст. Аргиров, Ст. Младенов, А. Теодоров-Балан, and Б. Цонев, editors. 1927. *Български тълковен речник*, volume I. София: Печатница „Художник". [S. Argirov, S. Mladenov, A. Teodorov-Balan, and B. Tsonev, editors. 1927. *Bâlgarski tâlkoven rechnik*, volume I. Sofia: Pechatnitsa „Hudozhnik"].

И. М. Богуславский. 1996. *Сфера действия лексических единиц*. Москва: Языки русской культуры. [I. M. Boguslavskij. 1996. *Sfera dejstvija leksičeskix edinic*. Moskva: Jazyki russkoj kul'tury].

В. Й. Горобець. 1980. ЩЕ, ІЩЄ́. In С. І. Головащук, editor, *Словник української мови*, volume XI: Х–Ь, pages 575–576. Київ: Наукова думка. [V. J. Horobec′. 1980. ŠČE, IŠČÉ. In S. I. Holovaščuk, editor, *Slovnyk ukrajins′koji movy*, volume XI: Х–′, pages 575–576. Kyjiv: Naukova dumka].

И. Держански and О. Сирук. 2025. Българската служебна дума *си* и нейните украински съответствия. In *Доклади от Международната годишна конференция на Института за български език „Проф. Любомир Андрейчин“ (София, 2025 година)*, pages 176–186, София. Издателство на БАН „Проф. Марин Дринов“. [I. Derzhanski and O. Siruk. 2025. Bâlgarskata sluzhebna duma *si* i neynite ukrainski sâotvetstviya. In *Dokladi ot Mezhdunarodnata godishna konferentsiya na Instituta za bâlgarski ezik „Prof. Lyubomir Andreychin“ (Sofia, 2025 godina)*, pages 176–186, Sofia. Izdatelstvo na BAN „Prof. Marin Drinov“].

А. Мустайоки. 1988. О семантике русского темпорального *ещё*. *Studia Slavica Finlandensia*, V:99–142. [A. Mustajoki. 1988. O semantike russkogo temporal’nogo *eščë*. *Studia Slavica Finlandensia*, V:99–142].

С. Павлова and З. Гунова. 2002. ÒЩЕ. In Ю. Балтова and М. Чоролеева, editors, *Речник на българския език*, volume XI: О, pages 1235–1237. София: Акад. изд–во „Проф. Марин Дринов”, ET „ЕМАС”. [S. Pavlova and Z. Gunova. 2002. ÒSHTE. In Yu. Baltova and M. Choroleeva, editors, *Rechnik na bâlgarskiya ezik*, volume XI: O, pages 1235–1237. Sofia: Akad. izd–vo „Prof. Marin Drinov”, ET „EMAS”].

Є. М. Радченко. 1979. УЖÉ (ВЖЕ). In А. А. Бурячок and Г. М. Гнатюк, editors, *Словник української мови*, volume X: Т–Ф, pages 400–401. Київ: Наукова думка. [Je. M. Radčenko. 1979. UŽÉ (VŽE). In A. A. Burjačok and H. M. Hnatjuk, editors, *Slovnyk ukrajins′koji movy*, volume X: T–F, pages 400–401. Kyjiv: Naukova dumka].

М. Чоролеева. 1979. ВÈЧЕ. In Кр. Чолакова, editor, *Речник на българския език*, volume II: В, pages 128–129. София: Изд–во на БАН. [M. Choroleeva. 1979. VÈCHE. In K. Cholakova, editor, *Rechnik na bâlgarskiya ezik*, volume II: V, pages 128–129. Sofia: Izd–vo na BAN].

Appendix A Words for ‘still’ and ‘already’

English	*still*	*already*
Bulgarian	*още*; *все още*	*вече*
French	*encore*	*déjà*
Finnish	(*vielä*)	*jo*
German	*noch*	*schon*
Italian	*ancora*	*già*
Polish	*jeszcze*	*już*
Russian	*ещё*	*уже*
Spanish	(*aún*, *todavía*)	*ya*
Ukrainian	*ще/іще*; (*все ще*)	*вже/уже*

Syntactic Patterns in Political Speech: Dependency-Based Ad Hominem Detection in Bulgarian Social Media

Nevena Grigorova
Gate Institute
Sofia, Bulgaria
nevena.grigorova@gate-ai.eu

Stanislav Penkov
Gate Institute
Sofia, Bulgaria
stanislav.penkov@gate-ai.eu

Keith Kiely
Gate Institute
Sofia, Bulgaria
keith.kiely@gate-ai.eu

Dobromir Tsolyov
Gate Institute
Sofia, Bulgaria
dobromir.tsolyov@gate-ai.eu

Iglika Ivanova
Gate Institute, Sofia University
Sofia, Bulgaria
iglika.ivanova@gate-ai.eu

Abstract

This paper explores the automatic detection of ad hominem attacks in Bulgarian political social media discourse by leveraging dependency-based structural features. Based on a linguistic analysis of ad hominem constructions, we derive structural patterns using Universal Dependencies relations. Grounded in a corpus of Bulgarian political discourse annotated for logical fallacies, we compare three classification settings: structural features alone, lexical features based on TF–IDF representations, and a combined model integrating both. Experiments with logistic regression, decision trees, and support vector machines show that dependency-based structural features achieve competitive detection performance and provide complementary signal to lexical TF–IDF baselines, with performance estimates sensitive to evaluation protocol in this small-data setting. Overall, ad hominem discourse exhibits identifiable syntactic configurations that can complement lexical cues in its automatic detection. The study demonstrates the value of linguistically informed structural features for analyzing political rhetoric and contributes computational methods for studying argumentative strategies in a low-resource language such as Bulgarian.

Keywords: ad hominem, political speech, social media, NLP methods, dependency parsing

1 Introduction

Research on offensive and manipulative language in social media has become a key focus across multiple disciplines, including media studies and linguistics, highlighting the need for a multidisciplinary approach that integrates methods from computer science, particularly Natural Language Processing (NLP).

Advances in NLP have significantly improved the ability to analyze and generate human language, enabling more sophisticated approaches to detecting problematic discourse. One area of growing interest is the identification of logical fallacies as a tool for fostering critical thinking and limiting the spread of misinformation. In the context of "errors in reasoning that can undermine the validity of an argument" (Tindale, 2007), an important subtype is ad hominem, that is, arguments that target an individual rather than their position. Such attacks are especially prevalent in politically sensitive contexts and often contribute to misleading or manipulative communication.

This study models the syntactic structure of ad hominem attacks in a systematic way using tools from computational linguistics. A defining feature of such arguments is the presence of negative personal targeting, which can be identified through recurring syntactic patterns.

The investigation relies on a linguistic analysis of selected examples, for which UD-style annotations were inspected using UDPipe (Straka, 2018) for parse visualization, and on dataset-wide UD parses produced with Stanza for feature extraction and modelling. This methodology enables the extraction and generalization of structural patterns characteristic of ad hominem discourse.

The experimental evaluation assesses whether these syntactic features provide meaningful signals for detecting such arguments. Improved performance of models incorporating structural information, compared to those relying solely on lexical features, would support the hypothesis

that dependency-based representations encode distinctive and informative patterns for identifying ad hominem attacks.

2 Related Work

The detection of offensive and manipulative language in online platforms has become a major topic in NLP, with surveys highlighting both the societal urgency and the technical complexity of modelling abusive, hateful, and toxic discourse (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018). Early work in this area largely relied on surface-level lexical features such as dictionaries, bag-of-words, and n-grams, which proved effective for explicit slurs and profanity but struggled with more implicit or context-dependent forms of abuse (Fortuna and Nunes, 2018: pp. 2–4); (Wenjie and Zubiaga, 2021). This limitation has motivated a shift towards richer representations that incorporate syntax, discourse structure, and pragmatic cues in order to better capture the underlying strategies of offensive and manipulative communication. Within this broader line of research, logical fallacies and related manipulations have recently been proposed as a useful lens for understanding how arguments mislead or manipulate audiences, and several projects now aim to operationalise fallacy theory in computational models (Lawrence and Reed, 2019; Cruz et al., 2023).

3 Ad Hominem in Political Speech

Ad hominem attacks occupy a central place in the argumentation-theory literature on manipulation techniques because they directly target an interlocutor's character or credibility rather than addressing the merits of their claims, thus functioning as a powerful mechanism of delegitimisation and polarisation in political debate (Walton, 1998; Tindale, 2007). From an argumentation perspective, ad hominem is typically treated as a violation of the "freedom rule" in pragma-dialectics or as a derailment of the dialogical norms that should govern critical discussion (Walton, 1998: pp. 218–225).

Whereas many formal fallacies are inherently invalid, in political contexts, an ad hominem move is not always fallacious, creating unique hurdles for automated NLP detection. In political and legal contexts, questioning a speaker's credentials, past misconduct, or conflicts of interest can be a "reasonable move" if it bears on their reliability as a source of testimony. An attack only becomes a fallacy when the personal attribute (e.g., a candidate's appearance or a past speeding ticket) is logically irrelevant to the truth of the policy claim they are defending. This distinction between an attack assessing reliability and dismissing a claim entirely is central to Tindale (2007). These properties of ad hominem in political discourse make it significantly harder for computational models to detect than other fallacies.

The conceptual distinction between a fallacy and an insult as a rhetorical strategy is prominently developed by Kišiček (2013) in her analysis of parliamentary discourse. Her work emphasizes that many personal attacks in politics are not errors in reasoning (fallacies) but are pure slanging or rhetorical devices intended to rattle an opponent.

In their study "Breaking Down the Invisible Wall of Informal Fallacies in Online Discussions," Sahai et al. (2021) demonstrated that identifying fallacies like ad hominem is highly context-dependent, and neural models achieve higher accuracy when they can "see" the preceding dialogue rather than analyzing a single comment in isolation. Similarly, Habernal et al. (2018) investigated how much debate context is required for both humans and machine learning systems to accurately recognize an ad hominem.

In political communication, empirical studies show that such attacks are pervasive and can blur the line between legitimate scrutiny of a politician's ethos and purely abusive, diversionary discourse (Kišiček, 2018). These observations have motivated specific computational work on ad hominem detection, treating it as a fine-grained subtype of abusive language that requires modelling both content and argumentative role.

In NLP, ad hominem has been addressed both as part of broader datasets on argumentative quality and as a dedicated classification task. Habernal et al. (2018) introduce resources from online debates that annotate various fallacies and argumentative weaknesses, including ad hominem, enabling supervised learning approaches that go beyond coarse "toxic vs. non-toxic" labels. These studies collectively suggest that identifying ad hominem requires attention to how negative personal targeting is realised in context, not only which words are used.

Parallel developments in propaganda and argumentation mining further support the move towards structurally informed models. Shared

tasks on fine-grained propaganda detection, such as the SemEval-2020 Task 11, treat techniques like name-calling and "loaded language" as structured rhetorical devices and show that systems combining lexical information with contextual and syntactic features achieve strong performance (Da San Martino et al., 2020: pp. 1377–1380). Work on argumentation structure parsing in persuasive essays demonstrates that dependency-based and discourse-level representations can effectively capture relations such as support and attack between argument components (Stab and Gurevych, 2017: pp. 619–623). In a similar spirit, recent fallacy-detection frameworks argue that robust models must account for semantics, discourse structure, pragmatics, and world knowledge, and they often rely on manually annotated corpora where specific fallacy types are marked at the span or argument level (Lawrence and Reed, 2019; Cruz et al., 2023).

Despite these advances, most existing resources and systems focus on English or a small set of high-resource languages, and they are trained on relatively large, domain-specific corpora. Multilingual surveys underline that offensive language detection in low-resource and morphologically rich languages remains challenging, and that leveraging linguistic knowledge and structural features is especially valuable when data is scarce (Fortuna and Nunes, 2018; Wenjie and Zubiaga, 2021). This creates a clear gap for work that combines dependency-based syntactic patterns with traditional lexical features to detect ad hominem attacks in low-resource political settings, such as Bulgarian social media, thereby testing whether the recurring configurations identified in prior studies generalise beyond English and high-resource domains.

4 Data

We use a Bulgarian political social media corpus collected via CrowdTangle and annotated for logical fallacies by six annotators. To ensure label reliability, we restrict the present study to high-agreement items, i.e., texts for which annotators show at least 85% agreement on the ad hominem label (majority-consistent subset), rather than reporting a separate inter-annotator agreement statistic.

For this study, we focus on ad hominem and formulate a binary sentence classification task (1 = `ad_hominem`, 0 = `control`). Positive instances were collected from the annotator sheets by selecting items labeled `ad_hominem` under the agreement threshold and segmenting the corresponding texts into sentences. Negative instances were drawn from the same domain and curated as non-`ad_hominem` controls. The resulting dataset contains 177 sentences in total (53 positives, 124 negatives; positive rate: 29.9%).

All sentences were parsed with Stanza (Qi et al., 2020) using its Bulgarian UD pipeline, aligned with the UD Bulgarian-BTB treebank (Universal Dependencies; Universal Dependencies, GitHub; Simov et al., 2002) for tokenization, POS tagging, lemmatization, and dependency parsing. For each sentence, we retain token-level UD annotations and derive sentence-level structural features from these parses. Evaluation uses the stratified protocols described in Section 6.

5 Linguistic Analysis and Patterns

5.1 Dependency-Based Representation

The present study adopts the Universal Dependencies (UD) framework as its formal basis for syntactic representation, specifically relying on the UD Bulgarian-BTB treebank. All dependency patterns are defined using standard UD relations such as `nsubj`, `amod`, `conj`, `ccomp`, `xcomp`, `obl`, `nmod`, `cop`, and `aux`, consistent with the Universal Dependencies framework and its annotation conventions (Nivre et al., 2020; Kübler et al., 2009).

Because the analysis of political statements containing ad hominem attacks requires modelling relations between words within the sentence, Part-of-Speech (POS) tags alone are insufficient for our purposes. However, POS and related morphological features can support the automatic detection of other markers associated with disinformation and manipulative language, such as the Bulgarian evidential renarrative (Temnikova et al., 2025). In the present study, we include UPOS as additional features while focusing primarily on dependency relations.

Unlike constituency-based approaches, which represent syntax through phrase-level structure, dependency parsing captures direct relations between individual words within a sentence. This makes it particularly suitable for identifying relational patterns associated with ad hominem attacks.

Based on the analyzed data, five recurring

dependency-based configurations were identified. They were selected because they represented the most frequent syntactic realizations of ad hominem attacks in the dataset.

- `amod` → `nsubj` → `root` (qualified subject)
- `nsubj + cop + NEG_PRED` (predicative insult)
- `nsubj` → `root` → `ccomp/xcomp` (attack in complement clause)
- `amod + conj` (coordinated discrediting)
- `nsubj:pass + NEG in obl/nmod` (passive discrediting)

Building on these observations, the identified configurations can be abstracted into generalized dependency templates. Moving from individual annotated instances to such templates enables the systematic formalization of recurring structures and supports their application in automatic detection. Accordingly, these configurations provide a foundation for pattern-based modelling of ad hominem.

5.2 Structural Patterns of Ad Hominem Attacks

The five patterns proposed below are not standard rhetorical categories. Rather, they are dependency-based syntactic realizations of ad hominem, derived from the analyzed data and informed by classical argumentation theory (Walton, 1998) as well as research on abusive language detection (Wenjie and Zubiaga, 2021).

The classification is intended as an operational taxonomy for feature extraction rather than a linguistic typology of argumentation. Consequently, some closely related constructions are represented as separate patterns when they correspond to distinct dependency structures that can be reliably identified and encoded as computational features.

Complement clause constructions often instantiate the same underlying attributive or predicative configurations as matrix clauses. However, within the proposed taxonomy they are represented as separate dependency-based patterns because clausal embedding (e.g., via ccomp or xcomp) introduces distinct dependency relations that are captured by different computational features during dependency parsing and feature extraction.

In the present study, the notion of ad hominem is operationalized to include attacks directed not only at individual persons but also at collective political entities, such as political parties, institutions, and social groups. In political discourse, such entities frequently function as proxies for their members, representatives, or supporters and therefore constitute legitimate targets of personal discrediting strategies. Across the data, the most recurrent configuration involves a negative lexical item targeting a person or group in a syntactic position associated with either the subject or the predicate.

To capture this phenomenon, the analysis introduces a set of higher-level semantic abstractions, including labels such as `NEG_PRED`, `NOUN_person/group`, and `NEG`. These labels are not part of the UD annotation scheme but serve as analytical categories that encode evaluative and referential properties relevant to the identification of ad hominem attacks. In this way, the proposed patterns operate at the interface between syntactic structure and semantic interpretation: UD relations provide the formal structure, while the additional labels capture the evaluative dimension of the utterance. Importantly, the proposed dependency patterns are not assumed to be unique to ad hominem discourse. Each of the identified syntactic configurations may also occur in semantically neutral or non-argumentative contexts. Their relevance lies not in the syntactic structure alone, but in the interaction between structural organization and negative evaluative lexical content targeting a person or social group. Consequently, the patterns should be interpreted as recurring structural realizations of ad hominem attacks rather than as sufficient conditions for their identification.

The analysis does not include automatic coreference resolution. When examples contain pronouns such as *he* or *they*, these are treated only as locally interpretable elements of the selected sentence or as part of the human linguistic interpretation of the example. The computational features extracted for classification do not resolve antecedents across preceding turns or wider discourse context. Consequently, attacks whose target can only be recovered from previous discourse remain outside the scope of the present sentence-level setup.

For the qualitative linguistic analysis, we inspected UD parses for a small set of representative

examples using UDPipe (Straka, 2018) as an annotation and visualization aid. For the quantitative experiments reported below, all dataset sentences were parsed with Stanza, and all structural features were extracted from the Stanza-generated UD annotations.

Attributive Insult (Qualified Subject). In this pattern, a negative adjective or insulting noun modifies a noun denoting a person or group, and the modified noun serves as the clausal subject.

Structure:

```
[ADJ_neg | NOUN_insult] → amod
→ [NOUN_person/group]
[NOUN_person/group] → nsubj →
[VERB_root]
```

Schema:

```
(amod* → nsubj) → root
```

Example:

„Днес обаче **престъпният режим на Радев** надмина своите мутренски предшественици.“ (bg)
*“Today, however, **Radev’s criminal regime** has surpassed its thug-like predecessors.”* (eng)

Predicative Insult. Here, the target person or group is realized as the subject, while the negative evaluation is expressed in the predicate via a copular construction.

Structure:

```
[NOUN_person/group] →
nsubj → [cop] → [ADJ_neg |
NOUN_insult]
```

Schema:

```
nsubj + cop + NEG_PRED
```

Example:

„500 делегати казаха, че не искат **Витанов** да е член на БСП, защото **е хулиган**. . . “ (bg)
*“500 delegates said that they do not want **Vitanov** to be a member of the BSP because **he is a hooligan**. . . ”* (eng)

Attack within Complement Clauses. In this configuration, the negative characterization is embedded in a complement clause rather than expressed directly in the matrix clause.

Structure:

```
[NOUN_person/group] → nsubj →
[VERB_root]
[VERB_root] → ccomp/xcomp →
[PRED]
```

Within `PRED`**:**

```
(amod → NOUN_person/group)
or
(nsubj → cop →
ADJ/NOUN_insult)
```

Schema:

```
nsubj → root → (ccomp |
xcomp)
```

In other words, the negative characterization is realized inside an embedded predicative structure.

Example:

„. . . твърди, че **този министър е некадърен**.“ (bg)
*“. . . claims that **this minister is incompetent**.”* (eng)

Coordinated Discrediting. This pattern uses coordination to stack multiple negative modifiers for the same target, creating an accumulative discrediting effect.

Structure:

```
(amod → NOUN_person)
cc/conj
(amod → NOUN_person)
```

Schema:

```
amod + conj(+)
```

Example:

„**Те са злобни, завистливи и дребни души**.“ (bg)
*“**They are spiteful, envious, and petty people**.”* (eng)

Passive Discrediting. In the passive pattern, the target appears as a passive subject, while the negative evaluation is contributed indirectly by a dependent phrase.

Structure:

```
[NOUN_person] → nsubj:pass →
[VERB_root]
(obl / nmod contains a negative
qualifier)
```

Schema:

```
nsubj:pass + NEG
(obl/nmod/xcomp)
```

Example:
„. . . партията заедно с коалиционните си партньори **да бъдат обявени за антидемократични**“. (bg)
“. . . that the party, together with its coalition partners, ***be declared anti-democratic****.”* (eng)

6 Methodology and Experiments

6.1 Feature Representation and Classification

The linguistic analysis identifies five recurring dependency configurations that realize ad hominem attacks. In the qualitative description, labels such as `ADJ_neg`, `NOUN_insult`, and `NEG_PRED` are used only as explanatory abstractions for evaluative content in representative examples. They are not implemented as manually curated lexical resources in the classification experiments. The structural models therefore do not rely on a hand-built negative lexicon or rule-based matching of insult words. Instead, the constructions are operationalized through dependency-derived features extracted from the Stanza-generated UD parses and represented with three feature groups:

- **UD relation features:** counts and length-normalized ratios of selected dependency relations (e.g., `amod`, `nsubj`, `cop`, `nsubj:pass`, `ccomp/xcomp`, `nmod`, `obl`, `case`, `conj/cc`).
- **UPOS distribution features:** counts and ratios of UPOS tags (e.g., `ADJ`, `NOUN`, `PROPN`, `VERB`, `AUX`, `ADV`, `PRON`, `PUNCT`), capturing coarse grammatical composition.
- **Template-shape indicators:** binary features approximating the structural frames of the five constructions (attributive, predicative, complement-clause embedding, coordination, passive) without lexical constraints, i.e., based on the presence and configuration of the relevant dependency relations rather than membership in a negativity lexicon. This reduces brittleness due to lexicon coverage, morphological variation, and lemma/wordform mismatch, while preserving the linguistic motivation of the templates.

For classification, we compare three feature settings: (i) structural features only, (ii) lexical features using TF–IDF over word unigrams and bigrams, and (iii) a combined representation concatenating TF–IDF and structural features.

We evaluate three model families—Logistic Regression (LR), Decision Trees (DT), and Support Vector Machines (SVM)—to cover both linear and non-linear decision boundaries while keeping the analysis interpretable.

6.2 Evaluation Protocols

Because the dataset is small and class-imbalanced, we report results using two evaluation protocols.

Protocol A: Preliminary single-split evaluation (illustrative). We use a stratified 80/20 train–test split (`stratify=y`) to preserve the original class distribution. Hyperparameters are selected via 5-fold stratified `GridSearchCV` conducted only on the training partition. After selecting the best configuration, the model is retrained on the full training set and evaluated once on the held-out test set. This provides an interpretable snapshot but is sensitive to the particular split. Results from the preliminary single-split protocol are reported in Table 1.

Protocol B: Main repeated nested cross-validation (primary). To reduce variance and avoid optimistic bias from hyperparameter tuning on small data, we use repeated stratified nested cross-validation with an outer evaluation loop (5 folds $\times$ 20 repeats) and an inner model-selection loop (5-fold stratified `GridSearchCV`) applied to each outer training fold, optimizing F1 for the positive class. We report mean $\pm$ standard deviation across outer folds for F1, precision, recall, and ROC–AUC. Our primary performance estimates are reported in Table 2 (mean $\pm$ std across outer folds).

7 Results

Table 1 and Table 2 report results for three feature settings—structural, TF–IDF, and combined—across Logistic Regression (LR), Decision Tree (DT), and Support Vector Machine (SVM). We report F1, precision, recall (positive class), and ROC–AUC.

7.1 Preliminary Single-Split Evaluation (Illustrative)

Under the single stratified 80/20 split (Table 1), structural features appear highly competitive and in several cases outperform TF–IDF in F1. For example, DT and SVM with structural features yield the highest F1 values in this protocol (DT:

Feature set	Model	F1	Prec.	Rec.	ROC-AUC
Structural only	Logistic Regression	0.4390	0.3000	0.8182	0.6400
	Decision Tree	0.5185	0.4375	0.6364	0.6455
	SVC	0.5385	0.4667	0.6364	0.6145
TF–IDF only	Logistic Regression	0.2222	0.2857	0.1818	0.6036
	Decision Tree	0.3158	0.3750	0.2727	0.5364
	SVC	0.3158	0.3750	0.2727	0.6255
Combined	Logistic Regression	0.3810	0.4000	0.3636	0.7200
	Decision Tree	0.3810	0.4000	0.3636	0.5764
	SVC	0.5385	0.4667	0.6364	0.6618

Table 1: Illustrative results from a single stratified 80/20 train–test split (`stratify=y`). Hyperparameters are selected via 5-fold stratified `GridSearchCV` on the training partition only; evaluation is performed once on the held-out test set. Due to small test size, results are split-sensitive; the primary evaluation is reported in Table 2.

F1 = 0.5185; SVM: F1 = 0.5385), while TF–IDF alone is lower (DT/SVM: F1 = 0.3158). For LR, structural features produce markedly higher recall than TF–IDF (0.8182 vs. 0.1818), though at the expense of specificity, consistent with a tendency to flag positives aggressively. In the same setting, the combined representation improves LR ROC–AUC (0.7200) relative to both structural (0.6400) and TF–IDF (0.6036), suggesting that lexical and structural signals can interact beneficially in at least some configurations.

These results support the plausibility that dependency-derived structure captures discriminative information; however, because the test set is small, this protocol is vulnerable to split-dependent variance, and the magnitude of apparent gains should not be treated as definitive.

7.2 Main Evaluation: Repeated Stratified Nested Cross-Validation (Primary)

The primary results are those from repeated stratified nested cross-validation in Table 2, which reduces variance and controls for optimistic bias during hyperparameter tuning. Under this protocol, TF–IDF provides the strongest overall baseline, particularly for Logistic Regression (F1 = 0.548±0.074; ROC–AUC = 0.698±0.093), exceeding the corresponding structural-only LR model (F1 = 0.489±0.099; ROC–AUC = 0.655±0.081). This indicates that, across many train/test partitions, lexical information provides a consistently strong signal in the present dataset.

Structural features remain competitive: across LR/DT/SVM, structural-only F1 lies in the 0.474–0.489 range, with ROC–AUC around 0.63–0.66. The combined representation does not yield a consistent F1 improvement over TF–IDF alone. For LR, combining TF–IDF with structural features produces similar F1 (0.491±0.105) and slightly lower ROC–AUC (0.688±0.088) compared to TF–IDF alone, but it changes the error profile: it increases precision relative to the TF–IDF LR baseline (0.470±0.121 vs. 0.414±0.060) while lowering recall (0.561±0.184 vs. 0.818±0.130). This pattern suggests that structural cues can contribute complementary information, but in this dataset their contribution is expressed more as a precision–recall rebalancing than a clear increase in average F1.

7.3 Interpreting the Discrepancy Between Protocols

A key observation is that the single-split evaluation suggests larger structural advantages than the repeated nested cross-validation. This discrepancy is expected in small datasets: a single held-out test set can, by chance, align well (or poorly) with particular feature types, inflating apparent differences. In contrast, repeated nested cross-validation averages performance over many distinct train/test partitions and thus provides a more reliable estimate of expected generalization. The attenuation of "structural > lexical" under the main protocol therefore indicates that variance from sampling is substantial in this regime, and conclusions should be drawn from the repeated nested results.

7.4 Summary of Findings

Across models, the main evaluation shows that lexical TF–IDF features are a strong and robust baseline for ad hominem detection in Bulgarian political social media sentences. Dependency-derived structural features are competitive and encode syntactic regularities motivated by the linguistic analysis, but they do not consistently outperform lexical features when evaluated under repeated stratified nested cross-validation. The

Feature set	Model	F1	Prec.	Rec.	ROC-AUC
Structural only	Logistic Regression	0.489 ± 0.099	0.435 ± 0.098	0.597 ± 0.182	0.655 ± 0.081
	Decision Tree	0.474 ± 0.140	0.468 ± 0.131	0.502 ± 0.181	0.632 ± 0.098
	SVC	0.478 ± 0.091	0.407 ± 0.085	0.615 ± 0.171	0.639 ± 0.082
TF–IDF only	Logistic Regression	0.548 ± 0.074	0.414 ± 0.060	0.818 ± 0.130	0.698 ± 0.093
	Decision Tree	0.418 ± 0.123	0.454 ± 0.124	0.410 ± 0.165	0.604 ± 0.083
	SVC	0.396 ± 0.157	0.350 ± 0.150	0.570 ± 0.340	0.509 ± 0.222
Combined	Logistic Regression	0.491 ± 0.105	0.470 ± 0.121	0.561 ± 0.184	0.688 ± 0.088
	Decision Tree	0.446 ± 0.124	0.455 ± 0.135	0.454 ± 0.154	0.612 ± 0.084
	SVC	0.477 ± 0.120	0.460 ± 0.132	0.538 ± 0.185	0.674 ± 0.097

Table 2: Primary evaluation: repeated stratified nested cross-validation (outer: 5 folds $\times$ 20 repeats; inner: 5-fold stratified `GridSearchCV` within each outer training fold; mean $\pm$ std over 100 outer folds). Metrics are reported for the positive class (ad hominem).

combined representation yields performance comparable to the best single-view setting and can shift the precision–recall balance in interpretable ways, supporting the view that syntactic structure contributes information that is not purely lexical, even if its benefit is modest in the current dataset.

8 Conclusion and Future Work

This paper investigated whether linguistically motivated dependency-based structure can support ad hominem detection in Bulgarian political social media beyond standard lexical baselines. Based on a linguistic analysis, we operationalised five recurring UD configurations via three structural feature groups (UD relation ratios, UPOS distributions, and template-shape indicators) and compared them to TF–IDF and a combined representation across LR, DT, and SVM classifiers. Under the primary repeated stratified nested cross-validation protocol, TF–IDF provides the strongest and most stable baseline, while dependency-derived structural features remain competitive and capture syntactic regularities consistent with the linguistic patterns. Although combining structural and lexical features does not yield a consistent average F1 improvement, it can shift the precision–recall profile in an interpretable way, indicating that syntactic cues contribute complementary information rather than merely duplicating lexical signal.

Future work will (i) extend the study to larger and more diverse Bulgarian political datasets and move from sentence-only classification to context-aware modelling, (ii) refine the structural representation from coarse relation counts to dependency paths and small subgraph motifs that more directly encode the five constructions, and (iii) incorporate large language models in complementary roles: as data-efficient baselines under the same strict evaluation, as annotation and validation assistants that propose candidate ad hominem spans or subtypes with uncertainty for human review, and as structured explanation models constrained to output target–predicate–frame summaries aligned with UD evidence, enabling more systematic error analysis and more reliable template refinement.

9 Limitations

The experiments operate in a small-data regime (177 sentences; 53 positives), which makes estimates sensitive to sampling; we mitigate this by reporting repeated nested cross-validation, but variance remains inherent in this setting. The task is formulated at sentence level and does not incorporate discourse context, including preceding turns, target coreference, speaker identity, or pragmatic intent, even though ad hominem is often context-dependent. No automatic coreference resolution is performed; therefore, examples involving pronouns are interpreted only within the local sentence-level analysis, and cases where the attacked target is recoverable only from prior discourse are likely to remain difficult for the present models. Some errors are therefore attributable to missing discourse context rather than purely lexical or syntactic ambiguity. Structural features depend on UD parses produced by general Bulgarian parsing models, and social-media style (elliptical syntax, unconventional punctuation, slang) can introduce parse noise that directly affects dependency-derived features. Finally, our template-shape indicators intentionally approximate the five linguistic configurations without enforcing explicit polarity or insult-lexicon constraints; this improves robustness to morphology and vocabulary variation, but it also allows structurally similar non-attacks, which can increase false positives and weaken the one-to-one correspondence between the linguistic patterns and their computational operationalisation.

At the same time, the strength of TF–IDF should not be interpreted as evidence that lexical features alone provide a robust general solution to ad hominem detection. In a small and imbalanced political dataset, TF–IDF may capture dataset-specific lexical regularities, topic markers, named entities, or highly frequent evaluative expressions associated with the positive class. Repeated stratified nested cross-validation reduces split-dependent variance and tuning bias, but it does not eliminate possible lexical or topical bias in the corpus itself. For this reason, the TF–IDF results are treated as a strong in-domain baseline rather than as evidence of reliable generalization to broader political discourse.

Acknowledgments

This research is part of the GATE project funded by the Horizon 2020 WIDESPREAD-2018-2020 TEAMING Phase 2 Programme under grant agreement no. 857155, the Programme "Research, Innovation and Digitalization for Smart Transformation"2021-2027 (PRIDST) under grant agreement no. BG16RFPR002-1.014-0010-C01 and, the project BROD (Bulgarian-Romanian Observatory of Digital Media) funded by the Digital Europe Programme of the European Union under contract number 101226153.

References

Fermín Cruz, José A. Troyano, Fernando Enríquez, and F. Javier Ortega. 2023. Automatic detection and generation of fallacies using large language models based on deep learning. In *Proceedings of the DEFALAC Workshop on Fallacy Detection*, pages 1–10.

Giovanni Da San Martino, Alberto Barrón-Cedeño, and Preslav Nakov. 2020. Semeval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation (SemEval-2020)*, pages 1377–1414.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, 51(4):85:1–85:30.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2018)*, pages 386–396.

Gabrijela Kišiček. 2013. The Political Discourse on Croatia's EU Accession: A Rhetorical Analysis of the Presentation of the European Union among Supporters and Opponents of the EU. In Gabrijela Kišiček and Igor Ž. Žagar, editors, *What Do We Know about the World? Rhetorical and Argumentative Perspectives*, pages 181–199. Educational Research Institute, Ljubljana. Accessed: 2026-04-15.

Gabrijela Kišiček. 2018. Ad hominem arguments in political discourse. In *Proceedings of the Conference on Argumentation and Language*, pages 55–70.

Sandra Kübler, Ryan McDonald, and Joakim Nivre. 2009. *Dependency Parsing*. Morgan & Claypool Publishers, San Rafael, CA.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 4034–4043, Marseille, France. European Language Resources Association.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. 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*.

Saumya Sahai, Oana Balalau, and Roxana Horincar. 2021. Breaking down the invisible wall of informal fallacies in online discussions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10.

Kiril Simov, Petya Osenova, Milena Slavcheva, Sia Kolkovska, Elisaveta Balabanova, Dimitar Doikoff, Krassimira Ivanova, Alexander Simov, and Milen Kouylekov. 2002. Building a linguistically interpreted corpus of Bulgarian: the BulTreeBank. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Milan Straka. 2018. 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*, pages 197–207.

Irina Temnikova, Ruslana Margova, Stefan Minkov, Tsvetina Stefanova, Nevena Grigorova, Silvia Gargova, and Venelin Kovatchev. 2025. Automatic Detection of the Bulgarian Evidential Renarrative. *Computational Linguistics in Bulgaria*.

Christopher W. Tindale. 2007. *Fallacies and Argument Appraisal*. Cambridge University Press.

Universal Dependencies. UD Bulgarian BTB. Treebank webpage. Accessed: 2026-04-15.

Universal Dependencies, GitHub. UD_Bulgarian-BTB. GitHub repository. Accessed: 2026-04-15.

Douglas Walton. 1998. *Ad Hominem Arguments*. University of Alabama Press.

Yin Wenjie and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: A review on abusive language detection. *PeerJ Computer Science*, 7:e598.

Thematic Dominants of German-Language Catholic Hymns: Automatic Analysis using Word Embeddings

Maria Khokhlova
St Petersburg State University
Russia
m.khokhlova@spbu.ru

Mikhail Koryshev
St Petersburg State University
Russia
m.koryshev@spbu.ru

Abstract

The study analyzes the thematic dominants of German Catholic hymns using methods of distributional semantics. The aim is to automatically identify the semantic structure of a corpus of hymns and to determine its key thematic fields. The material consists of an annotated corpus of German hymns compiled by the authors, including liturgical songs from the „Gotteslob“, from which a list of frequent nouns represented across different editions of the hymnal was extracted. To model semantic relationships and group lexical units, a series of procedures was carried out. First, word embeddings were generated using a multilingual transformer-based model. Second, nonlinear dimensionality reduction was performed using the UMAP algorithm with varying parameter settings (number of neighbors, dimensionality of the space, minimum distance between points, and distance metric). Third, clustering was conducted using the HDBSCAN algorithm, which made it possible to obtain sets of clusters of varying sizes for each model configuration. Model evaluation was carried out using three metrics: the Silhouette Score, the Davies–Bouldin Index, and the Calinski–Harabasz Index. The best-performing model configuration identified 17 thematic clusters of varying sizes, which were subsequently subjected to expert analysis. The findings demonstrate that the semantic space of Catholic hymns is well-structured as a combination of clearly defined thematic cores, which can be grouped into broader categories.

Keywords: word embeddings, German Catholic hymnography, distributional semantics, clustering.

1 Introduction

In recent years, neural network technologies have been increasingly employed in addressing a wide range of linguistic tasks, including the identification of semantically similar words, that is, the detection of groups into which lexical units may be clustered. The previously utilized classical distributional models are based on the frequency characteristics of word co-occurrence; but their principal limitation, however, lies in the fact that the vectors thus generated are, firstly, of high dimensionality and, secondly, characterized by sparsity (owing to zero co-occurrence frequencies) (Landauer and Dumais, 1997).

The subsequent generation of models comprises word vector representations, or embeddings, which likewise enable the capture of semantic relations between lexical units by representing the latter through their contexts, yet are characterized by considerably reduced dimensionality. The transformer architecture introduced in (Vaswani et al., 2017) constitutes one of the most prevalent frameworks in contemporary machine learning and is applied to a broad spectrum of tasks. Thus, the potential applications of the BERT model, which is based upon this architecture, are examined in numerous studies (see, for example, (Gardazi et al., 2025)); a number of investigations have focused specifically on the analysis of particular materials (for instance, the semantic analysis of reviews (Karabašević et al., 2026)). At the same time, deep learning algorithms are employed for the analysis and classification of texts of diverse genres.

A number of studies are devoted to the application of neural network models to the analysis of ecclesiastical chant, for instance, recordings of hymns from the Greek Orthodox Church (Iliadis et al., 2023), to the comparative examination of the Bible, the Qur'an, and the Bhagavad Gita utilizing approaches such as topic modeling and sentiment analysis (Mahit Nandan et al., 2025), as well as to the broader discussion

concerning the potential of processing religious texts by means of computational technologies (Hutchinson, 2024). General issues pertaining to the deployment of artificial intelligence systems within theological inquiry are addressed in (Graves, 2026), whereas (Lima et al., 2025) presents a comprehensive overview of the principal tasks in computational linguistics—including machine translation, authorship attribution, and automatic annotation—that have been undertaken using biblical texts as the primary material.

Although recent studies have demonstrated the usefulness of neural language models for analyzing religious texts, most existing research has focused on text classification, sentiment analysis, topic modeling, or the processing of biblical corpora. Comparatively little attention has been paid to the automatic exploration of semantic structures in hymns. In particular, the identification of thematic domains through distributional semantic representations has not yet been systematically investigated for German-language Catholic hymnography.

To address this gap, the present work studies whether transformer-based word embeddings combined with nonlinear dimensionality reduction and clustering can automatically identify semantically coherent thematic domains in a corpus of German Catholic hymns. The contributions of this paper are threefold. First, the study proposes a pipeline for identifying thematic clusters in a corpus of German Catholic hymns using multilingual contextual embeddings. Second, we evaluate 128 combinations of UMAP (McInnes et al., 2018) parameters using three clustering metrics. Third, we provide an expert interpretation of the resulting semantic clusters and discuss their correspondence to the theological structure of the corpus.

2 Materials and Methods

The corpus used for the study consists of the common (Stammteil) sections of the German Catholic hymnals *Gotteslob* (1975) and *Gotteslob* (2013); the process of its compilation was described in our study (Khokhlova and Koryshev 2024). These hymnals represent the principal official collections of German-language Catholic hymns used in practice. The corpus contains over 41,000 words and includes hymns originating from different historical periods, ranging from medieval texts to contemporary works. Although the corpus reflects the historical development of German Catholic hymnography, the texts were analyzed in their modern editorial form as presented in the two editions of *Gotteslob*. For the purposes of the present study, the corpus was divided into three subcorpora: (1) hymns common to both editions, (2) hymns included only in GL 1975, and (3) hymns introduced in GL 2013.

Since nouns represent the principal part of thematic information in hymnographic discourse, we focused on nouns for further analysis and built a list of lexemes occurring across the three aforementioned subcorpora. After removing duplicate lemmas and manually correcting annotation inconsistencies, the final list comprised 291 nouns (for example, *Gott*, *Leben*, *Jesus*, *Maria*, *Nacht*, *Liebe*, *Herz*, *Erde*, *Tag* and *Zeit*).

The identification of clusters united by a common semantic theme was carried out at the subsequent stage by means of several procedures. Firstly, word vector representations were obtained based on the contexts of word usage within the corpus, with L2 normalization applied so as to transform the vectors to unit length while preserving their direction. To accomplish this task, the pre-trained multilingual transformer-based model paraphrase-multilingual-mpnet-base-v2from the Sentence Transformers library for Python was employed (Reimers and Gurevych, 2019). The selected model provides multilingual sentence and word representations with strong semantic generalization across languages and has demonstrated competitive performance on semantic similarity tasks. Since the corpus contains historical and religious vocabulary, the use of multilingual transformer embeddings was expected to provide more robust semantic representations than static word embeddings.

Secondly, nonlinear dimensionality reduction was performed using the UMAP algorithm, taking into account a range of parameters and facilitated the conduction of a series of experiments. Thirdly, the HDBSCAN algorithm (Campello et al., 2013) was applied for the clustering of lexemes. The formal evaluation of the quality of the model outputs was carried out with reference to the silhouette score (Rousseeuw, 1987), the Davies–Bouldin index (Davies and Bouldin, 1979), and the Calinski–Harabasz index (Caliński and Harabasz, 1974). Since no single clustering metric fully captures cluster quality, three measures were

employed. The silhouette score estimates cluster separation, the Davies–Bouldin index measures cluster compactness and overlap, whereas the Calinski–Harabasz index evaluates the ratio between inter-cluster and intra-cluster variance.

In the course of the investigation, models were constructed with varying parameter settings for the algorithms employed. For the UMAP algorithm, which performs nonlinear dimensionality reduction, the following parameters were taken into consideration:

- Number of neighbors (n_neighbors): 5, 10, 15, 20.

Lower values facilitate the identification of more narrowly defined clusters, whereas higher values enable the description of broader thematic categories.

- Dimensionality of the embedding space, or number of components (n_components): 2, 3, 5, 10.
- Minimum distance between points (min_dist): 0, 0.05, 0.1, 0.15.

This parameter governs cluster density, that is, the extent to which points are positioned in close proximity to or at a distance from one another.

- Distance metric (metric): Euclidean, cosine.

To identify the optimal representation of the semantic space, we performed an exhaustive grid search over the UMAP parameter space. Four values of *n_neighbors*, four values of *min_dist*, four values of *n_components*, and two distance metrics were explored, resulting in 128 parameter configurations, each of which was subsequently evaluated according to the three aforementioned clustering measures.

3 Results

The purpose of the parameter search was not to maximize any single evaluation metric but rather to identify a configuration that provides the best balance between cluster compactness, separation, and interpretability. Consequently, the final model was selected based on the combined behaviour of the three clustering measures together with the semantic coherence of the resulting clusters. The models differ not only with respect to the manner in which lexemes are grouped into clusters and the number of clusters obtained, but also in terms of which lexical units, and in what proportion, were not assigned to any group (i.e., were allocated to the noise cluster, designated by the algorithm as -1). The substantial proportion of such lexemes (which, for individual models, ranges from 40 to 47 percent) may be attributed to the limited size of the corpus employed in the present study.

Table 1 presents the data for the top-10 models that achieved the highest scores according to the aforementioned indices and are ranked by the silhouette metric. The results of the clustering evaluation produced by the models may be interpreted as follows. The silhouette coefficient ranges from -1 to 1: negative values may indicate clustering errors, whereas values approaching 1 imply a clear separation of clusters by the model. Higher values of the Calinski–Harabasz index correspond to models that perform clustering more successfully. For the Davies–Bouldin metric, conversely, lower values are indicative of optimally selected algorithm parameters.

n_neighbours	min_dist	n_components	distance	number of clusters	noise	silhouette score	Davies–Bouldin index	Calinski–Harabasz index
5	0.05	2	cos	17	0.24	**0.60**	**0.50**	**527.57**
20	0.00	3	cos	14	0.41	0.55	0.61	311.60
10	0.10	2	cos	17	0.37	0.55	0.54	408.30
15	0.05	2	euc	17	0.38	0.54	0.53	455.46
5	0.00	5	cos	15	0.18	0.54	0.60	237.80
5	0.00	10	cos	18	0.18	0.54	0.58	216.11
5	0.00	2	euc	16	0.21	0.53	**0.50**	475.74
5	0.00	10	euc	18	0.17	0.53	0.62	218.64
10	0.05	2	cos	16	0.30	0.53	0.53	392.03
10	0.00	3	cos	16	0.28	0.52	0.60	265.72

Table 1: Results for top-10 models.

The highest metric scores were obtained in cases where dimensionality was reduced to two- and three-dimensional spaces (with two exceptions), as well as with the minimum distance between points set to lower values. The majority of the top-10 models employ the cosine measure as the distance metric and distinguish between 15 and 18 clusters. At the same time, the substantial proportion of lexemes designated as noise by Models 2, 3 and 4 merits particular attention.

Model 1 (n_neighbors=5, min_dist=0.05, n_components=2, metric=cosine) proved to be the most successful according to all three evaluation

criteria. It combined a comparatively high Silhouette score with one of the lowest Davies–Bouldin indices while maintaining a high Calinski–Harabasz value resulting in the best overall balance between cluster separation, compactness, and interpretability. More importantly, expert evaluation demonstrated that the resulting clusters were semantically coherent and could be consistently interpreted as thematic domains within the corpus. This model identified 17 clusters, of which the noise cluster accounted for approximately 24% of the lexical units. The latter include, for example, *Sünder, Land, Weg, Angesicht, Bund, Hut, Gericht, Hoffnung, Schmerz, Dorn, Speer* and *Stärke*. The remaining clusters are, on the whole, clearly separated from one another within the resulting embedding space, and semantic homogeneity is likewise observed within them. Although some alternative models achieved slightly better values for individual metrics, Model 1 consistently produced semantically coherent clusters that could be reliably interpreted by experts. It was therefore adopted as the primary model for the subsequent qualitative analysis.

The high performance exhibited by Model 1 may be attributed to the following factors. The selection of a low value for the n_neighbors parameter (equals to 5) enables the UMAP algorithm to preserve the fine-grained local structure of the embedding space, thereby facilitating the identification of compact, semantically homogeneous clusters while avoiding the excessive smoothing that would result in the conflation of distinct themes. The reduction of the embedding space to two dimensions (n_components=2), although entailing some loss of global variance, proves sufficient in this instance to capture the dominant axes of semantic variation within the relatively constrained lexical domain of paraliturgical hymnography. The choice of the cosine distance metric (metric=cosine) appears well-founded: in high-dimensional, sparse vector spaces characteristic of word embeddings, cosine similarity demonstrates greater robustness compared to Euclidean distance. The selected parameter configuration ensures that the resulting clusters reflect genuine contextual synonymy and thematic coherence, rather than constituting artifacts of superficial word co-occurrence.

The subsequent paragraphs examine the remaining clusters in greater detail; thematic labels corresponding to their semantic content have been proposed by the authors for each of these groupings.

1) Marian Emblems and Motherhood: *Mutter, Maria, Kind, Kindlein, Jungfrau, Frau*.

The lexemes pertaining to Mariological themes are, on the whole, successfully delineated, and the inclusion of the lexemes *Kind* and *Kindlein* within this cluster appears harmonious. Nevertheless, the cluster lacks lexemes describing the Nativity thematic complex, which is metonymically associated with the Mariological theme (e.g., *Krippe*, *Stall*).

2) Salvation History in its Temporal Dimension: *Ende, Fürst, Stunde, Anfang, End, Verheißung, Grund, Eva*.

The cluster successfully outlines the temporal perspective of the biblical history of salvation. At first glance, one might expect Adam to appear within this sequence (cf. the doctrine of Christ as the New Adam); however, within the corpus of texts, Adam is mentioned only once, in the hymn “O selger Urgrund allen Seins” (GL 2013, No. 359), which is borrowed from the Liturgy of the Hours.

3) Light and Day Metaphorics: *Stern, Schein, Tag, Licht, Nacht, Finsternis, Morgen, Kyrieleison, Gesicht, Morgenstern, Glanz, Tor, Sonne, Lichte, Aufgang, Morgenröte, Nebel, Mond, Dunkel, Schlaf, Pforte*.

The identification of lexemes associated with the theme of light and day, which subsequently function within metaphorical designations (cf., for example, the image of Mary as the Morning Star in the so-called Litany of Loreto), is complemented by the lexemes *Finsternis*, *Schlaf* and *Dunkel*, which are justifiably assigned to a contrasting cluster based on the principle of opposition. However, this same cluster also includes *Gesicht* ('face'), a lexeme whose connection to this thematic field is merely indirect, as well as *Pforte* ('gate'), which might more appropriately have been assigned to the Mariological cluster.

4) The Realm of Creation as Cosmos and Nature: *Kerub, Feld, Krippe, Pfad, Kreis, Abgrund, Meer, Lauf, Berg, Tal, Ferne, Schnee, Schatten, Paradies, Mantel, Strom, Flut, Wüste, Korn, Wind, Quell, Tier*.

The lexemes describing the created world are delineated with considerable success; even *Kreis* ('circle') appears appropriate, insofar as it expresses the notion of a boundary and the finite completeness of the realm of creation, which stands in opposition to the uncreated celestial world. It is precisely for this reason that *Paradies* ('paradise') ought to be excluded from this cluster, since Paradies in the texts under examination denotes not the created Garden of Eden but rather a locus of the contemplation of the glory of the Almighty.

5) Christ and the Church as the Body of Christ: *Jesus, Christus, Tod, Sterben, Christenheit, Kirche, Todesnot, Retter, Tote*.

This cluster raises no particular objections; the sole refinement that may be warranted concerns the lexeme *Retter* ('savior'), which would more appropriately be assigned to the cluster united by the notion of the messianic dignity of Christ.

6) Eucharistic Meal and Sacrifice Metaphorics: *Fleisch, Abendmahl, Blut, Gast, Mahl, Tisch, Speise, Hunger, Osterlamm, Herd, Wein, Brot, Opfer, Trank, Kelch, Bild, Traube*.

On the whole, this cluster is semantically coherent, uniting lexemes that express the notion of Eucharistic sacrifice and meal. The inclusion of the lexeme *Bild* ('picture, image') within this cluster is, however, infelicitous, given that the concept of the figure, or image, continually accompanies ecclesiastical teaching on the matter of the sacrament as the visible image of a supernatural, transcendent content.

7) Passion, Suffering, Distress, Penance and Sin: *Knie, Schoß, Kreuz, Spott, Stall, Last, Joch, Betlehem, Grab, Hölle*.

This cluster successfully unites lexemes that delineate the notion of Christ's death as a sacrifice for the sins of the world (cf. the versicle from the Good Friday liturgy: “Ecce lignum Crucis, in quo salus mundi pependit” — “Behold the wood of the Cross, on which hung the Savior of the world”) and the idea of penitential adoration as the path toward the reception of this salvation (cf. the responsory to the aforementioned versicle: “Venite, adoremus” — "Come, let us adore").

8) The Kingship of Christ and Messianic Titles: *Himmel, Sohn, Himmelreich, Thron, Volk, Sieg, Heiland, König, Hirt, Reich, Name, Lamm, Stamm, David, Held, Tempel, Himmelsbrot, Hirte, Herrscher, Friede, Josef, Ebenbild, Sieger, Jünger*.

The lexemes conveying the notion of Christ as the reigning Messiah, King of the Universe, are correctly identified. The sole observation that may be made concerns the proper name *Josef*, given that the hymns in question refer not to the Old Testament figure of Joseph the All-Comely—the son of Jacob and a prefiguration of Christ—but rather to Joseph, the Betrothed of the Virgin Mary.

9) Praise and Heavenly Joy: *Lob, Preis, Halleluja, Engel, Dank, Herrlichkeit, Wunder, Freude, Freund, Freud, Wonne, Schar, Gut, Seele, Ruhm, Lobgesang, Zebaot, Gesang, Glück, Loblied, Ehre, Lied, Schöpfung, Staunen, Jubel*.

The set of lexemes included within this cluster raises no objections, with the exception of the lexeme *Schöpfung* ('creation'), which ought to be reassigned to the fourth cluster, the latter comprising lexemes that thematize concepts pertaining to the created world. It should be noted, however, that the lexeme *Schöpfung* is actively employed in constituting contexts of doxological praise addressed to God (cf. Ps 150:6 (Einheitsübersetzung): “Alles, was atmet, lobe den Herrn” — "Let everything that breathes praise the Lord").

10) Faith, Truth, and Spiritual Knowledge: *Macht, Wahrheit, Sinn, Flügel, Weise, Schall, Verstand, Glaube, Geheimnis*.

From the standpoint of Christian theology, the cluster that unites faith, rational cognition, and sensory perception is delineated with considerable success (cf. the Augustinian notion, articulated in his Tractates on the Gospel of John, concerning the act of faith as a necessary step toward the act of genuine knowledge).

11) Prayer, Trust and Divine Help: *Güte, Huld, Heil, Zeichen, Hand, Erbarmen, Treue, Güt, Schutz, Geduld, Demut, Erlöser, Erlösung, Maß*.

This cluster may be merged with the preceding one, given that it merely elucidates particular aspects of an individual's “standing” before God while seeking His countenance (see above).

12) Grace, Consolation and Divine Care: *Trost, Gnade, Dienst, Segen*.

This is a small yet notably “pure” cluster, whose constituent lexemes may likewise be assigned to the thematic field uniting words that describe the relationship between God and humankind. This field encompasses both designations of human actions and deliberately cultivated virtues within this relationship (e.g., *Dienst* 'service, ministry') and the fruits thereof

(e.g., *Gnade* 'grace', *Segen* 'blessing', *Trost* 'consolation').

13) Divine Justice and Judgment: *Gerechtigkeit, Recht, Menschenkind, Urteil, Antwort*.

The clustering procedure successfully recovered the notion of divine justice—specifically, the inevitability of human accountability before God. The grouping is semantically coherent and reflects the theological opposition between justice and mercy frequently discussed in Christian doctrine.

14) Threat, Temptation, and Evil: *Feind, Gefahr, List, Gesetz, Streit, Böse, Plage*.

This cluster unites lexemes designating the forces of evil, their attributes, and the perils that confront the individual along the path of life. The noun *Gesetz* ('law') thematizes, within this cluster, its secondary status for the believer in the context of the Apostle Paul's polemic against proponents of a theology of the law, as opposed to the New Testament conception of salvation by grace.

15) Suffering, Distress and Existential Anxiety: *Pein, Not, Verlangen, Elend, Krankheit, Arm, Angst, Furcht*.

This cluster unites lexemes that delineate the notion of existential dread (*Angst*) experienced by the believer in the face of suffering—a believer who is filled with longing (*Verlangen*) in anticipation of the saving action of the Almighty's right hand (Arm*)* and with reverent, filial awe and veneration (*Furcht*) toward Him.

16) Action, Will and Moral Conduct: *Tun, Rat, Werk, Wille, Mut, Eigentum, Arbeit, Buße, Plan*.

The obtained semantic grouping comprises lexemes that designate being as a purposive (*Plan* 'plan, intention') activity (*Handeln* 'action, agency'), encompassing both volitional (*Wille* 'will', *Mut* 'courage, fortitude of spirit') and moral-ethical aspects (*Buße* 'repentance, penance'), the latter conceived as an attribute of the readiness to assume responsibility for errors committed, to rectify them, and to continue along the path of life in search of the paradise lost by our first parents.

17) Anthropology of Perception and Divine Communication: *Herz, Gnad, Wort, Stimme, Ohr, Mund, Ruf, Zeugnis, Seite, Auge, Hauch, Schweigen, Zunge, Frage*.

This field encompasses lexemes that designate the conceptual sphere of New Testament theological anthropology of perception in its communicative dimension. Within this framework, contextual synonyms include *Wort* ('word') and *Zeugnis* ('testimony'), as well as *Herz* ('heart') and *Gnade* ('grace'). The significance of listening is underscored in contradistinction to speaking, as evidenced by the inclusion of such lexemes as *Schweigen* ('silence'), *Ohr* ('ear'), *Frage* ('question') or *Mund* ('mouth').

The resulting clusters exhibit a high degree of correspondence with thematic domains traditionally distinguished by theological scholarship, suggesting that the proposed workflow successfully recovers meaningful semantic structures without predefined thematic labels. On the whole, a clearly defined cluster structure of the thematic fields within the hymns may be observed, a finding corroborated by the analysis of the results obtained from alternative models. The aggregation into clusters is not of a merely formal character; rather, semantic intersections are discernible in the outcomes.

To evaluate the robustness of the obtained thematic structure, we compared the selected model with two alternative high-ranking configurations (Models 7 and 9). Although these models recovered many of the same thematic domains, they produced either less coherent semantic groupings or a higher proportion of isolated lexical items. This observation suggests that the identified thematic structure is largely stable with respect to moderate changes in UMAP parameters. Models 7 and 9 nevertheless identified several meaningful thematic groups, including “Faith and Cognition”, “Christ the King”, and “Sin and Repentance”. However, the overall thematic organization produced by Model 1 proved to be the most consistent and interpretable. Some clusters correspond directly to established theological concepts (e.g., Mariology or Eucharistic theology), whereas others represent broader conceptual associations that emerge only through the interaction of several lexical items. It may be observed that the clusters are heterogeneous with respect to the number of constituent units they contain; nevertheless, they may, in turn, be aggregated into broader thematic fields.

Overall, the experiments demonstrate that transformer-based semantic embeddings combined with nonlinear dimensionality reduction and clustering constitute an effective approach to exploring the thematic organization of historical religious corpora. The automatically identified

clusters largely correspond to expert interpretations while also revealing previously unnoticed semantic associations, thereby illustrating the potential of distributional semantics for further research.

4 Conclusion

The findings obtained have confirmed the feasibility of automatic cluster identification based on vector representations for religious texts; these clusters, however, necessitate further analysis. The conducted study has demonstrated that the semantic space of German-language Catholic hymns is organized as a set of thematically dense clusters which are, on the one hand, surrounded by a substantial periphery of lexical units lacking robust connections, and, on the other, subsumed within larger groupings. At the same time, the difficulty encountered in assigning thematic labels to specific lexeme groups indicates that, firstly, a comparison with results obtainable through alternative models is required, and secondly, the incorporation of a larger corpus of a genre-similar nature is necessary to achieve greater precision. Thus, the applied approach made it possible to identify clusters and outline the range of themes addressed in the hymns.

References

Tadeusz Caliński and Joachim Harabasz. 1974. A dendrite method for cluster analysis. *Communications in Statistics*, 3(1): 1–27.

Ricardo J.G.B. Campello, Davoud Moulavi, and Joerg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining (PAKDD 2013)*, volume 7819 of *Lecture Notes in Computer Science*, pages 160–172. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-37456-2_14

D. L. Davies and D. W. Bouldin. 1979. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227.

Nadia Mushtaq Gardazi, Ali Daud, Muhammad Kamran Malik, Amal Bukhari, Tariq Alsahfi, and Bader Alshemaimri. 2025. BERT applications in natural language processing: a review. *Artificial Intelligence Review*, 58: 166. https://doi.org/10.1007/s10462-025-11162-5

Mark Graves. 2026. Generative AI and theology: a three-year retrospective. *Theology and Science*, 24(1): 1–7. https://doi.org/10.1080/14746700.2025.2592326

Ben Hutchinson. 2024. Modeling the sacred: considerations when using religious texts in natural language processing. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1029–1043. Mexico City.

Lazaros Alexios Iliadis, Sotirios P. Sotiroudis, Nikolaos Tsakatanis, Achilles D. Boursianis, Konstantinos-Iraklis D. Kokkinidis, George K. Karagiannidis, and Sotirios K. Goudos. 2023. Low complexity deep learning framework for Greek Orthodox Church hymns classification. *Applied Sciences*, 13(15): 8638. https://doi.org/10.3390/app13158638

Darjan Karabašević, Aleksandra Vujko, Vuk Mirčetić, Gabrijela Popović, and Dragiša Stanujkić. 2026. A transformer-based semantic encoding framework for quantitative analysis of large-scale textual reviews. *Axioms*, 15(3): 175. https://doi.org/10.3390/axioms15030175

T. K. Landauer and S. T. Dumais. 1997. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104: 211–240.

Bruno Cesar Lima, Nizam Omar, Israel Avansi and Leandro Nunes de Castro. 2025. Artificial intelligence applied to the analysis of biblical scriptures: a systematic review. *Analytics*, 4(2): 13. https://doi.org/10.3390/analytics4020013

Leland McInnes, John Healy, and James Melville. 2018. UMAP: uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*. https://arxiv.org/abs/1802.03426

Gotteslob. 1975. *Katholisches Gebet- und Gesangbuch*. Ausgabe für das Bistum Trier. Trier: Paulinus Verlag. 1054 S.

Gotteslob. 2013. *Katholisches Gebet- und Gesangbuch*. Ausgabe für das Bistum Trier. Trier. 1296 S.

Maria Khokhlova and Mikhail Koryshev. 2024. A corpus of liturgical texts in German: Towards multilevel text annotation. In *Proceedings of the Sixth International Conference on Computational Linguistics in Bulgaria (CLIB 2024)*, pages 201–205. Bulgarian Academy of Sciences.

Mahit Nandan A. D., Ishan Godbole, Pranav M Kapparad, and Shrutilipi Bhattacharjee. 2025. Comparative analysis of religious texts: NLP approaches to the Bible, Quran, and Bhagavad Gita. In *Proceedings of the New Horizons in Computational Linguistics for Religious Texts*, pages 1–10. Association for Computational Linguistics, Abu Dhabi.

Nils Reimers and Iryna Gurevych. 2019. 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)*, pages 3982–3992. Hong Kong, China.

Peter J. Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20: 53–65.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NeurIPS 2017)*, pages 5998–6008.

Corpus-based Study of Word Formation: Anti-agentive Nominals in English and Japanese

Junya Morita
Kinjo Gakuin University
Japan
morita@kinjo-u.ac.jp

Abstract

This study reveals the characteristics of anti-agentive nouns in English and Japanese (cf. *instructee*, *hi-shidoo-sha*) and then proposes the underlying word formation mechanism. Related derivatives extracted from two large-scale corpora are examined in detail from grammatical and functional perspectives. The result exposes four generalizations: (i) both languages have core devices for their formation that involve high productivity values (*P*), anaphoric conceptualization, and the compositional core meaning 'one who is V-ed,' (ii) there exist unproductive subsidiary types that deviate in part from the core meaning (see *attendee* without a passive reading and *enhancee* without a personal reading), (iii) Japanese has a one-to-one correspondence between semantic units and morphemes, but English does not (cf. ***hi***-*shidoo*-***sha*** [passive-instruct-person] and *instruct**ee*** [instruct-passive.person]), and (iv) Japanese has two types of anti-agentive nouns, *hi-shidoo-sha* and *taiho-sha* 'arrestee.' We then show that an "anti-lexical" theory reconciles the common core aspects with the morphological difference, each deviation in the subsidiary types is related to "grammaticalization," and pragmatic strategies are at work in the selection of competing anti-agentive forms.

Keywords: anti-agentive nouns, corpus research, generalizations, word-formation systems, English, Japanese

1 Introduction

This study investigates in detail the nominal expressions seen in (1).

(1) Unlike many, Martin was able to get through her prepared remarks, whereas more often than not, *acceptees'* thank-yous were interrupted mid-phrase by "the orchestra from Hell," which functioned as a sort of musical hook if *awardees* went one second past their allotted thirty seconds per speech. (*Plays and Players*, July 1993, p. 45)

The derived nouns *acceptee* and *awardee* share the meaning of 'one who receives an action' and involve a covert passive voice. They are therefore called "anti-agentive nouns." Lexically, they are very low frequency words and are hence involved in a creative word formation that generates words on the spot.

The aim of this study is to collect extensive data on anti-agentive nominals in English and Japanese using large-scale corpora, to clarify their similarities and differences based on this data, and then to theoretically explain their origins. Based on extensive data collected from large corpora, we disclose the morphosyntactic, semantic, and functional properties of *-ee* derivatives (§2), other type of anti-agentive nouns (§3), and their Japanese counterparts (§4). Section 5 compares anti-agentive nominals in English and Japanese, presents several empirical generalizations, and then show their theoretical implications.

2 The Properties of *-ee* Nominals

2.1 Lexical properties

This section deals with two lexical properties of *-ee* derivatives. First, they exhibit lexical attributes in that they conceptualize "person receiving an action." A good example of this is provided in (2).

(2) He's Paul Klemchalk, "*laid off* at 43," one of the thousands felled by AT&T's downsizing ax. … Klemchalk's timely resuscitation may be part of a big AT&T PR campaign to soften its image as a heartless job-slasher—an effort that included full-page newspaper ads last week touting its job bank for *lay-off-ees*. (*Newsweek*, March 25, 1996, p. 42)

The context of discourse (2) is that a person was laid off and then a certain measure was taken against those who were laid off. In response to the antecedent verb phrase, a situation arises in which "the people felled by AT&T" are conceptualized. A word, rather than a phrase, is then used to conceptualize these people. Consequently, the anti-agentive noun *lay-off-ee* is innovated on the spot. This word is a neologism because it is not found in the British National Corpus (BNC), a 100-million-word corpus, or large dictionaries like *The Oxford English Dictionary* (2nd ed.) (*OED*). Similarly, in example (1) observed above, the neologism *acceptee* is coined to label a contextually established concept.

The second lexical property concerns productivity. *Acceptees* and *awardees* in (1) are BNC hapaxes—token frequency 1[1]. A complex word of very low frequency is quite likely to be created on demand rather than stored in the lexicon (Hay, 2003). Accordingly, the degree to which words of this kind can be produced is an indicator of the productivity of a word formation device (Baayen and Renouf, 1996; Plag, 1999). A productivity measure has been proposed: *Productivity (P)*=n_1/V, where n_1 is the number of hapaxes and V is the total number of word types. The productivity value (P) can then indicate the probability that a word type in *-ee* is a neologism when it is formed. Our BNC survey has detected 22 hapaxes and 94 word types of *-ee* derivatives, giving its productivity value of 0.234. The productivity of *-ee* means that one out of four word types in *-ee* used is a new word, thus indicating the high potential of *-ee* suffixation to generate new words.

In summary, *-ee* derivation is highly creative: it can coin a relevant word whenever the recipient of a particular action needs to be conceptualized in an appropriate context.

2.2 Semantic properties

Previous studies on *-ee* nouns have tended to simply list their various meanings (Barker, 1998; Marchand, 1969; Huddleston and Pullum, 2002). Below, by scrutinizing related words collected from a large corpus, we will perform an "elegant" semantic classification of *-ee* derivatives and demonstrate that their meanings are divided into a single basic meaning (Patterns 1-3) and meanings that partially deviate from it (Pattern 4). The semantic features of each pattern are exemplified, while also mentioning the productivity of each pattern.

2.2.1 V_{tr} + *-ee* (Pattern 1)

Pattern 1 consists of a transitive verb and the suffix *-ee*, which can be divided into three groups: A, B, and C. Type A is one in which the base verb participates in the passive voice and its direct object represents the referent of *-ee* derived nouns, as in *interviewee* and *examinee*. This type of *-ee* derivatives can be paraphrased as 'one who is V-ed,' and Marchand (1969: 268) calls this type "direct passive nouns." Our BNC research has found 62 word types of this sort out of a total of 94 word types in *-ee*; Type A accounts for 66.0 % of the total. Moreover, 19 hapaxes have been detected among the 62 word types of Type A *-ee* nouns, giving its productivity value of 0.306. This type then is judged to be the prototype of *-ee* derivatives.

In Type B *-ee* derivatives, the underlying dative verb is passively involved and its indirect object is the referent of *-ee* nouns. This type then is interpreted as 'one who is V-ed something.' In our BNC survey, five word types of this sort have been discerned: *allotee*, *assignee*$_2$, *legatee*, *presentee*, *providee*[2].

In addition, Type C is related to "prepositional passive," in which the object of the preposition represents the referent of *-ee* nouns. Our BNC survey has detected only one word type of this kind, *conferee*, which is paraphrased by 'one who is conferred with.'

A word needs to be said here about the productivity of Type B and Type C. In addition to the productivity measure presented in the last section (P), we assume "the global productivity P*": P* of affixation is defined in terms of its coordinates in the P-V interaction region, with productivity (P) on the horizontal axis and word types (V) on the vertical axis; productive affixation occupies a central position in the region (Baayen and Lieber, 1991: 818-819). According to this definition, cases with a very small number of word

[1] I am indebted to the research engine of www.english-corpora.org/bnc/. By repeatedly using the "wild card" function of the research engine, words ending in *-ee* are extracted from BNC.

[2] If the underlying verb is interpreted in two ways, it is separately counted: *assignee* means 'one who is assigned' and 'one to whom something is assigned,' and hence it belongs to both Type A (*assignee*$_1$) and Type B (*assignee*$_2$).

types, such as Type B and Type C *-ee* derivations, fall outside the domain of productive processes.

2.2.2 N + *-ee* (Pattern 2)

It has been pointed out that a noun can be used as the base of an *-ee* derivative other than a verb (Barker, 1998: 705; Huddleston and Pullum, 2002: 1697). Contrary to the previous analyses, we show below that the base of this type of *-ee* derivative is not actually a noun. It is well known that one of the distinctive features of English word formation is that conversion from noun to verb produces an abundance of verbs. Quirk et al. (1985: 1561) classify the conversion from N to V into seven and point out that most of these verbs are transitive. One of the seven semantic classes is 'ornative': the verb *coat* is interpreted as 'to give a coat (of paint) to,' the base noun being converted referring to an entity given to something or somebody. This process is quite productive, giving rise to various verbs such as *poison*, *lipstick*, *bribe*, *lecture*, and *money*[3].

We claim that the *-ee* nouns at issue are derived via this type of conversion: the noun *patent* is converted into a verb ('to give a patent') and then the suffix *-ee* is added to it to derive the noun *patentee*. Thus, the suffix *-ee* that appears to be attached to a noun is actually connected to a verb. There are six related word types found in BNC, and according to "the global productivity *P**" (cf. §2.2.1), Pattern 2 is not productive. Interestingly, of these six word types, *off-licencee*, *therapee*, and *awardee* are hapaxes, and furthermore, *off-licence* and *therapy* are not listed as verbs in major dictionaries. This fact indicates that these *-ee* nouns are not created directly from verbs, but rather through the process described above[4].

2.2.3 Stem + *-ee* (Pattern 3)

Let us go on to a case in which *-ee* is superficially added to a verbal stem. Four relevant examples have been detected in BNC: *evacuee*, *nominee*, *amputee*$_{1}$, and *rehabilitee*. It can be argued that this type of *-ee* derivative is produced by dropping the internal suffix *-ate* when *-ee* is added to an *-ate* final verb. First, there is a semantic reason for this view. The readings of *evacuee* and *nominee* are 'one who is evacuated' and 'one who is nominated,' respectively, and if the suffix *-ate* is underlyingly involved, the reading of each derivative is easily and naturally defined. The truncation at issue is also justified by a context-dependent neologism, as exemplified in (3).

(3) In particular, the ERCs fail in some cases to *rehabilitate* the men and women who pass through them … Their routine is quite demanding and in some ways it expects a level of determination on the part of *rehabilitee* that would probably be high enough to get him back to work without a course of formal *rehabilitation*. (BNC FS6: 761)

What is of interest here is that the noun *rehabilitee* is a hapax, thus demonstrating that a new *-ee* derivative for 'one who is rehabilitated' is composed from the prior verb *rehabilitate*. It is shown that when an *-ee* noun is created, the internal suffix *-ate* is generally removed, thus lending support to *-ate* truncation. This process is unique to *-ee* suffixation, since it does not occur in the related "process nominalization," as evidenced in *rehabilitation*[5].

2.2.4 Others (Pattern 4)

There are a set of *-ee* nouns that lack one of the definitional features of *-ee* nouns. They are divided into four classes depending on what feature is missing. The first class is one in which the underlying verb has to do with active voice and its subject represents the referent of *-ee* noun. Our BNC research has detected four related *-ee* word types: *absentee*, *escapee*, *returnee*, and *standee*. Interestingly, Barker (1998: 707, 718) points out that the referent of this kind of *-ee* noun performs an active action and the action is forced or constrained by an external force. Accordingly, *standee* is better interpreted as 'one who is forced to stand' (cf. *OED*). Nouns such as *escapee* and *returnee* can be analyzed similarly; the senses of *escapee* and *returnee* are, respectively, something like 'one who is forced to escape (from an

[3] Semantic dependence can be pointed out as a criterion for the direction of conversion from N to V: if defining one (*to net* 'put into a net') is difficult without using the other (*net*), then the latter is the source of conversion (Quirk et al., 1985: 1558-1559).

[4] For more on how noun-to-verb conversion can generate a wealth of unlisted verbs, see Clark and Clark (1979).

[5] There is a case where no base verb can be found that combines with *-ee* (cf. *appellee*, *donee*, *lessee*). This type of *-ee* nouns were borrowed entirely from French and are not considered to be formed in English soil.

oppressive or irksome condition)' and 'one who is ordered to return (from abroad to one's native land).' Consequently, the anomaly of this class is that the base verb is not in the passive voice.

The second one is devoid of passive sense. Three related word types are taken from BNC: *attendee*, *cohabitee*, *signee*. According to *OED*, *signee* is defined as 'one who has signed a contract or register.' In contrast to the corresponding agentive noun *signer* 'one who signs,' this *-ee* derivative emphasizes that a person is in a resulting state of an action. In addition, while *attender* 'one who regularly goes to an event' implies an intentional act, *attendee* foregrounds one's state like 'one who is at an event.' *Cohabitee* can be treated in the same way as *attendee*.

The third class of atypical *-ee* noun is one whose referent is non-human. Five related word types are found in BNC: *adoptee*$_2$, *enhancee*, *investee*, *selectee*$_2$, *touchee*. *Adoptee* means not only 'person who is adopted' but also 'thing which is adopted.' The last class is one whose referent is a part of a body. The only relevant *-ee* noun, *amputee*$_2$, is discerned in BNC, which has the reading 'one who has had a limb amputated': what receives the action is a body part. The four classes of *-ee* derivations are unproductive because of the small number of related word types (cf. §2.2.1). It comes from the fact that each of them deviates from the prototypical *-ee* derivation and lacks one of its properties.

To summarize Section 2, based on objective data from a large corpus, we have presented evidence supporting conversion from noun to verb (§2.2.2) and the removal of *-ate* (§2.2.3), thus demonstrating that *-ee* is added only to verbs. As a result, we have confirmed that the meanings of *-ee* derivatives are divided into two categories: the meaning of 'one who is V-ed' and the meanings that partially deviate from that meaning. Furthermore, by quantifying the productivity of *-ee* derivatives (§2.1) and their subclasses (§2.2), we have confirmed that the former is the primary meaning, while the latter is secondary. It should be emphasized that this study is "corpus-based": rather than simply listing various meanings of *-ee* derivatives collected from a large corpus, it "elegantly" classifies them. In doing so, it focuses on "*-ee* hapaxes" and uses them importantly as a measure of productivity, for detecting context-dependent creative word formation, and for arguments in favor of conversion and the removal of an intraword suffix. Thus, this study exemplifies the significance of "corpus-based research" in morphological studies.

3 The Characteristics of Personal Substantivized Adjectives

The preceding section has been concerned with the characterization of the anti-agentive *-ee* nouns. There is another type of personal nominals involving a passive element, as shown in "cohabitees have more liberal views towards marriage and divorce than *marrieds*" (BNC CB8: 1868) and "the differences between the conquerors and *the conquered* began to disappear" (BNC B2V: 13). The italicized expressions are complex nominals signifying 'people who receive an action,' and neither of them actualizes a semantic unit representing people. This type of nominals always has a generic meaning and has traditionally been treated as a subtype of "conversion" from adjective to noun (Jespersen, 1949: 231-234). Poutsma (1914: 365-368) calls *the* + *Adj* formation "partial conversion," to distinguish it from "total conversion" of *V-eds* formation[6]. While nouns like *marrieds* can rarely be used in the singular, expressions like *the conquered* require a definite article, cannot take the plural -s, and always refer to a class of referents (Quirk et al., 1985: 42). The latter are considered a type of anti-agentive nominals because they exhibit lexical properties: (i) coordinating with nouns (cf. the interest of *women, minorities* and *the disabled*) and (ii) conceptualizing people's socioeconomic or physical conditions (cf. *the oppressed*/*the wounded*)[7].

[6] Regarding *the* + *Adj* nominals, there is also a syntactic analysis that assumes an empty noun after the participial adjective, rather than partial conversion (Kester, 1996).

[7] Since only six related word types are discerned in BNC, *V-eds* formation is not productive (§2.2.1). The reason is that functional categories such as inflections and prepositions generally do not appear inside words (cf. **[booksstore]*$_N$, **[**in**-city traffic]*$_N$), but *V-eds* nouns do not meet this restriction (cf. *[[[marri**ed**]*$_N$*] s]*$_N$). As for *the* + *Adj* formation, it is currently difficult to comprehensively extract relevant expressions from BNC, making it difficult to quantify its productivity.

4 Japanese Anti-agentive Nominals

4.1 *hi-VN-sha*

We turn now to the characterization of anti-agentive nouns in Japanese [8]. A detailed observations of Japanese anti-agentive nouns using the method presented in Section 2 has revealed that they can be divided into four types. The characteristics of each type are described below. The prototypical type is *hi-VN-sha* noun, which consists of the passive prefix *hi-*, verbal noun (VN), and the person-denoting suffix *-sha*[9]. Accordingly, the noun *hi-koyoo-sha* has the same meaning as the *-ee* noun *employee*: 'one who is employed.' The internal structure is $[[\text{hi-VN}]_A\ \text{-sha}]_N$, where the prefix *hi-* is added to a verbal noun (VN) to form an adjective and then the suffix *-sha* is added to it to make a noun. VN is defined as a stem that has a predicate function and combines with the light verb *suru* 'do' to yield a complex verb, although the light verb is omitted within a word. Evidence that the *hi-VN* part is an adjective is that it allows the addition of the suffix *-sei* '-ity,' which combines with adjectives.

Additionally, we can point out the constraints on the historical origin of the base. Japanese words are divided into three kinds: (i) native (Japanese) words, (ii) Sino-Japanese words, and (iii) words of foreign origin. The prefix *hi-* only attaches to Sino-Japanese VNs, and hence we can say *hi-koyoo-sha* 'employee,' but not **hi-yatoi-nin*, where the prefix is added to the native VN *yatoi*.

Turning to the communicative function of *hi-VN-sha* derivatives, they may serve as anaphoric substitutes, as is exemplified in (4).

(4) nyuushosha karano hiyoo-chooshuu-joogengaku-o hikiageru totomoni … resident from fee-collection-upper.limit-Acc raise and ... ***hi-sochi-sha***-chooshuu-gendo no … pas-treat-person-collection-limit of

'in addition to raising the upper limit of fees that are collected from residents … in the collection limit for **those who are so treated** …' (BCCWJ[10])

The content of the VN *sochi* 'treat' in the italicized derivative is given in the first sentence. This propositional content is then incorporated to create the personal passive noun *hi-sochi-sha*. Importantly, the internal element of the derivative (*sochi*) can only be properly interpreted in relation to the preceding propositional expression[11].

Let us move on to examining the creativity of *hi-VN-sha* derivation. The productivity value is calculated, following the measure presented in section 2.1. Our BCCWJ survey has detected 91 word types of the nouns including 35 hapaxes, and so its productivity value is 0.385. The productivity shows the potentiality of creating 35 kinds of new words when 91 kinds of *hi-VN-sha* derivatives are used, and hence *hi-VN-sha* formation is very productive. Like *-ee* derivation in English, *hi-VN-sha* derivation is central to the formation of personal passive nouns in Japanese.

4.2 *VN-sha*

The second type of personal passive nouns consists of Sino-Japanese VN and *-sha*. Personal nouns in *VN-sha* can be divided into agentive ($shiyoo\text{-}sha_1$ 'user') and passive ($taiho\text{-}sha_2$ 'arrestee') nouns. The reason for this is that a Sino-Japanese VN combines with *suru* 'do' to make a verb, and when it is followed by *-sha*, *suru* is omitted, and further *-sha* itself is a personal suffix rather than an agentive suffix. Consequently, the personal passive nouns concerned are formally indistinguishable from agentive nouns.

In what circumstances are nouns in *-sha* used as personal passive nouns? There are two pragmatic determinants. First, $VN\text{-}sha_2$ nouns are acceptable when the agents of the verbal nouns are self-evident. For example, in the case of $taiho\text{-}sha_2$ 'arrestee,' the party carrying out arrests (the

[8] While fragmentary observations have been made regarding anti-agentive nouns in Japanese (Kageyama, 2011; Morita et al., 1989), no systematic analysis has been conducted.

[9] There are two endings for personal passive nouns based on Sino-Japanese stems: *-sha* and *-nin*. The former is more widely used, so *-sha* is used as the representative.

[10] For collecting anti-agentive nominals in Japanese, I am indebted to the research engine of *Chunagon* (BCCWJ): https://chunagon.ninjal.ac.jp/auth/login?service=https%3A%2F%2Fchunagon.ninjal.ac.jp%2Fj_spring_cas_security_check.

[11] An anonymous reviewer asks if there is an English equivalent for *hi-sochi-sha* in (4). The English equivalent would be *treatee*, which is listed as "rare" in *OED* and not found in BNC. The verb *treat* is well known to require an adverbial (cf. *he treated John *(well/badly)*). Therefore, *treatee* 'one who is treated' is generally not acceptable, although it may be acceptable if some related adverbial is given or inferred from the context, as in (4)

police) is clear, so that the informational focus is on the party being detained. Under this circumstance, *taiho-sha*$_2$ is only given the anti-agentive sense. The same applies to, *kiso-sha*$_2$ 'one who is accused.' The second case is when common sense and world knowledge allow us to anticipate the "doing" side, while the informational focus is on the "done" side. In the case of *kyuujo-sha*$_2$ 'rescued person,' the rescuers are anticipated and emphasis in communication is placed on those who are rescued, as seen in news reports. Consequently, *kyuujo-sha*$_2$ is most likely to be interpreted as an anti-passive noun[12].

4.3 *yoo-VN-sha*

The third type of personal passive nouns is *yoo-VN-sha*, which is formally the second type with the prefix *yoo-* 'required' added. Thus, *yoo-kyuujo-sha* represents 'one who is required to be rescued.' The internal structure is *[[yoo-VN]*$_A$ *-sha]*$_N$, of which the *yoo-VN(-no)* part is an adjectival passive, similar to *hi-VN(-no)*. This is illustrated by the BCCWJ compound *yoo-shien-hi-hoken-sha* (required-support-pas-insure-person, 'those who are required to be supported and are insured'). In this compound, the adjectives *yoo-shien(-no)* and *hi-hoken(-no)* are coordinated. In principle, coordinates are syntactically of the same kind, which provides evidence that not only the adjective *hi-shien(-no)* but also the adjective *yoo-shien(-no)* implies the passive.

In terms of productivity, 14 word types of *yoo-VN-sha* nouns including 4 hapaxes are discerned in BCCWJ, giving its productivity value of 0.286. This indicates a considerable productivity of this derivation: one out of four word types in *yoo-VN-sha* used is a new word. The reason for the lower number of word types compared to *hi-VN-sha* derivation is likely due to the additional element *yoo-* 'required,' which limits their use.

4.4 *V-passive.suffix-nin*

Japanese also has personal passive nouns based on native verbs: *tora-wa-re-nin* (arrest-verb.suf-pas-person) and *kira-wa-re-mono* (hate-verb.suf-pas-person). The key point is that they contain inflectional suffixes. For instance, in *tora-wa-re-nin*, the native verb stem *tora-* is followed by the verb inflection *-wa* and the passive inflection *-re*, and then the person-denoting suffix *-nin*. (*-Mono* is another suffix denoting person.) This contrasts sharply with Sino-Japanese-based personal passive nouns, which have no internal inflectional suffixes (cf. *taiho-sha* 'arrestee'). *V-pas.suf-nin* nouns have connotations of scorn or mockery and therefore take the place of the compared complex nouns in such situations. Thus, *wara-wa-re-mono* (ridicule-verb.suf-pas-person 'ridiculed person') and *kira-wa-re-mono* (hate-verb.suf-pas-person 'one who is hated') are preferable to *??hi-chooshoo-sha* (pas-ridicule-person) and *??hi-keno-sha* (pas-hate-person)[13].

5 Comparison of English and Japanese Anti-agentive Nominals

5.1 Empirical generalizations

Based on the observations made so far, this section identifies two major similarities and two key differences between English and Japanese anti-agentive nominals (§5.1) and then indicates their theoretical implications (§5.2). The first commonality concerns semantics. As laid out in section 2, only *-ee* derivation is productive in the formation of English personal passive nouns. Furthermore, among the four types of *-ee* derivation, Pattern 1 (e.g. *employee*) is the most productive core device and therefore the meaning 'one who is V-ed' is the compositional core meaning of *-ee* derivation. Patterns 2 and 3 (*patentee*, *evacuee*) basically have the same structure as Pattern 1. In comparison, each of the four subtypes of Pattern 4 conveys meaning that lacks part of the primary meaning (*standee*, *signee*, *enhancee*, *amputee*$_2$) (cf. §2.2.4). Because of their low productivity, each meaning can be considered a secondary meaning of *-ee* derivation. It is important to clearly distinguish between these two kinds of meanings. As discussed in section 4.1, the Japanese counterpart of Pattern 1 is *hi-VN-sha* derivation: it is highly productive and has the prototypical meaning mentioned above. In

[12] Note that, given an appropriate context, the *VN-sh*$_2$ noun in the second case (*kyuujo-sha*$_2$) can be used as an agentive noun. As for the productivity of *VN-sha*$_2$ derivation, it is difficult to apply the above productivity measure, so it is unclear whether the derivation is productive or not. One reason for the difficulty is that it is not always clear from the context whether a given *VN-sha*$_2$ noun meets the above formation conditions.

[13] Six word types of *V-pas.suf-nin* derivatives are detected in BCCWJ, so this derivation is not productive due to the restriction mentioned in Note 7.

Japanese, the other three patterns also express the central meaning, although the *yoo-VN-sha* pattern has the added meaning of 'be required' (e.g. ***yoo-kyuujo-sha*** 'one who **is required** to be rescued'). Thus, in English and Japanese, the most productive anti-agentive nominalization provides a consistent core and compositional meaning.

Second, the main class of *-ee* derivation and its Japanese counterpart are highly creative (cf. §2.2.1, §4.1). The productivity values for each derivation are 0.306 and 0.385, respectively, implying that one out of three or four word types in *V-ee* or *hi-VN-sha* is a new word. The anaphoric and dynamic facets of the word composition are particularly noteworthy, as in examples (1) and (4), where an appropriate label can only be given to an entity by referring to a larger unit in discourse.

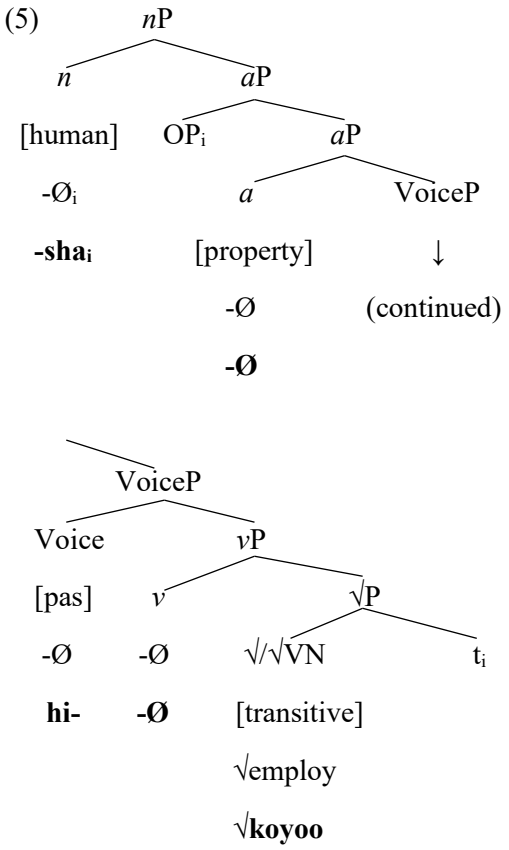
Let us move on the differences between English and Japanese personal passive derivatives. The observations in sections 2.2 and 4.1 identify a morphological difference between the main classes of personal passive nouns. Specifically, in the Japanese derivative ***hi-koyoo-sha*** 'employ**ee**,' the passive import is given to the prefix *hi-* and the person import is given to the suffix *-sha*. In its English equivalent, however, the suffix *-ee* represents 'one who is V-ed,' which involves the passive and personal elements. In addition, unlike English, Japanese has the isomorph *VN-sha* for both agentive (*tsuiseki-sha*$_1$ 'pursuer') and anti-agentive nouns (*taiho-sha*$_2$ 'arrestee'), so that it has two kinds of anti-agentive nouns, *hi-koyoo-sha* and *taiho-sha*$_2$ (cf. §4.2).

5.2 Theoretical implications

The above generalizations have three theoretical implications. The first is what kind of theory is needed to grasp the core part of personal passive nominalization. *-Ee* and *hi-VN-sha* derivations can generate numerous nouns expressing a compositional meaning. Their generative and compositional nature is characteristic of sentence construction. It is therefore quite natural to capture this aspect in syntax. As for the morphological difference—unlike English, Japanese has a one-to-one correspondence between semantic units and morphemes—it is precisely what morphology should address: "fusion" is applied in English but not in Japanese, realizing the two semantic units of passive and person with a single morpheme, *-ee*. The dual nature of the derivatives supports an "anti-lexical" model: the derivative's core structure is constructed in syntax, which is then materialized in the post-syntactic morphology module (cf. Matushansky and Marantz, 2013).

The specific derivation process is shown below. Given that affixes are the heads of phrases, the underlying structure of the noun *employee* and the Japanese equivalent *hi-koyoo-sha* is constructed by the merging of a root (√*employ*/√*koyoo*) and category-defining heads like *v*, *a*, and *n*, as depicted in (5). The lexical items in bold are for the *hi-VN-sha* nominal and the non-bold lexical items are for the *-ee* nominal[14].

(5)



Structure (5) consists of the noun $n_{[human]}$ and the adjectival passive phrase *a*P. The embedded adjectival passive construction is based on Bruening (2014). In this structure, the null operator (OP_i) occupies the internal argument and is therefore given the Theme role by the predicative head √*employ*/√*koyoo*. It is then moved to the specifier of *a*P in the passive environment and is linked to the noun $n_{[human]}$, which is external to the

[14] Root (√) is defined as bound morpheme that becomes the core of a word and -Ø signifies a bound zero-morpheme.

adjectival passive. Structure (5) signifies that the base verb is involved in the passive voice and its object expresses the referent of the entire derivative. Thus, the core meaning of both nominals, 'one who is V-ed,' is derived from a common underlying structure depicted in (5).

The syntactic output is then sent to the morphology module, where it is required to be constructed into a word form: the "boundness" nature of the inserted lexical items triggers morphological operations that combine them to form a word. A particularly important operation here is "fusion," which is defined as the process of coalescing two adjacent terminal nodes dominated by a given higher node into a single terminal node (Calabrese 1996: 5). In English, Voice$_{[pas]}$ and $a_{[prop]}$ in (5), which are adjacent terminal nodes under the higher *a*P node, are coalesced into a single terminal node ($a_{[prop, pas]}$), and then the resulting *a* node and the *n* head ($n_{[human]}$) are fused into a single terminal node, $n_{[human, prop, pas]}$. Since the fused node is assigned a new set of features, a new lexical item is required. Vocabulary—a set of lexical items—contains the item *-ee* with the features "[N][human][prop][pas]," which signify '*-ee* is a nominalizer representing a person involved in the adjectival passive.' Accordingly, the morpheme *-ee*, retrieved from Vocabulary, is correctly inserted into the fused node. In Japanese, on the other hand, fusion is not applied to the nodes Voice$_{[pas]}$, $a_{[prop]}$, and $n_{[human]}$, so each is given an independent morpheme[15].

Second, the peripheral aspect of *-ee* nominalization may be associated with a kind of "grammaticalization," which is generally defined as "the increase of the range of a morpheme advancing from a lexical to a grammatical or from a less grammatical to a more grammatical status" (Heine et al., 1991a: 3). For example, the Ewe noun *megbé* expands from a noun referring to person's back to a noun referring to the back of an object, and further to a spatial adverb ('behind') (Heine et al., 1991b: 163). Here we define it simply as extending the grammatical functions of a morpheme. Its prime motivation is to extend the use of an existing form in order to label a new concept, rather than creating a new label (Heine et al., 1991a: 27). The formation of the expressions at issue is motivated by this. In the first peripheral *-ee* noun, *standee*, the passive element of *-ee* is not linked to the base verb but to a hidden causative verb. In order to conceptualize "one who is forced to stand" in contrast to "one who stands intentionally," *standee* is formed by *-ee* affixation, where the morpheme *-ee* functions as a 'forced' marker. In the second noncentral *signee*, the 'state' import is extracted from the "stative passive," while the passive element itself completely disappears. Consequently, the suffix *-ee* comes to signify 'one who is in a resulting state of an action.' Thus, it functions as a 'state' marker in order to conceptualize "one who has signed" in contrast to "one who signs" (*signer*). In the third and fourth peripheral types, *enhancee* and *amputee*$_2$, the 'person' marker *-ee* becomes grammaticalized to acquire the additional functions of an 'object' marker and a 'body part" marker. We can see that the extension of *-ee* exhibits the nature of grammaticalization: *-ee* extends the function by changing one of its properties while adhering to or retaining the remaining properties ("persistence," Hopper and Traugott, 1993: 90)[16].

And finally, the fact that Japanese has two kinds of anti-agentive nouns provides interesting data on an important morphological concept, "competition" in lexical insertion (cf. Rainer et al., 2019): why can only one of the synonyms, *hi-VN-sha* or *VN-sha*$_2$, be selected? Two pragmatic strategies are at work here, leading to the choice between them. The first one is the strategy of "Avoid ambiguity," which prioritizes *hi-VN-sha* nouns over *VN-sha*$_2$ nouns, since the former explicitly indicate a passive sense, whereas the latter can show both passive and active senses. Contrastively, in the two situations where *VN-sha*$_2$ nouns are disambiguated (cf. §4.2), they win out over their more complex rivals, due to the "Avoid synonymy" strategy where a well-established simpler form preempts a complex semantic equivalent (Kiparsky, 1982: 13). Thus, two pragmatic strategies play a central role in the selection between two competing noun forms.

[15] Marantz (1997) and Harley and Noyer (2000) are studies that support anti-lexicalism using examples of process event nominals.

[16] This kind of grammaticalization does not occur in Japanese, and the reason for this seems closely related to the English-Japanese morphological difference mentioned above: in Japanese, unlike English, the passive and personal elements basically are manifested separately, so their respective meanings are preserved.

6 Conclusion

In forming anti-agentive nominals in English and Japanese, it is significant to distinguish between core and peripheral parts. While the core patterns, Pattern 1 *-ee* derivation and *hi-VN-sha* derivation, exhibit the phrasal attributes of "free combination" and compositional meaning, there is a morphological difference between them in terms of the combination of semantic elements and morphemes. To capture this dual nature of the primary nominalization, an "anti-lexical" model would be required. Meanwhile, in English, four minor types of *-ee* derivation are recognized, each of which deviates in part from the prototypical pattern. It would be appropriate to address these deviations within a grammaticalization framework. Furthermore, in Japanese, there are two types of anti-agentive nouns, *hi-VN-sha* and *VN-sha*$_2$, and competition for selection occurs. Two pragmatic strategies are strongly involved in this process. We have revealed the characteristics of anti-agentive nominals in English and Japanese using large-scale corpora, and have proposed the word formation systems behind them. Further corpus-based exploration will be needed to elaborate on the proposed systems.

Acknowledgments

I would like to express my gratitude to three anonymous reviewers for their valuable comments and suggestions on an earlier draft of this paper.

References

Harald R. Baayen and Rochell Lieber. 1991. Productivity and English derivation: a corpus-based study. *Linguistics*, 29:801-843.

Harald R. Baayen and Antoinette Renouf. 1996. Chronicling *the Times*: productive lexical innovations in an English newspaper. *Language*, 72:69-96.

Chris Barker. 1998. Episodic *-ee* in English: a thematic role constraint on new word formation. *Language*, 74:695-727.

Benjamin Bruening. 2014. Word formation is syntactic: adjectival passives in English. *Natural Language & Linguistic Theory*, 32:363-422.

Andrea Calabrese. 1996. Some remarks on the Latin case system and its development in Romance. ms., Harvard University.

Eve V. Clark and Herbert H. Clark. 1979. When nouns surface as verbs. *Language*, 55:767-811.

Heidi Harley and Ralf Noyer. 2000. Formal versus encyclopedic properties of vocabulary: evidence from nominalisations. In B. Peeters, editor. *The Lexicon-Encyclopedia Interface*, pages 349-374. Elsevier, Amsterdam.

Bernd Heine, Ulrike Claudi, and Friederike Hünnemeyer. 1991a. *Grammaticalization: A Conceptual Framework*. University of Chicago Press, Chicago.

Bernd Heine, Ulrike Claudi, and Friederike Hünnemeyer. 1991b. From cognition to grammar: evidence from African languages. In E. C. Traugott and B. Heine, editors. *Approach to Gammaticalization*, volume 1, pages 149-187. John Benjamins, Amsterdam.

Paul J. Hopper and Elizabeth Closs Traugott. 1993. *Grammaticalization*. Cambridge University Press, Cambridge.

Rodney Huddleston and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge University Press, Cambridge

Otto Jespersen. 1949. *A Modern English Grammar on Historical Principles*, volume 2. George Allen and Unwin, London.

Taro Kageyama, editor. 2011. *Nichiei Taisho Meishi-no Imi-to Kobun* (Meanings and Constructions of Nouns: A Contrastive Study of Japanese and English). Taishukan, Tokyo.

Ellen-Petra Kester. 1996. Adjectival inflection and the licensing of empty categories in DP. *Journal of Linguistics*, 32:57-78.

Paul Kiparsky. 1982. Word-formation and the lexicon. *Proceedings of the 1982 Mid-America Linguistic Conference*, 1:26.

Alec Marantz. 1997. No escape from syntax: don't try morphological analysis in the privacy of your own lexicon. *University of Pennsylvania Working Papers in Linguistics*, 4:201-225.

Hans Marchand. 1969. *The Categories and Types of Present-Day English Word-Formation: A Synchronic-Diachronic Approach*, 2nd ed. C. H. Beck, München.

Ora Matushansky and Alec Marantz, editors. 2013. *Distributed Morphology Today: Morphemes for Morris Halle*. MIT Press, Cambridge, MA.

Yoshiyuki Morita, Shinjiro Muraki, and Masao Aizawa. 1989. *Kesu Sutadi Nihongo-no Goi* (Case Study: Japanese Vocabulary). Ofusha, Tokyo.

Ingo Plag. 1999. *Morphological Productivity: Structural Constraints in English Derivation*. Mouton de Gruyter, Berlin.

H. Poutsma. 1914. *A Grammar of Late Modern English*, part II, IA. P. Noordhoff, Groningen.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman, London.

Franz Rainer, Francesco Gardani, Wolfgang U. Dressler, and Hans Christian Luschützky, editors. 2019. *Competition in Inflection and Word-Formation*. Springer, Cham, Switzerland.

LIST OF AUTHORS

Author	Affiliation	Country
Mullosharaf Arabov	Kazan Federal University	Russia
Atanas Atanasov	Sofia University “St. Kliment Ohridski”, Faculty of Slavic Studies	Bulgaria
Öner Aytaş	Işık University	Turkey
Karina Azarevich	HSE University	Russia
Abdelouahab Baflah	CRSTDLA	Algeria
Marina Bagi	Institute for the Serbian Language, SASA	Serbia
Mehmet Ali Bayram	Yıldız Technical University	Turkey
Göksel Biricik	Yıldız Technical University	Turkey
Nikola Blajev	Identrics; Experian	Bulgaria
Pavel Boytchev	Sofia University “St. Kliment Ohridski”, Faculty of Mathematics and Informatics	Bulgaria
Svetla Boytcheva	Sofia University “St. Kliment Ohridski”, Faculty of Mathematics and Informatics	Bulgaria
Ioana Bucsa	Gheorghe Sincai National College	Romania
Radek Čech	Masaryk University, Faculty of Arts	Czech Republic
Jan Černý	Charles University, Faculty of Mathematics and Physics	Czech Republic
Sofiia Chepovetskaia	HSE University – Saint Petersburg	Russia
Madalina Chitez	West University of Timişoara	Romania
Silvie Cinková	Charles University, Faculty of Mathematics and Physics	Czech Republic
Mihaela Cristescu	University of Bucharest, Faculty of Letters	Romania
Karla Csuros	West University of Timişoara	Romania
Natalia Dankova	Université du Québec en Outaouais	Canada
Viktoria Deborina	HSE University – Saint Petersburg	Russia
Ivan Derzhanski	Institute of Mathematics and Informatics, Bulgarian Academy of Sciences	Bulgaria
Stanimir Dimitrov	Big Data for Smart Society Institute (GATE), Sofia University	Bulgaria
Tsvetana Dimitrova	Institute for Bulgarian Language, Bulgarian Academy of Sciences	Bulgaria
Banu Diri	Yıldız Technical University	Turkey
Marcello Ferro	Institute for Computational Linguistics “A. Zampolli”, National Research Council (CNR)	Italy
Silvia Gaftandzhieva	University of Plovdiv “Paisii Hilendarski”	Bulgaria
Dimitar Georgiev	Institute for Bulgarian Language, Bulgarian Academy of Sciences	Bulgaria
Georgi Goranov	Big Data for Smart Society Institute (GATE), Sofia University	Bulgaria
Mikus Grasmanis	Institute of Mathematics and Computer Science, University of Latvia	Latvia
Nevena Grigorova	Big Data for Smart Society Institute (GATE), Sofia University	Bulgaria
Karomatullo Habibullozoda	Bokhtar State University “Nosiri Khusrav”	Tajikistan
Jakob Horsch	Slovak National Corpus, Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences	Slovakia
Radu Ion	Research Institute for AI “Mihai Drăgănescu”, Romanian Academy	Romania

Author	Affiliation	Country
Elena Irimia	Research Institute for AI "Mihai Drăgănescu", Romanian Academy	Romania
Iglika Ivanova	Big Data for Smart Society Institute (GATE), Sofia University	Bulgaria
Iryna Karamysheva	Lviv Polytechnic National University	Ukraine
Ouassim Karrakchou	Université Internationale de Rabat	Morocco
Bogomil Katanov	Identrics	Bulgaria
Maria Khokhlova	Saint Petersburg State University	Russia
Keith Kiely	Big Data for Smart Society Institute (GATE), Sofia University	Bulgaria
Todor Kiriakov	Identrics	Bulgaria
Svetla Koeva	Institute for Bulgarian Language, Bulgarian Academy of Sciences	Bulgaria
Alexander Komarov	Big Data for Smart Society Institute (GATE), Sofia University	Bulgaria
Mikhail Koryshev	Saint Petersburg State University	Russia
Michaela Košcová	Slovak Academy of Sciences	Slovakia
Boryana Kostadinova	Identrics	Bulgaria
Devora Kotseva	Identrics	Bulgaria
Jordan Kralev	Technical University of Sofia; Institute for Bulgarian Language, Bulgarian Academy of Sciences	Bulgaria
Cvetana Krstev	Language Resources and Technologies Society (JeRTeh)	Serbia
Hristina Kukova	Institute for Bulgarian Language, Bulgarian Academy of Sciences	Bulgaria
Ivana Kvapilíková	Charles University, Faculty of Mathematics and Physics	Czech Republic
Amina Laggoun	Université de Lorraine, LORIA	France
Gaël Lejeune	Sorbonne Université	France
Svetlozara Leseva	Institute for Bulgarian Language, Bulgarian Academy of Sciences	Bulgaria
Ilze Lokmane	University of Latvia	Latvia
Ján Mačutek	Mathematical Institute, Slovak Academy of Sciences	Slovakia
Mila Marcheva-Nash	University of Cambridge	United Kingdom
Ruslana Margova	GATE Institute, Sofia University	Bulgaria
Aleksandra Marković	Institute for the Serbian Language, SASA	Serbia
Claudia Marzi	Institute for Computational Linguistics "A. Zampolli", National Research Council (CNR)	Italy
Maria Sampedro Mella	UCLouvain	Belgium
Aleksei Melnik	HSE University-Saint Petersburg	Russia
Brandon Caillahua Mendoza	Universidad Nacional de Educación a Distancia	Spain
Catalin Mihaila	Research Institute for AI "Mihai Drăgănescu", Romanian Academy	Romania
Stefan Minkov	Big Data for Smart Society Institute (GATE), Sofia University	Bulgaria
Verginica Barbu Mititelu	Research Institute for AI "Mihai Drăgănescu", Romanian Academy	Romania
Maria Mitrofan	Research Institute for AI "Mihai Drăgănescu", Romanian Academy	Romania
Junya Morita	Kinjo Gakuin University	Japan
Mihaela Moskova	Institute for Bulgarian Language, Bulgarian Academy of Sciences	Bulgaria
Youness Moukafih	Université Internationale de Rabat	Morocco

Author	Affiliation	Country
Andrea Nadalini	Institute for Computational Linguistics "A. Zampolli", National Research Council (CNR)	Italy
Milica Ikonić Nešić	University of Belgrade, Faculty of Philology	Serbia
Gunta Nešpore-Bērzkalne	Institute of Mathematics and Computer Science, University of Latvia	Latvia
Iglika Nikolova-Stoupak	Sorbonne Université	France
Michaela Nogolová	University of Ostrava	Czech Republic
Petya Osenova	Sofia University "St. Kliment Ohridski"; Institute of Information and Communication Technologies, Bulgarian Academy of Sciences	Bulgaria
Vasile Păiş	Research Institute for AI "Mihai Drăgănescu", Romanian Academy	Romania
George Pashev	University of Plovdiv "Paisii Hilendarski"	Bulgaria
Stanislav Penkov	Big Data for Smart Society Institute (GATE), Sofia University	Bulgaria
Irina Petrova	HSE University – Saint Petersburg	Russia
Yolina Petrova	New Bulgarian University; Identrics	Bulgaria
Vito Pirrelli	Institute for Computational Linguistics "A. Zampolli", National Research Council (CNR)	Italy
Simina Popa	University of Bucharest	Romania
Lea Radu	The Saint Sava National College	Romania
Roxana Rogobete	West University of Timişoara	Romania
Eva Schaeffer-Lacroix	Sorbonne Université	France
Tuğçe Şen	Işık University	Turkey
Tatiana Sherstinova	HSE University – St Petersburg	Russia
Nurali Shirinov	Bokhtar State University "Nosiri Khusrav"	Tajikistan
Olena Siruk	Institute of Mathematics and Informatics, Bulgarian Academy of Sciences	Bulgaria
Kamel Smaili	Université de Lorraine, LORIA	France
Ranka Stanković	University of Belgrade, Faculty of Mining and Geology	Serbia
Tsvetelina Stefanova	Angel Kanchev University of Ruse	Bulgaria
Valentina Stefanova	Institute for Bulgarian Language, Bulgarian Academy of Sciences	Bulgaria
Ivelina Stoyanova	Institute for Bulgarian Language, Bulgarian Academy of Sciences	Bulgaria
Weiwei Sun	University of Cambridge	United Kingdom
Ekaterina Tarpomanova	Sofia University "St. Kliment Ohridski", Faculty of Slavic Studies	Bulgaria
Irina Temnikova	Big Data for Smart Society Institute (GATE), Sofia University	Bulgaria
Maria Todorova	Institute for Bulgarian Language, Bulgarian Academy of Sciences	Bulgaria
Dobromir Tsolyov	Big Data for Smart Society Institute (GATE), Sofia University	Bulgaria
Miloš Utvić	University of Belgrade, Faculty of Philology	Serbia
Sylvia Vassileva	Sofia University "St. Kliment Ohridski", Faculty of Mathematics and Informatics	Bulgaria

Organised by:



Department of Computational Linguistics
Institute for Bulgarian Language
Bulgarian Academy of Sciences

With the support of:

















ISSN: 2367-5675