

SIXTH  
INTERNATIONAL  
CONFERENCE



**COMPUTATIONAL  
LINGUISTICS  
IN BULGARIA  
CLIB 2024**

**9 – 10 September 2024**  
**Sofia, Bulgaria**

**PROCEEDINGS**

Department of Computational Linguistics  
Institute for Bulgarian Language  
BULGARIAN ACADEMY OF SCIENCES

The Sixth International Conference *Computational Linguistics in Bulgaria* (CLIB 2024) is organised with the support of the National Science Fund of the Republic of Bulgaria under Grant Agreement No. KP-06-MNF/5 from 23.05.2024.



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 2024 is organised by:



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

## PUBLICATION AND CATALOGUING INFORMATION

|                            |   |
|----------------------------|---|
| Title:                     | Proceedings of the Sixth International Conference<br><i>Computational Linguistics in Bulgaria</i> (CLIB 2024)   |
| ISSN:                      | 2367 5675 (online)  |
| Published and distributed: | Institute for Bulgarian Language<br>Bulgarian Academy of Sciences   |
| Editorial address:         | Institute for Bulgarian Language<br>Bulgarian Academy of Sciences<br>52 Shipchenski Prohod Blvd., Bldg. 17<br>Sofia 1113, Bulgaria<br>+3592/ 872 23 02  |
| Copyright:                 | Copyright of each paper stays with the respective authors.<br><br>The works in the Proceedings are licensed under a Creative Commons Attribution 4.0 International Licence (CC BY 4.0).<br><br>License details:<br><a href="http://creativecommons.org/licenses/by/4.0">http://creativecommons.org/licenses/by/4.0</a><br>Copyright © 2024 |

Proceedings of the  
Sixth International Conference  
COMPUTATIONAL LINGUISTICS IN BULGARIA



9 – 10 September 2024  
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

**Iana Atanassova** – University of Burgundy, Centre for Interdisciplinary and Transcultural Research, France

**Verginica Barbu Mititelu** – Research Institute for Artificial Intelligence, Romanian Academy, Romania

**Khalid Choukri** – Evaluations and Language Resources Distribution Agency, France

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

**Tsvetana Dimitrova** – Institute for Bulgarian Language, Department of Computational Linguistics, Bulgarian Academy of Sciences, Bulgaria

**Milena Dobрева** – University of Strathclyde, United Kingdom

**Radovan Garabík** – Ľudovít Štúr Institute of Linguistics, Slovak Academy of Sciences, Slovakia

**Maria Gavriliidou** – Institute for Language and Speech Processing, Natural Language Processing and Knowledge Extraction Department, Greece

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

**Momchil Hardalov** – Amazon AWS AI Labs, Spain / Sofia University St. Kliment Ohridski, Bulgaria

**Radu Ion** – Research Institute for Artificial Intelligence, Romanian Academy, Romania

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

**Cvetana Krstev** – University of Belgrade, Faculty of Philology, Serbia

**Eric Laporte** – University of Paris-Est Marne-la-Vallée, France

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

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

**Ana Ostroski Anic** – Institute of the Croatian Language, Croatia

**Vasile Pais** – Research Institute for Artificial Intelligence, Romanian Academy, Romania

**Bolette Pedersen** – University of Copenhagen, Centre for Language Technology, Denmark

**Maciej Piasecki** – Wrocław University of Technology, Poland

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

**Ewa Rudnicka** – Wrocław University of Technology, Poland

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

**Milena Slavcheva** – Institute of Information and Communication Technologies, Department

of Linguistic Modelling and Knowledge Processing, Bulgarian Academy of Sciences, Bulgaria  
**Kresimir Sojat** – University of Zagreb, Faculty of Humanities and Social Sciences, Croatia  
**Ranka Stankovic** – University of Belgrade, Serbia  
**Ivelina Stoyanova** – Institute for Bulgarian Language, Department of Computational Linguistics, Bulgarian Academy of Sciences, Bulgaria  
**Stan Szpakowicz** – University of Ottawa, Canada  
**Hristo Tanev** – Joint Research Centre of the European Commission, Italy  
**Irina Temnikova** – Big Data for Smart Society Institute (GATE), Bulgaria  
**Tinko Tinchev** – Sofia University, Faculty of Mathematics and Informatics, Bulgaria  
**Maria Todorova** – Institute for Bulgarian Language, Department of Computational Linguistics, Bulgarian Academy of Sciences, Bulgaria  
**Cristina Vertan** – University of Hamburg, Germany  
**Shuly Wintner** – University of Haifa, Department of Computer Science, Israel  
**Piek Vossen** – Free University of Amsterdam, the Netherlands  
**Katerina Zdravkova** – University St Cyril and Methodius in Skopje, North Macedonia

## ORGANISING COMMITTEE

### Chair:

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

**Atanas Atanasov** – Sofia University, Faculty of Slavic Studies, Bulgaria

**Rositsa Dekova** – Plovdiv University, Faculty of Philology, Department of English Studies, Bulgaria

**Nevena Grigorova** – Gate Institute, Sofia University, Bulgaria

**Dimitar Hristov** – Cleversoft, Bulgaria

**Nikolaos-Digenis Karagiannis** – Identrics, Bulgaria

**Hristina Kukova** – Institute for Bulgarian Language, Department of Computational Linguistics, Bulgarian Academy of Sciences, Bulgaria

**Todor Lazarov** – New Bulgarian University, Bulgaria

**Viktoria Petrova** – A1 Bulgaria, Bulgaria

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

**Ekaterina Tarpomanova** – Sofia University, Faculty of Slavic Studies, Bulgaria

# Table of Contents

|  |     |
|--|-----|
| <b>PLENARY TALKS</b> . . . . .   | 1   |
| Dr. Veselin Stoyanov <i>Large Language Models for the Real World: Explorations of Sparse, Cross-lingual Understanding and Instruction-Tuned LLMs</i> . . . . .   | 2   |
| Prof. Joakim Nivre <i>Ten Years of Universal Dependencies</i> . . . . .  | 3   |
| Prof. Vito Pirrelli <i>Written Text Processing and the Adaptive Reading Hypothesis</i> . . . . .   | 4   |
| <b>MAIN CONFERENCE</b> . . . . .   | 5   |
| Radu Ion, Verginica Barbu Mititelu, Vasile Păiș, Elena Irimia, Valentin Badea <i>A Cross-model Study on Learning Romanian Parts of Speech with Transformer Models</i> . . . . .  | 6   |
| Ekaterina Goliakova, David Langlois <i>What do BERT Word Embeddings Learn about the French Language?</i> . . . . .   | 14  |
| Camille Lavigne, Alex Stasica <i>Whisper-TAD: A General Model for Transcription, Alignment and Diarization of Speech</i> . . . . .   | 33  |
| Iglika Nikolova-Stoupak, Gaël Lejeune, Eva Schaeffer-Lacroix <i>Contemporary LLMs and Literary Abridgement: An Analytical Inquiry</i> . . . . .  | 39  |
| Milica Ikonić Nešić, Saša Petalinkar, Mihailo Škorić, Ranka Stanković, Biljana Rujević <i>Advancing Sentiment Analysis in Serbian Literature: A Zero and Few-Shot Learning Approach Using the Mistral Model</i> . . . . .                          | 58  |
| Lyuboslav Karev, Ivan Koychev <i>Generating Phonetic Embeddings for Bulgarian Words with Neural Networks</i> . . . . .   | 71  |
| Nelda Kote, Rozana Rushiti, Anila Çepani, Alba Haveriku, Evis Trandafil, Elinda Kajo Meçe, Elsa Skënderi Rakiplari, Lindita Khanari, Albana Deda <i>Universal Dependencies Treebank for Standard Albanian: A New Approach</i> . . . . .            | 80  |
| Verginica Barbu Mititelu, Tudor Voicu <i>Function Multiword Expressions Annotated with Discourse Relations in the Romanian Reference Treebank</i> . . . . .  | 90  |
| Atanas Atanasov <i>Dependency Parser for Bulgarian</i> . . . . .   | 98  |
| Madalina Chitez, Ana-Maria Bucur, Andreea Dinca, Roxana Rogobete <i>Towards a Romanian Phrasal Academic Lexicon</i> . . . . .  | 106 |
| Laura Rituma, Gunta Nešpore-Bērzkalne, Agute Klints, Ilze Lokmane, Madara Stāde, Pēteris Paikens <i>Classifying Multi-Word Expressions in the Latvian Monolingual Electronic Dictionary Tēzaurus.lv</i> . . . . .                                  | 113 |
| Laura Occhipinti <i>Complex Word Identification for Italian Language: A Dictionary-based Approach</i> . . . . .  | 119 |
| Ivana Brač, Matea Birtić <i>Verbal Multiword Expressions in the Croatian Verb Lexicon</i>  | 130 |
| Alessandro Lento, Andrea Nadalini, Marcello Ferro, Claudia Marzi, Vito Pirrelli, Tsvetana Dimitrova, Hristina Kukova, Valentina Stefanova, Maria Todorova, Svetla Koeva <i>Assessing Reading Literacy of Bulgarian Pupils with Finger-tracking</i> | 140 |

|  |     |
|--|-----|
| Denitza Kurshumova <i>Educational Horizons: Mapping the Terrain of Artificial Intelligence Integration in Bulgarian Educational Settings</i> . . . . .                         | 150 |
| Ekaterina Tarpomanova <i>Evidential Auxiliaries as Non-reliability Markers in Bulgarian Parliamentary Speech</i> . . . . .   | 157 |
| Iglika Nikolova–Stoupak, Eva Schaeffer–Lacroix, Gaël Lejeune <i>Extended Context at the Introduction of Complex Vocabulary in Abridged Literary Texts</i> . . . . .            | 166 |
| Junya Morita <i>Corpus-based Research into Derivational Morphology: A Comparative Study of Japanese and English Verbalization</i> . . . . .                                    | 178 |
| Ivan Derzhanski, Olena Siruk <i>The Verbal Category of Conditionality in Bulgarian and its Ukrainian Correspondences</i> . . . . .   | 187 |
| Natalia Dankova <i>Lexical Richness of French and Quebec Journalistic Texts</i> . . . . .  | 196 |
| Maria Khokhlova, Mikhail Koryshev <i>A Corpus of Liturgical Texts in German: Towards Multilevel Text Annotation</i> . . . . .  | 201 |
| Valentin Zmiycharov, Todor Tsonkov, Ivan Koychev <i>EurLexSummarization – A New Text Summarization Dataset on EU Legislation in 24 Languages with GPT Evaluation</i> . . . . . | 206 |
| Petya Osenova <i>On a Hurltlex Resource for Bulgarian</i> . . . . .  | 214 |
| Fabio Maion, Tsvetana Dimitrova, Andrej Bojadziev <i>Unified Annotation of the Stages of the Bulgarian Language. First Steps</i> . . . . .                                     | 220 |
| Amal Haddad Haddad, Damith Premasiri <i>ChatGPT: Detection of Spanish Terms Based on False Friends</i> . . . . .   | 227 |
| Jordan Kralev <i>Deep Learning Framework for Identifying Future Market Opportunities from Textual User Reviews</i> . . . . .   | 241 |
| Ruslana Margova, Bastiaan Bruinsma <i>Look Who’s Talking: The Most Frequently Used Words in the Bulgarian Parliament 1990-2024</i> . . . . .                                   | 249 |
| Sabrina Mennella, Maria Di Maro, Martina Di Bratto <i>Estimating Commonsense Knowledge from a Linguistic Analysis on Information Distribution</i> . . . . .                    | 257 |
| Georgi Pashev, Silvia Gaftandzhieva <i>Pondera: A Personalized AI-Driven Weight Loss Mobile Companion with Multidimensional Goal Fulfillment Analytics</i> . . . . .           | 264 |
| Stanislav Penkov <i>Mitigating Hallucinations in Large Language Models via Semantic Enrichment of Prompts: Insights from BioBERT and Ontological Integration</i> . . . . .     | 272 |
| Maria Todorova <i>Commercially Minor Languages and Localization</i> . . . . .  | 277 |
| <b>SPECIAL SESSION ON WORDNETS, FRAMENETS AND ONTOLOGIES</b> . . . . .   | 285 |
| Ivelina Stoyanova <i>Semantic features in the automatic analysis of verbs of creation in Bulgarian and English</i> . . . . .   | 286 |
| Svetlozara Leseva <i>A ‘Dipdive’ into Motion: Exploring Lexical Resources towards a Comprehensive Semantic and Syntactic Description</i> . . . . .                             | 296 |
| Ivelina Stoyanova, Hristina Kukova, Maria Todorova, Tsvetana Dimitrova <i>Multilingual Corpus of Illustrative Examples on Activity Predicates</i> . . . . .                    | 309 |
| Svetla Koeva <i>Large Language Models in Linguistic Research: the Pilot and the Copilot</i>  | 319 |
| <b>LIST OF AUTHORS</b> . . . . .   | 332 |

---

## PLENARY TALKS

---

# Large Language Models for the Real World: Explorations of Sparse, Cross-lingual Understanding and Instruction-Tuned LLMs

Dr. Veselin Stoyanov (Tome AI, USA)

---

Large language models (LLMs) have revolutionized NLP and the use of Natural Language in products. Nonetheless, there are challenges to the wide adoption of LLMs. In this talk, I will describe my explorations into addressing some of those challenges. I will cover work on sparse models addressing high computational costs, multilingual LLMs addressing the need to handle many languages, and work on instruction finetuning addressing the alignment between model outputs and human needs.

# Ten Years of Universal Dependencies

Prof. Joakim Nivre (Uppsala University and RISE, Sweden)

---

Universal Dependencies (UD) is a project developing cross-linguistically consistent treebank annotation for many languages, with the goal of facilitating multilingual parser development, cross-lingual learning, and parsing research from a language typology perspective. Since UD was launched almost ten years ago, it has grown into a large community effort involving over 500 researchers around the world, together producing treebanks for 148 languages and enabling new research directions in both NLP and linguistics. In this talk, I will review the history and development of UD and discuss challenges that we need to face when bringing UD into the future.

# Written Text Processing and the Adaptive Reading Hypothesis

Prof. Vito Pirrelli (NRC, Institute for Computational Linguistics, Pisa, Italy)

---

Oral reading requires the fine coordination of eye movements and articulatory movements. The eye provides access to the input stimuli needed for voice articulation to unfold at a relatively constant rate, while control on articulation provides internal feedback to oculomotor control for eye movements to be directed when and where a decoding problem arises.

A factor that makes coordination of the eye and the voice particularly hard to manage is their asynchrony. Eye movements are faster than voice articulation and are much freer to scan a written text forwards and backwards. As a result, given a certain time window, the eye can typically fixate more words than the voice can articulate.

According to most scholars, readers compensate for this functional asynchrony by using their phonological buffer, a working memory stack of limited temporal capacity where fixated words can be maintained temporarily, until they are read out loud. The capacity of the phonological buffer thus puts an upper limit on the distance between the position of the voice and the position of the eye during oral text reading, known as the eye-voice span.

In my talk, I will discuss recent reading evidence showing that the eye-voice span is the “elastic” outcome of an optimally adaptive viewing strategy, interactively modulated by individual reading skills and the lexical and structural features of a text. The voice span not only varies across readers depending on their rate of articulation, but it also varies within each reader, getting larger when a larger structural unit is processed. This suggests that skilled readers can optimally coordinate articulation and fixation times for text processing, adaptively using their phonological memory buffer to process linguistic structures of different size and complexity.

---

## MAIN CONFERENCE

---

# A cross-model study on learning Romanian parts of speech with Transformer models

**Radu Ion**  
Institute for AI,  
“Mihai Drăgănescu”  
radu@racai.ro

**Verginica Barbu  
Mititelu**  
Institute for AI,  
“Mihai Drăgănescu”  
vergi@racai.ro

**Vasile Păiș**  
Institute for AI, “Mihai  
Drăgănescu”  
vasile@racai.ro

**Elena Irimia**  
Institute for AI, “Mihai  
Drăgănescu”  
elena@racai.ro

**Valentin Badea**  
Institute for AI, “Mihai  
Drăgănescu”  
valentin.badea@racai.ro

## Abstract

This paper will attempt to determine experimentally if POS tagging of unseen words produces comparable performance, in terms of accuracy, as for words that were rarely seen in the training set (i.e. frequency less than 5), or more frequently seen (i.e. frequency greater than 10). To compare accuracies objectively, we will use the odds ratio statistic and its confidence interval testing to show that odds of being correct on unseen words are close to odds of being correct on rarely seen words. For the training of the POS taggers, we use different Romanian BERT models that are freely available on HuggingFace.

**Keywords:** BERT, POS tagging, Romanian, odds ratio, POS learning.

## 1 Introduction

Transformer models (Vaswani et al., 2023) and Deep Learning have changed the face of Natural Language Processing (NLP) domain, with a huge number of papers reporting superior performances of any conceivable task of NLP, including machine translation, question answering (which is now handled almost flawlessly by generative Large Language Models), and language analysis (POS tagging, dependency parsing, word sense disambiguation, etc.)

Transformer models are very good at any NLP task, provided they are pre-trained on very large corpora (billions of words) at supervised tasks such as masked language modeling or next sentence prediction (Devlin et al., 2019) and then,

fine-tuned to the task at hand, e.g. POS tagging. Central to the Transformer models’ remarkable ability to learn syntagmatic information about words is the attention mechanism (Vaswani et al., 2023), which encodes co-occurrence information in a large window of tokens (typically 512 tokens) for a large vocabulary of tokens (typically 50K tokens).

Comparatively, the number of papers dealing with the subject of how the Transformer model is learning a language (or multiple languages at once), which presumably makes them so good at any language processing task, is very small with respect to the number of papers presenting extensions of the model, accuracy improvements, applications, and so on.

With this paper, we want to contribute to the set of papers tapping into the learning mechanisms of the Transformer models, and we present a study on if and how the BERT models (a type of Transformer models) learn the grammatical categories (e.g. noun, verb, article, determiner, etc.) of a word in its context (i.e. POS tagging with a smaller tagset). We focus on Romanian, and we use Romanian-specific BERT models for the job. We will try to experimentally prove that BERT models have about the same accuracy on unseen (during training) words as on words that were rarely seen. Furthermore, the accuracy of frequently seen words is not that much higher than the accuracy of unseen words. To quantify these comparisons, we will use the odds ratio statistic.

## 2 Related work

Experiments on how POS taggers work on words not seen during training were performed more than 20 years ago, at the time when POS taggers were actively developed using e.g. Hidden Markov Models. An example in this regard is the work by Dematas and Kokkinakis (1995), which addresses the POS tagging of unseen words with enhanced HMMs. At that time, the best tagging accuracy on these words was about 66% for English.

Kim and Smolensky (2021) investigated the ability of pre-trained Transformer models (i.e. BERT-large, Devlin et al., 2019) to perform grammatical category-based generalization of novel words, after being finetuned on limited contexts (without categorization-specific training). Inspired by an experimental design in which infants were familiarized to contexts containing novel words and then tested with new sentences that either obeyed or violated category-based co-occurrence restrictions, the authors assumed that a Masked Language Model’s (MLM) ability to assign a higher probability to a word in a novel context that obeys the co-occurrence restriction for that category (over a word that does not) means the MLM makes a valid grammatical category inference about a novel word.

The two-step method involves finetuning the MLM on two signal contexts that unambiguously mark the novel words ( $w_1$  and  $w_2$ ) grammatical categories and testing the fine-tuned model by comparing the probabilities of  $w_1$  and  $w_2$  on multiple test contexts (higher probability to the new word in the correct test context meaning accurate category inference).

The signal and test contexts are based on MNLI corpus (Williams et al., 2018), that had different sources from the model’s pre-training data. A finetuning set contains two signal contexts with one unseen word each ( $w_1$  and  $w_2$ ) and 400 test contexts, 200 for each grammatical category (MNLI-sampled sentences in which words with grammatical categories of interest are masked out). Six English datasets that test for the binary classification between the four open-class grammatical categories (noun, verb, adjective and adverb) were constructed.

To use “unknown” words and make the BERT-large model to “forget” learned words, random weights were used for the unknown words’ embeddings. The BERT-large model was finetuned

for 70 epochs and accuracy was tested at a significance level of  $p < 5\%$  with a one proportion z-test. Conclusions were as follows:

1. accuracy largely varied between category pairs, from 67.3% for noun vs. adverb to 88.1% for noun vs. verb.
2. category inference was quite slow in comparison to competent speakers’ performances who often can solve the task from a single example.

In another study targeted at “what contextual representations encode that conventional embeddings do not?”, Tenney et al. (2019) compare conventional word embeddings to Transformer-generated word embeddings, which they call “contextual embeddings”. For this purpose, they propose to probe a contextual embeddings model based on a simple architecture employing span representations and binary classifiers. In their approach, a span corresponds to a word or a sequence of words and the classifiers are trained to predict specific labels. The authors call this approach “edge probing”. For part-of-speech (POS) tagging, the OntoNotes (Weischedel et al., 2013) corpus is used (even though the authors investigate other tasks as well, making use of OntoNotes or additional corpora). The span for which a prediction is made corresponds to a single word. The classifiers are trained to predict individual part of speech tags (such as noun, verb, adjective, etc.) for the current word. The authors explore 4 contextual encoder models: CoVe (McCann et al., 2017), ELMo (Peters et al., 2018), OpenAI GPT (Radford et al., 2018), and BERT (Devlin et al., 2019). The models’ weights are not fine-tuned. For BERT and GPT, contextual word vectors are obtained using two methods: concatenation of the subword embeddings with the activations of the top layer, or a linear combination of layer activations (including embeddings) using learned task-specific scalars. The authors compare the results of the entire model with so-called “lexical baselines” in which the probing model is trained only on the most closely related context-independent word representations (for example in the case of ELMo, only the activations of the context-independent character-CNN layer (layer 0) are used). Regarding POS tagging, the BERT models outperform the other models (with BERT-base, using a concatenation approach, achieving the highest F1 score). In all cases, using the full models outperform the “lexical baselines”, while

ELMo “lexical baseline” outperforms the others. The authors consider that the results suggest that ELMo encodes local type information. In addition, the authors try to estimate how much information is derived from long distance tokens. Their original architecture is extended with a CNN layer of width 3 (considering one token to the left and one to the right). This addition significantly reduces the gap between a “lexical baseline” ELMo and the full model, indicating that ELMo improvements are due to the encoding of long-range information.

Metheniti et al. (2022) report on experiments in which several Transformer models (BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2019), ALBERT (Lan et al., 2019)) are tested with respect to their ability to capture information about telicity and duration of verbs. While these semantic features are not directly related to grammatical category learning, their study represents another very good example of analyzing how Transformer models learn the language.

An action is telic if it has an end point and atelic otherwise. Durative verbs describe an action, while stative verbs describe states. The authors work with English and French, and in one experiment, they fine-tune the transformer models for binary sequence classification of telicity and duration (separately), and of testing their accuracy on predicting these features. For fine-tuning they use a set of sentences annotated for these features.

In another experiment, no fine-tuning was performed. Instead, a logistic regression to the contextual embeddings of each layer is applied, as provided by the pre-trained models. Contextual word embeddings for the annotated verbs are extracted from each layer of the transformer model and a logistic regression model is trained to classify telicity and duration, to understand how much such information was learned by each layer.

For classifying telicity, all systems obtain an accuracy above 80% and it improved when training the models with the extra information of verb position in the sentence. BERT (both base and large) had the best results.

For classifying duration, the results are even better (higher than 93%), despite using a smaller dataset. No improvement could be noticed when

training the models with the extra information of verb position in the sentence. BERT was also the best performing.

An error analysis showed that conflicting characteristics of the linguistic context prevent the correct analysis: e.g., sentences where the verb or the verbal phrase would be considered (a)telic, but part of the context defines the temporal aspect of the sentence in the opposite way.

For French, the results are not as good as for English, probably because of the characteristics of the French verbal system.

Contextual embeddings proved to be an efficient way to encode the aspectual information of a verb and its interaction with its context, and this knowledge is probably already learned in the pre-training process.

### 3 RoBERT models

RoBERT<sup>1</sup> is a Romanian-only, pre-trained BERT model. Masala et al. (2020) developed this model to address the gap in pre-trained language models for languages other than English. The model was designed similarly to BERT with small, base, and large variants, having the same number of layers, hidden params, and attention heads. The training time in hours for each model was 28, 77, and 255, respectively, training for 40 epochs on a v3-8 TPU on two supervised tasks: masked language modeling and next sentence prediction. The Romanian dataset that was used for training was comprised of 3 sources, totaling 2.07B words.

Without dwelling into details, the RoBERT models outperform the competition in several tasks, namely mBERT (Devlin et al., 2019), XLM-RoBERTa (Conneau et al., 2020), and the only true Romanian BERT model at that time, BERT-base-ro (Dumitrescu et al., 2020).

## 4 Experiments

### 4.1 Preliminaries

In our experiments we use the RoRefTrees (RRT) Romanian UD corpus<sup>2</sup> (Barbu Mititelu, 2018), currently at version 2.13. The corpus is pre-split into the train, dev and test sets and we join the dev and test sets into a bigger test set, while only

<sup>1</sup> <https://huggingface.co/readerbench>

<sup>2</sup>

[https://github.com/UniversalDependencies/UD\\_Romanian-RRT](https://github.com/UniversalDependencies/UD_Romanian-RRT)

training on the train set. All accuracy results that are presented in the next sections are computed over this bigger test set.

Since, for the time being, we are interested in how BERT models learn the grammatical categories, and only the grammatical categories without any other morphosyntactic attributes such as number, gender, tense, etc., we selected the POS and its type to comprise the grammatical category label to target the training for. We ended up with 35 categories, as follows:

- Proper and common nouns and numerals.
- Main and auxiliary verbs.
- Adjectives and adverbs.
- Abbreviations (of different types, e.g. nominal, adjectival, adverbial, etc.).
- Pronouns and determiners of different types (e.g. personal, demonstrative, reflexive, indefinite, etc.)
- Articles (possessive, indefinite)
- Prepositions
- Conjunctions (coordinative, subordinative)
- Particles (infinitive, negative).

We further split the set of grammatical categories into two subsets: the *content words* set which contains all categories of “meaning bearing” words (proper and common nouns, main verbs, general adverbs, adjectives, numerals, and abbreviations) and the set of *functional words* which is the complement of the full set of categories with respect to the content words set. The reason we consider these subsets is that the rarely seen and the unseen words in the test set vastly belong to the content words set (see Table 2, below), and computing accuracies including functional words would yield an unfair advantage to the words that are seen frequently in the training set.

Table 1 presents statistics of the content and functional words in the RRT, in each split and Table 2 shows how different word types from dev plus test splits are distributed at  $F = 0$  (do not appear at all in the train split),  $F = 1$  (appear once in the train split) and  $F > 1$  (appear more than once in the train split).

|              | Cont.  | Func.  | Punc.  |
|--------------|--------|--------|--------|
| <b>train</b> | 92,694 | 68,740 | 23,691 |
| <b>dev</b>   | 8,633  | 6,217  | 2,223  |
| <b>test</b>  | 8,277  | 5,964  | 2,083  |

Table 1: RRT word type statistics

|                   | F = 0 | F = 1 | F > 1  |
|-------------------|-------|-------|--------|
| <b>Content</b>    | 3,185 | 1,702 | 12,023 |
| <b>Functional</b> | 39    | 19    | 12123  |

Table 2: Word count distribution by frequency for the dev plus test bigger test set

## 4.2 Testing methodology

The BERT models are fitted with a POS classification layer on top of the last hidden state of each token. The POS layer has 35 dimensions, one for each considered grammatical category, and it is trained with a softmax learning objective. We also update the BERT model’s parameters in the backward propagation stage. The starting learning rate parameter is set at  $10^{-5}$  and it is decreased by a factor of 0.9 every epoch, out of the 5 training epochs.

We will attempt to experimentally prove the following hypothesis: the POS tagging of unseen words (i.e. in the training set) is as accurate as POS training of words that were seen in the training set. We will measure the odds ratio (*OR*, Bland and Altman, 2000) of the odds of being correct vs. being incorrect when the frequency  $F$  of the targeted words is greater than 0 compared to when  $F$  is 0 in the training set. Thus, we compute the *OR* fraction from the following contingency table:

|                  | F > 0     | F = 0     |
|------------------|-----------|-----------|
| <b>Correct</b>   | $p_c$     | $q_c$     |
| <b>Incorrect</b> | $1 - p_c$ | $1 - q_c$ |

Table 3: *OR* contingency table

as

$$OR = \frac{\frac{p_c}{1 - p_c}}{\frac{q_c}{1 - q_c}} = \frac{p_c(1 - q_c)}{q_c(1 - p_c)}$$

and show that it is close to 1, in a confidence interval that forbids rejecting the null hypothesis of it being different than 1. In the above equation,  $p_c$  and  $q_c$  are the probabilities of being correct in the chosen sample (i.e. the ratio of correctly tagged words out of all tagged words in the sample).

We will only target words that belong to the chosen BERT model vocabulary, such that the evaluated word is not split into sub-words by the WordPiece tokenizer. We enforce this constraint for two reasons:

1. We do not want to average BERT representations of sub-words to obtain a representation for the full word, because the average of embeddings is not necessarily the equivalent of producing the true representation of the full word.
2. We are interested in a study targeting the specific dimensions of a word representation that mostly decide its grammatical category.

Finally, as previously mentioned, we only compute and compare odds ratios for content words, for the reasons explained above.

### 4.3 Results with the RoBERT models

We trained the `readerbench/RoBERT-small`, `readerbench/RoBERT-base` and `readerbench/RoBERT-large` models from HuggingFace the way we described previously. Table 4 presents an overview of the accuracies we obtained on the POS tagging task, with the 35 POS labels, on all words (content and functional), at different frequency thresholds, as shown in the table’s header.

|              | $F \geq 0$   | $F = 0$      | $F = 1$      | $F > 0$    |
|--------------|--------------|--------------|--------------|------------|
| <b>small</b> | 96.5%        | 92.6%        | 90.9%        | 96.6%      |
| <b>base</b>  | <b>97.9%</b> | 92.6%        | <b>94.8%</b> | <b>98%</b> |
| <b>large</b> | 96.5%        | <b>93.3%</b> | 93.1%        | 96.6%      |

Table 4: Accuracy on content and functional words

Table 5 below shows the same accuracy figures, but only for content words POS tagging.

|              | $F \geq 0$   | $F = 0$      | $F = 1$    | $F > 0$      |
|--------------|--------------|--------------|------------|--------------|
| <b>small</b> | 93.8%        | 92.9%        | 89.4%      | 93.9%        |
| <b>base</b>  | 96.4%        | 92.7%        | <b>94%</b> | <b>96.7%</b> |
| <b>large</b> | <b>96.5%</b> | <b>93.5%</b> | 92.4%      | 95.5%        |

Table 5: Accuracy on content words only

One thing we see from Tables 4 and 5 is that the large model is better at tagging unseen words while the base model is better at everything else. Comparing the values of the accuracy figure from Table 5 for  $F = 0$  and  $F > 0$ , we see differences of at least 1%. This could suggest that the model is not able to learn the grammatical categories of unknown words, but this conclusion is going to be amended when we plot odds ratios at different frequency bands.

Figures 1 to 3 show the odds ratios plot, for each of the RoBERT models, computed as in Table 3 for

content words only. Thus, we have the *OR* on the *Y* axis, while on the *X* axis we have a frequency step of 1 for which the Table 3  $F > 0$  condition holds: 1, 1 and 2, 1 to 3, ..., 1 to 10, ..., 1 to 20, etc. We show the current sample (dev plus test) *OR* variation in blue with dots, the low value of the confidence interval (CI) in orange with downward arrows and the high value of the CI in green with upward arrows, considering a 95% level of confidence.

RoBERT-small and RoBERT-large models show that the sample *OR* statistic is close to 1 when  $F < 3:1$  for the small model and 1.4 for the large model. That is, being correct on unseen words happens at about the same rate as being correct on rarely seen words. Going up the frequency range, the *OR* starts to increase in all cases: models learn to disambiguate the more frequently occurring words better, because they have seen more contexts of those words. Lastly, in all three plots we see that the sample *OR* statistic sits comfortably within the limits of its CI, meaning that the value is very likely to be correct, and not smaller or greater than what we got.

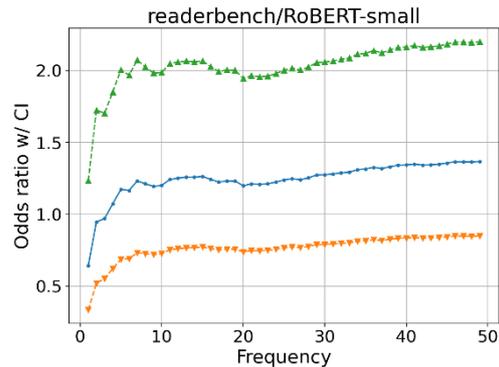


Figure 1: RoBERT-small OR variation with frequency

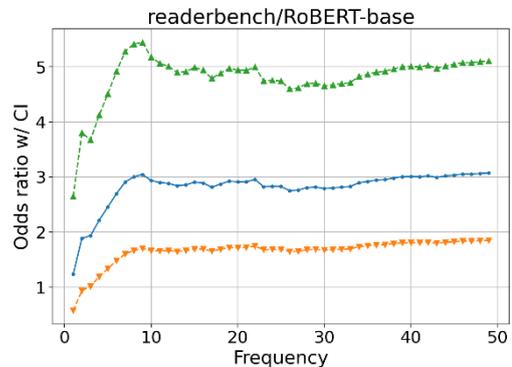


Figure 2: RoBERT-base OR variation with frequency

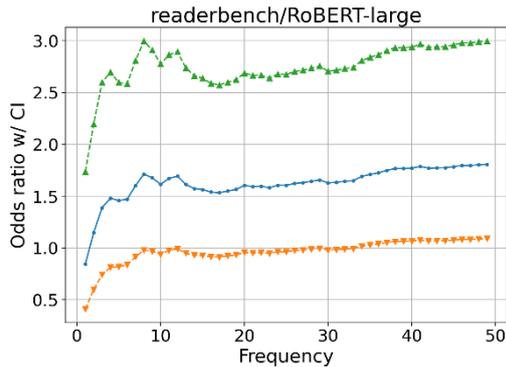


Figure 3: RoBERT-large OR variation with frequency

The RoBERT-base model is different, with the sample *OR* settling around 2 for  $F < 3$  indicating that this model is more likely to be correct when words have been barely seen in the training. By the time the frequency range gets to 10, the *OR* statistic is 3, more than twice the one from the other two models.

#### 4.4 Results with the CoRoLa BERT model

We previously trained a small BERT model (of approximately the same size as the RoBERT-small model) on the CoRoLa reference corpus for the contemporary Romanian language (Barbu Mititelu et al., 2019). We intended to use this model to study how a Transformer encoder learns the grammar of a language (in our case, Romanian). The model uses a vocabulary that is 13 times bigger than RoBERT’s, wishing to account for the inflected nature of Romanian. The CoRoLa train set had just over 760 million words, and the CoRoLa BERT model was trained with the Masked Language Modeling training objective.

Table 6 shows accuracies at different frequencies, for all words (content plus functional) and for content words only.

|              | $F \geq 0$ | $F = 0$ | $F = 1$ | $F > 0$ |
|--------------|------------|---------|---------|---------|
| <b>All</b>   | 93.8%      | 76.9%   | 89.6%   | 95.8%   |
| <b>Cont.</b> | 91.4%      | 77.3%   | 90.1%   | 94.4%   |

Table 6: CoRoLa BERT accuracy

We can compare these figures with the RoBERT-small’s accuracies (Tables 4 and 5), as CoRoLa BERT is about the same size, parameter-wise. While RoBERT-small outperforms CoRoLa BERT at all categories, except for the accuracy on content words when the frequency  $F > 0$ , the biggest difference is when  $F = 0$ : more than 15 percents in

favor of RoBERT-small. There are two explanations for this:

1. CoRoLa BERT has been under pre-trained for its massive vocabulary, which has 500K words vs. 38K words of RoBERT-small’s. We pre-trained on only 760M words while RoBERT-small model was pre-trained on 2B words.
2. We only evaluate on words from the model’s vocabulary, and thus, CoRoLa BERT is evaluated on many more words than RoBERT, at all frequency thresholds, because its vocabulary is much bigger. Just for the sake of comparison, RoBERT tokenizer recognizes 18K word occurrences in our test set while CoRoLa BERT tokenizer recognizes 27K word occurrences.

When we plot the variation of the *OR* statistic with the frequency, as we did for the RoBERT models, we see the picture of an undertrained BERT model (see Figure 4, below).

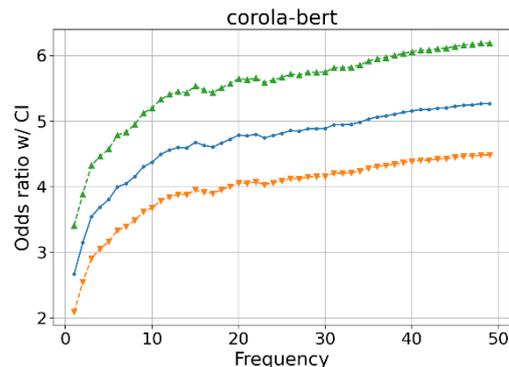


Figure 4: CoRoLa BERT OR variation with frequency

For  $F \leq 3$ , the *OR* statistic is already 3.5 and the function quickly increases, showing that this BERT model can do better POS tagging only on seen during training words.

#### 4.5 A dimension-by-dimension hidden state analysis for POS tagging

The RoBERT-small and the CoRoLa BERT models have the same size of the hidden state vector: 256 dimensions, counted from 0 to 255. We wanted to know if we can find a common subset of dimensions that are responsible for the correct classification of each grammatical category.

To obtain the subset of dimensions that contribute the most to the output layer’s highest probable grammatical category, we can use the output layer weight matrix from which we extract the row corresponding to the index of the most

probable grammatical category and multiply it, element-wise, with the hidden state of our target word. From the obtained vector, we extract the indexes of the top 10 largest elements, as the dimensions of the model hidden state that contribute the most to the correct grammatical category classification.

If we compute the most important 10 dimensions for each correctly classified word in the test set, we can derive a conditional probability distribution for each of the 35 grammatical categories, for both RoBERT-small and CoRoLa BERT. Table 7 shows which dimensions have been found as being common between RoBERT-small and CoRoLa BERT, for each grammatical category<sup>3</sup>, with their sum of conditional probabilities.

| cat | $\sum P(d cat)$ | $\sum Q(d cat)$ | Common d                 |
|-----|-----------------|-----------------|--------------------------|
| Af  | 0.120           | 0.184           | 255, 28, 138, 31         |
| Cc  | 0.037           | 0.066           | 77                       |
| Cr  | 0.068           | 0.112           | 213, 113                 |
| Cs  | 0.075           | 0.098           | 59, 10                   |
| Dd  | 0.043           | 0.012           | 52                       |
| Di  | 0.197           | 0.146           | 26, 65, 5, 84            |
| Ds  | 0.024           | 0.013           | 22                       |
| Dw  | 0.137           | 0.171           | 147, 196, 104            |
| Mc  | 0.152           | 0.239           | 88, 67, 218, 40, 197     |
| Mo  | 0.044           | 0.048           | 101, 113                 |
| Nc  | 0.024           | 0.017           | 160                      |
| Pd  | 0.142           | 0.230           | 134, 141, 1, 9, 213, 143 |
| Pp  | 0.072           | 0.160           | 234, 213, 11             |
| Pw  | 0.061           | 0.102           | 234, 31                  |
| Px  | 0.010           | 0.012           | 103                      |
| Pz  | 0.110           | 0.155           | 220, 146, 187            |
| Qn  | 0.105           | 0.110           | 5, 32                    |
| Qs  | 0.160           | 0.130           | 239, 249, 112, 128       |
| Qz  | 0.038           | 0.017           | 74                       |
| Rc  | 0.052           | 0.101           | 212, 241                 |
| Rg  | 0.015           | 0.066           | 167                      |
| Rp  | 0.057           | 0.236           | 119, 115, 104            |
| Sp  | 0.164           | 0.053           | 109, 134, 112            |
| Tf  | 0.117           | 0.096           | 94, 192, 62              |
| Ti  | 0.054           | 0.112           | 47, 22                   |
| Ts  | 0.098           | 0.113           | 36, 145                  |

<sup>3</sup> For an explanation of the grammatical category codes, one can consult the MSD definitions from <https://nl.ijs.si/ME/V6/msd/html/msd-ro.html>

|    |       |       |               |
|----|-------|-------|---------------|
| Va | 0.041 | 0.139 | 49, 0         |
| Vm | 0.095 | 0.070 | 157, 193      |
| Yn | 0.084 | 0.072 | 255, 102, 179 |

Table 7: CoRoLa BERT and RoBERT-small common dimensions per grammatical category

From the cumulative probabilities of CoRoLa BERT ( $\sum P(d|cat)$ ) and RoBERT-small ( $\sum Q(d|cat)$ ), we see that the common dimensions do not carry a lot of the whole probability mass for a category. If the sums of the probabilities had been higher for both models (say above 0.5), that would have been an indication that the common set of categories is important for both models, but this is not the case here. Thus, we can conclude that different BERT models do not assign the same importance to the same dimensions for a given grammatical category.

## 5 Conclusions

We have presented evidence that properly trained BERT models exhibit learning words’ grammatical categories, especially when the words were not seen during the training process. We drew this conclusion by measuring the odds ratio of POS tagging accuracy when the frequency of the test words (in the train set) is greater than 0 vs. when this frequency is 0. Thus, models RoBERT-small and RoBERT-large show an odds ratio that is less than 2 for the accuracy of tagging frequent words vs. tagging unseen words. We could not say that model CoRoLa BERT exhibits the same behavior due to its insufficient pre-training for its large vocabulary.

The model RoBERT-base shows a different behavior with respect to accuracy odds ratio vs. test word frequency: while the odds ratio of POS tagging accuracy is below 2 when comparing rare words ( $F \leq 3$ ) to unseen words ( $F = 0$ ), as in the case of the other two sibling models, when the frequency increases (e.g.  $F \geq 10$ ), the odds ratio settles at a bit over 3 (twice as much when compared to the other two models). While it is expected that the POS tagging accuracy increases with the test word frequency (in the train set), as more contexts of those words were seen during training, RoBERT-base does much better than the

other two sibling models when test words were seen during training. This hypothesis is supported by the top POS tagging accuracy of RoBERT-base compared to any other tested model (see Tables 4 and 5). One possible explanation for this situation is that RoBERT-base has the best number of parameters (not too few, nor too many) for our POS tagging task and this enables its accuracy odds ratio curve to increase more sharply than siblings' curves, but not that sharply as the odds ratio curve of CoRoLa BERT which indicates more of an overfit of the training data than a good performance.

## References

- Verginica Barbu Mititelu. 2018. Modern Syntactic Analysis of Romanian. In Ofelia Ichim, Luminița Botoșineanu, Daniela Butnaru, Marius-Radu Clim, Ofelia Ichim, Veronica Olariu (eds.), *Clasic și modern în cercetarea filologică românească actuală*, Iași, Publishing House of "Alexandru Ioan Cuza" University, 2018, pp. 67—78.
- Verginica Barbu Mititelu, Dan Tufiș, Elena Irimia, Vasile Păiș, Radu Ion, Nils Diewald, Maria Mitrofan, Mihaela Onofrei. 2019. Little Strokes Fell Great Oaks. Creating CoRoLa, The Reference Corpus of Contemporary Romanian. *Revue roumaine de linguistique*, Issue 3. 2019
- Bland, J., and Altman, D. 2000. Statistics notes: The odds ratio. *British Medical Journal*, 320 (7247), 1468.
- 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)*, pp. 4171–4186, Minneapolis, Minnesota.
- Ștefan Dumitrescu, Andrei-Marius Avram and Sampo Pyysalo. 2020. The birth of Romanian BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4324–4328.
- Evangelos Dermatas and George Kokkinakis. 1995. Automatic stochastic tagging of natural language texts. *Computational Linguistics, Volume 21, Issue 2*, pp. 137—163.
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942.
- 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.
- Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. Learned in translation: Contextualized word vectors. In *Proceedings of NIPS*, 2017.
- Eleni Metheniti, Tim Van De Cruys, and Nabil Hathout. 2022. About Time: Do Transformers Learn Temporal Verbal Aspect? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pp. 88–101, Dublin, Ireland.
- Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of NAACL*, 2018.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. <https://blog.openai.com/language-unsupervised>, 2018.
- Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. *OntoNotes release 5.0 LDC2013T19*. Linguistic Data Consortium, Philadelphia, PA, 2013.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*, 32:5753–5763.
- Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112—1122, New Orleans, Louisiana.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin. 2023. Attention is All You Need. arXiv:1706.03762v7 [cs.CL]

# What do BERT word embeddings learn about the French language?

**Ekaterina Goliakova**

Université de Lorraine

ekaterina.goliakova6@  
etu.univ-lorraine.fr  
edgolyakova@gmail.com

**David Langlois**

Université de Lorraine,

LORIA (Laboratoire Lorrain en Recherche  
en Informatique et ses Applications)  
david.langlois@loria.fr

## Abstract

Pre-trained word embeddings (for example, BERT-like) have been successfully used in a variety of downstream tasks. However, do all embeddings, obtained from the models of the same architecture, encode information in the same way? Does the size of the model correlate to the quality of the information encoding? In this paper, we will attempt to dissect the dimensions of several BERT-like models that were trained on the French language to find where grammatical information (gender, plurality, part of speech) and semantic features might be encoded. In addition to this, we propose a framework for comparing the quality of encoding in different models.

**Keywords:** interpretability, word embeddings, intrinsic evaluation, BERT.

## 1 Introduction

With over 95,000 citations (and counting) since its publication in 2019, BERT (Devlin et al., 2019) can be considered one of the most prominent Large Language Models (LLMs) architectures in the current state of the art, finding applications in fields ranging from text to image generation (Rombach et al., 2022) to protein structure prediction (Jumper et al., 2021), still showing competitive results (Samuel et al., 2023). We can attribute one of the reasons for such a successful and wide-range usage of BERT-like models to the *word embeddings* (multidimensional word representations) they produce.

However, with ever-growing model sizes, the interpretability of dimensions of the learned representations is still a complex task. While the number of parameters is constantly growing, the performance of the models is not improving as rapidly, and we are facing diminishing returns with the increased model and training data scale (Kaplan et al., 2020, van Schijndel et al., 2019). Additionally, the size of available data does not grow at the same

rate as the model hyperparameters and all possible training data for Language Models might be exhausted between 2030 and 2040 (Villalobos et al., 2022). Therefore, shifting the research focus from constantly increasing the training datasets towards understanding more about existing models, their parameters (and how they can be improved) might become necessary in the nearest future.

In this paper, we propose an intrinsic metric evaluating the quality of information encoding (InfEnc) and suggest a framework that allows the identification of the best dimension candidates of word embeddings that potentially encode target information (in our experiments: grammatical number and gender for French nouns and adjectives, part of speech (POS) for French nouns, adjectives, and verbs, and semantic information for French nouns).

## 2 Related works

The explainability and interpretability of LLMs have become a growing interest for researchers. One of the approaches to the problem is to learn an explainable distributional word embedding model linking each feature to a word (for example, Snidarov et al., 2019). In this work, the representations were learned from a co-occurrence matrix, which allowed for high interpretability of the embeddings. However, the representations were underperforming on similarity tasks such as WS-353 in comparison to other distributional representation models such as GloVe (Lee et al., 2020) and BERT-like embeddings (Chronis and Erk, 2020).

Another approach to interpretability is to probe pre-trained models (Alain and Bengio, 2016; Belinkov, 2022; Tjoa and Guan, 2020). Torroba Henigen et al., 2020 focus on intrinsic probing that aims not only to identify if a linguistic feature is encoded in representations but additionally how the information is encoded. According to this defi-

dition, our work remains in the scope of intrinsic evaluation. [Torroba Hennigen et al., 2020](#) propose a method to efficiently detect the most relevant subset overall features whereas we pre-select an initial subset and compute a score for each pre-selected feature.

[Miaschi et al., 2020](#) explore the usefulness of word embeddings to predict features at the sentence level (length, depth of the syntactic parsing tree, etc.). The sentence embeddings are obtained by merging word embeddings (by sum, maximum, minimum, and average operators). The authors do not explore which features encode linguistic information. For [Ravichander et al., 2021](#), the objective is also not to dissect word embeddings in order to find features encoding linguistic information, but to study the usefulness of linguistic information for a classification task at sentence level. Contrary to these works, our objective is not to explore linguistic features at the sentence level, but at the word level.

### 3 Models

In our work, we compared word embeddings of 10 different BERT-like models working with the French language, details about which can be found in [Table 1](#).

#### mBERT

Being the multilingual version of BERT ([Devlin et al., 2019](#)), mBERT can be considered as the baseline for multilingual word embeddings. mBERT is a bi-directional model trained with the Masked Language Modelling (MLM) and the Next Sentence Prediction (NSP) objectives. The data used for the training of mBERT was sourced from the dump multilingual Wikipedias for 104 different languages, including French<sup>1</sup>.

#### DistilBERT

DistilBERT ([Sanh et al., 2019](#)) was created in response to the growing sizes of models. It was shown that despite being 40% smaller than the original BERT, DistilBERT has comparable performance on downstream tasks ([Sanh et al., 2019](#), [Jia et al., 2021](#), [Abdaoui et al., 2020](#)).

#### XLM-R

XLM-RoBERTa (or XLM-R) was created to address the underperformance of the traditional BERT

architecture on low-resource languages ([Conneau et al., 2020](#)). The model was trained on 2.5TB of CommonCrawl corpus ([Wenzek et al., 2020](#)), out of 56.8GB (2.3%) were the French data. The model was trained with the MLM objective, using only monolingual data in each of the 100 languages.

#### CamemBERT

Unlike the models listed above, CamemBERT ([Martin et al., 2020](#)) is a monolingual BERT-like model, the architecture of which is similar to that of RoBERTa ([Liu et al., 2019](#)). The model was trained on the OSCAR dataset ([Ortiz Suárez et al., 2019](#)) (monolingual corpora retrieved from CommonCrawl dataset snapshots), comprising 138GB of raw French data.

#### FlauBERT

Similarly to CamemBERT, FlauBERT ([Le et al., 2020](#)) is a monolingual French BERT-like model. It was trained with the MLM objective on 71GB of diverse French data, out of which 43GB (60%) were obtained from CommonCrawl data, 9GB (13%) from NewsCrawl ([Li et al., 2019](#)) corpus, around 7GB (9%) from Wikipedia and Wikisource dumps, with the remaining part of the dataset being constituted from various data sources.

#### 3.1 Extracting word embeddings

Word embeddings of BERT-like models are contextual: if a certain word has multiple senses, its representation might change depending on the sense ([Miaschi et al., 2020](#), [Ethayarajh, 2019](#)). In our experiments, we encode every word from the obtained vocabulary using the models and retrieve weights of the last layer to obtain embeddings, assuming that following this procedure we can get the most frequent representation of the word by a given model<sup>2</sup>. Likewise, to avoid ambiguity, we only consider words that are uniquely a noun, an adjective, or a verb.

For similar reasons, if a word is tokenized into multiple pieces, we consider that the tokens might contain ambiguous meanings (compare: *simplement* ('simply') → [simple, ment] and *mentons* ('[we] lie') → [ment, ons]<sup>3</sup>). Therefore, to avoid this sort of uncertainty, we only consider embeddings of words that are directly in the learned vocabulary of models (tokenized as one token). This

<sup>2</sup>CLS and end of the string tokens were discarded and not included in the obtained representations.

<sup>3</sup>For *simplement*, *ment* is the suffix for adverbs whereas for *mentons*, *ment* is the root of the verb *mentir*.

<sup>1</sup><https://github.com/google-research/bert/blob/master/multilingual.md>

| Model                                     | Embedding size | # of parameters | # of layers | Vocabulary size | Tokenization       | Training objective |
|---|----------------|-----------------|-------------|-----------------|--------------------|--------------------|
| FlauBERT <sub>small</sub>                 | 512            | 54M             | 6           | 68K             | Byte-Pair Encoding | MLM                |
| DistilBERT                                | 768            | 66M             | 6           | 119K            | WordPiece          | MLM                |
| CamemBERT <sub>base</sub>                 | 768            | 110M            | 12          | 32K             | SentencePiece      | multilingual MLM   |
| mBERT <sub>base</sub> (cased and uncased) | 768            | 110M            | 12          | 105K            | WordPiece          | MLM, NSP           |
| FlauBERT <sub>base</sub> (uncased)        | 768            | 137M            | 12          | 67K             | Byte-Pair Encoding | MLM                |
| FlauBERT <sub>base</sub> (cased)          | 768            | 138M            | 12          | 68K             | Byte-Pair Encoding | MLM                |
| XLM-R <sub>base</sub>                     | 768            | 270M            | 12          | 250K            | SentencePiece      | multilingual MLM   |
| FlauBERT <sub>large</sub>                 | 1024           | 373M            | 24          | 68K             | Byte-Pair Encoding | MLM                |
| XLM-R <sub>large</sub>                    | 1024           | 550M            | 24          | 250K            | SentencePiece      | multilingual MLM   |

Table 1: Model sizes for different French BERT-like models (sorted by the number of parameters).

approach also allows us to focus only on the most frequent words that the models have learned during their training.

### 3.2 Vocabulary and grammatical information

Morphalou (Romary et al., 2004) was used to obtain the initial vocabulary of French words, alongside the grammatical information (gender, number, POS). Morphalou is a lexical resource, that follows the Lexical Markup Framework (LMF) format and contains over 99,000 nouns and their inflected forms, over 14,000 verbs, and over 36,000 adjectives. The corpus contains other POS as well, however, their investigation lies outside of the scope of our project.

For each noun, adjective, or verb entry of Morphalou, we attempt to get a word embedding<sup>4</sup> by a given model. If the word is in the model vocabulary (tokenized as one token), we store this word and the corresponding grammatical information. For each word in the vocabulary of the models we stored its POS, grammatical number and gender (if applicable), and word embedding, afterward applying min-max normalization to the obtained word embeddings for each model.

The sizes of obtained datasets can be found in Table 2, and it can be noticed that the multilingual models have significantly smaller French vocabulary sizes, which can be expected due to French data being only a part of the final vocabulary of the model.

### 3.3 Semantic information

Additionally, FrSemCor’s Sequoia corpus (Barque et al., 2020) was used in order to retrieve semantic information about the obtained words. The corpus contains 12,917 French nouns annotated with 24 "supersenses" (e.g. Act, Person, State, Institution, etc). In our work we focus only on the PER-

<sup>4</sup>For all listed models, we extracted hidden states of the last layer to treat them as the word embeddings of the model, without any fine-tuning.

| Model                     | Cased | Nouns  | Adjectives | Verbs |
|---------------------------|-------|--------|------------|-------|
| FlauBERT <sub>small</sub> | ✓     | 12,807 | 6,504      | 5,425 |
| DistilBERT                | ✓     | 3,858  | 925        | 1,079 |
| CamemBERT <sub>base</sub> | ✓     | 8,945  | 4,584      | 3,852 |
| mBERT <sub>base</sub>     |       | 6,065  | 2,353      | 2,163 |
| mBERT <sub>base</sub>     | ✓     | 3,858  | 925        | 1,079 |
| FlauBERT <sub>base</sub>  |       | 15,579 | 7,590      | 6,377 |
| FlauBERT <sub>base</sub>  | ✓     | 12,807 | 6,504      | 5,425 |
| XLM-R <sub>base</sub>     | ✓     | 3,401  | 895        | 1,233 |
| FlauBERT <sub>large</sub> | ✓     | 12,807 | 6,504      | 5,425 |
| XLM-R <sub>large</sub>    | ✓     | 3,401  | 895        | 1,233 |

Table 2: Number of word embedding obtained by different models by POS.

SON<sup>5</sup> and ACT<sup>6</sup> as being one of the most frequent senses in the corpora, as well as in the vocabulary of models. The number of word embeddings of each model associated with the supersenses can be found in Table 3.

| Model                     | Cased | Act | Person |
|---------------------------|-------|-----|--------|
| FlauBERT <sub>small</sub> | ✓     | 405 | 354    |
| DistilBERT                | ✓     | 157 | 139    |
| CamemBERT <sub>base</sub> | ✓     | 311 | 285    |
| mBERT <sub>base</sub>     |       | 182 | 156    |
| mBERT <sub>base</sub>     | ✓     | 157 | 139    |
| FlauBERT <sub>base</sub>  |       | 417 | 364    |
| FlauBERT <sub>base</sub>  | ✓     | 405 | 354    |
| XLM-R <sub>base</sub>     | ✓     | 110 | 70     |
| FlauBERT <sub>large</sub> | ✓     | 405 | 354    |
| XLM-R <sub>large</sub>    | ✓     | 110 | 70     |

Table 3: The number of word embeddings obtained by different models by supersense.

## 4 Experimental protocol

As mentioned above, our approach consisted of combining word embeddings with grammatical/semantic information, extracted from human-annotated sources where the final list of tested features consisted of grammatical gender for nouns and adjectives, grammatical number for nouns and adjectives, POS for nouns, adjectives and verbs, and semantic categories for nouns. In order to use the annotations, a decision was made to create a binary vector for each feature: 1 was assigned to if a

<sup>5</sup>Examples: *étudiants* ('students'), *neveu* ('nephew'), *femme* ('woman').

<sup>6</sup>Examples: *ablation* ('ablation'), *accueil* ('reception').

word possessed a certain feature and 0 - otherwise<sup>7</sup>.

Due to the relatively small sizes of the datasets of annotated word embeddings, we opted to incorporate 5-fold cross-validation (Fushiki, 2011) instead of the traditional train/test split.

#### 4.1 Intrinsic evaluation of information encoding quality

It can be presumed that the information is encoded either throughout all dimensions of a word embedding or in a smaller subset of dimensions. However, identifying the exact subset is essentially computationally impossible (the number of dimension combinations ranging from  $2^{512}$  for smaller models to  $2^{1024}$  for larger ones). Therefore, we propose an approach to identify dimension subsets' candidates that can potentially contain the target information using the following steps:

- Identify sets of dimensions that are not independent of the feature vector and their intersections;
- Identify sets of important dimensions that are likely to be important for encoding the feature and their intersections;
- Compare the accuracy among each found dimension subset and all dimensions of the word embedding.

##### 4.1.1 Dependent dimensions

The intuition is that if a dimension is classified as dependent from the feature vector, it can be a potential candidate that encodes the information. As a way to find dependent dimensions, we perform two types of tests: Mutual Information (MI) (Kraskov et al., 2004) test and one-way ANOVA test. Additionally, we find the intersection between the sets of dimensions found by the MI test and the ANOVA test, to get another subset of dimension candidates.

As for the MI test, for  $i = 1 \dots n$ , where  $n$  is the number of dimensions in an embedding,  $f$  is the feature vector for all words of the vocabulary, we check the condition  $MI(d_i, f) > 0$  and if it is fulfilled we add  $d_i$  to the list of dependent dimensions detected by the MI test.

The ANOVA test operates with the null hypothesis that both samples have the same population mean, which in our case would represent values of

<sup>7</sup>For gender, 1 = feminine and 0 = masculine. For number, 1 = plural and 0 = singular.

dimensions where  $f = 0$  and  $f = 1$  have the same distribution. Therefore, for  $i = 1 \dots n$ ,  $d_i$  is split into two samples ( $d_{i_{f=0}}$  (values of dimension  $d_i$  for words that have the associated value  $f = 0$ ) and  $d_{i_{f=1}}$  (generated similarly to  $d_{i_{f=0}}$ ), the one-way ANOVA test is performed using the samples and if  $p\text{-value} < 0.01$ ,  $d_i$  is added to the list of dependent dimensions detected by the ANOVA test.

##### 4.1.2 Important dimensions

Another assumption we operate under is that some dimensions are more important than others in encoding the feature information. For this, we performed a series of tests that allowed us to rank the dimensions and picked only the top  $\alpha\%$  of them as important dimensions. The first approach to identifying important dimensions involves training a Logistic Regression (LR) model using all dimensions of an embedding with the feature vector  $f$  being used as its target. The absolute weights of the trained LR classifier that are associated with each dimension are sorted in descending order, and top  $\alpha\%$  are selected as important dimensions found by the LR test. Similarly, we train a Perceptron classifier following exactly the same approach.

Finally, for  $i = 1 \dots n$  we computed  $corr(d_i, f)$  where  $n$  is the number of dimensions,  $f$  is the feature vector and  $corr$  is the Point-biserial correlation coefficient. The dimensions were then sorted by the absolute values of the associated correlation score and the top  $\alpha\%$  were selected as the ones highlighted by the correlation test. The decision to use the Point-biserial correlation was driven by the fact that it is possible to use the metric with continuous (dimension values) and discrete (the feature vector) and in our case is synonymous with easily computable Pearson correlation.

Moreover, for each  $\alpha$  we calculate the intersection between all groups of important dimensions in order to find additional subsets of dimension candidates. All the tests were repeated for the following values of  $\alpha$ : [1, 5, 10, 25, 50, 75].

##### 4.1.3 Computing predictions

Having identified a set  $S$  consisting of 28 dimension subsets for each model's embeddings (all dimensions, 3 subsets of dependent dimensions, 24 subsets of important dimensions), we use each subset of dimensions to predict values of  $f$  on the test set.

To do this for each  $s \in S$  we compute the median values of each  $d \in s$  associated with  $f = 0$

and for  $f = 1$  on the train set separately. After this process, we obtain two vectors:  $med_0$  and  $med_1$ . Following that, for each word embedding  $w_{test}$  we select only dimensions  $d_{d \in s}$  and compute Mean Absolute Error between  $w_{test_d}$  and  $med_0$  ( $mae_0$ ), as well as Mean Absolute Error between  $w_{test_d}$  and  $med_1$  ( $mae_1$ ). If  $mae_0 < mae_1$ , the predicted value of the feature vector  $f_{test}$  is 0, and 1 otherwise. Having obtained predictions for all words in the test set, we compute prediction accuracy. Finally, we selected the subset  $s$  with the highest accuracy to be the *best candidate*. Prediction accuracy for best candidates in each fold is averaged among 5 folds, and the final metric is considered to be *InfEnc*.

## 4.2 Stable dimensions

The process of 5-fold cross-validation additionally allowed us to validate if a certain dimension appears in the best candidate subset for feature  $f$  consistently throughout all 5 folds. If it does, we consider such dimension to be a *stable dimension* for the feature  $f$ .

## 5 Results

Following the protocol above, we calculated InfEnc for all listed models for encoding quality of grammatical gender (for adjectives and nouns), grammatical number (for adjectives and nouns), POS, and semantic supersenses. It is worth noting that for observed experiments, subsets of dimensions appear to achieve higher accuracy in the vast majority of cases.

### 5.1 Grammatical gender

We performed experiments for gender in 3 parts: nouns (N), adjectives (A), and nouns and adjectives (N+A) combined (the results can be found in Table 4).

| Model                     | Cased | N                | A               | N+A              |
|---------------------------|-------|------------------|-----------------|------------------|
| FlauBERT <sub>small</sub> | ✓     | 0.805 [2]        | <b>0.95 [1]</b> | 0.794 [2]        |
| DistilBERT                | ✓     | 0.605 [4]        | 0.682 [3]       | 0.626 [4]        |
| CamemBERT <sub>base</sub> | ✓     | 0.534 [9]        | 0.538 [9-10]    | 0.546 [8]        |
| mBERT <sub>base</sub>     |       | 0.516 [10]       | 0.538 [9-10]    | 0.537 [9-10]     |
| mBERT <sub>base</sub>     | ✓     | 0.552 [8]        | 0.612 [5]       | 0.537 [9-10]     |
| FlauBERT <sub>base</sub>  |       | 0.59 [5]         | 0.587 [8]       | 0.585 [6]        |
| FlauBERT <sub>base</sub>  | ✓     | 0.669 [3]        | 0.653 [4]       | 0.655 [3]        |
| XLM-R <sub>base</sub>     | ✓     | 0.557 [7]        | 0.589 [7]       | 0.565 [7]        |
| FlauBERT <sub>large</sub> | ✓     | <b>0.895 [1]</b> | 0.933 [2]       | <b>0.905 [1]</b> |
| XLM-R <sub>large</sub>    | ✓     | 0.575 [6]        | 0.6 [6]         | 0.6 [5]          |

Table 4: InfEnc results for the grammatical gender feature, the best results are bolded. The rank is added in square brackets.

We can notice that despite being described as

partially trained, only recommended for debugging by the authors<sup>8</sup>, and the smallest out of all models tested, FlauBERT<sub>small</sub> achieves the best score in encoding information about adjectives gender. Similarly, the second smallest model DistilBERT trained on multilingual data performs comparably to FlauBERT<sub>base</sub>, trained only on French data.

As could be expected, generally, multilingual models perform significantly worse in the quality of French gender encoding. Surprisingly, CamemBERT scores in InfEnc are as low as those of multilingual models which could be attributed to its tokenization method, further commented on in Section 6.1. Moreover, one could notice that uncased models appear to perform worse in information encoding than their cased variants (both for mBERT and FlauBERT<sub>base</sub>).

### 5.2 Grammatical number

Similarly to gender, the evaluation of the quality of encoding number information was conducted in 3 parts: nouns, adjectives, and nouns and adjectives combined (see Table 5). It can be seen that on aver-

| Model                     | Cased | N                | A                | N+A              |
|---------------------------|-------|------------------|------------------|------------------|
| FlauBERT <sub>small</sub> | ✓     | 0.951 [2]        | <b>0.957 [2]</b> | 0.943 [2]        |
| DistilBERT                | ✓     | 0.698 [4]        | 0.706 [4]        | 0.692 [4]        |
| CamemBERT <sub>base</sub> | ✓     | 0.518 [10]       | 0.551 [10]       | 0.539 [10]       |
| mBERT <sub>base</sub>     |       | 0.562 [9]        | 0.563 [9]        | 0.569 [9]        |
| mBERT <sub>base</sub>     | ✓     | 0.604 [6]        | 0.57 [7]         | 0.588 [8]        |
| FlauBERT <sub>base</sub>  |       | 0.651 [5]        | 0.597 [6]        | 0.645 [5]        |
| FlauBERT <sub>base</sub>  | ✓     | 0.709 [3]        | 0.736 [3]        | 0.697 [3]        |
| XLM-R <sub>base</sub>     | ✓     | 0.589 [8]        | 0.641 [8]        | 0.598 [7]        |
| FlauBERT <sub>large</sub> | ✓     | <b>0.956 [1]</b> | <b>0.959 [1]</b> | <b>0.953 [1]</b> |
| XLM-R <sub>large</sub>    | ✓     | 0.599 [7]        | 0.647 [5]        | 0.64 [6]         |

Table 5: InfEnc results for the grammatical number feature, the best results are bolded. The rank is added in square brackets.

age multilingual models appear to encode number information better than grammatical gender, one of the explanations for such phenomena could be the fact that the plural form of French nouns is formed similarly to plural forms in other languages, therefore, models could have more exposure to this vocabulary during training.

Similarly, as for gender, we can notice smaller models perform either on par (FlauBERT<sub>small</sub>) or better (DistilBERT performing better than non-distilled mBERT) than their bigger counterparts. Likewise, uncased models show lower InfEnc scores than the cased versions.

<sup>8</sup><https://github.com/getalp/Flaubert>

### 5.3 POS

For POS, the experiments were performed in 3 parts: nouns (N) encoded as 1s and non-nouns as 0s; adjectives (A) encoded as 1s and non-adjectives as 0s; verbs (V) encoded as 1s and non-verbs as 0s. The corresponding results can be found in Table 6. What can be noticed is that for most

| Model                     | Cased | N                | A                | V                |
|---------------------------|-------|------------------|------------------|------------------|
| FlauBERT <sub>small</sub> | ✓     | 0.893 [2]        | <b>0.896 [1]</b> | <b>0.938 [1]</b> |
| DistilBERT                | ✓     | 0.641 [5]        | 0.671 [3]        | 0.659 [5]        |
| CamemBERT <sub>base</sub> | ✓     | 0.548 [9]        | 0.573 [8]        | 0.579 [9]        |
| mBERT <sub>base</sub>     |       | 0.539 [10]       | 0.532 [10]       | 0.543 [10]       |
| mBERT <sub>base</sub>     | ✓     | 0.573 [8]        | 0.563 [9]        | 0.608 [8]        |
| FlauBERT <sub>base</sub>  |       | 0.689 [3]        | 0.639 [4]        | 0.718 [3]        |
| FlauBERT <sub>base</sub>  | ✓     | 0.643 [4]        | 0.611 [6]        | 0.695 [4]        |
| XLM-R <sub>base</sub>     | ✓     | 0.594 [6]        | 0.596 [7]        | 0.615 [7]        |
| FlauBERT <sub>large</sub> | ✓     | <b>0.901 [1]</b> | 0.889 [2]        | 0.937 [2]        |
| XLM-R <sub>large</sub>    | ✓     | 0.586 [7]        | 0.618 [5]        | 0.616 [6]        |

Table 6: InfEnc results for the encoding of POS information, the best results are bolded. The rank is added in square brackets.

models, except DistilBERT, the score for encoding verb information is the highest among all investigated POS. Also, interestingly, the uncased FlauBERT<sub>base</sub> model performs better than the cased one in POS information encoding, contrary to the gender and number information encoding. Similarly to previous results, FlauBERT<sub>small</sub> and DistilBERT show either better or comparable results to bigger models.

### 5.4 Semantic supersenses

| Model                     | Cased | Act              | Person          |
|---------------------------|-------|------------------|-----------------|
| FlauBERT <sub>small</sub> | ✓     | <b>0.809 [1]</b> | 0.868 [2]       |
| DistilBERT                | ✓     | 0.699 [3]        | 0.695 [3]       |
| CamemBERT <sub>base</sub> | ✓     | 0.564 [9]        | 0.581 [8]       |
| mBERT <sub>base</sub>     |       | 0.498 [10]       | 0.5 [10]        |
| mBERT <sub>base</sub>     | ✓     | 0.598 [8]        | 0.651 [6]       |
| FlauBERT <sub>base</sub>  |       | 0.63 [5]         | 0.659 [4]       |
| FlauBERT <sub>base</sub>  | ✓     | 0.666 [4]        | 0.653 [5]       |
| XLM-R <sub>base</sub>     | ✓     | 0.629 [6]        | 0.618 [7]       |
| FlauBERT <sub>large</sub> | ✓     | 0.806 [2]        | <b>0.87 [1]</b> |
| XLM-R <sub>large</sub>    | ✓     | 0.624 [7]        | 0.562 [9]       |

Table 7: InfEnc results for the encoding of supersense information, the best results are bolded. The rank is added in square brackets.

As we can see in Table 7 the results of InfEnc for the semantic features, the mBERT uncased score is the poorest through all experiments run. Even if FlauBERT<sub>large</sub> is still showing comparatively high results in the metric, we can notice a big drop in accuracy from the previous experiment results for the model; this could potentially be a sign of the complex nature of semantic features in comparison to grammatical ones.

### 5.5 Correlation with classification task

To validate if the obtained InfEnc scores are representative of the performance of the embeddings in downstream tasks, for each model, we trained 5 different classifiers: LR, Decision Tree (DT), Random Forest (RF), Naive Bayes (NB), and K-Nearest Neighbors (KNN). For the targets of classification, we used the encoded feature vectors  $f$  (gender, number, POS, semantic information). We calculated the mean accuracy for each classifier across 5 folds and additionally mean accuracy among all classifiers, and computed Pearson correlation between obtained accuracies and the InfEnc scores of the models. As can be seen in Figure 1, the obtained

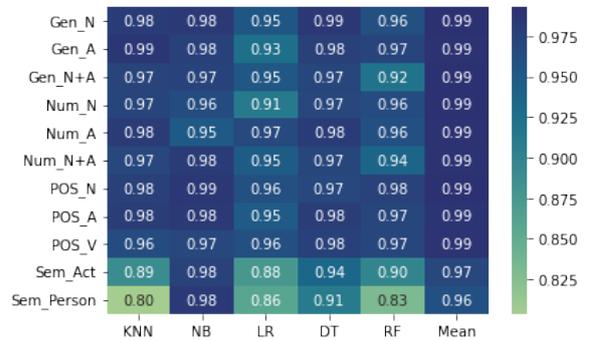


Figure 1: Correlation between accuracies of classifiers and InfEnc scores for each model. "Mean" stands for mean accuracy between all classifiers.

correlation is very high. It can be argued that it is linked to using LR weights as a way to extract meaningful dimensions, however, it can be noticed that the correlation between LR accuracies and InfEnc scores is among the lowest for multiple tasks. Hence, we can assume that InfEnc scores can be a good predictor of classification performance by different classification models. Moreover, it is worth remarking that the correlation with accuracies for classifying semantic features ("Person" and "Act") appears to be lower than for the grammatical features, which can be explained by a more subjective structure of such features.

### 5.6 Stable dimensions

As can be seen in Table 8, except for FlauBERT family models we could not find stable dimensions for all possible features for other models. For DistilBERT, we managed to obtain stable dimensions for all features except for the ones responsible for encoding the POS of adjectives. However, even finding a single stable dimension can be beneficial:

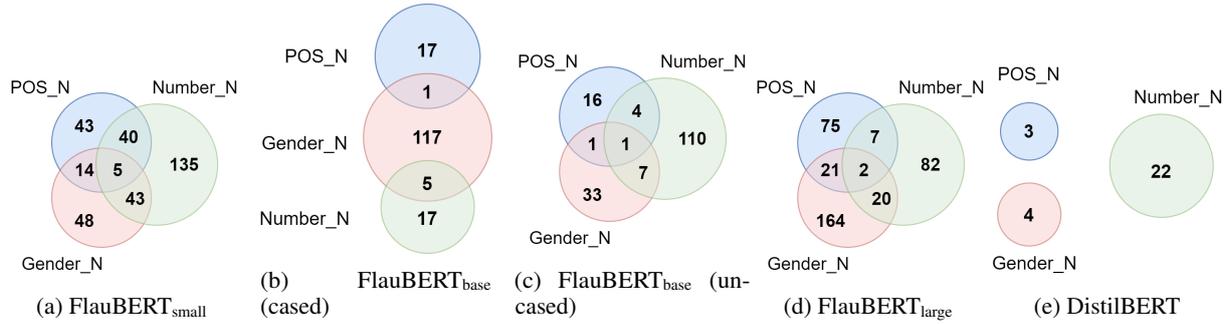


Figure 2: Number of overlapping dimensions for noun gender, number, and POS features. Note that when the three sets overlap (a, c, and d subgraphs), the number of overlapping dimensions given in the intersection of the three sets is also included in the other intersections of the two sets.

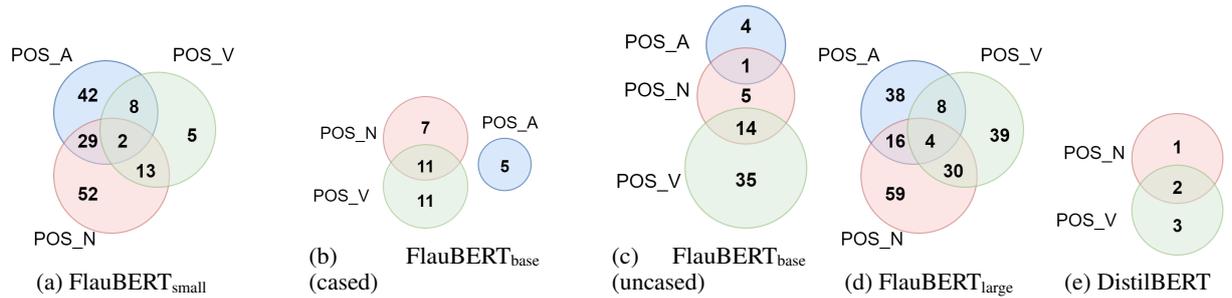


Figure 3: Number of overlapping dimensions for POS features. Note that when the three sets overlap (a and d subgraphs), the number of overlapping dimensions given in the intersection of the three sets is also included in the other intersections of two sets.

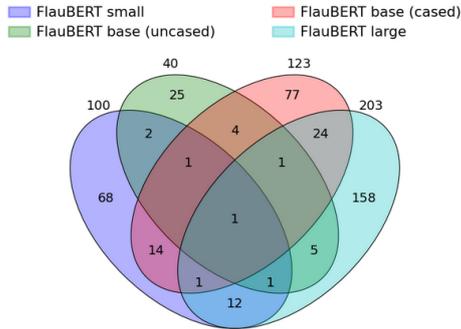


Figure 4: Number of overlapping stable dimensions for noun gender for all FlauBERT models.

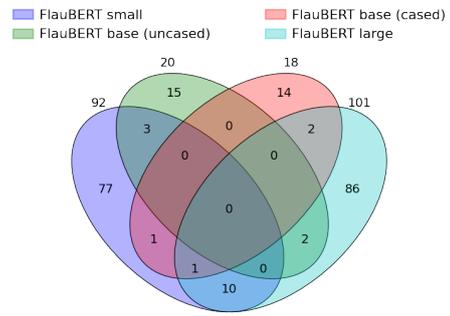


Figure 5: Number of overlapping stable dimensions for noun POS for all FlauBERT models.

for XLM-R<sub>base</sub> we found only one stable dimension for the adjective number ( $d_{467}$ ) that has InfEnc score of 0.628 (which is comparable to the score observed in Table 5).

Regarding grammatical information, we noticed that the patterns of encoding are different from model to model (can be seen in Figure 2): some models have dimensions that appear to encode number, gender, and POS simultaneously (like FlauBERT<sub>small</sub>, for example); others, have no overlap between such dimensions (e.g. DistilBERT). On the other hand, for all FlauBERT models and

DistilBERT we could find dimensions that appear to encode both information about a word being a noun and a word being a verb (see Figure 3).

Additionally, we could find that some stable dimensions that are shared between FlauBERT<sub>small</sub> and FlauBERT<sub>large</sub> (see Figures 4 and 5). This fact could be explained by the highest number of stable dimensions that we observed in these models. We discovered that all models of FlauBERT family share dimension  $d_{177}$  in stable dimension corresponding to the gender of a noun and addition-

| Model                     | Cased | Gen_N | Gen_A | Gen_N+A | Num_N | Num_A | Num_N+A | POS_N | POS_A | POS_V | Act | Person |
|---------------------------|-------|-------|-------|---------|-------|-------|---------|-------|-------|-------|-----|--------|
| FlauBERT <sub>small</sub> | ✓     | 100   | 175   | 74      | 213   | 107   | 127     | 92    | 77    | 24    | 48  | 95     |
| DistilBERT                | ✓     | 4     | 18    | 14      | 22    | 0     | 15      | 3     | 0     | 5     | 3   | 1      |
| CamemBERT <sub>base</sub> | ✓     | 0     | 1     | 0       | 0     | 0     | 9       | 1     | 0     | 0     | 1   | 0      |
| mBERT <sub>base</sub>     |       | 0     | 0     | 3       | 0     | 0     | 4       | 1     | 0     | 0     | 0   | 0      |
| mBERT <sub>base</sub>     | ✓     | 0     | 5     | 0       | 0     | 0     | 0       | 1     | 0     | 0     | 0   | 6      |
| FlauBERT <sub>base</sub>  |       | 40    | 26    | 47      | 120   | 4     | 91      | 20    | 5     | 49    | 2   | 5      |
| FlauBERT <sub>base</sub>  | ✓     | 123   | 22    | 68      | 22    | 4     | 7       | 18    | 5     | 22    | 14  | 1      |
| XLM-R <sub>base</sub>     | ✓     | 0     | 0     | 0       | 2     | 1     | 0       | 1     | 1     | 0     | 1   | 1      |
| FlauBERT <sub>large</sub> | ✓     | 203   | 134   | 227     | 107   | 151   | 195     | 101   | 58    | 73    | 50  | 63     |
| XLM-R <sub>large</sub>    | ✓     | 0     | 0     | 1       | 0     | 0     | 1       | 0     | 1     | 1     | 0   | 0      |

Table 8: The number of stable dimensions for each feature.

ally all cased FlauBERT models have  $d_{55}$  among dimensions corresponding to POS of nouns<sup>9</sup>.

## 6 Discussion

We believe that identifying dimensions with grammatical or semantic features can be used both during training models, combined with reinforcement training, ensuring a better quality of encoding of information, and as a way to find the best-suited model for the downstream task, as well as evaluating the effect of fine-tuning on the encoding of the target features. Due to time and resource limitations, our experiments were limited in terms of studied features and language which we hope to address in further work. We believe that the research could benefit from incorporating other semantic and grammatical features. However, additionally, we propose the following topics for discussion.

### 6.1 Differences in tokenization

Investigated models use different tokenization algorithms: BERT and DistilBERT use WordPiece embeddings (Wu et al., 2016), XLM-R and CamemBERT incorporate SentencePiece tokenization (Kudo and Richardson, 2018) and FlauBERT utilizes Byte-Pair Encoding (BPE) algorithm (Sennrich et al., 2016). As can be seen from Tables 1 and 2, the tokenization quite significantly affects the size of the extracted vocabulary. Precisely, CamemBERT is much less likely to tokenize a noun, adjective, or verb as one token than FlauBERT. For example, the tokenization *brunes* ('brown (plural)') → [brune, s] or *gaieté* ('cheerfulness, joy') → [ga, ie, té] signifies that despite the relative commonality of words, CamemBERT gives preference to subword tokenization. This makes the extraction of important dimensions a more complex task for models like this, therefore, additional research needs to be performed where

<sup>9</sup>Find the list of all retrieved stable dimensions in Appendix A and the corresponding accuracy of classifiers in Appendix B.

the effect of tokenization on the information encoding is studied.

It is also worth mentioning that in our work we compared the models' performance on vocabularies specific to the model, however, it is worth investigating how the models would perform on a certain basic vocabulary that all models share.

### 6.2 Changes in representation

Additionally, as was shown in our experiments, smaller models can learn word embeddings that encode target features as effectively as larger models, therefore, saving energy and computation capacity. Investigating dimensions alongside training a model can give a greater insight into how the target features are learned. The researchers have previously found that the F1 score on downstream tasks appears to plateau after a certain number of steps (Müller-Eberstein et al., 2023), which can be expanded in studying how a target feature encoding changes with the number of steps.

## 7 Conclusion

In this paper, we proposed an intrinsic metric In-fEnc and a framework allowing the extraction of stable dimensions that potentially encode grammatical or semantic information from word embeddings of BERT-like models trained on the French language. Our findings include:

1. For all tested features, subsets of dimensions appear to encode the information better than all dimensions of word embedding.
2. Smaller size models can encode the target information on par or better than larger models.
3. Gender information appears to be better encoded in cased models, than in their uncased counterparts.
4. Tokenization affects greatly the encoding of information in word embeddings.

5. There exist the same dimensions that appear to encode target information in different sizes of models (in our case, FlauBERT).
6. There are signs that noun-ness and verb-ness can be encoded in the same dimensions.
7. For multilingual models, one or a small subset of dimensions can encode information better than a large subset of dimensions.

We believe that understanding what information is encoded in each dimension can be beneficial for a multitude of applications: identifying dimensions encoding gender information can potentially help to mitigate gender bias; knowing what dimensions encode the information related to the downstream task can lead to reducing dimensionality and therefore computational cost; contrastive and reinforcement learning in combination with interpretable embedding dimensions can be advantageous for reduction of hallucination of LLMs.

## Acknowledgments

The topic of this work was initiated by the work of 3 students: Clémentine Bleuze, Ekaterina Goliakova, and Chun Yang.

## References

- Amine Abdaoui, Camille Pradel, and Grégoire Sigel. 2020. [Load what you need: Smaller versions of multilingual BERT](#). In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, pages 119–123, Online. Association for Computational Linguistics.
- Guillaume Alain and Yoshua Bengio. 2016. [Understanding intermediate layers using linear classifier probes](#). In *ArXiv*.
- Lucie Barque, Pauline Haas, Richard Huyghe, Delphine Tribout, Marie Candito, Benoit Crabbé, and Vincent Segonne. 2020. [FrSemCor: Annotating a French corpus with supersenses](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5912–5918, Marseille, France. European Language Resources Association.
- Yonatan Belinkov. 2022. [Probing classifiers: Promises, shortcomings, and advances](#). In *Computational Linguistics*, volume 48, pages 207–219, Cambridge, MA. MIT Press.
- Gabriella Chronis and Katrin Erk. 2020. [When is a bishop not like a rook? when it’s like a rabbi! multi-prototype BERT embeddings for estimating semantic relationships](#). In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 227–244, Online. Association for Computational Linguistics.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- 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.
- Kawin Ethayarajh. 2019. [How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.
- Tadayoshi Fushiki. 2011. [Estimation of prediction error by using K-fold cross-validation](#). In *Statistics and Computing*, volume 21, pages 137–146. Springer.
- Qinjin Jia, Jialin Cui, Yunkai Xiao, Chengyuan Liu, Parvez Rashid, and Edward F Gehring. 2021. [All-in-one: Multi-task learning bert models for evaluating peer assessments](#).
- John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. 2021. [Highly accurate protein structure prediction with alphafold](#). In *Nature*, volume 596, pages 583–589. Nature Publishing Group.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. [Scaling laws for neural language models](#).
- Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. 2004. [Estimating mutual information](#). In *Physical review E*, volume 69, pages 69–85. APS.
- Taku Kudo and John Richardson. 2018. [SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

- Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Al-lauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2020. [Flaubert: Unsupervised language model pre-training for french](#). In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.
- Yang-Yin Lee, Hao Ke, Ting-Yu Yen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020. [Combining and learning word embedding with wordnet for semantic relatedness and similarity measurement](#). In *Journal of the association for information science and technology*, volume 71, pages 657–670. Wiley Online Library.
- Xian Li, Paul Michel, Antonios Anastasopoulos, Yonatan Belinkov, Nadir Durrani, Orhan Firat, Philipp Koehn, Graham Neubig, Juan Pino, and Hassan Sajjad. 2019. [Findings of the first shared task on machine translation robustness](#). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 91–102, Florence, Italy. Association for Computational Linguistics.
- 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](#).
- Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. [CamemBERT: a tasty French language model](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.
- Alessio Miaschi, Dominique Brunato, Felice Dell’Orletta, and Giulia Venturi. 2020. [Linguistic profiling of a neural language model](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 745–756, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Max Müller-Eberstein, Rob van der Goot, Barbara Plank, and Ivan Titov. 2023. [Subspace chronicles: How linguistic information emerges, shifts and interacts during language model training](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13190–13208, Singapore. Association for Computational Linguistics.
- Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. [Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures](#). *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019*, Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.
- Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2021. [Probing the probing paradigm: Does probing accuracy entail task relevance?](#) In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3363–3377, Online. Association for Computational Linguistics.
- Laurent Romary, Susanne Salmon-Alt, and Gil Francopoulo. 2004. [Standards going concrete: from LMF to Morphalou](#). In *Proceedings of the Workshop on Enhancing and Using Electronic Dictionaries*, pages 22–28, Geneva, Switzerland. COLING.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. [High-resolution image synthesis with latent diffusion models](#). In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.
- David Samuel, Andrey Kutuzov, Lilja Øvrelid, and Erik Velldal. 2023. [Trained on 100 million words and still in shape: BERT meets British National Corpus](#). In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1954–1974, Dubrovnik, Croatia. Association for Computational Linguistics.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. [DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter](#).
- Marten van Schijndel, Aaron Mueller, and Tal Linzen. 2019. [Quantity doesn’t buy quality syntax with neural language models](#). 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 5831–5837, Hong Kong, China. Association for Computational Linguistics.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Neural machine translation of rare words with subword units](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.
- Lauro Snidaro, Giovanni Ferrin, and Gian Luca Foresti. 2019. [Distributional memory explainable word embeddings in continuous space](#). In *2019 22th International Conference on Information Fusion (FUSION)*, pages 1–7.
- Erico Tjoa and Cuntai Guan. 2020. [A survey on explainable artificial intelligence \(XAI\): Toward medical XAI](#). In *IEEE transactions on neural networks and learning systems*, volume 32, pages 4793–4813. IEEE.
- Lucas Torroba Hennigen, Adina Williams, and Ryan Cotterell. 2020. [Intrinsic probing through dimension selection](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*

(EMNLP), pages 197–216, Online. Association for Computational Linguistics.

Pablo Villalobos, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn, and Anson Ho. 2022. [Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning](#).

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. [CCNet: Extracting high quality monolingual datasets from web crawl data](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. [Google’s neural machine translation system: Bridging the gap between human and machine translation](#).

## Appendix A Stable dimensions

In this section, you will find the retrieved stable dimensions for tested linguistic features for each model.

### A.1 FlauBERT<sub>small</sub>

| Feature                  | Dimension list   |
|--------------------------|--|
| Gender: Noun             | 0, 5, 7, 20, 25, 28, 30, 36, 40, 42, 50, 54, 62, 74, 85, 88, 95, 96, 100, 113, 115, 117, 121, 123, 124, 130, 133, 141, 142, 147, 152, 156, 160, 162, 173, 177, 181, 186, 192, 193, 195, 198, 200, 202, 210, 213, 214, 220, 234, 237, 239, 245, 250, 255, 256, 261, 265, 269, 276, 279, 292, 293, 296, 306, 310, 312, 315, 316, 318, 320, 332, 335, 352, 362, 363, 374, 376, 377, 387, 390, 403, 417, 426, 432, 434, 436, 439, 443, 455, 466, 468, 470, 477, 488, 490, 495, 497, 499, 501, 507  |
| Gender: Adjective        | 0, 4, 5, 7, 8, 11, 12, 14, 26, 28, 30, 33, 36, 38, 39, 42, 45, 46, 50, 53, 54, 55, 56, 57, 59, 65, 67, 70, 72, 74, 75, 82, 83, 85, 88, 89, 94, 95, 96, 99, 114, 117, 121, 122, 123, 124, 133, 142, 144, 145, 147, 157, 160, 162, 167, 170, 175, 177, 178, 181, 184, 185, 187, 189, 192, 193, 195, 200, 202, 203, 206, 211, 213, 214, 222, 228, 230, 233, 234, 237, 245, 249, 250, 251, 255, 256, 260, 268, 274, 275, 276, 283, 284, 286, 287, 289, 290, 292, 293, 296, 300, 302, 304, 306, 309, 310, 313, 314, 316, 318, 320, 321, 331, 332, 333, 336, 339, 340, 341, 345, 348, 352, 353, 354, 357, 360, 362, 363, 364, 365, 366, 370, 372, 374, 376, 377, 379, 380, 385, 387, 389, 390, 396, 398, 399, 401, 409, 425, 426, 429, 430, 432, 436, 439, 443, 449, 450, 453, 461, 465, 466, 470, 471, 477, 478, 482, 486, 488, 489, 490, 500, 501, 503, 506, 507   |
| Gender: Noun & Adjective | 0, 7, 11, 25, 28, 36, 40, 42, 55, 62, 74, 88, 95, 100, 115, 117, 121, 124, 130, 144, 147, 149, 159, 160, 162, 175, 177, 181, 186, 192, 195, 198, 202, 210, 211, 237, 239, 245, 250, 256, 261, 265, 269, 270, 279, 292, 296, 300, 306, 309, 310, 315, 316, 318, 320, 332, 335, 363, 377, 387, 390, 403, 409, 432, 434, 443, 448, 455, 468, 470, 488, 490, 499, 507  |
| Number: Noun             | 1, 3, 5, 6, 7, 8, 9, 10, 13, 15, 19, 21, 24, 25, 26, 27, 31, 37, 42, 43, 48, 50, 51, 52, 53, 54, 55, 56, 57, 66, 67, 69, 70, 73, 74, 76, 77, 80, 81, 83, 84, 86, 88, 89, 92, 96, 97, 99, 101, 107, 110, 111, 112, 114, 115, 117, 118, 119, 128, 129, 130, 131, 132, 133, 137, 138, 142, 148, 149, 150, 151, 154, 155, 156, 158, 160, 161, 164, 165, 167, 170, 171, 172, 175, 177, 181, 182, 183, 185, 187, 192, 198, 200, 205, 207, 208, 209, 210, 212, 213, 214, 220, 223, 224, 229, 238, 243, 244, 246, 250, 251, 252, 254, 255, 257, 259, 268, 269, 270, 273, 277, 278, 281, 282, 285, 286, 288, 289, 291, 295, 296, 297, 299, 302, 303, 305, 306, 307, 308, 309, 310, 311, 313, 317, 322, 326, 328, 330, 336, 337, 338, 342, 343, 351, 352, 353, 355, 356, 359, 360, 372, 373, 374, 376, 378, 381, 382, 384, 391, 399, 403, 404, 405, 413, 414, 416, 419, 420, 422, 423, 430, 431, 432, 433, 435, 438, 441, 442, 445, 451, 452, 453, 454, 455, 457, 458, 461, 462, 470, 473, 475, 477, 481, 485, 488, 489, 490, 492, 495, 506, 507, 510, 511 |
| Number: Adjective        | 1, 3, 9, 15, 21, 24, 25, 50, 54, 56, 67, 69, 73, 74, 81, 83, 84, 96, 109, 112, 115, 125, 129, 131, 138, 149, 154, 156, 158, 159, 161, 165, 167, 172, 175, 181, 182, 183, 185, 191, 192, 198, 200, 205, 208, 210, 220, 224, 238, 250, 251, 252, 254, 257, 278, 285, 288, 289, 297, 303, 306, 310, 311, 313, 317, 337, 340, 342, 347, 351, 352, 353, 356, 359, 360, 372, 374, 378, 381, 384, 399, 403, 405, 410, 419, 420, 430, 445, 453, 454, 455, 458, 461, 462, 474, 475, 477, 479, 481, 483, 484, 485, 490, 495, 497, 499, 501   |
| Number: Noun & Adjective | 1, 3, 7, 8, 9, 10, 15, 21, 37, 42, 43, 50, 51, 52, 54, 55, 57, 62, 67, 69, 73, 74, 81, 83, 92, 96, 101, 110, 111, 115, 118, 119, 129, 131, 132, 138, 148, 149, 150, 154, 155, 156, 158, 160, 161, 165, 167, 172, 175, 177, 182, 183, 185, 192, 198, 200, 205, 208, 209, 210, 212, 213, 214, 220, 223, 246, 250, 251, 252, 254, 255, 257, 277, 278, 285, 288, 289, 295, 296, 297, 302, 303, 306, 308, 310, 311, 317, 322, 328, 330, 337, 338, 342, 351, 352, 355, 356, 359, 360, 372, 373, 374, 376, 378, 384, 399, 405, 419, 420, 422, 430, 432, 438, 445, 451, 454, 455, 461, 473, 475, 477, 485, 489, 495, 501, 506, 507   |
| POS: Noun                | 11, 15, 29, 31, 33, 36, 37, 38, 41, 51, 53, 55, 57, 63, 65, 72, 73, 78, 82, 83, 92, 102, 103, 106, 112, 117, 122, 128, 130, 134, 141, 154, 155, 159, 164, 166, 168, 176, 178, 183, 192, 197, 198, 205, 207, 212, 224, 229, 233, 260, 275, 278, 281, 285, 286, 300, 303, 305, 314, 320, 339, 341, 346, 362, 378, 382, 387, 393, 395, 401, 404, 405, 409, 423, 424, 426, 434, 450, 452, 455, 458, 460, 462, 464, 465, 466, 468, 478, 480, 504, 508, 509  |
| POS: Adjective           | 4, 9, 15, 31, 37, 39, 41, 44, 53, 56, 72, 73, 82, 84, 87, 92, 100, 111, 117, 119, 130, 134, 138, 139, 149, 153, 158, 159, 176, 199, 205, 213, 222, 239, 246, 250, 260, 265, 275, 276, 284, 288, 297, 299, 301, 309, 310, 314, 315, 330, 332, 337, 339, 346, 348, 350, 387, 405, 417, 428, 429, 432, 435, 439, 445, 450, 451, 455, 461, 462, 464, 465, 467, 468, 478, 485, 499  |
| POS: Verb                | 0, 36, 56, 89, 103, 139, 158, 159, 192, 198, 233, 282, 297, 303, 310, 318, 341, 378, 410, 432, 462, 480, 504, 508  |
| Semantic: Act            | 25, 26, 33, 39, 41, 102, 113, 114, 119, 128, 146, 166, 170, 172, 173, 177, 182, 187, 201, 202, 218, 227, 245, 255, 273, 274, 281, 284, 294, 313, 339, 357, 366, 380, 381, 399, 400, 406, 417, 423, 455, 461, 466, 468, 485, 489, 499, 500  |
| Semantic: Person         | 0, 5, 6, 9, 16, 20, 24, 30, 32, 33, 39, 41, 57, 58, 59, 70, 75, 78, 84, 86, 90, 94, 101, 112, 113, 121, 139, 140, 142, 144, 146, 153, 160, 162, 174, 176, 177, 184, 197, 204, 207, 220, 223, 225, 228, 235, 243, 263, 268, 273, 275, 276, 283, 287, 295, 296, 303, 304, 316, 329, 337, 339, 355, 357, 358, 359, 364, 369, 375, 377, 387, 388, 389, 396, 401, 404, 405, 407, 417, 421, 423, 447, 448, 450, 456, 458, 472, 475, 478, 483, 489, 494, 495, 499, 505  |

### A.2 DistilBERT

| Feature                  | Dimension list  |
|--------------------------|---|
| Gender: Noun             | 302, 653, 713, 727  |
| Gender: Adjective        | 9, 37, 51, 76, 185, 278, 301, 425, 526, 531, 551, 633, 641, 676, 716, 737, 747, 749                         |
| Gender: Noun & Adjective | 70, 116, 161, 262, 301, 302, 353, 488, 499, 563, 656, 727, 728, 761   |
| Number: Noun             | 81, 106, 148, 169, 177, 184, 186, 275, 333, 343, 354, 367, 376, 392, 403, 441, 498, 522, 535, 573, 663, 710 |
| Number: Adjective        | -   |
| Number: Noun & Adjective | 152, 184, 186, 216, 354, 363, 413, 440, 518, 535, 584, 649, 668, 745, 761                                   |
| POS: Noun                | 52, 501, 700  |
| POS: Adjective           | -   |
| POS: Verb                | 52, 189, 327, 700, 740  |
| Semantic: Act            | 398, 446, 621   |
| Semantic: Person         | 346   |

### A.3 CamemBERT

| Feature                  | Dimension list                             |
|--------------------------|--|
| Gender: Noun             | -  |
| Gender: Adjective        | 215  |
| Gender: Noun & Adjective | -  |
| Number: Noun             | -  |
| Number: Adjective        | -  |
| Number: Noun & Adjective | 24, 138, 176, 213, 303, 386, 482, 493, 562 |
| POS: Noun                | 696  |
| POS: Adjective           | -  |
| POS: Verb                | -  |
| Semantic: Act            | 343  |
| Semantic: Person         | -  |

### A.4 mBERT (uncased)

| Feature                  | Dimension list    |
|--------------------------|-------------------|
| Gender: Noun             | -                 |
| Gender: Adjective        | -                 |
| Gender: Noun & Adjective | 274, 445, 447     |
| Number: Noun             | -                 |
| Number: Adjective        | -                 |
| Number: Noun & Adjective | 74, 194, 556, 654 |
| POS: Noun                | 74                |
| POS: Adjective           | -                 |
| POS: Verb                | -                 |
| Semantic: Act            | -                 |
| Semantic: Person         | -                 |

### A.5 mBERT (cased)

| Feature                  | Dimension list               |
|--------------------------|------------------------------|
| Gender: Noun             | -                            |
| Gender: Adjective        | 9, 223, 519, 540, 575        |
| Gender: Noun & Adjective | -                            |
| Number: Noun             | -                            |
| Number: Adjective        | -                            |
| Number: Noun & Adjective | -                            |
| POS: Noun                | 529                          |
| POS: Adjective           | -                            |
| POS: Verb                | -                            |
| Semantic: Act            | -                            |
| Semantic: Person         | 143, 319, 447, 504, 585, 715 |

### A.6 FlauBERT<sub>base</sub> (uncased)

| Feature                  | Dimension list  |
|--------------------------|---|
| Gender: Noun             | 2, 17, 21, 47, 81, 98, 130, 132, 138, 149, 177, 180, 185, 197, 198, 244, 299, 307, 309, 310, 314, 382, 433, 456, 505, 507, 508, 546, 572, 596, 597, 604, 644, 662, 671, 696, 698, 735, 755, 757   |
| Gender: Adjective        | 17, 32, 45, 55, 57, 136, 149, 169, 203, 215, 277, 314, 315, 374, 382, 396, 508, 516, 546, 574, 585, 607, 609, 671, 680, 698   |
| Gender: Noun & Adjective | 0, 21, 24, 32, 81, 119, 120, 138, 156, 171, 177, 185, 186, 187, 196, 197, 198, 212, 223, 244, 263, 310, 314, 322, 333, 341, 382, 432, 435, 456, 480, 507, 508, 519, 545, 546, 574, 585, 587, 594, 604, 610, 667, 680, 696, 719, 766   |
| Number: Noun             | 0, 1, 5, 11, 17, 19, 27, 28, 29, 31, 40, 49, 50, 56, 58, 66, 79, 81, 85, 96, 98, 113, 127, 133, 148, 150, 182, 186, 191, 192, 193, 197, 202, 207, 210, 221, 234, 242, 250, 252, 255, 274, 294, 296, 299, 301, 306, 309, 322, 327, 329, 345, 353, 354, 365, 367, 374, 381, 384, 396, 402, 412, 414, 415, 425, 431, 435, 443, 451, 465, 470, 476, 477, 480, 487, 490, 496, 499, 512, 519, 528, 532, 535, 539, 543, 573, 577, 580, 598, 600, 604, 614, 615, 617, 623, 626, 632, 635, 638, 647, 650, 652, 673, 678, 687, 699, 706, 715, 719, 730, 736, 737, 738, 743, 744, 745, 751, 752, 757, 762, 763 |
| Number: Adjective        | 11, 687, 730, 737   |
| Number: Noun & Adjective | 0, 11, 19, 27, 29, 40, 53, 56, 58, 77, 79, 80, 85, 97, 127, 133, 167, 181, 186, 191, 192, 197, 202, 221, 234, 242, 250, 252, 255, 274, 288, 294, 299, 301, 306, 309, 319, 329, 353, 356, 365, 366, 372, 384, 391, 402, 412, 415, 419, 421, 431, 435, 438, 465, 474, 476, 477, 486, 487, 496, 499, 510, 512, 535, 542, 583, 590, 600, 604, 609, 614, 615, 617, 623, 626, 633, 650, 687, 706, 730, 733, 736, 737, 738, 743, 744, 751, 752, 757, 762, 763  |
| POS: Noun                | 32, 95, 107, 108, 112, 133, 170, 186, 224, 238, 299, 383, 390, 405, 406, 435, 545, 585, 649, 672  |
| POS: Adjective           | 32, 320, 449, 680, 746  |
| POS: Verb                | 11, 54, 56, 58, 89, 95, 96, 107, 108, 112, 119, 132, 169, 170, 179, 188, 216, 217, 224, 250, 276, 292, 299, 307, 335, 336, 357, 383, 390, 405, 406, 424, 450, 498, 505, 530, 553, 558, 585, 612, 614, 615, 649, 672, 698, 717, 720, 744, 748  |
| Semantic: Act            | 48, 752   |
| Semantic: Person         | 52, 303, 360, 399, 687  |

**A.7 FlauBERT<sub>base</sub> (cased)**

| Feature                  | Dimension list   |
|--------------------------|--|
| Gender: Noun             | 7, 10, 17, 22, 26, 28, 33, 41, 43, 46, 51, 54, 62, 70, 71, 80, 83, 85, 110, 117, 118, 130, 136, 141, 155, 157, 158, 160, 161, 170, 172, 175, 177, 179, 187, 189, 190, 196, 209, 212, 216, 223, 224, 246, 261, 271, 272, 274, 277, 283, 289, 291, 296, 302, 319, 325, 331, 334, 337, 340, 355, 356, 357, 359, 363, 372, 386, 387, 388, 395, 398, 399, 401, 406, 425, 426, 429, 434, 446, 459, 461, 466, 484, 491, 493, 494, 500, 513, 517, 526, 529, 537, 538, 544, 554, 563, 573, 574, 577, 580, 587, 589, 597, 611, 621, 623, 625, 638, 656, 663, 664, 677, 687, 693, 696, 698, 701, 702, 743, 748, 749, 752, 755 |
| Gender: Adjective        | 71, 85, 118, 130, 155, 170, 189, 225, 309, 331, 337, 340, 359, 398, 425, 484, 491, 611, 621, 661, 663, 698   |
| Gender: Noun & Adjective | 30, 33, 43, 46, 62, 68, 70, 83, 85, 106, 117, 130, 155, 157, 170, 172, 177, 187, 189, 209, 216, 223, 224, 272, 274, 277, 283, 296, 319, 331, 337, 349, 357, 359, 363, 372, 395, 398, 406, 425, 429, 435, 484, 493, 494, 500, 506, 512, 517, 526, 537, 544, 570, 571, 573, 580, 587, 589, 597, 600, 611, 655, 663, 698, 725, 746, 749, 752  |
| Number: Noun             | 34, 53, 125, 127, 176, 180, 196, 205, 238, 268, 279, 290, 343, 398, 449, 466, 500, 571, 594, 623, 672, 760   |
| Number: Adjective        | 163, 176, 238, 253   |
| Number: Noun & Adjective | 34, 176, 180, 238, 325, 466, 594   |
| POS: Noun                | 55, 81, 87, 106, 119, 162, 243, 248, 345, 413, 417, 454, 564, 569, 602, 688, 698, 720  |
| POS: Adjective           | 163, 331, 485, 571, 764  |
| POS: Verb                | 0, 55, 80, 81, 87, 90, 106, 136, 162, 240, 243, 248, 324, 412, 417, 443, 454, 569, 687, 688, 705, 752  |
| Semantic: Act            | 1, 11, 49, 62, 87, 333, 397, 417, 470, 561, 601, 725, 729, 765   |
| Semantic: Person         | 611  |

**A.8 XLM-R<sub>base</sub>**

| Feature                  | Dimension list |
|--------------------------|----------------|
| Gender: Noun             | -              |
| Gender: Adjective        | -              |
| Gender: Noun & Adjective | -              |
| Number: Noun             | 440, 484       |
| Number: Adjective        | 467            |
| Number: Noun & Adjective | -              |
| POS: Noun                | 593            |
| POS: Adjective           | 100            |
| POS: Verb                | -              |
| Semantic: Act            | 690            |
| Semantic: Person         | 741            |

### A.9 FlauBERT<sub>large</sub>

| Feature                  | Dimension list   |
|--------------------------|--|
| Gender: Noun             | 2, 12, 13, 15, 24, 33, 39, 40, 56, 61, 64, 66, 67, 68, 75, 80, 83, 88, 95, 99, 108, 118, 123, 131, 136, 141, 148, 153, 165, 166, 169, 171, 172, 175, 177, 180, 182, 193, 194, 197, 199, 203, 205, 207, 211, 215, 224, 227, 228, 235, 241, 242, 243, 251, 255, 256, 259, 268, 272, 273, 274, 277, 284, 286, 290, 291, 292, 299, 311, 319, 320, 324, 335, 341, 342, 343, 346, 351, 355, 357, 359, 367, 368, 369, 378, 383, 390, 399, 401, 407, 412, 430, 444, 451, 454, 464, 465, 467, 480, 481, 493, 495, 506, 507, 511, 518, 524, 532, 539, 543, 548, 552, 575, 577, 584, 617, 621, 633, 635, 639, 647, 650, 662, 680, 683, 690, 698, 701, 707, 708, 714, 715, 721, 725, 734, 736, 740, 743, 749, 759, 760, 766, 771, 774, 775, 779, 781, 788, 793, 795, 796, 800, 802, 803, 806, 813, 816, 819, 823, 839, 841, 843, 844, 855, 856, 862, 868, 871, 882, 886, 888, 890, 898, 899, 902, 907, 909, 916, 917, 922, 927, 929, 940, 947, 952, 953, 960, 970, 972, 974, 976, 979, 983, 994, 996, 999, 1005, 1007, 1010, 1016, 1018, 1022, 1023  |
| Gender: Adjective        | 13, 21, 35, 44, 46, 55, 68, 72, 73, 84, 118, 119, 129, 131, 141, 143, 148, 149, 153, 160, 165, 169, 175, 180, 197, 205, 220, 227, 241, 248, 251, 255, 256, 258, 262, 283, 290, 310, 311, 313, 325, 333, 342, 351, 359, 380, 393, 398, 419, 427, 440, 442, 451, 472, 474, 478, 480, 494, 495, 497, 500, 503, 506, 508, 531, 552, 585, 586, 591, 600, 601, 610, 616, 621, 633, 647, 649, 655, 667, 680, 702, 707, 714, 717, 718, 725, 730, 737, 749, 759, 760, 768, 774, 785, 787, 795, 796, 800, 806, 807, 811, 814, 817, 818, 833, 834, 841, 844, 848, 849, 858, 860, 862, 880, 882, 886, 888, 897, 899, 900, 907, 917, 920, 929, 931, 934, 936, 943, 958, 972, 996, 1014, 1017, 1022  |
| Gender: Noun & Adjective | 3, 12, 13, 15, 24, 33, 35, 39, 40, 44, 55, 56, 61, 64, 66, 72, 75, 80, 83, 88, 95, 99, 106, 108, 111, 118, 120, 131, 136, 137, 148, 153, 165, 169, 172, 175, 178, 180, 182, 184, 193, 194, 197, 199, 204, 205, 207, 215, 224, 225, 227, 228, 232, 235, 241, 242, 243, 251, 255, 256, 257, 259, 262, 268, 272, 274, 277, 284, 287, 290, 291, 292, 294, 299, 311, 318, 320, 324, 336, 341, 342, 343, 347, 351, 354, 355, 357, 364, 367, 368, 382, 383, 390, 399, 401, 415, 421, 428, 430, 444, 445, 454, 455, 459, 464, 465, 467, 472, 478, 480, 493, 496, 502, 511, 518, 539, 548, 552, 566, 567, 568, 570, 575, 577, 584, 585, 586, 591, 597, 599, 600, 615, 617, 621, 624, 632, 635, 639, 647, 649, 650, 655, 662, 665, 683, 685, 698, 701, 707, 708, 714, 715, 721, 725, 736, 743, 749, 754, 759, 760, 766, 767, 771, 774, 775, 779, 781, 788, 793, 795, 800, 802, 806, 809, 813, 814, 816, 823, 839, 841, 843, 844, 848, 856, 862, 868, 882, 886, 890, 892, 899, 902, 907, 916, 919, 922, 926, 927, 929, 939, 940, 942, 947, 954, 960, 966, 970, 972, 976, 979, 983, 985, 986, 990, 993, 994, 996, 998, 999, 1002, 1005, 1007, 1010, 1014, 1016, 1022, 1023 |
| Number: Noun             | 7, 28, 29, 34, 55, 59, 93, 103, 121, 123, 136, 138, 139, 147, 150, 171, 184, 185, 191, 194, 209, 223, 234, 246, 250, 259, 262, 278, 281, 305, 315, 330, 334, 352, 357, 358, 370, 373, 387, 389, 398, 404, 421, 435, 436, 469, 476, 480, 485, 486, 488, 491, 497, 505, 508, 517, 532, 545, 546, 554, 556, 561, 565, 576, 583, 587, 602, 606, 634, 638, 641, 653, 660, 678, 680, 690, 691, 707, 721, 724, 729, 730, 774, 775, 783, 787, 814, 851, 874, 877, 898, 911, 921, 927, 928, 930, 937, 939, 947, 967, 978, 991, 994, 996, 1015, 1020, 1022   |
| Number: Adjective        | 15, 24, 26, 28, 34, 35, 55, 59, 62, 72, 80, 86, 87, 93, 94, 103, 113, 123, 127, 133, 137, 147, 148, 150, 161, 170, 171, 174, 179, 184, 190, 191, 193, 194, 198, 206, 209, 223, 226, 237, 250, 252, 259, 262, 266, 268, 272, 296, 298, 306, 315, 330, 336, 339, 352, 356, 360, 370, 386, 389, 392, 398, 403, 404, 414, 417, 436, 451, 452, 464, 469, 480, 485, 486, 489, 490, 491, 494, 497, 505, 508, 516, 517, 538, 539, 556, 560, 565, 570, 576, 583, 591, 606, 609, 613, 629, 634, 636, 638, 641, 652, 681, 682, 691, 707, 709, 716, 717, 719, 726, 730, 731, 748, 761, 774, 775, 783, 787, 796, 802, 805, 806, 814, 821, 830, 851, 863, 865, 871, 877, 880, 895, 911, 920, 921, 927, 928, 939, 943, 946, 956, 963, 978, 993, 996, 1006, 1010, 1015, 1020, 1021, 1022   |
| Number: Noun & Adjective | 7, 9, 10, 20, 28, 29, 34, 39, 47, 55, 56, 59, 74, 85, 86, 87, 93, 96, 98, 103, 110, 111, 119, 121, 123, 133, 136, 137, 139, 140, 147, 150, 153, 161, 179, 184, 185, 191, 193, 194, 198, 203, 206, 209, 219, 223, 234, 238, 243, 246, 250, 259, 262, 266, 278, 293, 294, 296, 301, 305, 306, 315, 320, 330, 345, 352, 357, 365, 370, 373, 386, 387, 389, 397, 398, 404, 416, 417, 419, 421, 422, 428, 435, 438, 439, 451, 459, 469, 476, 477, 480, 485, 486, 489, 497, 505, 508, 514, 516, 521, 539, 545, 546, 554, 556, 565, 567, 572, 575, 576, 583, 584, 585, 587, 591, 602, 606, 608, 609, 634, 636, 638, 641, 644, 648, 649, 651, 655, 672, 680, 683, 690, 691, 703, 707, 709, 712, 716, 721, 724, 726, 729, 730, 731, 732, 748, 772, 774, 775, 783, 787, 793, 802, 808, 814, 818, 830, 842, 851, 858, 861, 873, 874, 877, 881, 883, 895, 901, 911, 920, 921, 925, 927, 928, 930, 937, 939, 947, 955, 956, 966, 967, 978, 988, 990, 991, 994, 996, 1000, 1004, 1015, 1016, 1017, 1020, 1022  |
| POS: Noun                | 2, 4, 21, 31, 38, 44, 54, 55, 65, 78, 88, 92, 98, 107, 137, 151, 171, 172, 176, 205, 206, 207, 220, 232, 252, 253, 267, 271, 292, 299, 308, 317, 325, 362, 366, 375, 380, 394, 408, 413, 415, 418, 420, 430, 437, 458, 461, 479, 481, 482, 485, 494, 495, 497, 506, 517, 524, 525, 526, 549, 562, 566, 569, 571, 597, 599, 631, 656, 671, 680, 685, 686, 689, 699, 722, 765, 790, 793, 800, 809, 848, 864, 867, 878, 888, 893, 907, 914, 925, 932, 945, 956, 961, 966, 968, 974, 978, 988, 993, 1010, 1016   |
| POS: Adjective           | 43, 49, 55, 61, 80, 88, 126, 131, 175, 206, 207, 292, 304, 323, 349, 351, 357, 361, 406, 430, 437, 459, 490, 510, 545, 559, 563, 597, 599, 631, 650, 656, 686, 689, 697, 726, 749, 758, 764, 787, 795, 804, 822, 830, 839, 852, 890, 907, 908, 911, 915, 955, 968, 976, 977, 978, 1005, 1013   |
| POS: Verb                | 2, 22, 31, 44, 88, 98, 99, 106, 172, 182, 207, 208, 232, 233, 249, 252, 253, 266, 267, 271, 274, 278, 299, 333, 362, 374, 375, 404, 406, 416, 424, 427, 436, 437, 439, 441, 473, 479, 481, 490, 506, 529, 539, 580, 664, 667, 683, 689, 724, 748, 765, 775, 793, 800, 803, 809, 848, 853, 864, 886, 895, 904, 909, 912, 915, 925, 927, 932, 974, 976, 988, 991, 1006   |
| Semantic: Act            | 16, 49, 53, 62, 75, 88, 163, 174, 175, 195, 210, 216, 254, 263, 265, 267, 278, 334, 391, 400, 418, 452, 527, 528, 556, 601, 604, 617, 639, 652, 654, 655, 683, 693, 697, 702, 726, 737, 769, 776, 798, 827, 873, 878, 890, 924, 941, 953, 980, 1013  |
| Semantic: Person         | 4, 27, 74, 77, 88, 153, 162, 186, 209, 213, 241, 257, 285, 291, 307, 321, 334, 342, 345, 385, 435, 477, 486, 503, 505, 520, 534, 553, 598, 603, 622, 634, 635, 638, 655, 657, 660, 671, 677, 678, 695, 696, 705, 718, 730, 736, 750, 770, 781, 799, 810, 813, 814, 855, 872, 880, 891, 892, 931, 948, 985, 994, 998  |

### A.10 XLM-R<sub>large</sub>

| Feature                  | Dimension list |
|--------------------------|----------------|
| Gender: Noun             | -              |
| Gender: Adjective        | -              |
| Gender: Noun & Adjective | 122            |
| Number: Noun             | -              |
| Number: Adjective        | -              |
| Number: Noun & Adjective | 50             |
| POS: Noun                | -              |
| POS: Adjective           | 130            |
| POS: Verb                | 42             |
| Semantic: Act            | -              |
| Semantic: Person         | -              |

## Appendix B Classification accuracy

We trained several classifiers to predict the values of feature vectors using all dimensions of word embeddings and the obtained stable dimensions associated with the feature (if any). In this section, you will find the accuracies achieved using stable dimensions only and how they compare to the accuracies of the same classifiers trained using all dimensions of word embeddings.

The used classifiers were K-Nearest Neighbors (KNN), Naive Bayes (NB), Logistic Regression (LR), Random Forest (RF), and Decision Tree (DT), in the implementation of `scikit-learn`. All accuracies were obtained using 5-fold cross-validation.

In many cases despite sharp decrease in the number of used dimensions, the achieved accuracies are comparable to the whole word embedding vector. However, it is worth noting that for the LR classifier (which was used in the initial setup to retrieve best candidates) the observed accuracies are notably lower in a lot of cases.

### B.1 Gender

#### B.1.1 Gender: Noun

| Model                     | Cased | KNN             | NB              | LR              | RF              | DT              | Number of stable dimensions |
|---------------------------|-------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------------------|
| FlauBERT <sub>small</sub> | ✓     | 0.763 (↑ 0.066) | 0.827 (↓ 0.035) | 0.921 (↑ 0.03)  | 0.835 (↑ 0.092) | 0.635 (↑ 0.014) | 100                         |
| DistilBERT                | ✓     | 0.537 (↓ 0.03)  | 0.59 (↑ 0.035)  | 0.59 (↓ 0.116)  | 0.542 (↓ 0.067) | 0.521 (↓ 0.006) | 4                           |
| CamemBERT <sub>base</sub> | ✓     | -               | -               | -               | -               | -               | 0                           |
| mBERT <sub>base</sub>     | -     | -               | -               | -               | -               | -               | 0                           |
| mBERT <sub>base</sub>     | ✓     | -               | -               | -               | -               | -               | 0                           |
| FlauBERT <sub>base</sub>  | -     | 0.561 (↓ 0.001) | 0.6 (↓ 0.0308)  | 0.607 (↑ 0.011) | 0.603 (↑ 0.015) | 0.543 (↑ 0.003) | 123                         |
| FlauBERT <sub>base</sub>  | ✓     | 0.616 (↑ 0.031) | 0.68 (↑ 0.041)  | 0.718 (↓ 0.041) | 0.683 (↑ 0.019) | 0.56 (↑ 0.008)  | 140                         |
| XLM-R <sub>base</sub>     | ✓     | -               | -               | -               | -               | -               | 0                           |
| FlauBERT <sub>large</sub> | ✓     | 0.851 (↑ 0.072) | 0.904 (↑ 0.018) | 0.921 (↓ 0.014) | 0.886 (↑ 0.005) | 0.717 (↑ 0.018) | 203                         |
| XLM-R <sub>large</sub>    | ✓     | -               | -               | -               | -               | -               | 0                           |

Table 9: Accuracies of classifiers trained to predict gender of nouns using only stable dimensions. The number in the brackets signifies absolute difference with the accuracies of the same classifiers trained on all dimensions. ↑ marks an increase in accuracy when using only stable dimensions, ↓ marks the decrease.

#### B.1.2 Gender: Adjective

| Model                     | Cased | KNN             | NB              | LR              | RF              | DT               | Number of stable dimensions |
|---------------------------|-------|-----------------|-----------------|-----------------|-----------------|------------------|-----------------------------|
| FlauBERT <sub>small</sub> | ✓     | 0.718 (↓ 0.023) | 0.791 (↓ 0.144) | 0.911 (↓ 0.076) | 0.804 (↓ 0.126) | 0.622 (↓ 0.0976) | 175                         |
| DistilBERT                | ✓     | 0.549 (↓ 0.063) | 0.597 (↑ 0.01)  | 0.6 (↓ 0.191)   | 0.598 (↓ 0.035) | 0.537 (↓ 0.0168) | 18                          |
| CamemBERT <sub>base</sub> | ✓     | 0.494 (↓ 0.011) | 0.518 (↓ 0.003) | 0.514 (↓ 0.035) | 0.501 (↓ 0.033) | 0.499 (↑ 0.008)  | 1                           |
| mBERT <sub>base</sub>     | -     | -               | -               | -               | -               | -                | 0                           |
| mBERT <sub>base</sub>     | ✓     | 0.543 (↑ 0.01)  | 0.5 (↓ 0.026)   | 0.489 (↑ 0.019) | 0.532 (↓ 0.076) | 0.504 (↓ 0.041)  | 5                           |
| FlauBERT <sub>base</sub>  | -     | 0.529 (↓ 0.017) | 0.571 (↓ 0.033) | 0.582 (↓ 0.076) | 0.575 (↓ 0.016) | 0.528 (↑ 0.003)  | 26                          |
| FlauBERT <sub>base</sub>  | ✓     | 0.58 (↓ 0.025)  | 0.617 (↑ 0.05)  | 0.625 (↓ 0.197) | 0.624 (↓ 0.066) | 0.546 (↓ 0.036)  | 22                          |
| XLM-R <sub>base</sub>     | ✓     | -               | -               | -               | -               | -                | 0                           |
| FlauBERT <sub>large</sub> | ✓     | 0.815 (↑ 0.042) | 0.863 (↓ 0.044) | 0.896 (↓ 0.071) | 0.859 (↓ 0.053) | 0.685 (↓ 0.051)  | 134                         |
| XLM-R <sub>large</sub>    | ✓     | -               | -               | -               | -               | -                | 0                           |

Table 10: Accuracies of classifiers trained to predict gender of adjectives using only stable dimensions. The number in the brackets signifies absolute difference with the accuracies of the same classifiers trained on all dimensions. ↑ marks an increase in accuracy when using only stable dimensions, ↓ marks the decrease.

### B.1.3 Gender: Noun & Adjective

| Model                     | Cased | KNN            | NB             | LR             | RF             | DT             | Number of stable dimensions |
|---------------------------|-------|----------------|----------------|----------------|----------------|----------------|-----------------------------|
| FlauBERT <sub>small</sub> | ✓     | 0.775 (↑0.071) | 0.844 (↑0.111) | 0.908 (↓0.054) | 0.848 (↑0.056) | 0.715 (↑0.104) | 74                          |
| DistilBERT                | ✓     | 0.59 (0)       | 0.561 (↓0.01)  | 0.617 (↓0.088) | 0.626 (↓0.003) | 0.522 (↓0.043) | 14                          |
| CamemBERT <sub>base</sub> | ✓     | -              | -              | -              | -              | -              | 0                           |
| mBERT <sub>base</sub>     | ✓     | 0.531 (↑0.003) | 0.55 (↑0.035)  | 0.561 (↓0.033) | 0.558 (↓0.005) | 0.542 (↑0.021) | 3                           |
| mBERT <sub>base</sub>     | ✓     | -              | -              | -              | -              | -              | 0                           |
| FlauBERT <sub>base</sub>  | ✓     | 0.548 (↓0.012) | 0.588 (↑0.021) | 0.592 (↓0.055) | 0.593 (↑0.007) | 0.540 (↑0.007) | 47                          |
| FlauBERT <sub>base</sub>  | ✓     | 0.646 (↑0.054) | 0.606 (↑0.055) | 0.72 (↓0.043)  | 0.693 (↑0.023) | 0.595 (↑0.038) | 48                          |
| XLM-R <sub>base</sub>     | ✓     | -              | -              | -              | -              | -              | 0                           |
| FlauBERT <sub>large</sub> | ✓     | 0.84 (↑0.036)  | 0.92 (↑0.022)  | 0.964 (↑0.024) | 0.909 (↑0.019) | 0.744 (↑0.057) | 227                         |
| XLM-R <sub>large</sub>    | ✓     | 0.467 (↓0.117) | 0.509(↓0.057)  | 0.471 (↓0.152) | 0.502 (↓0.102) | 0.503 (↓0.027) | 1                           |

Table 11: Accuracies of classifiers trained to predict the gender of nouns and adjectives using only stable dimensions. The number in the brackets signifies absolute difference with the accuracies of the same classifiers trained on all dimensions. ↑ marks an increase in accuracy when using only stable dimensions, ↓ marks the decrease.

## B.2 Number

### B.2.1 Number: Noun

| Model                     | Cased | KNN            | NB             | LR             | RF             | DT             | Number of stable dimensions |
|---------------------------|-------|----------------|----------------|----------------|----------------|----------------|-----------------------------|
| FlauBERT <sub>small</sub> | ✓     | 0.861 (↑0.07)  | 0.969 (↑0.033) | 0.991 (0)      | 0.947 (↑0.011) | 0.769 (↑0.013) | 213                         |
| DistilBERT                | ✓     | 0.616 (↓0.004) | 0.615 (↑0.01)  | 0.713 (↓0.108) | 0.676 (↓0.005) | 0.564(↓0.002)  | 22                          |
| CamemBERT <sub>base</sub> | ✓     | -              | -              | -              | -              | -              | 0                           |
| mBERT <sub>base</sub>     | ✓     | -              | -              | -              | -              | -              | 0                           |
| mBERT <sub>base</sub>     | ✓     | -              | -              | -              | -              | -              | 0                           |
| FlauBERT <sub>base</sub>  | ✓     | 0.602 (↑0.032) | 0.666 (↑0.041) | 0.693 (↓0.017) | 0.674 (↑0.032) | 0.565 (↑0.018) | 120                         |
| FlauBERT <sub>base</sub>  | ✓     | 0.717 (↑0.033) | 0.72 (↑0.115)  | 0.749 (↓0.065) | 0.746 (↑0.004) | 0.642 (↑0.003) | 22                          |
| XLM-R <sub>base</sub>     | ✓     | 0.572 (↑0.036) | 0.59 (↑0.107)  | 0.602 (↓0.139) | 0.553 (↓0.042) | 0.563 (↓0.029) | 2                           |
| FlauBERT <sub>large</sub> | ✓     | 0.95 (↑0.09)   | 0.954 (↑0.005) | 0.972 (↓0.01)  | 0.943905 (0)   | 0.827 (↑0.012) | 107                         |
| XLM-R <sub>large</sub>    | ✓     | -              | -              | -              | -              | -              | 0                           |

Table 12: Accuracies of classifiers trained to predict the number of nouns using only stable dimensions. The number in the brackets signifies absolute difference with the accuracies of the same classifiers trained on all dimensions. ↑ marks an increase in accuracy when using only stable dimensions, ↓ marks the decrease.

### B.2.2 Number: Adjective

| Model                     | Cased | KNN            | NB             | LR             | RF             | DT             | Number of stable dimensions |
|---------------------------|-------|----------------|----------------|----------------|----------------|----------------|-----------------------------|
| FlauBERT <sub>small</sub> | ✓     | 0.889 (↑0.148) | 0.957 (↑0.021) | 0.986 (↓0.002) | 0.943 (↑0.017) | 0.773 (↑0.06)  | 107                         |
| DistilBERT                | ✓     | -              | -              | -              | -              | -              | 0                           |
| CamemBERT <sub>base</sub> | ✓     | -              | -              | -              | -              | -              | 0                           |
| mBERT <sub>base</sub>     | ✓     | -              | -              | -              | -              | -              | 0                           |
| mBERT <sub>base</sub>     | ✓     | -              | -              | -              | -              | -              | 0                           |
| FlauBERT <sub>base</sub>  | ✓     | 0.549 (↑0.003) | 0.585 (↓0.019) | 0.583 (↓0.075) | 0.567 (↓0.027) | 0.541 (↑0.018) | 4                           |
| FlauBERT <sub>base</sub>  | ✓     | 0.655 (↑0.05)  | 0.689 (↑0.122) | 0.693 (↓0.129) | 0.671 (↓0.022) | 0.593 (↑0.015) | 4                           |
| XLM-R <sub>base</sub>     | ✓     | 0.5 (↓0.059)   | 0.526 (↓0.003) | 0.523 (↓0.137) | 0.53 (↓0.061)  | 0.51 (↓0.052)  | 1                           |
| FlauBERT <sub>large</sub> | ✓     | 0.936 (↑0.163) | 0.955 (↑0.049) | 0.974 (↑0.007) | 0.94 (↑0.027)  | 0.831 (↑0.096) | 151                         |
| XLM-R <sub>large</sub>    | ✓     | -              | -              | -              | -              | -              | 0                           |

Table 13: Accuracies of classifiers trained to predict the number of adjectives using only stable dimensions. The number in the brackets signifies absolute difference with the accuracies of the same classifiers trained on all dimensions. ↑ marks an increase in accuracy when using only stable dimensions, ↓ marks the decrease.

### B.2.3 Number: Noun & Adjective

| Model                     | Cased | KNN            | NB             | LR              | RF             | DT             | Number of stable dimensions |
|---------------------------|-------|----------------|----------------|-----------------|----------------|----------------|-----------------------------|
| FlauBERT <sub>small</sub> | ✓     | 0.888 (↑0.103) | 0.966 (↑0.035) | 0.989 (↓0.003)  | 0.945 (↑0.016) | 0.774 (↑0.028) | 127                         |
| DistilBERT                | ✓     | 0.606 (↑0.012) | 0.653 (↑0.068) | 0.674 (↓0.125)  | 0.662 (↓0.013) | 0.577 (↑0.004) | 15                          |
| CamemBERT <sub>base</sub> | ✓     | 0.518 (↓0.009) | 0.511 (↓0.002) | 0.518 (↑0.009)  | 0.522 (↓0.012) | 0.518 (↑0.011) | 9                           |
| mBERT <sub>base</sub>     |       | 0.534 (↓0.008) | 0.555 (↑0.007) | 0.554 (↓0.097)  | 0.555 (↓0.035) | 0.539 (↓0.002) | 4                           |
| mBERT <sub>base</sub>     | ✓     | -              | -              | -               | -              | -              | 0                           |
| FlauBERT <sub>base</sub>  |       | 0.601 (↑0.025) | 0.661 (↑0.05)  | 0.683 (↓0.0195) | 0.668 (↑0.032) | 0.553 (↑0.011) | 91                          |
| FlauBERT <sub>base</sub>  | ✓     | 0.682 (↑0.02)  | 0.698 (↑0.104) | 0.7 (↓0.105)    | 0.702 (↓0.036) | 0.615 (↓0.014) | 7                           |
| XLM-R <sub>base</sub>     | ✓     | -              | -              | -               | -              | -              | 0                           |
| FlauBERT <sub>large</sub> | ✓     | 0.943 (↑0.074) | 0.957 (↑0.014) | 0.974 (↓0.007)  | 0.946 (↑0.004) | 0.833 (↑0.027) | 195                         |
| XLM-R <sub>large</sub>    | ✓     | 0.517 (↓0.084) | 0.566 (↓0.027) | 0.58 (↓0.087)   | 0.545 (↑0.08)  | 0.545 (↓0.005) | 1                           |

Table 14: Accuracies of classifiers trained to predict the number of nouns and adjectives using only stable dimensions. The number in the brackets signifies absolute difference with the accuracies of the same classifiers trained on all dimensions. ↑ marks an increase in accuracy when using only stable dimensions, ↓ marks the decrease.

## B.3 POS

### B.3.1 POS: Noun

| Model                     | Cased | KNN            | NB             | LR             | RF              | DT                | Number of stable dimensions |
|---------------------------|-------|----------------|----------------|----------------|-----------------|-------------------|-----------------------------|
| FlauBERT <sub>small</sub> | ✓     | 0.934 (↑0.01)  | 0.897 (↑0.037) | 0.937 (↓0.027) | 0.914 (↓0.0003) | 0.794041 (↑0.008) | 92                          |
| DistilBERT                | ✓     | 0.561 (↓0.046) | 0.608 (↓0.001) | 0.603 (↓0.106) | 0.565 (↓0.076)  | 0.535 (↓0.029)    | 3                           |
| CamemBERT <sub>base</sub> | ✓     | 0.507 (↓0.068) | 0.531 (↑0.008) | 0.532 (↓0.048) | 0.492 (↓0.079)  | 0.493 (↓0.037)    | 1                           |
| mBERT <sub>base</sub>     |       | 0.506 (↓0.033) | 0.54 (↑0.005)  | 0.547 (↓0.043) | 0.503 (↓0.075)  | 0.501 (↓0.019)    | 1                           |
| mBERT <sub>base</sub>     | ✓     | 0.526 (↓0.052) | 0.502 (↓0.002) | 0.468 (↓0.1)   | 0.505 (↓0.14)   | 0.503 (↓0.079)    | 1                           |
| FlauBERT <sub>base</sub>  |       | 0.726 (↑0.008) | 0.727 (↑0.09)  | 0.734 (↓0.051) | 0.76 (↓0.006)   | 0.657 (↑0.011)    | 20                          |
| FlauBERT <sub>base</sub>  | ✓     | 0.637 (↓0.025) | 0.637 (↑0.06)  | 0.678 (↓0.107) | 0.677 (↓0.028)  | 0.577 (↓0.01)     | 18                          |
| XLM-R <sub>base</sub>     | ✓     | 0.507 (↓0.053) | 0.549 (↑0.011) | 0.551 (↓0.119) | 0.512 (↓0.106)  | 0.507 (↓0.046)    | 1                           |
| FlauBERT <sub>large</sub> | ✓     | 0.916 (↑0.039) | 0.903 (↑0.063) | 0.932 (↓0.026) | 0.902 (↑0.008)  | 0.768 (↑0.02)     | 101                         |
| XLM-R <sub>large</sub>    | ✓     | -              | -              | -              | -               | -                 | 0                           |

Table 15: Accuracies of classifiers trained to predict the POS of noun vs non-nouns using only stable dimensions. The number in the brackets signifies absolute difference with the accuracies of the same classifiers trained on all dimensions. ↑ marks an increase in accuracy when using only stable dimensions, ↓ marks the decrease.

### B.3.2 POS: Adjective

| Model                     | Cased | KNN            | NB             | LR             | RF             | DT             | Number of stable dimensions |
|---------------------------|-------|----------------|----------------|----------------|----------------|----------------|-----------------------------|
| FlauBERT <sub>small</sub> | ✓     | 0.93 (↑0.042)  | 0.862 (↓0.025) | 0.927 (↓0.019) | 0.909 (↑0.007) | 0.783 (↑0.031) | 77                          |
| DistilBERT                | ✓     | -              | -              | -              | -              | -              | 0                           |
| CamemBERT <sub>base</sub> | ✓     | -              | -              | -              | -              | -              | 0                           |
| mBERT <sub>base</sub>     |       | -              | -              | -              | -              | -              | 0                           |
| mBERT <sub>base</sub>     | ✓     | -              | -              | -              | -              | -              | 0                           |
| FlauBERT <sub>base</sub>  |       | 0.598 (↓0.063) | 0.593 (↓0.009) | 0.607 (↓0.111) | 0.62 (↓0.055)  | 0.562 (↓0.002) | 5                           |
| FlauBERT <sub>base</sub>  | ✓     | 0.531 (↓0.075) | 0.555 (↓0.003) | 0.556 (↓0.186) | 0.541 (↓0.120) | 0.516 (↓0.045) | 5                           |
| XLM-R <sub>base</sub>     | ✓     | 0.524 (↓0.028) | 0.562 (↓0.007) | 0.562 (↓0.007) | 0.522 (↓0.024) | 0.519 (↑0.046) | 1                           |
| FlauBERT <sub>large</sub> | ✓     | 0.839 (↓0.018) | 0.826 (↓0.039) | 0.863 (↓0.08)  | 0.843 (↓0.051) | 0.718 (↓0.026) | 58                          |
| XLM-R <sub>large</sub>    | ✓     | 0.52 (↑0.004)  | 0.503 (↓0.059) | 0.498 (↓0.096) | 0.504 (↓0.056) | 0.504 (↑0.011) | 1                           |

Table 16: Accuracies of classifiers trained to predict the POS of adjectives vs non-adjectives using only stable dimensions. The number in the brackets signifies absolute difference with the accuracies of the same classifiers trained on all dimensions. ↑ marks an increase in accuracy when using only stable dimensions, ↓ marks the decrease.

### B.3.3 POS: Verb

| Model                     | Cased | KNN            | NB             | LR             | RF             | DT               | Number of stable dimensions |
|---------------------------|-------|----------------|----------------|----------------|----------------|------------------|-----------------------------|
| FlauBERT <sub>small</sub> | ✓     | 0.875 (↓0.05)  | 0.792 (↓0.111) | 0.835 (↓0.138) | 0.848 (↓0.096) | 0.754 (↓0.089)   | 24                          |
| DistilBERT                | ✓     | 0.596 (↓0.001) | 0.618 (↑0.012) | 0.621 (↓0.075) | 0.616 (↓0.038) | 0.545162 (↓0.04) | 5                           |
| CamemBERT <sub>base</sub> | ✓     | -              | -              | -              | -              | -                | 0                           |
| mBERT <sub>base</sub>     | ✓     | -              | -              | -              | -              | -                | 0                           |
| FlauBERT <sub>base</sub>  | ✓     | 0.736 (↓0.033) | 0.723 (↑0.041) | 0.74 (↓0.092)  | 0.762 (↓0.044) | 0.659 (↓0.02)    | 49                          |
| FlauBERT <sub>base</sub>  | ✓     | 0.623 (↓0.088) | 0.645 (↑0.05)  | 0.673 (↓0.163) | 0.675 (↓0.086) | 0.578 (↓0.045)   | 22                          |
| XLM-R <sub>base</sub>     | ✓     | -              | -              | -              | -              | -                | 0                           |
| FlauBERT <sub>large</sub> | ✓     | 0.89 (↑0.022)  | 0.854 (↓0.023) | 0.894 (↓0.075) | 0.875 (↓0.056) | 0.752 (↓0.029)   | 73                          |
| XLM-R <sub>large</sub>    | ✓     | 0.506 (↓0.081) | 0.576 (↓0.02)  | 0.579 (↓0.077) | 0.496 (↓0.136) | 0.496 (↓0.07)    | 1                           |

Table 17: Accuracies of classifiers trained to predict the POS of verbs vs non-verbs using only stable dimensions. The number in the brackets signifies absolute difference with the accuracies of the same classifiers trained on all dimensions. ↑ marks an increase in accuracy when using only stable dimensions, ↓ marks the decrease.

## 7.4 Semantic supersenses

### 7.4.1 Act

| Model                     | Cased | KNN            | NB             | LR             | RF             | DT                | Number of stable dimensions |
|---------------------------|-------|----------------|----------------|----------------|----------------|-------------------|-----------------------------|
| FlauBERT <sub>small</sub> | ✓     | 0.801 (↑0.083) | 0.795 (↑0.022) | 0.831 (↓0.009) | 0.809 (↑0.02)  | 0.701 (↑0.055)    | 48                          |
| DistilBERT                | ✓     | 0.5 (↓0.03)    | 0.542 (↓0.007) | 0.493 (↓0.157) | 0.526 (↓0.062) | 0.523 (↓0.032)    | 3                           |
| CamemBERT <sub>base</sub> | ✓     | 0.512 (↓0.023) | 0.515 (↓0.013) | 0.498 (↓0.027) | 0.487 (↓0.058) | 0.487 (↓0.035)    | 1                           |
| mBERT <sub>base</sub>     | ✓     | -              | -              | -              | -              | -                 | 0                           |
| FlauBERT <sub>base</sub>  | ✓     | 0.55 (↓0.014)  | 0.58 (↓0.05)   | 0.587 (↓0.091) | 0.518 (↓0.114) | 0.533 (↓0.025)    | 2                           |
| FlauBERT <sub>base</sub>  | ✓     | 0.615 (↓0.001) | 0.616 (↑0.033) | 0.638 (↓0.082) | 0.626 (↓0.022) | 0.557 (↑0.020070) | 14                          |
| XLM-R <sub>base</sub>     | ✓     | 0.433 (↓0.086) | 0.49 (↓0.009)  | 0.481 (↓0.19)  | 0.49 (↓0.076)  | 0.49 (↓0.052)     | 1                           |
| FlauBERT <sub>large</sub> | ✓     | 0.766 (↑0.019) | 0.799 (↑0.016) | 0.819 (↓0.019) | 0.798 (↑0.01)  | 0.667 (↑0.029)    | 50                          |
| XLM-R <sub>large</sub>    | ✓     | -              | -              | -              | -              | -                 | 0                           |

Table 18: Accuracies of classifiers trained to classify nouns into 2 categories: Act vs non-Act. The number in the brackets signifies absolute difference with the accuracies of the same classifiers trained on all dimensions. ↑ marks an increase in accuracy when using only stable dimensions, ↓ marks the decrease.

### 7.4.2 Person

| Model                     | Cased | KNN            | NB             | LR             | RF             | DT             | Number of stable dimensions |
|---------------------------|-------|----------------|----------------|----------------|----------------|----------------|-----------------------------|
| FlauBERT <sub>small</sub> | ✓     | 0.81 (↑0.052)  | 0.905 (↑0.04)  | 0.915 (↑0.010) | 0.877 (↓0.002) | 0.74 (↑0.015)  | 95                          |
| DistilBERT                | ✓     | 0.524 (↓0.057) | 0.545 (↓0.052) | 0.443 (↓0.292) | 0.5 (↓0.134)   | 0.5 (↓0.093)   | 1                           |
| CamemBERT <sub>base</sub> | ✓     | -              | -              | -              | -              | -              | 0                           |
| mBERT <sub>base</sub>     | ✓     | -              | -              | -              | -              | -              | 0                           |
| FlauBERT <sub>base</sub>  | ✓     | 0.634 (↑0.017) | 0.508 (↓0.004) | 0.468 (↑0.008) | 0.675 (↓0.033) | 0.618 (↓0.008) | 6                           |
| FlauBERT <sub>base</sub>  | ✓     | 0.581 (↑0.067) | 0.638 (↑0.047) | 0.646 (↓0.008) | 0.568 (↑0.003) | 0.529 (↓0.022) | 5                           |
| FlauBERT <sub>base</sub>  | ✓     | 0.555 (↓0.048) | 0.567 (↓0.023) | 0.55 (↓0.188)  | 0.547 (↓0.133) | 0.545 (↓0.03)  | 1                           |
| XLM-R <sub>base</sub>     | ✓     | 0.532 (↑0.022) | 0.547 (↑0.023) | 0.445 (↓0.095) | 0.563 (↑0.001) | 0.563 (↑0.118) | 1                           |
| FlauBERT <sub>large</sub> | ✓     | 0.798 (↑0.06)  | 0.843 (↓0.005) | 0.861 (↑0.002) | 0.815 (↓0.007) | 0.675 (↑0.05)  | 63                          |
| XLM-R <sub>large</sub>    | ✓     | -              | -              | -              | -              | -              | 0                           |

Table 19: Accuracies of classifiers trained to classify nouns into 2 categories: Person vs non-Person. The number in the brackets signifies absolute difference with the accuracies of the same classifiers trained on all dimensions. ↑ marks an increase in accuracy when using only stable dimensions, ↓ marks the decrease.

# Whisper-TAD: A general model for Transcription, Alignment and Diarization of speech

**Camille Lavigne**

Université de Nancy

lavignecamille37@gmail.com

**Alex Stasica**

Utrecht University

a.stasica@uu.nl

## Abstract

Currently, there is a lack of a straightforward implementation of diarization-augmented speech transcription (DAST), ie. implementation of transcription, diarization and alignment to the audio within one model. These tasks typically require distinct models, necessitating to stack them together for complete processing. In this study, we advocate for leveraging the advanced capabilities of the Whisper models, which already excels in automatic transcription and partial alignment. Our approach involves fine-tuning the model’s parameters on both transcription and diarization tasks in a SOT-FIFO (Serialized Output Training-First In First Out) manner. This comprehensive framework facilitates the creation of orthographic transcriptions, identification of speakers, and precise alignment, thus enhancing the efficiency of audio processing workflows. While our work represents an initial step towards a unified transcription and diarization framework, the development of such a model demands substantial high-quality data augmentation and computational resources beyond our current scope. Consequently, our focus is narrowed to the English language. Despite these limitations, our method demonstrates promising performance in both transcription and diarization tasks. Comparative analysis between pre-trained models and fine-tuned TAD (Transcription, Alignment, Diarization) versions suggests that incorporating diarization into a Whisper model doesn’t compromise transcription accuracy. Our findings hint that deploying our TAD framework on the largest Whisper model could potentially yield state-of-the-art performance across all mentioned tasks.

**Keywords:** Diarization, automatic speech recognition, Whisper

## 1 Introduction

Speaker diarization (SD) endeavors to ascertain “*who spoke when*” (Tranter and Reynolds, 2006).

Various methodologies have been employed to annotate audio data for the purpose of identifying speakers within it. Conventionally, this task was compartmentalized into distinct sub-modules (Park et al., 2022), ranging from voice activity detection (VAD) to clustering speech segments and assigning speaker labels. However, the optimization of each module in isolation restricted overall optimization. With the advent of deep learning techniques, neural networks have been leveraged to improve the performance of these sub-modules by extracting speaker embedding (Variani et al., 2014; Heigold et al., 2016), thereby rendering models easier to train, more resilient to speaker variability, and robust under varying acoustic conditions (Zhang et al., 2019). A recent breakthrough is the adoption of fully end-to-end Neural Diarization (EEND; Fujita et al. (2019a,b)), wherein all sub-modules are replaced by a single neural network. This promising approach enables the joint optimization of model components, potentially enabling the handling of multi-speaker audio and overlapping speech. Initially implemented using bi-directional long short-term memory architectures (Fujita et al., 2019a), these models swiftly transitioned to self-attention-based networks (Fujita et al., 2019b). Nevertheless, challenges persist, including the model’s limited capacity to handle a large number of speakers, the difficulty in achieving online processing, and the tendency for models to overfit the training data distribution (Park et al., 2022).

Recent advancements have demonstrated that the concurrent modeling of SD and automatic speech recognition (ASR) can enhance the performance of both tasks, as exemplified in various models (Silovsky et al., 2012; Huang et al., 2007). This integration allows SD to use both acoustic and linguistic information, resulting in superior performance compared to models relying solely

on acoustic information. Furthermore, it enables not only to determine "who spoke when" but also discerning "what" was spoken. As discussed in Park et al. (2022), various approaches have been explored, including the introduction of speaker tag roles in transcripts (Shafey et al., 2019), MAP-based joint decoding frameworks (Kanda et al., 2019), and the emergence of End-to-End Speaker Attribution ASR (E2E SA-ASR, Kanda et al. (2020a)), which facilitates speaker counting, multi-talker ASR, and speaker identity determination without limitations on the number of speakers.

Our aim in this research is to unify the diarization and transcription task in one model. We achieved this by fine-tuning existing Whisper models, (Radford et al., 2023), which already transcribe speech with state of the art performance and align the transcription to the audio. Our fine-tuning enables the recognition of distinct speakers within the speech audio. By focusing on fine-tuning rather than extensive pre-training, we achieve transferable results even with limited data, making our model applicable to languages with minimal available resources. Thus, we introduce Whisper-TAD (Transcribe, Align, Diarize), an initial version of a versatile model that streamlines the DAST pipeline.

Our article is structured as follows; in section 2 we present our methodology, then in section 3 our experimental setup, in section 4 our experimental results and we finally discuss possible further works in section 5.

## 2 Methodology

### 2.1 Foundation model

As a foundation model we use Whisper (Radford et al., 2023). Whisper models already reach state of the art performance in orthographic transcription task. As highlighted by the authors, these models were designed in a multi-task format, also solving: translation, VAD, partial alignment, and language identification tasks. Although diarization was cited as a desirable task to solve in an ASR pipeline, the authors didn't address this in their original publication. In order to add this ability to the Whisper models, we add special tokens to the tokenizer as well as new randomly initialized embeddings for these new tokens. The new tokens are up to five speaker tokens as well as a noSpk token for VAD. We then fine-tune the models on both ASR and diarization tasks jointly.

### 2.2 Fine-tuning task

For the fine tuning, we used the SOT FIFO framework. SOT (Serialized Output Training) as been first introduced in Kanda et al. (2020b). It allows to train an attention-based neural network on both transcription and diarization using only one output. It is usable on data that contains multiple speakers and overlapping speech. When there is multiple speakers to classify, there are different ways to output the result of the deep neural network in one output. We choose FIFO (First In First Out) as it is the most used variant of SOT. In the FIFO approach, a distinct speaker ID is incrementally assigned to each newly detected speaker in the audio. For instance, the initial speaker detected is labeled as "spk1," the subsequent one as "spk2," and so forth. Consequently, there is no correspondence between the speaker IDs assigned to two different segments of audio, even if they contain the same speaker or speakers. For this reason our framework is a local E2E DAST model. We have not yet implemented a clustering of the speakers to recognize when speakers in different chunks of a same audio have the same identity.

One of the limitation of the SOT method that we use is that we cannot classify more than five speakers in one chunk of thirty seconds. However, the cases in which more than five speakers talk in one 30s chunk are pretty rare. It would therefore require a large amount of data augmentation to achieve decent accuracy on more than 5 speakers.

Figure 1 illustrates The SOT FIFO framework where all necessary tokens for each 30-second segment of audio (chunk) are generated by the autoregressive decoder. For every utterance of an audio chunk, the model initiates by outputting a speaker token (depicted in blue) alongside a timestamp token (depicted in green) to mark the beginning of the utterance. Following this, tokens outputted by the Whisper byte-pair encoding tokenizer (depicted in orange) are employed to transcribe the utterance. Once transcription is complete, a final timestamp token is appended to signify the end of the speaker's utterance. Furthermore, if the same speaker contributes multiple times within a single chunk, they are assigned a consistent speaker ID (ranging from 1 to 5).

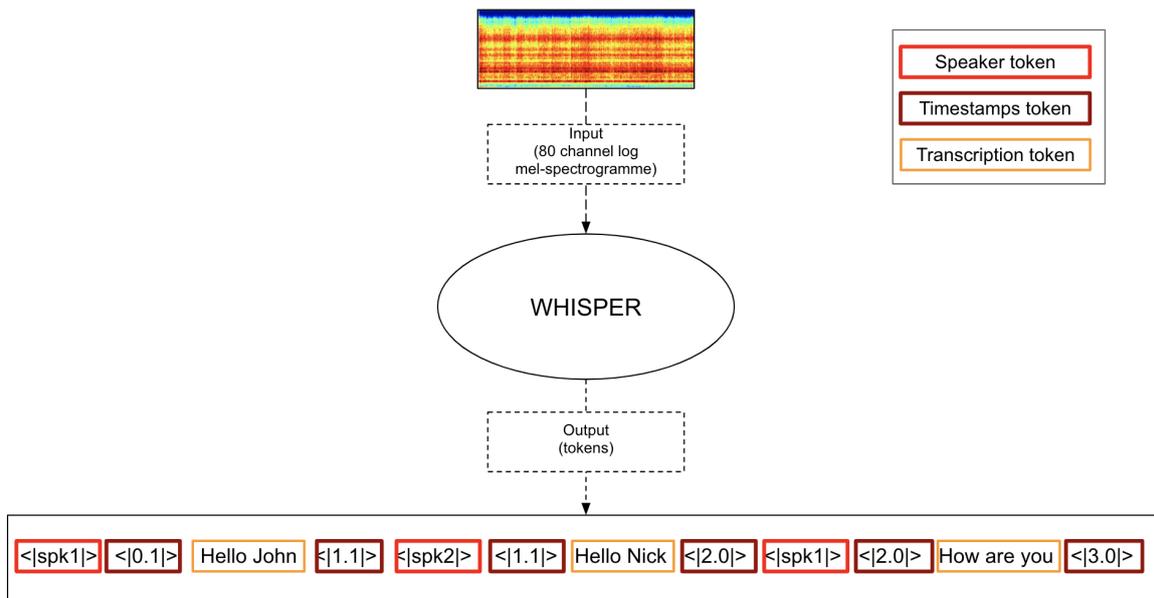


Figure 1: WHISPER-TAD framework

### 3 Experimental setup

#### 3.1 Data

This study makes use of the AMI (Kraaij et al., 2005) and ICSI (Janin et al., 2003) benchmark datasets for our experiments.

The ICSI corpus regroups 75 meetings, with 4 different types of meetings with up to 10 participants, and the AMI corpus comprises 100 hours of audio from 171 meetings coming from multiple sites in which 3 to 5 participants are present. Both datasets provide the meetings transcriptions, word level alignment, and speaker labels. These datasets are suitable for the fine-tuning task as we plan on evaluating the performance of SD along with an ASR module. As we didn't find clear guidelines to split the ICSI corpus we used the full ICSI for training and validation of our model, not for testing. For the AMI corpus guidelines diverges (Landini et al., 2022). We decided to split it in train, validation and test sets as described by the official suggestions on the website of the corpus<sup>1</sup> as it seems to be a reliable, efficient split of the data.

Note that a few hours of audio from the AMI and ICSI corpus contains audio speech without transcription which increase the probability of hallucination at inference if the model was trained on these data. For ICSI corpus, these parts without

<sup>1</sup><https://groups.inf.ed.ac.uk/ami/corpus/datasets.shtml>

|                   | AMI | ICSI | Total |
|-------------------|-----|------|-------|
| <i>Training</i>   | 78  | 58   | 136   |
| <i>Validation</i> | 10  | 12   | 22    |
| <i>Test</i>       | 9   | 0    | 9     |

Table 1: Share of the AMI and ICSI corpus in the training - validation - test sets. Shares are given in hours of audio.

transcription are parts where the speakers are asked to pronounce random numbers all together. These parts were removed from the training - validation - test datasets.

#### 3.2 Hyper-parameters

Due to limited computational power for this experiment, we only fine-tuned the base, small and medium Whisper models, but could not fine-tune the larger versions. We used a 0.05 dropout with a learning rate of  $1e^{-5}$  and a batch size of 100. The optimizer used is Adam. We had access to one (24 GB ram) RTX 6000 GPU.

#### 3.3 Metrics

Our models are trained on three distinct tasks, each requiring specific metrics for evaluation.

**For the speaker diarization task**, we chose the Diarization Error Rate (DER), which quantifies the accuracy of speaker diarization systems by measur-

ing the alignment between the predicted speaker segments and the ground truth. DER accounts for missed speakers, false alarms, and speaker misalignment. Specifically, we utilized its Python implementation from `pyannote` (Bredin et al., 2020). We do not use collar, as recommended by `pyannote` guidelines.

**For the transcription task**, we employed the Concatenated minimum-Permutation Word Error Rate (cpWER) (Watanabe et al., 2020). Unlike traditional Word Error Rate (WER), cpWER gathers all the speech productions from a same speaker and calculate the WER per speaker. This is particularly useful in scenarios where the speech stream is segmented in a "Diarization-style" manner, i.e., segmented by speaker.

Additionally, we employed traditional WER to compare the performance of the pre-trained models with those fine-tuned on both transcription and diarization. This comparison allows us to assess whether multitasking improves or hampers the performance of the models on their primary task.

**For the Voice Activity Detection (VAD)**, we utilized the Equal Error Rate (EER) metric. EER represents the point where the false acceptance rate (ie. falsely classifying non-speech as speech) equals the false rejection rate (ie. falsely classifying speech as non speech), providing a balanced measure of VAD performance across different operating conditions.

#### 4 Experimental Results

The results of our fine-tuning task on the different Whisper models are illustrated in Table 2. As to be expected, the larger the model, the better the metrics. Another notable observation is that the performance difference between the Base and the Small models is more significant than the one between the Small and the Medium. One explanation for this phenomenon is the lack of data, and the fact that we didn't do any data augmentation to mitigate this.

|               | DER   | EER   | cpWER |
|---------------|-------|-------|-------|
| <i>Base</i>   | 0.498 | 0.655 | 0.548 |
| <i>Small</i>  | 0.202 | 0.120 | 0.345 |
| <i>Medium</i> | 0.189 | 0.151 | 0.313 |

Table 2: Results of the fine-tuning task on the Base, Small and Medium Whisper models

As explained in 3.3, we also calculated the stan-

| Base  |       | Small |       | Medium |       |
|-------|-------|-------|-------|--------|-------|
| PT    | TAD   | PT    | TAD   | PT     | TAD   |
| 0.621 | 0.346 | 0.466 | 0.279 | 0.403  | 0.269 |

Table 3: WER comparison for three different sizes of the Whisper model. The models labeled as PT denote the pre-trained models, while those labeled as TAD indicate models fine-tuned for diarization

dard WER in order to demonstrate that even while adding the diarisation task, the performance of the models on the initial task they have been trained on does not decrease, but even increases as depicted by Table 3, showing that the fine-tuning on another task is also useful for the initial task, and a joint pipeline can only increases the performance for both tasks.

#### 5 Further work

This study serves as a proof of concept, with further investigations required to fully evaluate the methodology's feasibility.

Firstly, a crucial step is transitioning from local E2E processing to global E2E processing. This entails enabling the model to consistently assign the same speaker ID to speakers across different audio chunks, rather than assigning new speaker IDs for each chunk as done in prior research by Cornell et al. (2024) using Wav2Vec (Schneider et al., 2019). Various approaches can be explored, such as incorporating a classification head by clustering all speakers across the entire audio, thus necessitating an additional output head for the model.

Secondly, larger versions of Whisper need to be fine-tuned to ascertain the maximum performance achievable using this methodology.

Thirdly, in order to train larger models for the diarization task, data augmentation is indispensable. Leveraging datasets like LibriSpeech (Panayotov et al., 2015) for data augmentation can enhance the training process. Additionally, data augmentation can facilitate the fine-tuning of these models for diarization tasks in languages with limited accessible resources.

#### 6 Conclusion

This study introduces Whisper-TAD, a preliminary investigation into a versatile model designed to integrate transcription, sentence-level alignment, and diarization tasks within a unified pipeline. Employ-

ing a SOT FIFO method, special tokens are incorporated for speaker identification, enabling recognition of up to 5 speakers per 30-seconds audio chunk. Our experiments conducted on the ISCI and AMI corpora yield promising outcomes, suggesting potential applicability across languages with limited resources. Notably, our approach achieves competitive performance, even in the absence of data augmentation and without the exploration of larger models. These findings underscore the robustness and effectiveness of Whisper-TAD, offering valuable insights for future research directions in multi-task audio processing.

## References

- 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 *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7124–7128. IEEE.
- Samuele Cornell, Jee-weon Jung, Shinji Watanabe, and Stefano Squartini. 2024. One model to rule them all? towards end-to-end joint speaker diarization and speech recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11856–11860. IEEE.
- Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Kenji Nagamatsu, and Shinji Watanabe. 2019a. End-to-end neural speaker diarization with permutation-free objectives. *arXiv preprint arXiv:1909.05952*.
- Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Yawen Xue, Kenji Nagamatsu, and Shinji Watanabe. 2019b. End-to-end neural speaker diarization with self-attention. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 296–303. IEEE.
- Georg Heigold, Ignacio Moreno, Samy Bengio, and Noam Shazeer. 2016. End-to-end text-dependent speaker verification. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5115–5119. IEEE.
- Jing Huang, Etienne Marcheret, Karthik Visweswariah, and Gerasimos Potamianos. 2007. The ibm rt07 evaluation systems for speaker diarization on lecture meetings. In *International Evaluation Workshop on Rich Transcription*, pages 497–508. Springer.
- Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. 2003. The icsi meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 1, pages I–I. IEEE.
- Naoyuki Kanda, Yashesh Gaur, Xiaofei Wang, Zhong Meng, Zhuo Chen, Tianyan Zhou, and Takuya Yoshioka. 2020a. Joint speaker counting, speech recognition, and speaker identification for overlapped speech of any number of speakers. *arXiv preprint arXiv:2006.10930*.
- Naoyuki Kanda, Yashesh Gaur, Xiaofei Wang, Zhong Meng, and Takuya Yoshioka. 2020b. Serialized output training for end-to-end overlapped speech recognition. *arXiv preprint arXiv:2003.12687*.
- Naoyuki Kanda, Shota Horiguchi, Yusuke Fujita, Yawen Xue, Kenji Nagamatsu, and Shinji Watanabe. 2019. Simultaneous speech recognition and speaker diarization for monaural dialogue recordings with target-speaker acoustic models. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 31–38. IEEE.
- Wessel Kraaij, Thomas Hain, Mike Lincoln, and Wilfried Post. 2005. The ami meeting corpus. In *Proc. International Conference on Methods and Techniques in Behavioral Research*.
- Federico Landini, Ján Profant, Mireia Diez, and Lukáš Burget. 2022. Bayesian hmm clustering of x-vector sequences (vbx) in speaker diarization: theory, implementation and analysis on standard tasks. *Computer Speech & Language*, 71:101254.
- Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.
- Tae Jin Park, Naoyuki Kanda, Dimitrios Dimitriadis, Kyu J Han, Shinji Watanabe, and Shrikanth Narayanan. 2022. A review of speaker diarization: Recent advances with deep learning. *Computer Speech & Language*, 72:101317.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.
- Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.
- Laurent El Shafey, Hagen Soltau, and Izhak Shafran. 2019. Joint speech recognition and speaker diarization via sequence transduction. *arXiv preprint arXiv:1907.05337*.

- Jan Silovsky, Jindrich Zdansky, Jan Nouza, Petr Cerva, and Jan Prazak. 2012. Incorporation of the asr output in speaker segmentation and clustering within the task of speaker diarization of broadcast streams. In *2012 IEEE 14th International Workshop on Multimedia Signal Processing (MMSP)*, pages 118–123. IEEE.
- Sue E Tranter and Douglas A Reynolds. 2006. An overview of automatic speaker diarization systems. *IEEE Transactions on audio, speech, and language processing*, 14(5):1557–1565.
- Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez. 2014. Deep neural networks for small footprint text-dependent speaker verification. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4052–4056. IEEE.
- Shinji Watanabe, Michael Mandel, Jon Barker, Emmanuel Vincent, Ashish Arora, Xuankai Chang, Sanjeev Khudanpur, Vimal Manohar, Daniel Povey, Desh Raj, et al. 2020. Chime-6 challenge: Tackling multi-speaker speech recognition for unsegmented recordings. *arXiv preprint arXiv:2004.09249*.
- Aonan Zhang, Quan Wang, Zhenyao Zhu, John Paisley, and Chong Wang. 2019. Fully supervised speaker diarization. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6301–6305. IEEE.

# Contemporary LLMs and Literary Abridgement: An Analytical Inquiry

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

Within the framework of this study, several contemporary Large Language Models (ChatGPT, Gemini Pro, Mistral-Instruct and BgGPT) are evaluated in relation to their ability to generate abridged versions of literary texts. The analysis is based on 'The Ugly Duckling' by H. C. Andersen as translated into English, French and Bulgarian. The different scenarios of abridgement experimented with include zero-shot, one-shot, division into chunks and crosslingual (including chain-of-thought) abridgement. The resulting texts are evaluated both automatically and via human evaluation. The automatic analysis includes ROUGE and BERTScore as well as the ratios of a selection of readability-related textual features (e.g. number of words, type-to-token ratio) as pertaining to the original versus automatically abridged texts. Professionally composed abridged versions are regarded as gold standard. Following the automatic analysis, six selected best candidate texts per language are then evaluated by volunteers with university education in terms of textual characteristics of a more qualitative nature, such as coherence, consistency and aesthetic appeal.

**Keywords:** LLMs, literary abridgement, multilingual text generation

## 1 Introduction

The current work seeks to provide an overview of the ability of contemporary Large Language Models (LLMs) to generate abridged versions of literary works. As per the Merriam-Webster dictionary, 'abridged' means 'shortened or condensed, especially by the omission of words or passages'. Notably, abridgement makes literary texts accessible to audiences that would find it hard to read or work with the corresponding original texts, such as young children,

foreign language learners or people with learning disabilities. The term will not be used as synonymous to 'summary' due to both its particular relevance to the literary domain and its focus on overall simplification rather than merely reduction in size.

## 2 Background

Although literary abridgement by LLMs is not yet an established research topic, it implies several sets of abilities pertaining to the technology that are currently of marked academic interest, notably the use of long context, summarisation, and creative/literary writing.

### 2.1 LLMs and Long Context

A major limitation of contemporary LLMs is their imperfect ability to receive and make sense of large amounts of text. Through the tasks of multi-document question answering and key-value retrieval, Liu et al. (2023) evaluate LLM's general ability to use long contexts, revealing drawbacks such as position bias, i.e. the tendency of models to work better with information situated toward the beginning or end of a document (a.k.a the 'lost-in-the-middle' problem). They note that even extended-context models, such as LongChat-13B, are not characterised with better use of long context. In contrast, instruction fine-tuned models use contexts more efficiently. Different techniques to extend models' context window have been proposed, such as position interpolation, a type of minimal fine-tuning, in which position indices provided to transformer models are scaled down to accommodate the additional context (Chen et al., 2023). In relation to the task of machine translation (MT), Du et al. (2023) note that its quality diminishes significantly as document size increases, GPT-4 receiving

the highest BLUE scores among contemporary LLMs when long context is involved.

## 2.2 LLMs and Summarisation

One of the Natural Language Processing (NLP) tasks that most directly benefit from the use of extensive context is document summarisation (in particular, abstractive summarisation as it pertains to a text’s transformation rather than its mere reduction in size). [Chang et al. \(2024\)](#) divide a long document (over 100k tokens) into chunks and then merge them to derive full summaries. They experiment with merging the chunks hierarchically and incrementally and use textual coherence to evaluate the resulting summaries, thereby proposing an automatic metric of coherence. They attribute highest scores to GPT-4 and Claude 2 and to the practice of hierarchical merging. [Wu et al. \(2024\)](#) work around the aforementioned ‘lost-in-the-middle’ problem through an ‘extract-then-evaluate’ approach, in which they incrementally extract and concatenate key sentences from a document that result in the highest ROUGE score of the achieved summary.

Most state-of-the-art practices related to the task of summarisation pertain to the reduction in size of long, mostly news-based texts for the purpose of time efficiency whilst key information is preserved. [Xiao and Chen \(2023\)](#) focus on informativeness when applying evolutionary fine-tuning to news summarisation. [Zhang et al. \(2023\)](#) test ten LLMs’ summarisation performance and compare it to that of humans, concluding that instruction tuning provides a significant benefit and that human summaries tend to be more abstractive in nature (i.e. use paraphrasing rather than direct extraction) than LLM-generated counterparts. [Pu et al. \(2023\)](#)’s bold statement that (human-based) "summarization is (almost) dead" is based on the results of five discrete summarisation tasks, including crosslingual summarisation (CLS). Pairwise human evaluation rates LLMs as markedly better at the tasks than both humans and fine-tuned neural models, and particularly strong in terms of fluency and coherence. In contrast, LLM’s current performance in CLS is also tested by [Wang et al. \(2023a\)](#) and not left uncriticised. Based on CLS datasets and the ROUGE and BERTScore metrics, contemporary models such as GPT-4 and ChatGPT are evaluated as reach-

ing competitive but still worse zero-shot performance compared to a BART model that has been fine-tuned for the task. Open-source models such as Vicuna-13B are judged to outright lack zero-shot CLS ability. Additional experiments reveal that a chain-on-thought method of asking the model to first translate and then summarise (and vice-versa) a text helps improve performance.

## 2.3 LLMs and Creative Writing/Literature

The involvement of LLMs in creative writing as much as gives rise to philosophical questions about the nature of creativity. [Franceschelli and Musolesi \(2023\)](#) apply Margaret Boden’s theories of value, novelty and surprise to the function of LLMs, concluding that their creativity is by definition limited in nature and scope. Both due to ethical reasons and to a general opinion that LLMs’ current abilities are still lacking, their role in creative writing is often limited to subtasks such as plot outlines or character development. [Kreminski and Martens \(2022\)](#) systematise the potential of current LLMs to provide support for writers, providing guidelines for their effective use in the overcoming of ‘writer’s block’. User-friendly tools like *Story Centaur* ([Swanson et al., 2021](#)), which is based on LLMs’ few-shot abilities, have been developed to aid creative writers in their work by fulfilling narrowly framed tasks, such as the provision of a next sentence given the previous one and a ‘magic word’ to be incorporated. Also viewing LLMs as potential assistants in the creative writing process, [Shanahan and Clarke \(2023\)](#) use elaborate prompting strategies combined with fine-tuning of the temperature setting to collect textual samples from GPT-4 that they then evaluate qualitatively, basing themselves on an array of literary concepts including characterisation, imagery and use of idioms. They discuss the creativity of LLMs as analysable and multi-faceted albeit tightly dependent on the quality of underlying prompts. Other comprehensive studies on the topic include [Gómez-Rodríguez and Williams \(2023\)](#)’s evaluation of the creative writing abilities of a number of contemporary LLMs. The authors provide LLM models and several human writers with an identical creative writing task, and they apply to the issuing stories hu-

man evaluation based on established criteria in the domain, including coherence and the use of humour. They conclude that commercial LLMs perform comparably to human writers but do not match the latter in originality, and that the understanding of humour can be considered an emerging ability of LLMs.

Prior to the advancement to LLMs, translation of literary texts was seen as "the greatest challenge for MT" (Toral and Way, 2018) as it implies the reader's overall experience as opposed to a limited number of automatisable measures. Recently, Tencent AI Lab and China Literature Ltd. organised a shared task on discourse-level literary translation, thereby releasing a Chinese-English web novel corpus. Among the tested baseline systems, LLMs performed best by a significant margin based on both automatic and human evaluation (Wang et al., 2023b).

### 3 Methods

#### 3.1 Texts

The utilised source texts are published translations of 'The Little Duckling' (H. C. Andersen) into English<sup>1</sup>, French<sup>2</sup>, and Bulgarian<sup>3</sup>. An original work written in a language that is not discussed (Danish) is deliberately opted for in order to avoid the presence of both original and translated texts in the following experiments. In addition, up to four published abridged versions per language are used in the context of automatic experiments in order to define reference ratios of textual features between an original and abridged version. For the one-shot scenario, an original and abridged version of 'The Little Match Girl' (H. C. Andersen) in each language are utilised<sup>4</sup>. In order for the

<sup>1</sup>Andersen, Hans Christian. The Ugly Duckling. 1843. <https://pinkmonkey.com/dl/library1/tale120.pdf>.

<sup>2</sup>Andersen, Hans Christian. Le vilain petit canard. 1843. <https://touslescontes.com/biblio/contes.php?idcontes=158>.

<sup>3</sup>Andersen, Hans Christian. Groznoto patentse. Translated by Svetoslav Minkov, Chitanka, 1977. <https://chitanka.info/text/4819>.

<sup>4</sup>English full: Andersen, Hans Christian. *The Little Match Girl*. Short Story America, 1845. [https://shortstoryamerica.com/pdf\\_classics/andersen\\_little\\_match\\_girl.pdf](https://shortstoryamerica.com/pdf_classics/andersen_little_match_girl.pdf).

English abridged: Andersen, Hans Christian. *The Little Match Girl*. <https://fliphtml5.com/mcbeq/hrvp/basic>.

French full: Andersen, Hans Christian. *La petite fille*

relative impact of possible recognition of the text by LLMs to be tested, an alternative, non-published story, 'The Gift under the Bush'<sup>5</sup> is also used both in its original Bulgarian version and the author's own translations into the additional languages. Some models' context size restrictions did not allow for particular scenarios (typically, zero-shot) to be fulfilled on the respective full text. In this case, abridged versions were used as source texts (see Appendix C for details about the derivation of specific abridged versions by LLMs).

#### 3.2 Models

The models experimented with are Mistral-Instruct, BgGPT, Gemini Pro and ChatGPT (as based on GPT-3.5). Mistral-Instruct (7B) is an open-source model, developed by Mistral AI as a fine-tuned version of the original Mistral model, whose main characteristics include high inference speed and a sliding window attention mechanism (Jiang et al., 2023). Its context window comes at 32k tokens. BgGPT-7B-Instruct by INSAIT is based on Mistral-7B and fine-tuned with large amounts of textual data for the purpose of better understanding and production of Bulgarian text (INSAIT, 2024). Gemini Pro (600B) is a user-friendly version of the state-of-the-art Gemini model by Google DeepMind, which is documented to outperform GPT-4 in 30 out of 32 language benchmarks (Anil et al., 2023). It has a context window of 128k tokens. OpenAI's GPT-3.5 is the model behind the free and most commonly used version of ChatGPT in the moment of writing of this article. For the purpose of this project, Mistral-Instruct was deployed through the LM Studio interface<sup>6</sup>, Gemini through the Google AI Studio tool within the established free quota, and BgGPT and ChatGPT through their offi-

*aux allumettes*. <https://touslescontes.com/biblio/contes.php?idcontes=127>.

French abridged: Andersen, Hans Christian. *La petite fille aux allumettes*. <https://miladlh.com/wp-content/uploads/2020/11/La-Petite-Fille-aux-Allumettes.pdf>.

Bulgarian full: Andersen, Hans Christian. *Malkata kibritoprodavachka*. <https://chitanka.info/text/4826-malkata-kibritoprodavachka>.

Bulgarian abridged: Andersen, Hans Christian. 'Malkata kibritoprodavachka.' In *Prikazki ot tsyal svyat*, transl. Vasil Velchev, 2009.

<sup>5</sup>Stoupak, Stefan. The Gift under the Bush. Unpublished manuscript, 2024.

<sup>6</sup><https://lmstudio.ai/>

cial chatbot interfaces.

### 3.3 Abridgement Scenarios

This study seeks to test and compare the current ingrained capabilities of LLMs to generate abridged versions of literary texts. For the purpose, no extensive fine-tuning and prompt-engineering methods are applied. In addition, no definition of 'abridgement' is provided within prompts. The following discrete experimental settings are considered: zero-shot, chunking, one-shot and crosslingual. In the one-shot setting, an original and an abridged version of another text ('The Little Match Girl') are provided to the model as an example of the transformation it is expected to apply. In the chunking scenario, the original text is divided into several (typically, three) parts. Crosslingual experiments are conducted both in a zero-shot setting and via a simple chain-of-thought that asks the model to first translate and then provide an abridged version of the text (henceforth, 'chain-of-thought 1') and vice-versa (henceforth, 'chain-of-thought 2')<sup>7</sup>. Due to the possibility of the models having encountered 'The Little Duckling' during training, additional experiments are carried out using a text that has not been published before; which, however, is not coupled with a gold standard abridged version.

Experiments are carried out in English, French and Bulgarian. In the case of BgGPT, naturally only Bulgarian is used. The majority of Mistral experiments are discarded due to poor output quality<sup>8</sup>.

### 3.4 Evaluation

#### 3.4.1 Automatic Evaluation

A selection of ten automatic measures is applied to the generated texts. For each language, the range of ratios between a full text and its human-made abridged versions is taken as gold standard that the abridged versions are compared against. For instance, if the ratios between the number of words in the original English text and the four human-made

abridged English texts are between 2.0 and 10.0, the 'number of words' measure is marked positively for LLM-generated texts, for which it falls within this same range.

The ratios between the full and abridged versions of the text used in the one-shot setting are also included in the range. The same range of ratios is applied to the alternative text for the given language, as there is no professional abridged version of it. In the cases where abridged texts are derived from other abridged texts due to the models' context length restrictions, it is the ratios between the utilised human-made abridged text and the LLM-generated further-abridged text that are taken into account.

As both ROUGE and BERTScore inherently compare two texts, it is directly the scores that compare abridged to original texts that are calculated. ROUGE is a standard measure for automatically-generated textual summaries that typically considers the overlap between a newly generated and a gold standard summary (Lin, 2004). For the purpose of this work, ROUGE-1 recall is used to calculate the portion of individual words in an abridged version that are present in the associated original text.

BERTscore, often used as an improved alternative to ROUGE, compares two texts based on the cosine similarity of token embeddings, thus capturing closeness of meaning (Zhang et al., 2020). F1 values of the BERTScore comparing original and abridged texts are calculated, thus providing a balanced measure of the inclusion of relevant information in an abridged text and its conciseness.

Readability is a notion that refers to the general complexity of a given text and, by extension, to its potential modification or simplification, especially in view of a particular reader profile (traditionally, defined by grade level). Most established readability formulas make use of shallow characteristics that have proven to be good proxies of complexity, such as the average number of syllables per word or the average number of words per sentence, used within the Flesch Reading Ease Formula (DuBay, 2007). Recent studies, such as Feng et al. (2010) have sought to further systematise the atomic features used in readability measurement as well as to determine their in-

<sup>7</sup>For the full prompts used, please consult the following repository: <https://github.com/iglika88/Contemporary-LLMs-and-Literary-Abridgement/>

<sup>8</sup>including 'one-shot' and all experiments involving non-English languages, with the sole exception of crosslingual abridgement from Bulgarian to English

terconnectedness. For the purpose of this study, a set of readability-related features is used that aims at informativeness as well as balance between different textual aspects: length (total number of words, number of words per sentence, number of letters per word), vocabulary (type-to-token ratio, concreteness as per Brysbaert et al. (2014)<sup>9</sup>, words outside of a determined frequency list), syntax (ratio of content to function words) and discourse (presence of anaphora-denoting words).

### 3.4.2 Human Evaluation

Six LLM-generated texts per language are selected for the human evaluation survey. They are the texts rated most highly by the automatic evaluation process i.e. the ones with the highest number of characteristics that fall within the gold standard range. In cases of equal scores, a variety between models and generation scenarios is sought. Four versions of the survey per language were composed, each of them consisting of two texts to evaluate. One of the two texts was also present in another version, in order to allow for a calculation of agreement<sup>10</sup>. A minimum of one participant per version and per language (native or fluent speaker with a university background) completed the survey.

The general categories evaluated in the survey are: understandability, correctness, consistency, textual coherence and aesthetic appeal. The respondents were offered a scale of 4 ('no', 'mostly no', 'mostly yes' and 'yes') and also encouraged to leave comments in the form of free text. The protocol's overall form is adapted from Mousavi et al. (2022)<sup>11</sup>.

Cohen's Kappa coefficient was calculated for the texts that are present within two versions of the survey. The response values were taken as categorical. In cases of more than two participants, Fleiss' Kappa (Fleiss et al., 1969) was also calculated.

<sup>9</sup>applicable only to English text

<sup>10</sup>For a breakdown of the texts, please refer to Appendix B. For the full texts included in the survey, please consult <https://github.com/iglika88/Contemporary-LLMs-and-Literary-Abridgement/>

<sup>11</sup>For the entire protocol, please refer to <https://github.com/iglika88/Contemporary-LLMs-and-Literary-Abridgement/>

## 4 Results

### 4.1 Automatic Evaluation

A model's performance is defined as the percentage of examined textual characteristics that fall within the range defined by the human-made abridged texts, as elaborated in Section 3.4.1. For instance, the study contains 24 texts generated by ChatGPT. In total, they are evaluated in terms of 226 characteristics, out of which 93 fall within the defined range, thus giving ChatGPT a score of 41%.

Observable tendencies related to the discussed atomic textual characteristics include too short length in relation to crosslingual and zero-shot generation scenarios (an exception being the Bulgarian language, for which zero-shot generation renders excessively long text). Also, the process of crosslingual generation results in a high percentage of words not appearing in the respective language's frequency list. Some characteristics, particularly ROUGE, BERTScore and type-to-token ratio, score particularly weakly in relation to the French language.

Results are further summarised in the following subsections. For the detailed results of the automatic evaluation, please refer to Appendix C.

#### 4.1.1 Performance by language

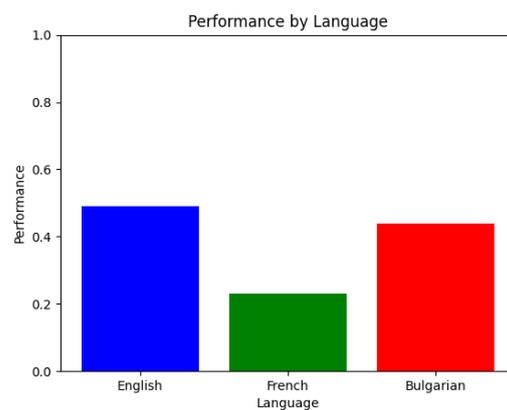


Figure 1: Performance by language

As shown in Figure 1, the highest performance is understandably attributable to English, somewhat surprisingly followed by the lower-resource language, Bulgarian.

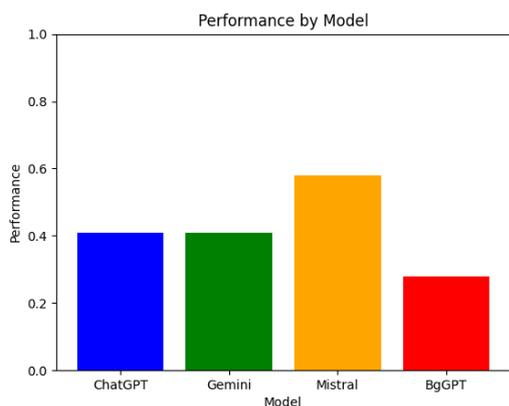


Figure 2: Performance by model

#### 4.1.2 Performance by model

The model that scores highest is Mistral. However, it is to be kept in mind that only a limited number of experiments were carried out using this model, and that they were all in the strongest performing language, English. ChatGPT and Gemini demonstrate equal global performance, and BgGPT comes last (see Figure 2).

#### 4.1.3 Performance by generation scenario

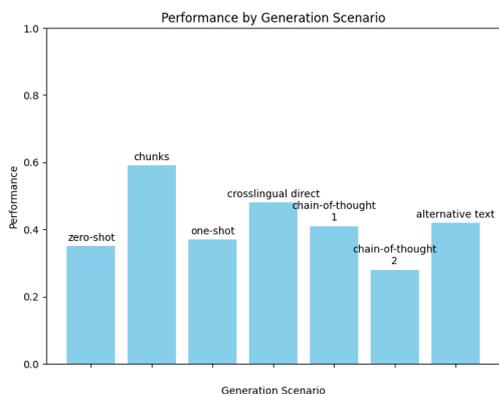


Figure 3: Performance by scenario

Top performance is exhibited by the 'chunks' and 'crosslingual: direct' abridgement scenarios (see Figure 3). In contrast, 'crosslingual: chain-of-thought 2' abridgement scores lowest. Zero-shot performance is in fact higher for the alternative text, showing that there is no significant influence of the text being present in training data on the models' performance.

#### 4.1.4 Performance by model and language

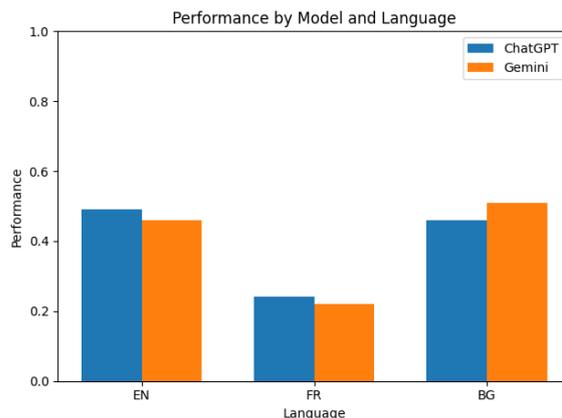


Figure 4: Performance by model and language

ChatGPT is observed to outperform Gemini in English and French, but not in Bulgarian (see Figure 4). The Mistral and bgGPT models are naturally excluded from this evaluation, as each of them addresses only a single language.

#### 4.1.5 Performance by scenario and language

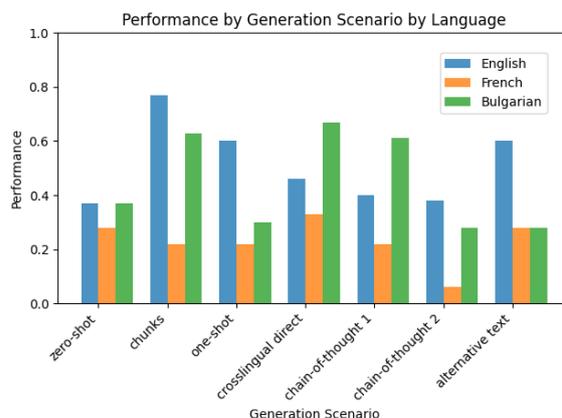


Figure 5: Performance by scenario and language

The one-shot setting in relation to both the primary and alternative text is high for the English language (see Figure 5). Crosslingual scenarios work best for Bulgarian, likely speaking of a benefit arising from use of the originally input English text.

#### 4.1.6 Performance by scenario and model

As seen in Figure 6, ChatGPT outperforms Gemini in relation to the 'chunks', 'one-shot'

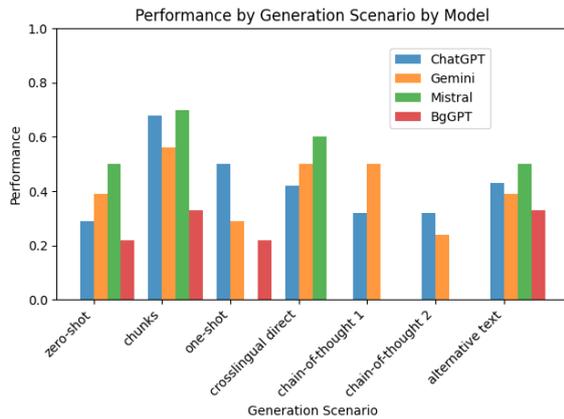


Figure 6: Performance by scenario and model

and 'crosslingual chain-of-thought 2' scenarios as well as with the alternative text. BgGPT's performance is the most uniform one between generation scenarios.

## 4.2 Human Evaluation

### 4.2.1 English Texts

Three participants responded to version 2, and one to each other version of the English survey. The first repeated text received low agreement per both Fleiss' Kappa and Cohen's Kappa, with the exception of the participant who responded to version one and the second participant who responded to version 2, who were in fair agreement. The second repeated text also received low Cohen's Kappa agreement.

Two texts received positive answers ('Yes' or 'Mostly Yes') for all categories of the survey: 'ChatGPT: crosslingual chain-of-thought 2' and 'ChatGPT: crosslingual direct'. 'Gemini: alternative' followed with 94.44%, 'Gemini: one-shot' with 81.48%, 'Mistral: chunks' with 55.56% and 'ChatGPT: chunks' with 50%.

Problems with understandability were noted in relation to 'Mistral: chunks' and 'ChatGPT: chunks'. The Mistral text was rated negatively for all aspects of correctness, whilst the 'ChatGPT: chunks' and 'Gemini: one-shot' ones were judged as having non-optimal structure. 'ChatGPT: crosslingual chain-of-thought 2' was seen as using awkward vocabulary (e.g. 'poultry yard') and unlikely parts of speech to render meaning. Inconsistencies in register and style were noted for the Mistral text and 'Gemini: one-shot'. In turn, 'ChatGPT: chunks' demonstrates inconsistency in the presented informa-

tion (e.g. an action taking place twice) and use of pronouns (the duckling being referred to as 'it' and 'he' in different parts of the story). 'ChatGPT: chunks' also received a fully negative rating for coherence. Aesthetic characteristics (notably, textual length, pacing and engagement) were commonly marked negatively for all texts except ChatGPT's two crosslingual ones.

### 4.2.2 French Texts

Two participants responded to version 3 of the survey and one to each of the other three. The first repeated text is associated with low agreement as per Cohen's Kappa, and the second one, which was evaluated by three people, received low Fleiss' Kappa as well as low Cohen's Kappa with the exception of the second participant who responded to version 3 and the participant who responded to version 4 (fair agreement).

The most highly rated text was 'Gemini: crosslingual direct' (100% positive answers), followed by 'ChatGPT: zero-shot' (94.12%), 'Gemini: zero-shot' (90.74%), 'ChatGPT: chunks' (88.89%), 'ChatGPT: alternative' (82.86%) and 'Gemini: crosslingual chain-of-thought 2' (69.44%).

The texts' understandability was rated fully positively, with a mention of occasional complex vocabulary ('Gemini: zero-shot') and grammar ('ChatGPT: alternative'). Marked issues pertaining to correctness included the type of text not resembling an abridged story but rather a 'fable' ('Gemini: zero-shot') or just a 'short story' ('ChatGPT: chunks') and wrong use of tenses ('Gemini: zero-shot'). The structure of 'ChatGPT: one-shot' was the only one marked negatively, whilst it was explicitly noted that in 'Gemini: crosslingual direct', "all the [ugly duckling's] adventures are present". Consistency of style was marked negatively for 'ChatGPT: one-shot' and 'Gemini: crosslingual chain-of-thought 1'. Within the latter, grammar was perceived to be too simple as compared to vocabulary. For 'ChatGPT: alternative', information was also marked as inconsistent. Problems with transitions were noted in 'ChatGPT: alternative', 'Gemini: zero-shot' and 'Gemini: crosslingual chain-of-thought 1', and the last was also claimed to include problems with anaphora use. When it comes to aes-

thetic qualities, 'ChatGPT: alternative', 'Gemini: zero-shot', 'ChatGPT: chunks' and 'Gemini: crosslingual chain of thought 1' received negative scores for engagement, comments referring to the texts as 'non-fluid' and 'frustratingly' weakly developed. Problems of pacing and textual length (particularly, texts being too short) were also brought forward.

### 4.2.3 Bulgarian Texts

Four participants responded to version 1 of the survey and one person each for the other three versions. Cohen's Kappa for the repeated text in versions 1 and 2 is fair between participants 1 and 5 and low for the rest; Fleiss' Kappa is low. Cohen's Kappa for the other repeated text is fair.

The 'Gemini: chunks' text was rated most highly, with 90.28% positive answers, followed by 'Gemini: crosslingual train-of-thought 1' and 'ChatGPT: zero-shot' (88.89%), 'ChatGPT: one-shot' (69.45%), 'ChatGPT: chunks' (54.45%) and 'Gemini: crosslingual direct' (44.45%).

The texts' understandability was generally rated highly. In contrast, correctness received a high number of negative answers, particularly in relation to vocabulary, grammar, and structure. For instance, vocabulary in 'Gemini: chunks' was judged to often be wrong, wrongly used or seemingly translated, the verbs in 'ChatGPT: chunks' were said to often be wrongly interpreted in terms of transitiveness, and 'Gemini: chunks' felt as if it were 'mixed with other stories'. Consistency was marked negatively for the 'ChatGPT: chunks' and 'ChatGPT: one-shot' texts. Underlined problems of coherence included excessive repetition, wrong use of anaphora and, in the case of 'ChatGPT: chunks', confusing transitions. Aesthetics was mostly rated positively; the most common problem being 'length' ('Gemini: crosslingual chain-of-thought 1' was the only text referred to as 'too long' rather than 'too short'). The 'ChatGPT: chunks' text was noted to be lacking descriptions and character interaction.

## 5 Discussion

Although Bulgarian texts received comparatively lower scores in the conducted human evaluation, they were shown to be mostly competitive to counterparts in more highly resourced

languages. Interestingly, they also tended to demonstrate different shortcomings compared with texts in English and French, such as excessive textual length.

Crosslingually derived texts were rated very highly by participants, notably occupying first place in the cases of French and English. Texts derived through the 'chunks' scenario were judged to have problems in relation to information and transitions, which leads us to hypothesise that an application of [Chang et al. \(2024\)](#)'s method of hierarchical merging would be of significant benefit.

The ChatGPT and Gemini models performed better than the smaller but instruction-tuned Mistral and BgGPT; however, the gap was not striking, Mistral-generated texts notably performing high in the conducted automatic evaluation.

The fact that agreement between participants in the survey is low speaks of high subjectivity, which in turn implies that the texts were mostly lacking obvious, objective drawbacks.

## 6 Conclusion and Future Directions

Four contemporary LLMs of different sizes and statuses of use were evaluated for their ability to provide abridged versions of a literary text. Three discrete languages were regarded: English, a relatively high-resourced language (French) and a relatively low-resourced language (Bulgarian).

Whilst English-language texts expectedly demonstrate superior quality, models such as ChatGPT and Gemini also perform competitively in other languages, whilst demonstrating different weaknesses in relation to different generation scenarios. Particularly, in a relatively low-resource language like Bulgarian, high quality text can be achieved if the models' limitations in terms of context length are overcome.

This study is an analytical inquiry into the current abilities of LLMs to generate abridged versions of literary texts on the basis of their original training data. These abilities are likely to be improved following additional training on relevant datasets as well as more elaborate prompting techniques.

A natural continuation of the presented study would be the exploration of abridgement by LLMs in relation to a variety of texts; this time

with a focus on the models and abridgement scenarios that proved strongest.

## 7 Limitations

It is important to note that depth rather than width was opted for in the present study and its conclusions are mostly based on a single literary text. Therefore, key characteristics of original literary texts such as length and genre are disregarded as variables.

In addition, abridgement is considered as a general term and is not further broken down, such as based on targeted audience (e.g. children of a certain age). It should also be noted that application of the study’s methods to additional texts is likely to necessitate refinement of the automatic evaluation metrics, such as type-to-token ratio, which is known to be highly dependent on a text’s size. Finally, one respondent to the survey brought forward a text’s similarity to a ‘short summary’ as a negative trait, whilst another one claimed that the text was a little ‘too vivid’ to be a ‘summary’; which leads us to conclude that the term ‘abridged version’ is highly open to interpretation and that the survey would have benefited from a short definition of what is meant by it.

## References

- R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, D. Silver, S. Petrov, M. Johnson, I. Antonoglou, J. Schrittwieser, A. Glaese, J. Chen, E. Pitler, and O. ... Vinyals. 2023. Gemini: A family of highly capable multimodal models. *arXiv*, 2312.11805.
- Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, 46(3):904–911.
- Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. Boookscore: A systematic exploration of book-length summarization in the era of llms.
- Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023. Extending context window of large language models via positional interpolation.
- Zefeng Du, Wenxiang Jiao, Longyue Wang, Chenyang Lyu, Jianhui Pang, Leyang Cui, Kaiqiang Song, Derek Wong, Shuming Shi, and Zhaopeng Tu. 2023. On extrapolation of long-text translation with large language models.
- William H. DuBay. 2007. *The Classic Readability Studies*. ERIC Clearinghouse.
- Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. 2010. A comparison of features for automatic readability assessment. In *23rd International Conference on Computational Linguistics (COLING 2010), Poster Volume*, pages 276–284.
- Joseph L Fleiss, Jacob Cohen, and Brian S Everitt. 1969. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 72(5):323–327.
- Giorgio Franceschelli and Mirco Musolesi. 2023. On the creativity of large language models.
- Carlos Gómez-Rodríguez and Paul Williams. 2023. A confederacy of models: a comprehensive evaluation of llms on creative writing.
- INSAIT. 2024. The latest advancements in llm: A comprehensive overview.
- A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, and W. ... El Sayed. 2023. Mistral 7b. *arXiv*, 2310.06825.
- Max Kreminski and Chris Martens. 2022. Unmet creativity support needs in computationally supported creative writing. In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*, pages 74–82, Dublin, Ireland. Association for Computational Linguistics.
- Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts.
- Seyed Mahed Mousavi, Gabriel Roccabruna, Michela Lorandi, Simone Caldarella, and Giuseppe Riccardi. 2022. Evaluation of response generation models: Shouldn’t it be shareable and replicable? In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 136–147, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead.
- Murray Shanahan and Catherine Clarke. 2023. Evaluating large language model creativity from a literary perspective.

- Ben Swanson, Kory Mathewson, Ben Pietrzak, Sherol Chen, and Monica Dinalescu. 2021. *Story centaur: Large language model few shot learning as a creative writing tool*. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 244–256, Online. Association for Computational Linguistics.
- Antonio Toral and Andy Way. 2018. *What level of quality can neural machine translation attain on literary text?*
- Jiaan Wang, Yunlong Liang, Fandong Meng, Beiqi Zou, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2023a. *Zero-shot cross-lingual summarization via large language models*.
- Longyue Wang, Zhaopeng Tu, Yan Gu, Siyou Liu, Dian Yu, Qingsong Ma, Chenyang Lyu, Lit-ing Zhou, Chao-Hong Liu, Yufeng Ma, Weiyu Chen, Yvette Graham, Bonnie Webber, Philipp Koehn, Andy Way, Yulin Yuan, and Shuming Shi. 2023b. *Findings of the wmt 2023 shared task on discourse-level literary translation: A fresh orb in the cosmos of llms*.
- Yunshu Wu, Hayate Iso, Pouya Pezeshkpour, Nikita Bhutani, and Estevam Hruschka. 2024. *Less is more for long document summary evaluation by llms*.
- Le Xiao and Xiaolin Chen. 2023. *Enhancing llm with evolutionary fine tuning for news summary generation*.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. *Bertscore: Evaluating text generation with bert*.
- Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2023. *Benchmarking large language models for news summarization*.

## Appendix A Automatically Evaluated Textual Features

| Textual feature                    | Definition   |
|------------------------------------|--|
| Total number of words              | The number of words within the given text  |
| Words per sentence                 | The average number of words per sentence in the text   |
| Letters per word                   | The average number of letters per word in the text   |
| Words not in frequency list        | The number of words in the text that are not part of a defined frequency list. For English, the Dale-Chall list is considered. For Bulgarian and French, respectively, the top 3000 words from the Open Subtitles <sup>1</sup> and the Leeds Internet-FR Corpus <sup>2</sup> are taken.  |
| Type-to-token ratio                | The word-based (as opposed to lemma-based) ratio of types and tokens in the text   |
| Concreteness                       | The average concreteness of the words found in Brysbaert's concreteness list   |
| Anaphora-denoting words            | The percentage of anaphora-related words in the text. For each language, these words are a defined set of definite articles, personal pronouns, demonstrative pronouns, relative pronouns, indefinite pronouns and adverbs of time and place   |
| Ratio of content to function words | The approximate ratio of the words that carry semantical significance and the words that denote grammatical features in the text. For English, the CMU Pronouncing Dictionary for function words as available in Python's <i>nltk</i> library is used. For French and Bulgarian, part-of-speech tagging is applied to set apart the two kinds of words. Determiners, pronouns, conjunctions and adpositions are considered to be function words. |
| ROUGE                              | The ROUGE-1 recall value between an original and abridged text are taken.  |
| BERTScore                          | The F1 BERTScore between an original and abridged text are considered.   |

<sup>1</sup>[https://en.wiktionary.org/wiki/Wiktionary:Frequency\\_lists/Bulgarian\\_wordlist](https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists/Bulgarian_wordlist)

<sup>2</sup><http://corpus.leeds.ac.uk/list.html>

## Appendix B Human Evaluation: Evaluated Texts

English

|               | Text 1  | Text 2                            |
|---------------|---|-----------------------------------|
| Participant 1 | ChatGPT: crosslingual crain-of-thought 2 (BG) | ChatGPT: chunks                   |
| Participant 2 | ChatGPT: crosslingual crain-of-thought 2 (BG) | Gemini: one-shot                  |
| Participant 3 | Mistral: chunks                               | ChatGPT: crosslingual direct (BG) |
| Participant 4 | Mistral: chunks                               | Gemini: alternative               |

French

|               | Text 1               | Text 2                                       |
|---------------|----------------------|--|
| Participant 1 | ChatGPT: alternative | ChatGPT: one-shot                            |
| Participant 2 | ChatGPT: alternative | Gemini: crosslingual direct (EN)             |
| Participant 3 | Gemini: zero-shot    | Gemini: crosslingual chain-of-thought 1 (EN) |
| Participant 4 | Gemini: zero-shot    | ChatGPT: chunks                              |

Bulgarian

|               | Text 1            | Text 2                                       |
|---------------|-------------------|--|
| Participant 1 | ChatGPT: chunks   | Gemini: chunks                               |
| Participant 2 | ChatGPT: chunks   | Gemini: crosslingual direct (EN)             |
| Participant 3 | ChatGPT: one-shot | Gemini: crosslingual chain-of-thought 1 (EN) |
| Participant 4 | ChatGPT: one-shot | ChatGPT: zero-shot                           |

## Appendix C Automatic Evaluation: Detailed Results

English

|                         | human-made 1 <sup>12</sup> | human-made 2 <sup>13</sup> | human-made 3 <sup>14</sup> | human-made 4 <sup>15</sup> | 'Little Match Girl' <sup>16</sup> | ChatGPT: zero-shot | ChatGPT: chunks <sup>17</sup> |
|-------------------------|----------------------------|----------------------------|----------------------------|----------------------------|-----------------------------------|--------------------|-------------------------------|
| total words             | 5.05                       | 17.15                      | 4.25                       | 3.83                       | 1.46                              | 23.3               | <b>10.01</b> <sup>18</sup>    |
| words per sentence      | 2.89                       | 3.15                       | 1.91                       | 2.77                       | 1.65                              | 1.64               | <b>1.67</b>                   |
| letters per word        | 1.02                       | 0.98                       | 0.99                       | 1.03                       | 1.04                              | 0.85               | 0.91                          |
| words not in freq. list | 5.01                       | 10.16                      | 3.75                       | 5.69                       | 1.9                               | 12.49              | <b>6.8</b>                    |
| TTR                     | 0.75                       | 0.52                       | 0.7                        | 1.29                       | 1.05                              | 0.49               | <b>0.58</b>                   |
| concreteness            | 0.94                       | 0.91                       | 0.97                       | 0.91                       | 0.91                              | 0.99               | <b>0.97</b>                   |
| anaphora words          | 0.8                        | 1.23                       | 1.01                       | 1.57                       | 0.91                              | <b>1.15</b>        | <b>1.05</b>                   |
| cont./funct. words      | 1.04                       | 0.53                       | 0.86                       | 0.74                       | 1.63                              | <b>1.58</b>        | <b>1.48</b>                   |
| ROUGE                   | 0.67                       | 0.58                       | 0.58                       | 0.64                       | 0.4                               | <b>0.6</b>         | <b>0.61</b>                   |
| BERTScore               | 0.84                       | 0.82                       | 0.83                       | 0.82                       | 0.82                              | 0.81               | <b>0.82</b>                   |

|             | ChatGPT: one-shot | ChatGPT: crossling. direct (FR) <sup>19</sup> | ChatGPT: crossling. direct (BG) <sup>20</sup> | ChatGPT: crossling. chain-of-thought 1 (FR) <sup>21 22</sup> | ChatGPT: crossling. chain-of-thought 1 (BG) <sup>23</sup> | ChatGPT: crossling. chain-of-thought 2 (FR) <sup>24 25</sup> | ChatGPT: crossling. chain-of-thought 2 (BG) <sup>26</sup> |
|-------------|-------------------|---|---|--|---|--|---|
| total words | 28.59             | 23.15   | <b>8.64</b>                                   | 22.6   | 53.15   | 37   | <b>9.53</b>   |

<sup>12</sup>Andersen, Hans Christian. The Ugly Duckling. Edited by Lynne Bradbury, Ladybird Books, adapted 1997.

<sup>13</sup>Andersen, Hans Christian. The Ugly Duckling. British Council. <https://learnenglishkids.britishcouncil.org/sites/kids/files/attachment/story-time-the-ugly-duckling-transcript.pdf>.

<sup>14</sup>Andersen, Hans Christian. The Ugly Duckling. [https://www.joliet86.org/assets/1/6/The\\_ugly\\_duckling.pdf](https://www.joliet86.org/assets/1/6/The_ugly_duckling.pdf).

<sup>15</sup>Andersen, Hans Christian. The Ugly Duckling. Edited by Maryann Dobeck, Parragon, 2009.

<sup>16</sup>Andersen, Hans Christian. The Little Match Girl. <https://fliphtml5.com/mcbeq/hrvp/basic>.

<sup>17</sup>The text was divided into 3 closely equal chunks. Depending on the language and model, some texts needed to be broken down into more chunks, in which case the number will be indicated.

<sup>18</sup>characteristics of the LLM-generated texts that fall within the gold standard range are marked in **bold**

<sup>19</sup>The text is directly abridged from the indicated language (here, French)

<sup>20</sup>The source text is not the full version but the abridged version 'human-made 1'

<sup>21</sup>The text is first translated from the source language (here, French) and then abridged using chain-of-thought prompts

<sup>22</sup>The source text is not the full version but the abridged version 'human-made 1'

<sup>23</sup>The source text is not the full version but the abridged version 'human-made 1'

<sup>24</sup>The source text is not the full version but the abridged version 'human-made 1'

<sup>25</sup>The text is first abridged in the source language (here, French) and then translated into the target language

<sup>26</sup>The source text is not the full version but the abridged version 'human-made 1'

|                         |             |             |             |             |             |             |             |
|-------------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| words per sentence      | <b>1.8</b>  | 1.61        | <b>1.87</b> | <b>1.73</b> | 1.46        | 1.35        | <b>1.97</b> |
| letters per word        | 0.94        | 0.81        | 0.94        | 0.92        | 0.83        | 0.89        | 0.92        |
| words not in freq. list | 20.59       | 12.49       | <b>5.82</b> | 16.21       | 26.28       | 29.31       | <b>6</b>    |
| TTR                     | 0.47        | 0.46        | <b>0.57</b> | 0.5         | 0.43        | 0.48        | <b>0.53</b> |
| concreteness            | <b>0.96</b> | 1.02        | 0.99        | <b>0.95</b> | 1.02        | 0.99        | 0.98        |
| anaphora words          | <b>1.14</b> | <b>1.34</b> | <b>0.83</b> | <b>1.24</b> | <b>1.11</b> | <b>0.84</b> | <b>0.87</b> |
| cont./funct. words      | <b>1.5</b>  | <b>1.4</b>  | <b>1.11</b> | <b>1.44</b> | <b>1.27</b> | 2.2         | <b>1</b>    |
| ROUGE                   | 0.71        | <b>0.51</b> | <b>0.6</b>  | <b>0.5</b>  | <b>0.45</b> | <b>0.66</b> | <b>0.6</b>  |
| BERTScore               | <b>0.82</b> | 0.81        | 0.81        | 0.81        | 0.8         | 0.8         | 0.81        |

|                         | ChatGPT: alternative <sup>27</sup> | Gemini: zero-shot | Gemini: chunks | Gemini: one-shot | Gemini: crossling. direct (FR) | Gemini: crossling. direct (BG) | Gemini: crossling. chain-of-thought 1 (FR) |
|-------------------------|------------------------------------|-------------------|----------------|------------------|--------------------------------|--------------------------------|--|
| total words             | <b>3.87</b>                        | 24.35             | <b>13.88</b>   | <b>16.41</b>     | <b>16.55</b>                   | 23.01                          | 24.99                                      |
| words per sentence      | 0.78                               | 1.56              | <b>1.96</b>    | <b>1.71</b>      | 1.39                           | 1.83                           | 1.23                                       |
| letters per word        | 0.92                               | 0.85              | 0.87           | 0.94             | 0.84                           | 0.81                           | 0.93                                       |
| words not in freq. list | <b>3.28</b>                        | 13.61             | <b>7.86</b>    | 11.21            | <b>10.03</b>                   | 12.49                          | 16.21                                      |
| TTR                     | <b>0.79</b>                        | 0.47              | <b>0.53</b>    | <b>0.55</b>      | 0.51                           | 0.51                           | 0.49                                       |
| concreteness            | 1.04                               | 1.05              | 1.02           | <b>0.93</b>      | 1.01                           | 1.03                           | <b>0.96</b>                                |
| anaphora words          | 1.67                               | <b>1.1</b>        | <b>1.29</b>    | <b>1.06</b>      | <b>0.91</b>                    | <b>1.51</b>                    | <b>1.07</b>                                |
| cont./funct. words      | <b>1.43</b>                        | <b>1.57</b>       | <b>1.15</b>    | <b>1.42</b>      | <b>1.3</b>                     | 1.95                           | <b>1.16</b>                                |
| ROUGE                   | <b>0.63</b>                        | <b>0.53</b>       | <b>0.55</b>    | <b>0.64</b>      | <b>0.53</b>                    | <b>0.46</b>                    | <b>0.57</b>                                |
| BERTScore               | 0.85                               | 0.64              | 0.7            | 0.65             | 0.8                            | 0.8                            | <b>0.82</b>                                |

<sup>27</sup>The alternative text ('The Gift under the Bush') is abridged in a zero-shot setting

|                            | Gemini:<br>crossling.<br>chain-of-<br>thought 1<br>(BG) | Gemini:<br>crossling.<br>chain-of-<br>thought 2<br>(FR) | Gemini:<br>crossling.<br>chain-of-<br>thought 2<br>(BG) | Gemini:<br>alterna-<br>tive | Mistral:<br>zero-shot | Mistral:<br>chunks <sup>28</sup> | Mistral:<br>crossling.<br>direct<br>(BG) <sup>29</sup> |
|----------------------------|---|---|---|-----------------------------|-----------------------|----------------------------------|--|
| total words                | 25.16   | 41.02   | 53.15   | <b>8.11</b>                 | <b>4.32</b>           | <b>8.56</b>                      | <b>2.31</b>  |
| words per<br>sentence      | 1.38  | 1.39  | 1.49  | 0.83                        | 0.64                  | 1.5                              | 0.42   |
| letters per<br>word        | 0.87  | 0.88  | 0.93  | <b>0.99</b>                 | 0.97                  | <b>1</b>                         | 0.89   |
| words not in<br>freq. list | 13.61   | 20.05   | 38.1  | <b>7.29</b>                 | <b>2.92</b>           | <b>6.4</b>                       | 1.88   |
| TTR                        | 0.47  | 0.44  | 0.42  | <b>0.78</b>                 | <b>0.68</b>           | <b>0.59</b>                      | <b>0.75</b>  |
| concreteness               | 1.01  | <b>0.96</b>   | 1.03  | 1                           | 0.99                  | 0.98                             | 1.04   |
| anaphora<br>words          | <b>1.23</b>   | <b>1.31</b>   | <b>1.11</b>   | <b>1.36</b>                 | 1.69                  | <b>1</b>                         | <b>1.24</b>  |
| cont./funct.<br>words      | <b>1.18</b>   | <b>0.96</b>   | <b>1.53</b>   | <b>1.43</b>                 | <b>1.48</b>           | <b>1.59</b>                      | <b>1.07</b>  |
| ROUGE                      | <b>0.55</b>   | <b>0.5</b>  | 0.74  | <b>0.63</b>                 | <b>0.61</b>           | <b>0.61</b>                      | <b>0.38</b>  |
| BERTScore                  | 0.81  | 0.8   | 0.81  | <b>0.83</b>                 | 0.86                  | 0.81                             | <b>0.83</b>  |

|                            | Mistral:<br>alterna-<br>tive |
|----------------------------|------------------------------|
| total words                | <b>5.53</b>                  |
| words per<br>sentence      | 0.84                         |
| letters per<br>word        | 0.95                         |
| words not in<br>freq. list | <b>5.02</b>                  |
| TTR                        | <b>0.7</b>                   |
| concreteness               | 1                            |
| anaphora<br>words          | <b>1.33</b>                  |
| cont.<br>/funct.<br>words  | <b>1.41</b>                  |

<sup>28</sup>The text was divided into 5 chunks.

<sup>29</sup>The source text is not the full version but the abridged version 'human-made 1'

|           |      |
|-----------|------|
| ROUGE     | 0.69 |
| BERTScore | 0.85 |

French

|                         | human-made <sup>1</sup> <sup>30</sup> | human-made <sup>2</sup> <sup>31</sup> | human-made <sup>3</sup> <sup>32</sup> | 'Little Match Girl' <sup>33</sup> | ChatGPT: zero-shot | ChatGPT: chunks | ChatGPT: one-shot |
|-------------------------|---------------------------------------|---------------------------------------|---------------------------------------|-----------------------------------|--------------------|-----------------|-------------------|
| total words             | 2.48                                  | 4.55                                  | 3.38                                  | 1.06                              | 32                 | 11.73           | 21.81             |
| words per sentence      | 2.22                                  | 0.96                                  | 0.92                                  | 0.95                              | 0.76               | 0.84            | <b>1.42</b>       |
| letters per word        | 1.1                                   | 0.96                                  | 0.96                                  | 1                                 | 0.95               | 0.95            | <b>0.97</b>       |
| words not in freq. list | 2.34                                  | 3.83                                  | 2.67                                  | 1.17                              | 23.42              | 8.6             | 14.05             |
| TTR                     | 0.87                                  | 0.71                                  | 0.74                                  | 1.03                              | 0.54               | 0.63            | 0.6               |
| anaphora words          | 1.08                                  | 1.13                                  | 1.09                                  | 0.92                              | 1.2                | <b>1.03</b>     | 0.81              |
| cont./funct. words      | 0.8                                   | 0.72                                  | 1.04                                  | 1.05                              | 0.44               | <b>0.91</b>     | <b>0.73</b>       |
| ROUGE                   | 0.55                                  | 0.48                                  | 0.53                                  | 0.59                              | <b>0.58</b>        | <b>0.54</b>     | <b>0.59</b>       |
| BERTScore               | 0.74                                  | 0.71                                  | 0.73                                  | 0.8                               | 0.62               | 0.63            | 0.66              |

|                    | ChatGPT: crossling. direct (EN) | ChatGPT: crossling. chain-of-thought 1 (EN) | ChatGPT: crossling. chain-of-thought 2 (EN) | ChatGPT: alternative | Gemini: zero-shot | Gemini: chunks | Gemini: one-shot |
|--------------------|---------------------------------|---|---|----------------------|-------------------|----------------|------------------|
| total words        | 0.78                            | 39.25                                       | 17.12                                       | 9.6                  | 22.3              | 13.89          | 3.71             |
| words per sentence | 0.64                            | 0.58  | 0.82  | 0.68                 | <b>1.12</b>       | 0.81           | 0.58             |
| letters per word   | 1.04                            | 0.93  | 0.87  | <b>0.97</b>          | <b>0.98</b>       | 0.87           | 0.93             |

<sup>30</sup>Andersen, Hans Christian. Le vilain petit canard. <https://data.over-blog-kiwi.com/1/11/17/78/20210801/obs730b8ie-vilain-petit-canard-tapuscrit.pdf>.

<sup>31</sup>Andersen, Hans Christian. Le vilain petit canard. BIGBEN Kids. <https://www.bigben.fr/wp-content/uploads/2021/10/HistoireLeVilainPetitCanard.pdf>.

<sup>32</sup>Andersen, Hans Christian. Le vilain petit canard. <https://bloc-note.ac-reunion.fr/9741309e/files/2020/03/0-conte-le-vilain-petit-canard.pdf>.

<sup>33</sup>Andersen, Hans Christian. La petite fille aux allumettes. <https://miladlh.com/wp-content/uploads/2020/11/La-Petite-Fille-aux-Allumettes.pdf>.

|                         |             |       |             |             |             |             |       |
|-------------------------|-------------|-------|-------------|-------------|-------------|-------------|-------|
| words not in freq. list | 0.91        | 25.55 | 10.81       | 7.2         | <b>2.93</b> | 10.81       | 25.54 |
| TTR                     | 1.18        | 0.47  | 0.54        | 0.66        | <b>0.79</b> | 0.54        | 0.47  |
| anaphora words          | <b>1.11</b> | 1.34  | <b>1.07</b> | <b>1</b>    | 1.33        | <b>1.07</b> | 1.34  |
| cont./funct. words      | 1.5         | 0.56  | 0.71        | <b>1.03</b> | 1.33        | 0.71        | 0.56  |
| ROUGE                   | 0.36        | 0.63  | 0.45        | 0.65        | 0.34        | 0.45        | 0.63  |
| BERTScore               | <b>0.77</b> | 0.61  | 0.63        | <b>0.71</b> | 0.69        | 0.63        | 0.61  |

|                         | Gemini: crossling. direct (EN) | Gemini: crossling. chain-of-thought 1 (EN) | Gemini: crossling. chain-of-thought 2 (EN) | Gemini: alternative |
|-------------------------|--------------------------------|--|--|---------------------|
| total words             | 13.26                          | 28.31                                      | 22.82                                      | 13.95               |
| words per sentence      | <b>1.03</b>                    | <b>1.29</b>                                | 0.44                                       | 0.77                |
| letters per word        | 0.94                           | <b>0.98</b>                                | 0.86                                       | <b>1</b>            |
| words not in freq. list | 9.47                           | 16.53                                      | 14.79                                      | 11.52               |
| TTR                     | 0.61                           | 0.53                                       | 0.56                                       | 0.62                |
| anaphora words          | <b>1.1</b>                     | <b>0.99</b>                                | 1.23                                       | 1.72                |
| cont./funct. words      | <b>0.75</b>                    | 0.56                                       | 0.64                                       | 0.47                |
| ROUGE                   | <b>0.58</b>                    | <b>0.52</b>                                | 0.44                                       | 0.69                |
| BERTScore               | 0.63                           | 0.68                                       | 0.62                                       | 0.68                |

Bulgarian

|  |                            |                            |                            |                            |                                   |                                |                             |
|--|----------------------------|----------------------------|----------------------------|----------------------------|-----------------------------------|--------------------------------|-----------------------------|
|  | human-made 1 <sup>34</sup> | human-made 2 <sup>35</sup> | human-made 3 <sup>36</sup> | human-made 4 <sup>37</sup> | 'Little Match Girl' <sup>38</sup> | BgGPT: zero-shot <sup>39</sup> | BgGPT: chunks <sup>40</sup> |
|--|----------------------------|----------------------------|----------------------------|----------------------------|-----------------------------------|--------------------------------|-----------------------------|

<sup>34</sup>Andersen, Hans Christian. Groznoto patentse. <https://roditel.bg/groznoto-patentse-prikazka-andersen/>.

<sup>35</sup>Andersen, Hans Christian. Groznoto patentse. Edited by Tanya Petkova, adapted 2020. <https://www.ourboox.com/books/грозното-патенце-2/>.

<sup>36</sup>Andersen, Hans Christian. Groznoto patentse. Prikazki s Dji Dji. <https://taleswithgigi.bg/the-ugly-duckling/>.

<sup>37</sup>Andersen, Hans Christian. Groznoto pate. Zlatnoto pate, adapted 2007.

<sup>38</sup>Andersen, Hans Christian. Malkata kibritoprodavachka. Prikazki ot tsyal svyat, transl. Vasil Velchev, 2009.

<sup>39</sup>The source text is not the full version but the abridged version 'human-made 1'

<sup>40</sup>The text was divided into 4 chunks.

|                         |      |      |      |      |      |             |             |
|-------------------------|------|------|------|------|------|-------------|-------------|
| total words             | 8.71 | 5.21 | 9.1  | 3.45 | 1.76 | 1.62        | 1.39        |
| words per sentence      | 1.01 | 0.99 | 0.94 | 1.16 | 2.4  | 0.89        | <b>1.02</b> |
| letters per word        | 0.9  | 0.92 | 1    | 0.95 | 0.93 | 1.04        | <b>0.98</b> |
| words not in freq. list | 6.01 | 4.2  | 7.46 | 2.86 | 1.58 | <b>1.83</b> | 1.33        |
| TTR                     | 0.65 | 0.72 | 0.69 | 0.77 | 0.86 | 0.92        | 0.91        |
| anaphora words          | 1.23 | 1.07 | 1.04 | 1.2  | 1.3  | 0.49        | 0.94        |
| cont./funct. words      | 0.68 | 0.4  | 0.71 | 1.07 | 1.8  | <b>1.58</b> | <b>0.94</b> |
| ROUGE                   | 0.45 | 0.47 | 0.56 | 0.38 | 0.33 | 0.29        | 0.7         |
| BERTScore               | 0.69 | 0.7  | 0.73 | 0.69 | 0.71 | 0.75        | 0.85        |

|                         | BgGPT: one-shot | BgGPT: alternative | ChatGPT: zero-shot <sup>41</sup> | ChatGPT: chunks <sup>42</sup> | ChatGPT: one-shot <sup>43</sup> | ChatGPT: crossing. direct (EN) | ChatGPT: crossing. chain-of-thought 1 (EN) <sup>44</sup> |
|-------------------------|-----------------|--------------------|----------------------------------|-------------------------------|---------------------------------|--------------------------------|--|
| total words             | 1.14            | 1.04               | <b>2.22</b>                      | <b>2.1</b>                    | <b>2.03</b>                     | <b>1.87</b>                    | 32.65  |
| words per sentence      | <b>1.03</b>     | <b>0.98</b>        | 0.75                             | <b>1.11</b>                   | 0.93                            | <b>1.16</b>                    | 0.54   |
| letters per word        | 1.03            | <b>1</b>           | 1.02                             | <b>0.95</b>                   | <b>0.97</b>                     | <b>0.92</b>                    | <b>0.96</b>  |
| words not in freq. list | 1.16            | 1.07               | <b>2.58</b>                      | <b>1.82</b>                   | <b>2.04</b>                     | 26.11                          | <b>1.83</b>  |
| TTR                     | 0.98            | 0.98               | 0.97                             | <b>0.83</b>                   | 0.96                            | 0.62                           | 0.92   |
| anaphora words          | 0.82            | 0.76               | 0.56                             | <b>1.06</b>                   | 0.53                            | 0.82                           | <b>1.05</b>  |
| cont./funct. words      | <b>1.13</b>     | <b>1.05</b>        | 2.19                             | <b>1.29</b>                   | 1.98                            | 2.47                           | 0.14   |
| ROUGE                   | 0.59            | 0.76               | <b>0.37</b>                      | 0.63                          | <b>0.44</b>                     | <b>0.49</b>                    | <b>0.37</b>  |
| BERTScore               | 0.81            | 0.83               | <b>0.71</b>                      | 0.78                          | <b>0.72</b>                     | 0.64                           | 0.78   |

<sup>41</sup>The source text is not the full version but the abridged version 'human-made 1'

<sup>42</sup>The text was divided into 4 chunks.

<sup>43</sup>The source text is not the full version but the abridged version 'human-made 1'

<sup>44</sup>The source text is not the full version but the abridged version 'human-made 4'

|                            | ChatGPT:<br>crossling.<br>chain-of-<br>thought 2<br>(EN) | ChatGPT:<br>alterna-<br>tive | Gemini:<br>zero-<br>shot <sup>45</sup> | Gemini:<br>chunks<br>The text<br>was di-<br>vided<br>into 4<br>chunks. | Gemini:<br>one-shot | Gemini:<br>crossling.<br>direct<br>(EN) <sup>46</sup> | Gemini:<br>crossling.<br>chain-of-<br>thought 1<br>(EN) |
|----------------------------|--|------------------------------|--|--|---------------------|---|---|
| total words                | 24.49  | 1.03                         | 1.4                                    | <b>2.39</b>  | 1                   | <b>5.4</b>  | <b>3.95</b>   |
| words per<br>sentence      | 0.92   | <b>0.97</b>                  | <b>1.23</b>                            | <b>1.34</b>  | 0.87                | <b>1.07</b>   | 0.83  |
| letters per<br>word        | 0.88   | <b>0.99</b>                  | <b>1</b>                               | <b>0.94</b>  | <b>1</b>            | <b>0.96</b>   | <b>0.91</b>   |
| words not in<br>freq. list | 15.27  | 1.03                         | 1.54                                   | <b>2.09</b>  | 1.17                | <b>4.53</b>   | <b>3.43</b>   |
| TTR                        | 0.57   | 0.99                         | 0.96                                   | <b>0.85</b>  | 1.04                | <b>0.76</b>   | <b>0.76</b>   |
| anaphora<br>words          | 0.66   | 0.69                         | 0.89                                   | 0.92   | 0.75                | <b>1.15</b>   | <b>1.27</b>   |
| cont./funct.<br>words      | <b>1.78</b>  | <b>1.02</b>                  | <b>1.35</b>                            | <b>1.04</b>  | <b>1.03</b>         | 0.39  | 0.55  |
| ROUGE                      | <b>0.46</b>  | 0.71                         | <b>0.51</b>                            | <b>0.49</b>  | 0.22                | <b>0.46</b>   | <b>0.42</b>   |
| BERTScore                  | 0.66   | 0.86                         | 0.81                                   | 0.74   | 0.74                | <b>0.72</b>   | 0.78  |

|                            | Gemini: crossling.<br>chain-of-thought 2<br>(EN) | Gemini:<br>alternative |
|----------------------------|--|------------------------|
| total words                | 17.53  | 1.07                   |
| words per<br>sentence      | 0.67   | <b>1</b>               |
| letters per<br>word        | 0.77   | 1.01                   |
| words not in<br>freq. list | 10.9   | 1.13                   |
| TTR                        | <b>0.65</b>                                      | 1.07                   |
| anaphora<br>words          | <b>1.15</b>                                      | 0.7                    |
| cont./funct.<br>words      | <b>1.42</b>                                      | <b>1.08</b>            |
| ROUGE                      | 0.31   | 0.7                    |
| BERTScore                  | 0.65   | 0.82                   |

<sup>45</sup>The source text is not the full version but the abridged version 'human-made 1'

<sup>46</sup>The source text is not the full version but the abridged version 'human-made 1'

# Advancing Sentiment Analysis in Serbian Literature: A Zero and Few-Shot Learning Approach Using the Mistral Model

**Milica Ikonić Nešić**

Faculty of Philology,

University of Belgrade, Serbia

milica.ikonik.nesic@fil.bg.ac.rs

**Saša Petalinkar**

University of Belgrade, Serbia

sasa5linkAr@gmail.com

**Mihailo Škorić** and

**Ranka Stanković** and

**Biljana Rujević**

Faculty of Mining and Geology

University of Belgrade, Serbia

mihailo.skoric@rgf.bg.ac.rs

ranka.stankovic@rgf.bg.ac.rs

biljana.rujevic@rgf.bg.ac.rs

## Abstract

This study presents the Sentiment Analysis of the Serbian old novels from the 1840-1920 period, employing the Mistral Large Language Model (LLM) to pioneer zero and few-shot learning techniques.

The main approach innovates by devising research prompts that include guidance text for zero-shot classification and examples for few-shot learning, enabling the LLM to classify sentiments into positive, negative, or objective categories. This methodology aims to streamline sentiment analysis by limiting responses, thereby enhancing classification precision. Python, along with the Hugging Face Transformers and LangChain libraries, serves as our technological backbone, facilitating the creation and refinement of research prompts tailored for sentence-level sentiment analysis. The results of sentiment analysis in both scenarios, zero-shot and few-shot, have indicated that the zero-shot approach outperforms, achieving an accuracy of 68.2%.

**Keywords:** zero-shot, few-shot, sentiment, Serbian, Mistral model

## 1 Introduction

Over the years, the need for sentiment analysis as one of the pivotal fields of Natural Language Processing (NLP) has significantly grown across various domains of interest, including but not limited to medicine (Ge et al., 2023), finance (Zhang et al., 2023), education (Altrabsheh et al., 2013), digital humanities (Stanković et al., 2022), politics and social media (Putra et al., 2023).

Previous research has mainly focused on a small number of languages that had a larger amount of training data available. Interest in languages with low resources such as Arabic (Alqarni and Rahman, 2023), Bangla (Hasan et al., 2023), African (Wang et al., 2023), and Serbian (Stanković et al., 2022;

Batanović, 2021) has grown over the years. Considering the insufficient resources for the Serbian language, the possibility of training large language models (LLM) without a large amount of training data represents an important step in sentiment analysis.

So far, various approaches have been employed for sentiment analysis over the Serbian language. The sentiment analysis on the Serbian Movie Review Dataset using by using unigram, bigram, and trigram features in a combination of Naïve Bayes (NB) and Support Vector Machines (SVM) (Batanović et al., 2016) showed the best accuracy of 85.5% for 2 classes and 62.2% for 3 classes. The sentiment analysis framework for a morphologically rich language (SAFOS) (Mladenović et al., 2016) had the best accuracy of 78.3% for movie reviews and 79.2% for newspapers using a combination of unigram and bigram features reduced by sentiment feature mapping. Within the same research, the sentiment lexicon and Serbian WordNet (SWN) synsets were integrated using sentiment polarity scores for feature selection and the lexicon derived from SWN was augmented by incorporating morphological variants of phrases and emotional terms from Serbian Morphological Electronic Dictionaries (Krstev, 2008). The lexicon-based approach using three existing lexicons: *NRC*, *AFFIN* and *Bing* with additional extensive corrections, using Multinomial Naïve Bayes (MNB) with Bag-of-Words approach combined with the features of the sentiment lexicon. This approach gave the accuracy of SA on the evaluation dataset of 82% for two classes, and 72% for 3 classes (Stanković et al., 2022).

The main motivation for this study lies in the fact that, to the best of our knowledge, sentiment analysis in Serbian literature utilizing the zero-shot and few-shot learning approach using the Mistral model has not been yet explored. Machine learning

has been highly successful in data-intensive applications but is often hampered when the data set is small, and this study offers a new approach to sentiment analysis in cases of smaller data sets.

The sentiment analysis was applied to the selected, annotated, and balanced sentences from the Serbian part of the ELTeC<sup>1</sup> multilingual corpus of novels. Novels written in the period 1840–1920 are built to test various distant reading methods among them sentiment analysis, presented in Section 2. Four human annotators performed careful checks of sentiment in sentences, yielding 1089 balanced sentences with three classes: positive, negative, and neutral.

Techniques used for automated classification were zero-shot and few-shot.

Zero-shot learning techniques, where the LLM is prompted without prior specific training on the task, rely solely on the general capabilities of the model [Romera-Paredes and Torr, 2015](#); [Xian et al., 2017](#); [Wang et al., 2019](#); [Brown et al., 2020](#).

Conversely, few-shot learning involves providing the LLM with a small number of examples before requesting it to perform the task. This method aims to prime the model with relevant context, enhancing its performance on specific sentiment classification tasks [Brown et al., 2020](#); [Wang et al., 2020](#).

The Mistral 7B-Instruct ([Jiang et al., 2023](#)) variant, specifically utilized in this work, has been fine-tuned to follow instructions with remarkable precision, thus providing an advantage in generating contextually relevant and accurate sentiment analysis. It achieves this by leveraging the base model’s architectural efficiencies without sacrificing performance on complex text inputs. This version of Mistral 7B outperforms comparative models in human and automated benchmarks, showcasing its utility in nuanced language tasks such as sentiment classification. Furthermore, further elaboration will be provided in Section 3.1.

In Section 3 the methodological approach is depicted through various prompts, while a detailed evaluation of the model on prepared sentences (with the findings and a thorough discussion) is given in Section 4. Finally, conclusions and plans for future work can be found in Section 5.

One of the main goals is to ascertain whether LLMs can provide a consistent, efficient, and potentially less biased means of sentiment annotation,

thereby overcoming some of the limitations associated with human annotators. Through this comparison, the feasibility and advantages of integrating LLMs into the sentiment analysis process are aimed to be illuminated, potentially revolutionizing how sentiment data is processed and interpreted in various applications.

This aspect is particularly significant for languages with limited linguistic resources, such as Serbian. These languages often lack comprehensive corpora with annotated sentiment, presenting a substantial challenge for traditional sentiment analysis techniques that rely heavily on such datasets. The scarcity of annotated corpora in these languages not only hinders the development of effective sentiment analysis models but also limits the applicability of these models in real-world scenarios.

## 2 Dataset

Serbian part of ELTeC corpus ([Krstev, 2021](#)), dubbed *SrpELTeC*, comprises 100 novels in the main collection and 20 in the extended collection. These novels are digitized and freely accessible, thus presenting no constraints on their usage. However, challenges arise concerning the analysis and extraction of information from such text collection, which consists of 5.886,528 tokens and 4.769,262 words. Novels are automatically annotated with part of speech, lemma, and named entity information, thereby paving the way for the application of advanced text analysis methods, in line with the distant reading paradigm. For sentiment analysis, a subset of this text collection is used in previous research. For evaluation, we will rely on a previously manually annotated dataset with 1089 sentences ([Stanković et al., 2022](#)). Figure 1 presents the distribution of sentence length, quantified by the number of words, which corresponds uniquely to each sentiment label. To evaluate the models and demonstrate the capabilities offered by zero-shot and few-shot methodologies compared to previous research, the same dataset was employed for evaluation purposes.

The process of annotating sentences occurred in several phases: 1) extraction of 30K sentences from *srpELTeC*; 2) manual evaluation by four annotators, where the annotation is conducted on a scale from -5 to -2 for negative sentiment gradation; -1, 0, 1 for neutral (objective) sentiment; and 2 to 5 for positive sentiment and 3) calculating inter-

<sup>1</sup>ELTeC: European Literary Text Collection

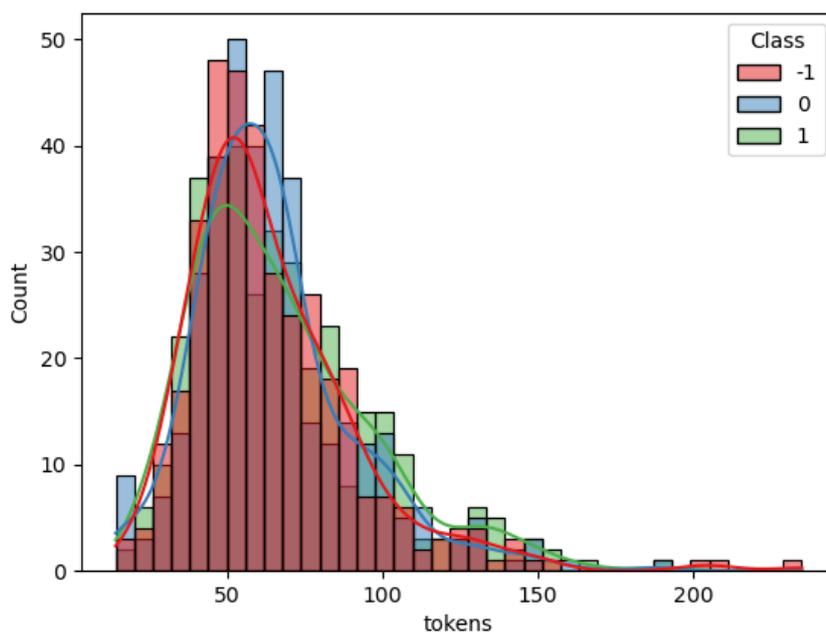


Figure 1: The sentence length (in number of words) distribution of in manually annotated sentiment dataset used for evaluation

annotator agreement was calculated using ReCal2 tool (Deen Freelon, 2011) that showed: Percent Agreement 82.5%, Scott’s Pi 0.737, Cohen’s Kappa 0.739, Krippendorff’s Alpha (nominal) 0.737.

The human annotator’s task in this context relied heavily on their intuition as a native speaker of the language. However, this approach had limitations, particularly when dealing with sentences that are sarcastic or express victory of one side over another, for example in sentence *”Kad su ga drugi dan iz crkve sa krštenja doneli, dodje i kršteni kum deteta, Sava Srbin, dobra duša ti je on bio, al ’ sav beše suzama poliven kad je u sobu ušao.”* (When they brought him back from the church the next day, the baptized godfather of the child, Sava the Serb, arrived, he was a good soul, but he was entirely bathed in tears when he entered the room.). In such cases, determining the polarity of the sentence became challenging without clear specifications on what constitutes positive, negative, and neutral sentiment.

To address these challenges, annotators may require additional context or guidelines to determine the intended sentiment accurately. Providing specifications on what constitutes positive, negative, and neutral sentiment can help standardize the annotation process and minimize subjective interpretation.

Moreover, leveraging advanced natural language processing techniques, such as sentiment analysis algorithms, can complement human annotation efforts by identifying sentiment patterns and detecting nuances in language that may be challenging for human annotators to discern alone.

In conclusion, while annotating sentences for sentiment analysis, relying on the annotator’s intuition as a native speaker is essential. However, to ensure accuracy and consistency, it is crucial to provide clear guidelines and consider contextual factors, especially when dealing with ambiguous or nuanced expressions like sarcasm or conflicting sentiments.

Manual annotation not only requires significant time investment but also heavily relies on the human annotators’ comprehension of the instructions and their proficiency in the native language of the text being analyzed. This dependency introduces a potential for variability and subjectivity in the annotations, which can influence the reliability of sentiment analysis outcomes.

### 3 Methodology

The research on sentiment analysis of ELTeC texts was performed using LLMs Mistral 7B model which will be briefly introduced in Section 3.1. The

methodology for this research employed a "Prompt and Response" technique (Amatriain, 2024), utilizing LLMs to analyze sentiment within a corpus. Prompts were generated from prompt templates. Prompt templates are crafted so that the prompts generated from them contain sentences or examples from the corpus, designed to elicit LLM responses that reflect a range of sentiments.

Prompt templates were prepared for both zero-shot and few-shot learning scenarios, with the former requiring no examples for the LLM to generate responses, and the latter incorporating specific examples to guide the model's output. Four templates were devised for the zero-shot learning approach, aiming to evaluate the model's innate understanding and response generation capabilities without prior context. Conversely, two templates were established for the few-shot approach, each including examples intended to orient the model toward the desired output, as will be detailed in Section 3.2. The LLMs responses to the prompts are parsed and classified into the same categories as those used for manual annotation within the corpus: positive, negative, and neutral, where the parsing process is crucial, given the LLM's potential to generate subtly nuanced responses. Finally, the LLM-generated sentiment classifications are compared to the manual annotations using accuracy and confusion matrices presented in Section 4.

The approach taken emphasizes minimizing extraneous elements in the LLM's responses. This was achieved by limiting the responses to specific instructions or grammatical structures, thereby simplifying the subsequent text-parsing process. To prepare the responses for classification, the following steps were systematically implemented in three steps: 1) Extraneous characters, including spaces, new lines, and punctuation, were removed from the LLM's responses. Additionally, all text was converted to lowercase to maintain consistency and eliminate any discrepancies caused by case sensitivity; 2) The cleaned text was then parsed to identify keywords that indicate sentiment. Specifically, the presence of words corresponding to "positive", "negative", "neutral", or variants thereof, such as "objective", was checked. 3) Based on the keywords identified, each response was classified into categories:

- 1 for "positive" responses,
- 0 for "neutral" responses,

- -1 for "negative" responses,
- 10 for any response that did not fit into these categories, labeled as an "error".

This method of response processing ensures that the textual responses from the Mistral model are efficiently classified, allowing for clear and quantifiable analysis of sentiment trends based on the LLM's outputs. Figure 2 outlines the systematic workflow for preparing the Mistral model for sentiment analysis.

In addition to quantitative analysis, this study also employed qualitative analysis to examine instances where LLMs may surpass human annotators in sentiment analysis accuracy. This qualitative examination focused on identifying specific cases within the corpus where the LLM's sentiment classification demonstrated a higher level of precision, nuanced understanding, or consistency compared to manual annotations.

This facet of analysis involved a detailed review of the LLM responses. Scenarios in which LLMs provided superior sentiment analysis were highlighted to uncover the potential advantages of integrating LLMs in areas requiring high levels of accuracy and objectivity in sentiment classification.

### 3.1 Mistral

In this study, mistralai/Mistral-7B-Instruct-v0.2 variant, a fine-tuned version of the Mistral 7B model is used. It was engineered for enhanced performance and efficiency in processing natural language instructions. Mistral 7B is distinguished by its 7-billion-parameter design, which has demonstrated very good performance across various benchmarks, outclassing even larger models such as the 13-billion-parameter Llama 2 and the 34-billion-parameter Llama 1, particularly in areas of reasoning, mathematics, and code generation. This model is released under the Apache 2.0 license as a part of MistralAI's open-source initiative, demonstrating a commitment to advancing NLP research and application. Its architecture facilitates easy fine-tuning across a wide array of tasks, underscoring its adaptability and superior performance in handling instructional datasets from public repositories like Hugging Face, without the need for proprietary data or complex training modifications (Jiang et al., 2023).

Employing the "mistralai/Mistral-7B-Instruct-v0.2" this study aims to explore its potential in

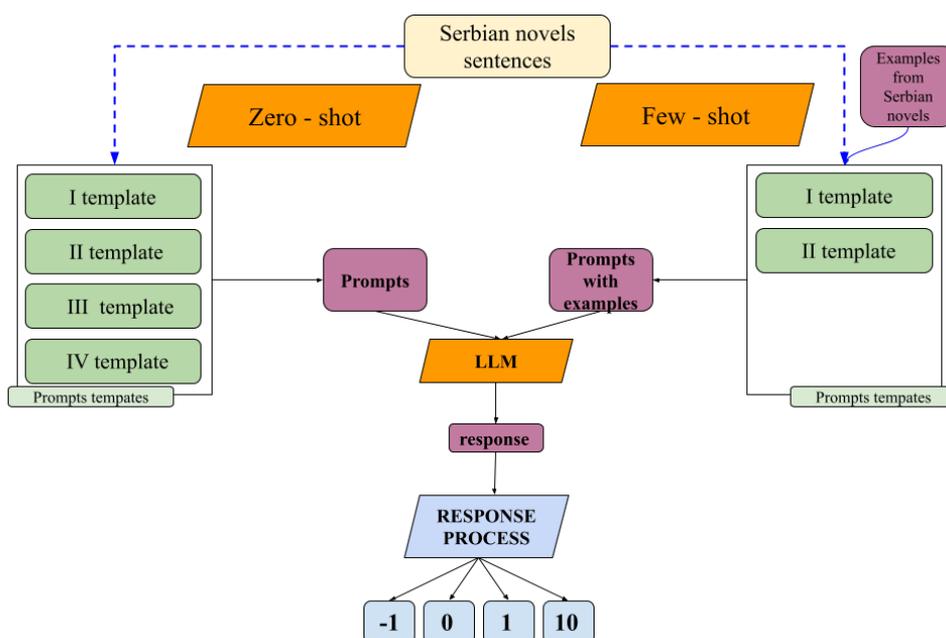


Figure 2: Workflow preparation for use Mistral model and getting response for sentiment analysis.

accurately parsing and classifying sentiment in Serbian novel sentences, offering insights into the advanced capabilities of modern LLMs in automating sentiment analysis with high efficiency and accuracy. The implementation of the model was carried out using Python, with a particular emphasis on leveraging the Langchain library (Chase, 2022). This choice facilitated a streamlined integration and application of the model for sentiment analysis tasks.

The computational experiments were conducted on a local machine equipped with an NVIDIA GeForce RTX 3060 GPU.

For the zero-shot prompts with the Mistral 7B-Instruct LLM, a strategic limitation was imposed on the output length, restricting it to seven tokens. This was done to favor the generation of concise responses, ideally single-word sentiments in Serbian. The aim was to simplify the parsing process and ensure the directness of sentiment classification.

However, the imposition of such token length restrictions was not feasible with implementations based on `llama.cpp`. To address this challenge and achieve consistency in the parsing of model outputs, an alternative strategy was adopted. Custom grammar rules were defined using Grammar-Based Normal Form (GBNF), effectively constraining the model’s responses to three specific, required formats. This approach significantly simplified the parsing process by rendering the structure of responses predictable and straightforward to inter-

pret.

### 3.2 Prompts Templates

#### 3.2.1 Zero-shot Prompts Templates

The first prompt template in the series designed for zero-shot learning scenarios is marked by its simplicity, tailored to elicit sentiment analysis on Serbian texts. This approach intentionally avoids giving the model elaborate instructions on conducting the analysis. As one of the simplest, the first template is presented as follows, while all other templates are presented in Appendix A.

#### Original Template in Serbian:

Kao ekspert za analizu sentimenta, analizirajte sledeći tekst na srpskom jeziku i odredite njegov sentiment. Sentiment treba da bude striktno klasifikovan kao "pozitivan", "negativan", ili "objektivan". Nijedan drugi odgovor neće biti prihvaćen! Tekst: text Sentiment:

#### English Translation:

As an expert in sentiment analysis, analyze the following text in Serbian and determine its sentiment. The sentiment should be strictly classified as "positive", "negative", or "objective". No other response will be accepted! Text: text Sentiment:

The template is segmented into three distinct parts (role play, clear instructions, and a specified response format) (Amatriain, 2024), each aimed at directing the model’s response straightforwardly:

1. **Role Play as Expert:** The prompt positions the LLM as an expert in sentiment analysis, priming it for task-specific responses.
2. **Instructions:** The model is given direct instructions to analyze the provided text and classify its sentiment within strict parameters, aimed at minimizing ambiguity in its responses.
3. **Expected Format of Response:** By clarifying the acceptable response format, the template simplifies the parsing process, facilitating straightforward sentiment classification.
4. **Placeholder for Dataset Sentences:** The ‘text’ placeholder signifies where sentences from the dataset are to be inserted, allowing for the template’s broad application across various texts.

This minimalist strategy is employed to assess how the Mistral model performs in interpreting and analyzing sentiment with only the most basic instructions. The design tests the model’s intrinsic sentiment analysis capabilities, offering insights into its performance when provided with just the essential task parameters and no further methodological guidance.

In the development of the second prompt template, a chain of thought (Amatriain, 2024) was incorporated, introducing a methodical approach to sentiment analysis. The chain of thought is described as a sequence of analytical steps that guides the model through a detailed examination of the text. It includes instructions for reading the entire text, identifying words that convey strong sentiment polarity, and noting instances of negation and sarcasm. This method facilitates a nuanced understanding of sentiment within the provided text.

For the third prompt template, a more specialized approach was adopted, aligning closely with the corpus’s characteristics. The model is positioned in the role of a professor of Serbian literature, with instructions emphasizing the differentiation between modern Serbian and the language found in old novels. This role-play, combined with

a chain of thought strategy, is aimed at encouraging the model to consider stylistic and linguistic variations when analyzing sentiment.

The fourth prompt template marks a return to simplicity, albeit with strategic emphasis on key instructions through the use of all-caps (Amatriain, 2024). While maintaining the role-play aspect as a professor of Serbian literature, detailed instructions were streamlined to exclude the notion of in-depth analysis. This approach emphasizes the importance of direct sentiment classification, with specific instructions highlighted in all-caps to ensure clarity and focus.

### 3.2.2 Few-shot Prompts Templates

In the progression toward the examination of few-shot templates, a cautionary note must be articulated. As previously discussed in the document, the classification tasks for the few-shot scenario were performed utilizing an 8-bit version of the Mistral model. This adaptation was necessitated by resource limitations, leading to a reduced context window of 512 tokens. Consequently, the length of the few-shot templates was constrained, resulting in the incorporation of only three examples within them, corresponding to each sentiment class. This limitation was pivotal in ensuring the feasibility of the few-shot classification under the specified computational constraints, albeit at the cost of a more extensive illustrative context.

In the deployment of few-shot templates within this investigation, a structured format was adhered to, consisting of a prefix, examples, and a suffix, following the established pattern of the Langchain library. This structured approach facilitated the systematic presentation of examples to the model.

The first few-shot template is an extension of the first zero-shot template. The prefix provides a simple clarification that examples will follow. This is succeeded by the examples themselves, and the instructions similar to the first zero-shot template, albeit slightly simplified and shortened due to the limited context window. This adaptation was necessary to fit within the computational constraints while maintaining the template’s instructional integrity.

The second template was an attempt to implement a chain-of-thought process. However, the limitations of the context window required significant pruning of the text. The language of instruction was simplified to minimize word count, reducing the instructions to the bare essentials. Despite these

adaptations, some sentences extended beyond the context window, ultimately impacting the effectiveness of this template in the experiment. This outcome highlighted the need for a larger context window to fully realize the potential of chain-of-thought processes in few-shot learning scenarios.

Thus, this part of the experiment was deemed a failure and no results were included. While it was possible to exclude those sentences containing over 150 tokens, it was deemed unnecessary due poor performance of the other few-shot template.

#### 4 Results/Discussion

The results given in this section represent sentiment analysis on the Serbian novels dataset by using responses generated by the Mistral model in both scenarios, i.e. zero-shot and few-shot learning. The accuracy values (acc.) depicted in Table 1 illustrate divergent performance across distinct prompt templates enumerated in the column labeled "prompt template" of zero and few-shot, underscoring the significance of template design on sentiment analysis accuracy.

The evaluation of the zero-shot templates reveals a varied range of accuracy, where the first template exhibited the highest result, suggesting that straightforward and direct prompts are most effective in eliciting accurate sentiment analysis from this model for Serbian sentiment. Figure 3 presents the confusion matrix for the first zero-shot template. In Appendix A is presented a confusion matrix for the rest zero-shot templates.

The first and fourth templates were most effective in identifying positive and negative sentiments. However, they struggled with objective sentences, showing a high rate of mislabeling. Interestingly, the first template, despite its higher accuracy in sentiment classification, also exhibited a higher number of errors where the LLM responses could not be classified into any of the categories. The fourth template utilized all-caps to emphasize key instructions, and also performed well, indicating that clarity in instruction plays a crucial role.

The second zero-shot template, which attempted a more complex chain-of-thought analysis, resulted in the lowest accuracy, highlighting the limitations of the model's processing capacity in its current configuration.

The third zero-shot template achieved the best accuracy in classifying objective sentiments. Nevertheless, it performed poorly with negative senti-

| Type      | Prompt Template | Acc.         |
|-----------|-----------------|--------------|
| zero-shot | 1               | <b>0.682</b> |
|           | 2               | 0.205        |
|           | 3               | 0.482        |
|           | 4               | 0.657        |
| few-shot  | 1               | 0.392        |

Table 1: Accuracy of SA on Serbian novels dataset for zero-shot and few-shot templates

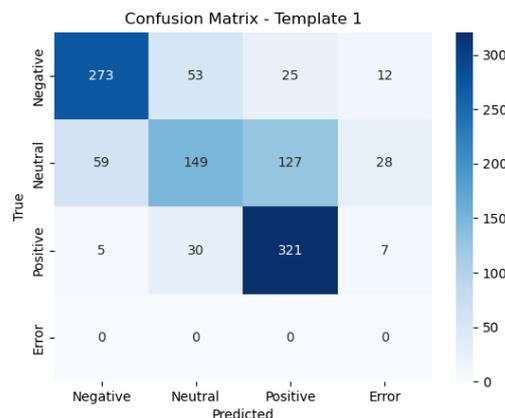


Figure 3: Confusion matrix for zero-shot first template

ments and was not very effective for positive sentiments. It is designed with a role-play scenario involving old literary Serbian, and showed moderate success, reflecting the added difficulty of interpreting historical and stylistic language nuances. The first zero-shot template recorded the highest accuracy but also the most unclassified responses, marked as errors at 47. In Table 2 are presented some of the examples where the model made mistakes. To illustrate the error, the template with the highest accuracy was chosen. It is important to note that in some cases of sarcasm, overuse of dashes (-) and presence of loanwords first template has attend to to classified as an error (10) as it is presented in the last sentence in Table 2. In contrast, the second template had no errors, while the third and fourth templates showed minimal errors with only two and one unclassifiable response, respectively. Notably, of the errors in the first template, 28 were attributed to objective sentiments, which correlates with a high number of misclassifications. This highlights the inherent difficulty in classifying objective sentiments, a challenge that is also evident among human annotators due to the subjectivity involved. It is important to note that templates 2 and 3 tended to detect sarcasm where it was not recognized by

| Example sentence   | Translation of sentence  | Annotators | Model |
|--|--|------------|-------|
| U Ivanu žilice se zaigraju, srce mu se stesni; ove dve tri reči, koje Mladen izusti, učine mu se proročanstvo koje ovaj govori iz magnetičnog sna. | In Ivan, his veins begin to throb, his heart tightens; these two or three words uttered by Mladen seem to him a prophecy spoken from a magnetic dream. | -1         | 1     |
| Sto je bio sav mokar, i s njegovih krajeva kapala je voda s mrvicama od duhana i pepelom od cigara, koji sam ja otresao na svećnjak.               | He was completely wet, and water dripped from his edges along with bits of tobacco and cigarette ash, which I shook off onto the candlestick.          | 0          | -1    |
| Oh, da znate vi, dragi prijatelju, kakva je to naslada prolivati suze na grudima vernog prijatelja il ' ljubavnika!...                             | Oh, if you only knew, dear friend, what a delight it is to shed tears on the chest of a faithful friend or lover!...                                   | 1          | 10    |

Table 2: Example sentences where the model recognized sentiment incorrectly.

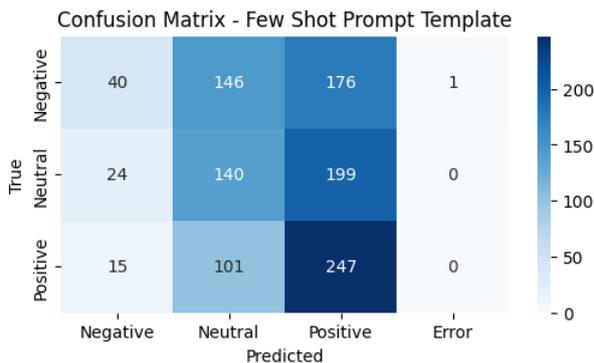


Figure 4: Confusion matrix for few-shot template

the annotators. Upon further examination, there have been instances where the LLM’s classification proved to be more accurate than human annotation. Notably, in many of these cases, the majority or at least half of the template responses were consistent with each other.

Furthermore, it is worth noting that the accuracy of the only few-shot template surpasses only the second zero-shot template, which displayed the lowest accuracy among zero-shot prompts. This outcome highlights the challenges associated with the few-shot scenario, especially given the limited context window. Although better than the lowest zero-shot results are still underwhelming, illustrating the inherent challenges of adapting few-shot learning strategies within a constrained computational environment. Figure 4 presents the confusion matrix for the few-shot first template. It is important that compared to previous sentiment analysis studies (Stanković et al., 2022), where an approach utilizing MNB solely with features derived from the sentiment lexicon achieved an accuracy of 65.7%, and MNB with a Bag-of-Words approach combined with sentiment lexicon features achieved an accuracy of 71.9%, tested on the same corpus

as this study, this approach demonstrates that employing zero-shot learning with the Mistral model can achieve a comparable accuracy of 68.2%, with a significant advantage being that the model does not require a training corpus.

## 5 Conclusion and Future Work

In this study, the simplification necessitated by using a quantified model with a limited context window appeared to strip away many of the benefits typically associated with the Mistral model. Despite its notable speed, the diminished performance suggests that such an approach may not be viable, particularly for less commonly studied languages like Serbian. It is important to mention that zero-shot prompts were not run on the quantified model in our study. Therefore, it remains unclear whether the quantification itself degrades performance for less commonly trained parts of the model (such as Serbian language processing), or if the limitations imposed by the reduced context window, especially when combined with the addition of examples, render the model unsuitable for this type of text. One potential method to further investigate these findings would be to run zero-shot prompts on quantified models. However, the value of such research remains uncertain. The findings of this study demonstrate that in literary texts of old Serbian novels, the zero-shot approach exhibits superior performance, particularly in the case of the simplest prompt, thereby leaving room for further exploration in this direction. Using all caps to highlight the part of instructions has proven useful in the elimination of unusable responses, but instructing LLM to detect sarcasm resulted in over-detection. Additionally, comparing this approach with fine-tuned XLM-R models will represent one of the future objectives.

## Acknowledgments

This research was supported by the Science Fund of the Republic of Serbia, #7276, Text Embeddings - Serbian Language Applications - TESLA.

## References

- Arwa Alqarni and Atta Rahman. 2023. Arabic tweets-based sentiment analysis to investigate the impact of covid-19 in ksa: A deep learning approach. *Big Data and Cognitive Computing*, 7(1):16.
- Nabeela Altrabsheh, Mohamed Medhat Gaber, Mihaela Cocea, et al. 2013. Sa-e: sentiment analysis for education. *Frontiers in Artificial Intelligence and Applications*, 255:353–362.
- Xavier Amatriain. 2024. Prompt design and engineering: Introduction and advanced methods. *arXiv preprint arXiv:2401.14423*.
- Vuk Batanović. 2021. Semantic similarity and sentiment analysis of short texts in serbian. In *2021 29th Telecommunications Forum (TELFOR)*, pages 1–7. IEEE.
- Vuk Batanović, Boško Nikolić, and Milan Milosavljević. 2016. Reliable baselines for sentiment analysis in resource-limited languages: The serbian movie review dataset. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 2688–2696.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Harrison Chase. 2022. *LangChain*.
- Deen Freelon. 2011. Reliability calculator for 2 coders.
- Yao Ge, Yuting Guo, Sudeshna Das, Mohammed Ali Al-Garadi, and Abeed Sarker. 2023. Few-shot learning for medical text: A review of advances, trends, and opportunities. *Journal of Biomedical Informatics*, page 104458.
- Md. Arid Hasan, Shudipta Das, Afiyat Anjum, Firoj Alam, Anika Anjum, Avijit Sarker, and Sheak Rashed Haider Noori. 2023. *Zero- and few-shot prompting with llms: A comparative study with fine-tuned models for bangla sentiment analysis*.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.
- Cvetana Krstev. 2008. *Processing of Serbian. Automata, Texts and Electronic Dictionaries*. Faculty of Philology of the University of Belgrade.
- Cvetana Krstev. 2021. *The serbian part of the eltec collection through the magnifying glass of metadata*. *Infotheca - Journal for Digital Humanities*, 21(2):26–42.
- Miljana Mladenović, Jelena Mitrović, Cvetana Krstev, and Duško Vitas. 2016. Hybrid sentiment analysis framework for a morphologically rich language. *Journal of Intelligent Information Systems*, 46(3):599–620.
- Diaz Tiyasya Putra, Erwin Budi Setiawan, et al. 2023. Sentiment analysis on social media with glove using combination cnn and roberta. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 7(3):457–563.
- Bernardino Romera-Paredes and Philip Torr. 2015. An embarrassingly simple approach to zero-shot learning. In *International conference on machine learning*, pages 2152–2161. PMLR.
- Ranka Stanković, Miloš Košprdić, Milica Ikonić Nešić, and Tijana Radović. 2022. *Sentiment analysis of Serbian old novels*. In *Proceedings of the 2nd Workshop on Sentiment Analysis and Linguistic Linked Data*, pages 31–38, Marseille, France. European Language Resources Association.
- Mingyang Wang, Heike Adel, Lukas Lange, Jannik Strötgen, and Hinrich Schütze. 2023. *Nlnde at semeval-2023 task 12: Adaptive pretraining and source language selection for low-resource multilingual sentiment analysis*. In *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics.
- Wei Wang, Vincent W Zheng, Han Yu, and Chunyan Miao. 2019. A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–37.
- Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34.
- Yongqin Xian, Bernt Schiele, and Zeynep Akata. 2017. Zero-shot learning-the good, the bad and the ugly. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4582–4591.
- Boyu Zhang, Hongyang Yang, and Xiao-Yang Liu. 2023. Instruct-fingpt: Financial sentiment analysis by instruction tuning of general-purpose large language models. *arXiv preprint arXiv:2306.12659*.

## Appendix A Prompt templates

### A.1 Second Zero-shot Template

#### Zero-shot Second (using chain of thought) template in Serbian:

Kao stručnjak za analizu sentimenta, analizirajte sledeći tekst na srpskom jeziku i odredite njegov sentiment. Sentiment treba klasifikovati strogo kao "pozitivan", "negativan" ili "objektivan". Neće biti prihvaćeni drugi odgovori.

1. Pročitajte i razumite dati tekst.
2. Identifikujte ključne reči ili fraze u tekstu koje ukazuju na sentiment. Posebnu pažnju obratite na pridjeve, priloge i bilo koje specifične glagole koji obično nose emotivnu težinu.
3. Razmotrite ukupni kontekst poruke. Ponekad, sentiment nije u vezi sa prisustvom specifičnih reči, već kako su te reči upotrebljene zajedno u rečenicama.
4. Odredite da li tekst primarno izražava pozitivna osećanja (kao što su sreća, zadovoljstvo ili nada), negativna osećanja (kao što su tuga, ljutnja ili frustracija), ili je primarno činjeničan ili neutralan, bez ikakvog emotivnog sadržaja.
5. Razmislite o prisustvu bilo kakvih negacija ili sarkazma jer to može značajno promeniti sentiment teksta.
6. Nakon analize teksta na osnovu gore navedenih koraka, klasifikujte sentiment kao "pozitivan", "negativan" ili "objektivan".
7. Samo vrednosti "pozitivan", "negativan" i "objektivan" će biti prihvaćene.
8. Ne treba objašnjavati svoj odgovor, već samo dati klasifikaciju sentimenta.
9. U odgovoru ne treba da bude novih redova, samo klasifikacija sentimenta.

Tekst: {text}

Sentiment teksta je

#### English Translation of Zero-shot Second (using chain of thought) template in Serbian:

As an expert in sentiment analysis, analyze the following text in Serbian and determine its sentiment. The sentiment should be strictly classified

as "positive", "negative", or "objective". No other response will be accepted.

1. Read and understand the given text.
2. Identify the key words or phrases in the text that indicate sentiment. Pay special attention to adjectives, adverbs, and any specific verbs that typically carry emotional weight.
3. Consider the overall context of the message. Sometimes, the sentiment is not about the presence of specific words, but how those words are used together in sentences.
4. Determine if the text primarily expresses positive feelings (such as happiness, satisfaction, or hope), negative feelings (such as sadness, anger, or frustration), or is primarily factual or neutral, without any emotional content.
5. Consider the presence of any negations or sarcasm as this can significantly change the sentiment of the text.
6. After analyzing the text based on the above steps, classify the sentiment as "positive", "negative", or "objective".
7. Only the values "positive", "negative", and "objective" will be accepted.
8. Do not explain your answer, but simply provide the sentiment classification.
9. The response should not include new lines, just the sentiment classification.

Text: {text}

The text's sentiment is

### A.2 Third Zero-shot Template

#### Third (advanced chain of thought) template in Serbian:

Kao profesor srpske literature, analizirajte sledeće rečenice izvadjene iz starih srpskih romana čija su autorska prava istekla. Zbog toga što su ti romani napisani pre mnogo godina, jezik može biti nešto zastareliji. Vaš zadatak je da odredite sentiment tih rečenica. Sentiment treba klasifikovati strogo kao "pozitivan", "negativan" ili "objektivan". Neće biti prihvaćeni drugi odgovori.

1. Pažljivo pročitajte i analizirajte dati tekst, uzimajući u obzir stil i kontekst u kojem je napisan.

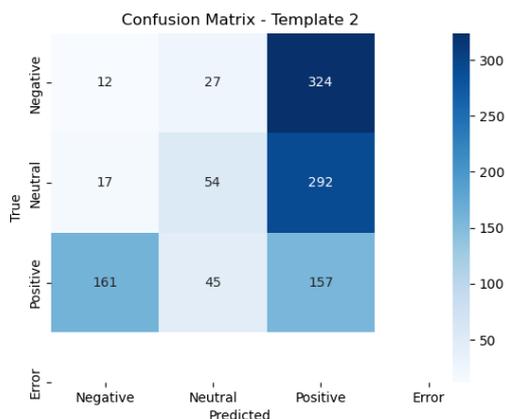


Figure 5: Confusion matrix for few-shot second template

- Identifikujte ključne reči ili fraze koje su karakteristične za period kada je delo napisano i koje mogu ukazivati na sentiment.
- Razmotrite kako zastareli izrazi ili konstrukcije utiču na izraženi sentiment i da li jezik tog vremena ima posebne načine izražavanja emocija.
- Analizirajte da li rečenice izražavaju pozitivne emocije (kao što su radost, zadovoljstvo ili očekivanje), negativne emocije (kao što su tuga, očajanje ili gubitak) ili su primarno deskriptivne i objektivne, bez izraženih emocija.
- Imajte na umu kontekst u kojem se rečenica nalazi unutar dela, jer to može promeniti percepciju sentimenta, naročito kada je jezik arhaičan.
- Klasifikujte sentiment rečenice kao "pozitivan", "negativan" ili "objektivan" nakon dublje analize uzete u obzir sve prethodne korake.
- Odgovor treba da se sastoji od samo od jedne reči: "pozitivan", "negativan" ili "objektivan".

Rečenica: {text}

Sentiment rečenice je

**English Translation of third (advanced chain of thought) template :**

As a professor of Serbian literature, analyze the following sentences extracted from old Serbian novels whose copyrights have expired. Since these novels were written many years ago, the language

may be somewhat outdated. Your task is to determine the sentiment of these sentences. The sentiment should be strictly classified as "positive", "negative", or "objective". No other responses will be accepted.

- Carefully read and analyze the given text, considering the style and context in which it was written.
- Identify key words or phrases characteristic of the period the work was written in that may indicate sentiment.
- Consider how outdated expressions or constructions affect the expressed sentiment and whether the language of that time has special ways of expressing emotions.
- Analyze whether the sentences express positive emotions (such as joy, satisfaction, or anticipation), negative emotions (such as sadness, despair, or loss), or are primarily descriptive and objective, without expressed emotions.
- Keep in mind the context in which the sentence is found within the work, as this can change the perception of sentiment, especially when the language is archaic.
- Classify the sentence's sentiment as "positive", "negative", or "objective" after a deeper analysis considering all the previous steps.
- The response should consist of only one word: "positive", "negative", or "objective".

Sentence: {text}

The sentence's sentiment is

**A.3 Fourth (All Caps) Zero-shot Template**

**Fourth (All Caps) template in Serbian:**

Kao PROFESOR SRPSKE LITERATURE, analizirajte sledeće rečenice izvadjene iz starih srpskih romana čija su autorska prava istekla. Jezik u tim delima može biti nešto zastareliji. VAŠ ZADATAK JE DA ODREDITE SENTIMENT REČENICA KORISTEĆI SAMO TRI MOGUĆE REČI: "POZITIVAN", "NEGATIVAN", ili "OBJEKTIVAN". VAŽNO JE! DOZVOLJENI SU SAMO TI ODGOVORI! BEZ IKAKVOG DODATNOG OPISA, RAZMATRANJA ILI DUGIH ODGOVORA!!!

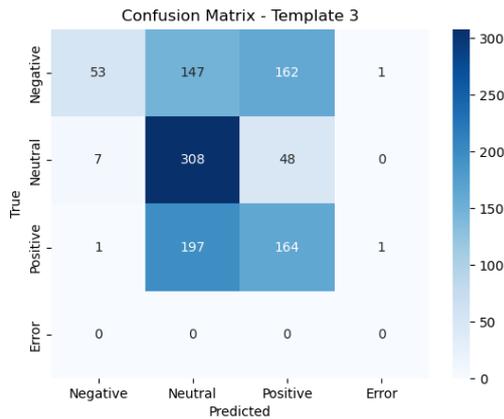


Figure 6: Confusion matrix for few-shot third template

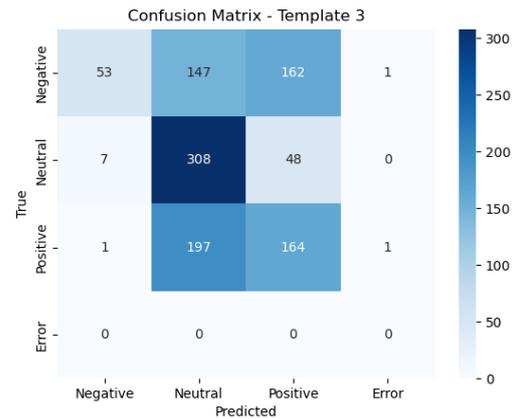


Figure 7: Confusion matrix for few-shot fourth template

1. PROČITAJTE DATI TEKST!
2. IDENTIFIKUJTE SENTIMENT BEZ DUBLJE ANALIZE!
3. ODGOVOR MORA BITI SAMO JEDNA OD TRI REČI: "pozitivan", "negativan", ili "objektivan"!!!
4. NEMA OBJAŠNJAVANJA, SAMO ODABERITE JEDNU OD TRI REČI!!!

Rečenica: {text}

Sentiment:

**English Translation of fourth (all caps) template:**

As a PROFESSOR OF SERBIAN LITERATURE, analyze the following sentences extracted from old Serbian novels whose copyrights have expired. The language in these works may be somewhat outdated. YOUR TASK IS TO DETERMINE THE SENTIMENT OF THE SENTENCES USING ONLY THREE POSSIBLE WORDS: "POSITIVE", "NEGATIVE", or "OBJECTIVE". IMPORTANT! ONLY THOSE RESPONSES ARE ALLOWED! WITHOUT ANY ADDITIONAL DESCRIPTION, CONSIDERATION, OR LONG ANSWERS!!!

1. READ THE GIVEN TEXT!
2. IDENTIFY THE SENTIMENT WITHOUT DEEP ANALYSIS!
3. THE RESPONSE MUST BE ONLY ONE OF THE THREE WORDS: "positive", "negative", or "objective"!!!
4. NO EXPLANATIONS, JUST CHOOSE ONE OF THE THREE WORDS!!!

Sentence: {text}

Sentiment:

#### A.4 Few-Shot Templates and Examples

Below are the examples and templates used for few-shot learning, presented separately for Serbian and English to ensure clarity.

**Examples for Few-Shot Learning:** The following Table 3 presents the examples utilized in the few-shot templates in Serbian, alongside their corresponding sentiment labels:

#### Prefix, Example Template, and Suffix in Serbian:

\Prefix in Serbian:

```
"Primeri sentiment analize na srpskom jeziku:
*****"
```

Example Template in Serbian:

```
"Tekst: {Text}
Sentiment je {Label}"
```

Suffix in Serbian:

```
"*****
Kao ekspert za analizu sentimenta, analizirajte sledeći tekst na srpskom jeziku i odredite njegov sentiment. Sentiment treba da bude klasifikovan kao "pozitivan", "negativan", ili "objektivan". Odgovor treba da bude u skladu sa primerima koje ste videli.
Tekst: {Text}
Sentiment:"
```

| Text in Serbian   | English Translation   | Label |
|---|---|-------|
| Kukavan mlad čovek; on bejaše tako dobar i veran drug,“ prihvati jedan drugi, kom su oči bile pune suza.  | A cowardly young man; he was such a good and faithful friend,” another accepted, his eyes full of tears.  | -1    |
| Juh, ala je to dobra žena, dobra kao dobar dan!   | Wow, what a good woman, as good as a good day!  | 1     |
| Nisam valjada ni pet puta udario, a pola avana odlete u stranu, a druga se polovina, koja je bila nešto veća, prevrte i rastok ode u prašinu... | I surely didn’t hit it five times, and half of the awn flew to the side, and the other half, which was slightly larger, turned over and crumbled into dust... | 0     |

Table 3: Examples of Serbian sentences for Few-Shot Learning

**English Translations:**

For accessibility, the examples and templates are also provided in English below:

Prefix in English:

```
"Examples of sentiment analysis
in Serbian language:
*****"
```

Example Template in English:

```
"Text: {Text}
The sentiment is {Label}"
```

Suffix in English:

```
"*****
As an expert in sentiment
analysis, analyze the following
text in Serbian and determine
its sentiment. The sentiment
should be classified
as "positive", "negative", or
"objective". The answer should
be in accordance with the
examples you have seen.
Text: {Text}
Sentiment:"
```

This structure provides a clear division between the Serbian texts and their English translations, aiding in comprehension for readers of both languages.

# Generating Phonetic Embeddings for Bulgarian Words with Neural Networks

**Lyuboslav Karev**

Faculty of Mathematics and Informatics  
Sofia University, "St. Kliment Ohridski"  
lkarev@uni-sofia.bg

**Ivan Koychev**

Faculty of Mathematics and Informatics  
Sofia University, "St. Kliment Ohridski"  
koychev@fmi.uni-sofia.bg

## Abstract

Word embeddings can be considered the cornerstone of modern natural language processing. They are used in many NLP tasks and allow us to create models that can understand the meaning of words. Most word embeddings model the semantics of the words. In this paper, we create phoneme-based word embeddings, which model how a word sounds. This is accomplished by training a neural network that can automatically generate transcriptions of Bulgarian words. We used the Jaccard index and direct comparison metrics to measure the performance of neural networks. The models perform nearly perfectly with the task of generating transcriptions. The model's word embeddings offer versatility across various applications, with its application in automatic paronym detection being particularly notable, as well as the task of detecting the language of origin of a Bulgarian word. The performance of this paronym detection is measured with the standard classifier metrics - accuracy, precision, recall, and F1.

**Keywords:** neural networks, word embeddings, transcriptions, phonemes, grapheme-to-phoneme

## 1 Introduction

In natural language processing, automatically generating transcriptions of words is a formidable challenge, particularly when confronted with languages like Bulgarian. Renowned for its near-perfect phonemic orthography, wherein each letter typically corresponds to a single sound, Bulgarian presents an intriguing paradox. While its orthographic structure promises clarity, exceptions within the system thwart the straightforward application of bijective mapping algorithms. These anomalies underscore the need for innovative approaches that seamlessly reconcile orthographic and phonetic representations.

In this work, we create a neural network that generates a phoneme transcription from a word. The

model is trained on five books, transforming each into a pair of words and transcriptions.

One particularly noteworthy application of our model lies in automatic paronym detection. By seamlessly integrating phonetic representations into the detection process, our model demonstrates good results for paronym detection.

Another application of phoneme embeddings is the detection of loanwords in the Bulgarian language. A classifier is trained using the embeddings as input, which classifies the language of origin of a given word.

### 1.1 Main concepts

We will define a few concepts that will be used throughout the paper.

- **Phoneme:** The smallest sound unit in a language that can differentiate words.
- **Transcription:** The process of representing spoken language in written form.
- **Grapheme:** The smallest unit of a writing system that represents a phoneme in the spelling of a word.
- **Syllable:** In the Bulgarian language, a syllable is a collection of sounds containing exactly one vowel.
- **International Phonetic Alphabet:** Standardized phonetic notation system representing spoken language sounds.

### 1.2 Rules for generating phonetic transcriptions

In the Bulgarian language, generating the phonetic transcription for a word is almost straightforward. We need to know the word and which syllables are stressed to get the transcription. We can check if the letter is a consonant or a vowel by going through each letter. We can replace the letter

with the matching phoneme if the letter is a consonant. If the letter is a vowel, we have to check if the vowel is stressed and then pick the correct phoneme based on that.

The result is a string written in the International Phonetic Alphabet (IPA). As an example, the Bulgarian word "здравей" (meaning "hello") will look something like this: "zdrʌvɛj"

### 1.3 Syllabic transposition rules

In the Bulgarian language, the structure of a syllable depends on the amount and placement of consonants around a vowel. From this, we can distinguish four types of syllables [Весела Кръстева \(2009\)](#):

- Syllables that have only a single vowel.
- Syllables that have one or more consonants followed by a vowel.
- Syllables that have a vowel followed by one or more consonants.
- Syllables that have one or more consonants, followed by a vowel, followed by one or more consonants.

This structure is not enough to split a word into syllables. To solve this problem, we can use the hyphenation rules in the Bulgarian language, described below. [BAH \(2011\)](#)

1. The consonant letter, which is between two vowels, is hyphenated on the next line.
2. Two or more consecutive consonant letters between two vowels are hyphenated such that at least one consonant is after the first vowel and at least one before the second vowel.
  - (a) Two repeated consonants are split equally on the first and second part of the hyphenation.
3. Two consecutive vowels are hyphenated such that one is on the first line while the second is on the second line.

A word can automatically be split into syllables with this set of rules.

## 2 Related work

Generating the phoneme transcriptions from the grapheme representation is not new, and multiple approaches have been proposed.

[Black and Lenzo \(2003\)](#) and [Elovitz et al. \(1976\)](#) both describe a rule-based approach where graphemes are directly replaced with their corresponding phonemes. On an English language dataset, [Elovitz et al. \(1976\)](#)'s approach achieves around 80% of correctly generated transcriptions, while [Black and Lenzo \(2003\)](#)'s approach results in around 70% of correctly generated transcriptions.

A statistical model is described in [Bisani and Ney \(2008\)](#), which relies on modeling the translation process as a linear sequence of operations. The model generates a phonemic transcription from the orthographic form of a word. This is achieved by approximating the sum of the joint probabilities of all possible grapheme sequences that match the given spelling.

[Rao et al. \(2015\)](#) describes an LSTM-based model for performing grapheme-to-phoneme conversion with a 25.8% error rate on a standard English dataset.

[Li et al. \(2022\)](#) describes multiple approaches that work on a single language - a joint n-gram model, a sequence-to-sequence LSTM model, and a transformer model. The models handle unknown languages by finding the closest language based on the language family tree and using the nearest k language models. The downside of this approach would be that if a model is not trained on Bulgarian data, it might not yield sufficient results.

[Yolchuyeva et al. \(2019\)](#) and [Engelhart et al. \(2021\)](#) propose using transformer-based models to automatically generate the transcriptions from words and apply the model to different tasks for the English language.

In this paper, a similar transformer-based network to [Yolchuyeva et al. \(2019\)](#) and [Engelhart et al. \(2021\)](#) will be used, with the main difference being that the network will be trained exclusively on Bulgarian words and transcriptions to solve the task of automatic transcription generation for the Bulgarian language.

### 3 Data gathering, analysis, and transformation

#### 3.1 Data gathering

The initial data set was collected from the website <https://chitanka.info/>. Five books, each in a separate text file, were downloaded from this website.

The website of "Chitanka" also provides an SQL database containing information about the stressed syllables in a word. All words and their stressed syllables are downloaded from the database. A small transformation was applied to the words. The stressed syllables are marked with the ' symbol. For this to be understood by the model, We replaced the ' symbol with the index of the stressed vowel. For example, the word "авиобранш" has two stressed vowels - the two 'a's. The entry in the database for that word would be "а'виобра'нш". After the transformation, the indices of the stressed syllables would be at indices 0 and 6.

#### 3.2 Transformation of the data

For each of the books, the following transformation is applied:

1. Read all lines of the book.
2. Split each line into separate sentences.
3. For each sentence, remove any special formatting from the website.
4. Split each sentence into words.
5. For each word, generate its transcription.
6. Each word, along with its transcription, is written in a file.

This transformation of the five books results in a list of 606,102 pairs of words and transcriptions.

#### 3.3 Analysis of the data

A short analysis of the data shows us that the amount of unique words is 39,405. The ten most common words are listed under Table 1

From this, we can see that the most common words are the conjunctions "на", "да", "и", "се", etc.

A word is seen 15.381 times on average in the dataset, making the dataset imbalanced in terms of words. However, it's vital to note that natural language is also imbalanced, so this dataset reflects real-life usages of the words. The average length

| Word | Count | % count |
|------|-------|---------|
| на   | 20369 | 3.36%   |
| да   | 20311 | 3.35%   |
| и    | 20048 | 3.31%   |
| се   | 16139 | 2.66%   |
| в    | 10114 | 1.67%   |
| от   | 9666  | 1.59%   |
| не   | 8151  | 1.34%   |
| си   | 7617  | 1.26%   |
| с    | 7160  | 1.18%   |
| че   | 6724  | 1.11%   |

Table 1: 10 most common words

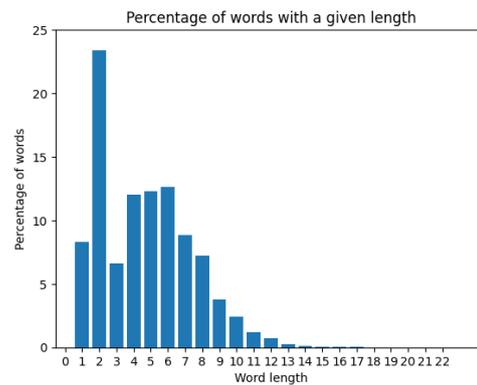


Figure 1: Word length count

of a word is 4.673 characters. The distribution of the lengths of the words can be seen in Figure 1.

Regarding the syllables, the dataset contains a total of 1,222,793 syllables. The unique syllables are 5,469. The 10 most common syllables are listed under Table 2. The average length of a syllable is 2.31 characters. The distribution of the lengths of the syllables can be seen in Figure 2. There, we can see that the 2-character syllables are the most common, with 61.482%, followed by 3-character syllables, with 23.695%.

#### 3.4 Data segmentation

The dataset is split into three subsets: one for training, one for validation, and one for testing. The subsets are split in the following way:

- Training set - 80% of the data (484,881 pairs)
- Validation set - 10% of the data (60,610 pairs)
- Testing set - 10% of the data (60,611 pairs)

## 4 Model overview and training

### 4.1 Transformer model

The idea is to train a neural model based on the *Transformer* architecture Vaswani et al. (2023).

We introduce the following notation:

| Syllable | Count | % count |
|----------|-------|---------|
| на       | 53107 | 8.76%   |
| то       | 34352 | 5.67%   |
| та       | 32000 | 5.28%   |
| и        | 30720 | 5.07%   |
| да       | 28596 | 4.72%   |
| ни       | 23501 | 3.88%   |
| те       | 22548 | 3.72%   |
| ка       | 22094 | 3.65%   |
| се       | 21358 | 3.52%   |
| е        | 20169 | 3.33%   |

Table 2: 10 most common syllables

- $s$  - syllable.
- $W = s^+$  - word, containing at least one syllable.
- $s^W$  - word  $W$ , split into its syllables.
- $l_W$  - the amount of syllables in the word  $W$ .
- $t_s$  - the transcription of the syllable  $s$ .
- $t_W$  - the transcription of the word  $W$ .
- $f(s) : s \mapsto \mathbb{N}$  - function, which maps a syllable to an index.
- $f'(s) : \mathbb{N} \mapsto t_s$  - function, which maps an index to a transcription of a syllable.
- $i_s$  - the index of the syllable  $s$ ,  $f(s) = i_s$ .
- $i_W$  - the indices of the syllables from the word  $W$ .
- $i_{t_s}$  - the index of the transcription of the syllable  $t_s$ .
- $i_{t_W}$  - the indices of the transcription of the word  $W$ .
- $v_{i_W}$  - the input vector of the model.
- $o_{i_W}$  - the output vector of the model.
- Embedding - a vector representation of a word, or in this case, a syllable.
- $emb$  - the size of the resulting embeddings from the model.
- $\langle bos \rangle$  - tag for the start of a word.
- $\langle eos \rangle$  - tag for the end of a word.

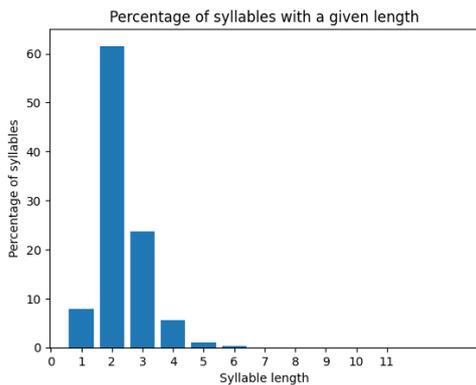


Figure 2: Syllable length count

The neural model accepts a vector  $v_{i_W}$  with size  $l_W$  and returns a new vector  $o_{i_W}$  with size  $l_W$ .

Before each word  $W$  can be sent to the model, it must be transformed. This transformation is described in Section 4.2

The model contains an embedding layer, an encoder layer, a decoder layer, and a linear layer. The

input vector is transformed into an input embedding augmented with positional encodings, which provide information about the order of the syllables in a word. From there, the positional encoded input is passed to the encoder and decoder parts of the Transformer model. The encoder encodes the input into a series of hidden representations, which pass through the layers of the encoder. The encoder outputs a series of embeddings, which are then fed to the decoder, which generates the transcription of the word. The final part of the model is a linear layer, which transforms the output of the transformer model into unnormalized probabilities for each transcription token.

#### 4.2 Transforming a word $W$

The transformations applied over a given word  $w$  so that it can be used as input for the model are as follows:

The word  $W$  is split into its syllables  $s^W$ . For each syllable  $s$ ,  $f(s)$  is applied, resulting in a vector of indices of the syllables of  $W$ ,  $i_W$ . At the start and end of this vector, the unique tags  $\langle \text{bos} \rangle$  and  $\langle \text{eos} \rangle$  are added. As a result, the vector  $v_{i_W}$  is created and can be used by the model.

The model's output is a vector  $o_{i_W}$ , which has a size of  $l_W$ . It contains the special symbols  $\langle \text{bos} \rangle$  and  $\langle \text{eos} \rangle$ . After their removal, the vector  $i_{t_W}$  is left. We apply  $f'$  for every index  $i_{t_s}$ , to get the transcription  $t_s$  for the syllable  $s$ . Once we have all the syllables, we get the transcription  $t_W$  of the word  $W$ .

#### 4.3 Training of the model

For training the module, the samples are passed in batches of 128.

The parameters of the model chosen during training are the following:

- Amount of Encoder layers: 3
- Amount of Decoder layers: 3
- Embedding size ( $E$ ): 512

The loss function used for training is the Cross entropy loss function. It is used when training models that solve multi-class classification problems. With this function, we can quantify how well the model performs. It evaluates the output vector of the model against the expected result vector and returns a scalar. The lower the number, the better the model performs.

The training is done for 25 epochs. Figure 3 shows the loss value change across each training epoch.

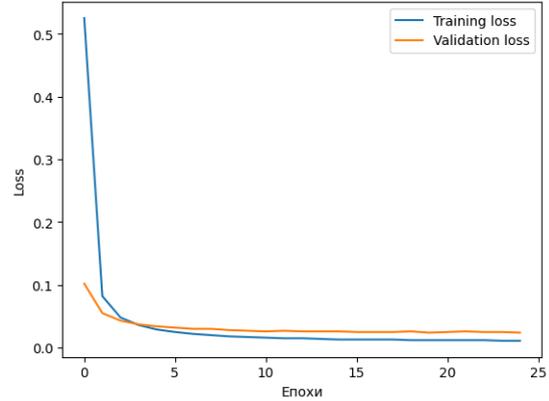


Figure 3: Results of the loss function on the train and validation datasets across the epochs

On the 25th epoch, the value of the loss function is as follows: Train set - 0.011, validation set - 0.024.

### 5 Usage of the phonetic embeddings

In this section, we will use the following notation:

- $E_W$  - the resulting vector from the encoder layer of the model. Size:  $(l_W, emb)$
- $wt_s(W)$  - the function that transforms the word  $W$  to the valid input for the model. This function is described in Section 4.2.

The embeddings result from using only the encoder part of the model. The user enters a word  $W$ , which is transformed (4.1) into the input vector for the model  $v_{i_W}$ . From there, the vector is passed to the encoder part of the model, which returns the vector  $E_W$ . As the model works on syllables, the returned embedding is not a single vector, but  $l_W$  vectors, each of size  $emb$  (which in this case is 512), containing floating-point numbers.

### 6 Experiments

Apart from the notation used in Section 4.1, we will introduce the following symbols:

- $t_W^p$  - the transcription of the word  $w$ , which results from the model working on the word  $W$ .
- $t_W^a$  - the transcription of the word  $W$ , taken from the dataset.

### 6.1 Result over random words

The model is run over a random set of Bulgarian words. The results of this experiment can be seen in Table 3.

| word      | generated trascription |
|-----------|------------------------|
| здравей   | zdrΛvej                |
| благодаря | blΛgodΛrjε             |
| лято      | ljεto                  |
| зима      | zimΛ                   |
| книга     | knigΛ                  |
| кафе      | kΛfε                   |
| синьо     | sinjo                  |
| часовник  | tʃΛjεjεsnik            |
| градина   | grΛdinΛ                |
| слънце    | slεntse                |
| месец     | mεsets                 |
| живот     | ʒivot                  |
| река      | rεkΛ                   |
| музика    | mozikΛ                 |
| храна     | xrΛnΛ                  |

Table 3: Results on a random set of Bulgarian words

As we can see, most of the words are correct. One of the examples that is wrong is the word ”часовник”, the transcription of which is generated as tʃΛjεjεsnik. Converting this transcription back to a written form would result in the word ”чаяясник”, which is incorrect.

### 6.2 Jaccard index

The Jaccard index is the first objective metric to evaluate the model’s performance. This metric evaluates the similarity between two sets and is defined over sets and multisets.

Let’s define the function  $J(A, B)$ , which measures the Jaccard index over two sets.

For two sets,  $A$  and  $B$ ,  $J(A, B)$  is defined as follows:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

As the transcription  $t_W$  of the word  $W$  can be considered an ordered multiset containing the transcriptions  $t_s$  of the syllables  $s$ , we can use the Jaccard index as an evaluation metric. This metric can evaluate what percentage of the generated syllables are correct (ignoring their ordering). As the transcription of a single syllable can appear multiple times in the transcription of the word, we must modify the Jaccard index to support multisets.

We can define a function  $J_m(A, B)$ , where  $A$  and  $B$  are multisets, in the following way:

$J_m(A, B) = \frac{|A \cap B|}{|A| + |B|}$ . It’s important to mention that  $J_m(A, B)$  returns values in the  $[0; 1/2]$  range. To normalize this interval, the result is multiplied by 2.

We can now define the metric Jaccard index for two transcriptions  $t_W^a$  and  $t_W^p$  in the following way:

$$J_t(t_W^a, t_W^p) = 2 * \left( \frac{|t_W^a \cap t_W^p|}{|t_W^a| + |t_W^p|} \right)$$

Evaluating the test set with the Jaccard index metric yields a result of 99.571% match

### 6.3 Direct comparison

The Jaccard index only evaluates whether the generated transcriptions of the syllables are correct, ignoring their ordering. Another metric must be defined to get a metric that includes the ordering. A good candidate is the direct comparison of the two transcriptions.

We can define a function that compares the syllables of two transcriptions  $t_W^a$  and  $t_W^p$  pairwise.

$$c(t_W^a, t_W^p) = \begin{cases} 1 & \text{If } |t_W^a| = |t_W^p| \\ & \text{and for } \forall i \in 0..|t_W^a| \\ & \text{the following is true: } t_{W_i}^a = t_{W_i}^p \\ 0 & \text{otherwise} \end{cases}$$

We are evaluating the test set with the direct comparison metric, which yields a result of 99.285% match.

Comparing this result to the Jaccard index metric, we can see that 0.286% of the generated transcriptions have all the correct syllables but in a different order than in the correct transcription.

## 7 Paronyms detection

### 7.1 Introduction

In the Bulgarian language, two words are paronyms if they are close in sound but different in meaning. As an example, the words ”статист” и ”статистик” sound close (in this case, they share syllables) but have vastly different meanings.

We hypothesize that the embeddings from the model described above can be used to detect if two words are paronyms. This can be done by classifying two words based on their ”phonemic distance”—the distance that would model how the words sound.

The metric used to calculate this ”phonemic distance” is the Cosine similarity metric. This metric gives us the level of similarity between two vectors. It’s defined the following way:

$S_C(A, B) = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$ . The cosine similarity returns values in the  $[-1; 1]$  range. A value of -1 means the vectors are opposite, while one means the vectors are proportional.

Based on this, we can define cosine similarity between two embeddings.

Let  $E_1$  and  $E_2$  be the embeddings of the words  $W_1$  and  $W_2$ . Let  $l_{W_1}$  and  $l_{W_2}$  be the amount of syllables in the two words and  $l = \min(l_{W_1}, l_{W_2})$ . Then:

$$S_E(E_1, E_2) = \frac{\sum_{i=0}^l S_C(E_{1i}, E_{2i})}{l}$$

$S_E$  again belongs to the  $[-1; 1]$  range. Result  $-1$  will be interpreted as the words don't sound alike, while  $1$  will mean that the words do sound alike.

## 7.2 Experiments and results

From [Върбанова and Весела Кръстева \(2009\)](#) 92 pairs of paronym words were collected. Ninety-three pairs of non-paronym words were added, resulting in a dataset of 185 pairs. This dataset was split into train and test sets in an 80:20 ratio—148 pairs for training and 37 for testing.

A logistic regression classifier is used to determine whether two words are paronyms. The model uses the cosine similarity between the embeddings of the two input words as input.

The following notation and metrics are used to evaluate the model: accuracy, precision, recall, and F1 score.

As a comparison, we used the word embeddings from the fastText library [Bojanowski et al. \(2016\)](#), [Joulin et al. \(2016\)](#) to train the same classifier on the same dataset. The results are shown in Table ??

| Metric    | Score    |          |
|-----------|----------|----------|
|           | Phonetic | fastText |
| accuracy  | 0.892    | 0.892    |
| precision | 0.850    | 0.938    |
| recall    | 0.944    | 0.833    |
| F1        | 0.895    | 0.882    |

Table 4: Results over the paronyms dataset comparing phonetic embeddings with fastText word embeddings

As we can see, our phonetic embeddings show the same accuracy as the fastText ones. fastText performs better on the precision metric, however loses a bit in the recall. Comparing the F1 scores shows that our phonetic embeddings perform around the same as the fastText embeddings.

## 8 Language of origin detection for Bulgarian words

### 8.1 Introduction

As the phoneme embeddings mentioned in Section 5 model the way a word sounds, we propose that the embeddings can be used to detect the language of origin of a Bulgarian word. We assume that loanwords in Bulgarian will sound differently than regular Bulgarian words. As part of this section, an RNN-based classifier is trained using data from loanword dictionaries. The dataset consists of pairs of words and their language of origin. Table 5 shows the amount of loanwords in the dataset. From there, three datasets are created, with different amounts of Bulgarian words - with 5 000 Bulgarian words, with 13 395 Bulgarian words, and with 30 875 Bulgarian words. Each model accepts Bulgarian words as input, gets its phoneme embedding, and returns the probability of the word belonging to a certain language.

| Language | Word count | % of total |
|----------|------------|------------|
| Latin    | 1504       | 29.10%     |
| Greek    | 984        | 19.04%     |
| French   | 958        | 18.51%     |
| Turkish  | 658        | 12.73%     |
| English  | 478        | 9.25%      |
| German   | 240        | 4.64%      |
| Italian  | 156        | 3.02%      |
| Russian  | 102        | 1.97%      |
| Spanish  | 33         | 0.64%      |
| Dutch    | 29         | 0.59%      |
| Hebrew   | 11         | 0.21%      |
| Arabic   | 7          | 0.14%      |
| Serbian  | 4          | 0.08%      |
| Persian  | 3          | 0.06%      |

Table 5: Words from a given language

From there, three models are trained, depending on the number of Bulgarian words used - we'll call these models *Phoneme-5k*, *Phoneme-13k* and *Phoneme-30k*.

### 8.2 Experiments and results

The standard metrics—accuracy, precision, recall, and F1—are used to evaluate the classifiers. For comparison, a classifier using the fastText word embeddings was trained on the same datasets. The fastText-based classifiers will be referred to as *fastText-5k*, *fastText-13k* and *fastText-30k*.

| Model              | Metric   |           |        |       |
|--------------------|----------|-----------|--------|-------|
|                    | Accuracy | Precision | Recall | F1    |
| <b>Phoneme-5k</b>  | 0.590    | 0.553     | 0.590  | 0.565 |
| <b>Phoneme-13k</b> | 0.736    | 0.727     | 0.736  | 0.730 |
| <b>Phoneme-30k</b> | 0.806    | 0.807     | 0.806  | 0.806 |
| fastText-5k        | 0.672    | 0.697     | 0.672  | 0.653 |
| fastText-13k       | 0.794    | 0.797     | 0.794  | 0.789 |
| fastText-30k       | 0.868    | 0.864     | 0.868  | 0.861 |

Table 6: Comparison between our phoneme embeddings and fastText embeddings for the language of origin task

The results from Table 6 show that the fastText embeddings perform a bit better than the phoneme embeddings, although it’s not a sizeable difference.

## 9 Limitations

While the results of the model are looking good, there are some limitations on it. The input of the model does not include information about the emphasis of the word. This information is only used to generate the proper transcription for the training data. This impacts the phonemes generated by the model.

The phonetic embeddings also don’t seem to exhibit any of the properties present in other word embeddings. Embeddings like word2vec Mikolov et al. (2013) represent the semantic and syntactic relationships between the words. For example, the distance between the words ”man” and ”woman” is similar to the distance between the words ”king” and ”queen”. The phonetic embeddings however don’t exhibit such connections. For example, if we have two words, which differ only at the suffix, the distance between their embeddings varies from small to a large, and is not consistent through different pairs of words.

## 10 Conclusion

A dataset containing 600,000+ words and their transcription was created. A transformer-based model was created to solve the Bulgarian language’s grapheme-to-phoneme task. The model performs with very high accuracy. Embeddings extracted from this model were used in a simple classifier that checks if two words are paronyms. The classifier performs also with a high accuracy percentage.

As a result, there is now an automatic system for paronym detection and automatic generation of phonemic transcriptions of Bulgarian words.

These approaches can be applied to other languages that are different from Bulgarian. In future research, we intend to use the embeddings to detect if a word is a loanword.

## 11 Acknowledgments

The work is partially financed by the European Union-NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project No BG-RRP-2.004-0008.

## References

- Maximilian Bisani and Hermann Ney. 2008. [Joint-sequence models for grapheme-to-phoneme conversion](#). *Speech Commun.*, 50:434–451.
- Alan Black and Kevin Lenzo. 2003. Issues in building general letter to sound rules.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.
- H. S. Elovitz, Rodney W. Johnson, Astrid McHugh, and John E. Shore. 1976. [Automatic translation of english text to phonetics by means of letter-to-sound rules \(nrl report 794\)](#).
- Eric Engelhart, Mahsa Elyasi, and Gaurav Bharaj. 2021. [Grapheme-to-phoneme transformer model for transfer learning dialects](#).
- Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, H erve J egou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.
- Xinjian Li, Florian Metze, David Mortensen, Shinji Watanabe, and Alan Black. 2022. [Zero-shot learning for grapheme to phoneme conversion with language ensemble](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2106–2115, Dublin, Ireland. Association for Computational Linguistics.

- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. [Efficient Estimation of Word Representations in Vector Space](#).
- Kanishka Rao, Fuchun Peng, Haşim Sak, and Françoise Beaufays. 2015. [Grapheme-to-phoneme conversion using Long Short-Term Memory recurrent neural networks](#) | IEEE Conference Publication | IEEE Xplore.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. [Attention Is All You Need](#).
- Sevinj Yolchuyeva, Géza Németh, and Bálint Gyires-Tóth. 2019. [Transformer based grapheme-to-phoneme conversion](#). In *Interspeech 2019*, pages 2095–2099.
- Институт за български език при БАН. 2011. [Правопис и пунктуация на българския език](#). [Institut za balgarski ezik. Pravopis i punktuatsiya na balgarskiya ezik. Prosveta, 2011.]
- Павлина Върбанова. 2021. [Списък на пароними. Как се пише](#). [Pavlina Varbanova. Spisak na paronimi. Kak se pishe.]
- Весела Кръстева. 2009. [Практическа граматика на съвременния български език. Кръгзор](#). [Vesela Krasteva. Prakticheska gramatika na savremenniya balgarski ezik. Kragozor, 2009.]

# Universal Dependencies Treebank for Standard Albanian: A new approach

|   |   |   |
|---|---|---|
| <b>Nelda Kote</b><br>Polytechnic University of Tirana<br>Tirana, Albania<br><i>nkote@fti.edu.al</i> | <b>Rozana Rushiti</b><br>University of Tirana<br>Tirana, Albania<br><i>rozana.rushiti@unitir.edu.al</i> | <b>Anila Çepani</b><br>University of Tirana<br>Tirana, Albania<br><i>anila.cepani@unitir.edu.al</i> |
|---|---|---|

|   |  |  |
|---|--|--|
| <b>Alba Haveriku</b><br>Polytechnic University<br>of Tirana, Tirana, Albania<br><i>alba.haveriku@fti.edu.al</i> | <b>Evis Trandafili</b><br>Polytechnic University of Tirana<br>Tirana, Albania<br><i>etrandafili@fti.edu.al</i> | <b>Elinda Kajo Meçe</b><br>Polytechnic University<br>of Tirana, Tirana, Albania<br><i>ekajo@fti.edu.al</i> |
|---|--|--|

|   |  |   |
|---|--|---|
| <b>Elsa Skënderi Rakiplari</b><br>University of Tirana<br>Tirana, Albania<br><i>elsa.skenderi@unitir.edu.al</i> | <b>Lindita Xhanari</b><br>University of Tirana<br>Tirana, Albania<br><i>lindita.latifi@unitir.edu.al</i> | <b>Albana Deda</b><br>University of Tirana<br>Tirana, Albania<br><i>albana.deda@unitir.edu.al</i> |
|---|--|---|

## Abstract

In this paper, we present a Universal Dependencies (UD) treebank for the Standard Albanian Language (SAL), annotated by expert linguistics supported by information technology professionals. The annotated treebank consists of 24,537 tokens (1,400 sentences) and includes annotation for syntactic dependencies, part-of-speech tags, morphological features, and lemmas. This treebank represents the largest UD treebank available for SAL. In order to overcome annotation challenges in SAL within the UD framework, we delicately balanced the preservation of the richness of SAL grammar while adapting the UD tagset and addressing unique language-specific features for a unified annotation.

We discuss the criteria followed to select the sentences included in the treebank and address the most significant linguistic considerations when adapting the UD framework conform to the grammar of the SAL. Our efforts contribute to the advancement of linguistic analyses and Natural Language Processing (NLP) in the SAL. The treebank will be made available online under an open license so that to provide the possibility for further developments of NLP tools based on the Artificial Intelligence (AI) models for the Albanian language.

**Keywords:** syntactic dependencies, UPOS, morphological features, Standard Albanian Language, manually annotated corpus

## 1 Introduction

The Albanian language is part of the Indo-European family and is spoken in Albania, Kosovo,

North Macedonia, Montenegro, and other diaspora communities. The language has several unique characteristics that distinguish it from other European languages. There are regional variations and dialects in the language, with the two main dialects being Gheg, which is spoken in northern Albania, and Tosk, which is spoken in southern Albania, as well as in diaspora communities in Greece and Italy. These two dialects have distinct lexical differences. The Standard Albanian Language (SAL) is based on the Tosk dialect (Hamp, 2023).

Universal Dependencies (UD) (Nivre et al., 2020) is a framework focused on the provision of research at a multilingual level for morphological and syntactic annotation. The currently available treebank for Albanian language in the UD framework consists of only 60 sentences, annotated with lemmas, morphological, and syntactic features (Toska et al., 2020).

This paper presents the Standard Albanian Language Treebank (SALT) annotated conform UD framework by expert linguistics with the support of information technology professionals, consisting of 24,537 tokens (1,400 sentences). The annotation includes sentence segmentation, word segmentation, universal part-of-speech (UPOS) tags, morphological and syntactic features, and lemmas, offering a new valuable resource for the study of the SAL.

**Contributions:** We discuss the methodology to select the sentences of our treebank and the inclusive criteria. Furthermore, after extensive work to identify special linguistic features in SAL, we present the decisions made by the expert linguistics

group for aspects such as verb form, noun declension, adjective agreement, and different syntactic problems, essential to create a proper standard for the annotation. The summary of our contributions is:

- Presenting the new SALT treebank with 24,537 tokens, 21 times larger than the existing treebank TSA (Toska et al., 2020).
- Emphasizing the most significant linguistic characteristics required to align the UD schema with the characteristics of the SAL by creating a valuable resource for interested researchers.

The rest of this paper is organized as follows: In Section 2, we review significant background and related research works. In Section 3, we discuss the treebank development, sentence collection, and the annotation process. Section 4 covers the treebank annotation schema and related discussions. Finally, in Section 5, we conclude our work and discuss directions for future research.

## 2 Background and Related Work

Research efforts for low-resource languages like Albanian have been historically constrained. Although there have been several attempts to develop annotated corpora in the Albanian language, they have either remained closed-source or proved to be too limited in size and lacking interconnection between them.

The largest existing corpora for the Albanian language are the “Albanian language corpus” (16.6 million tokens) created by the Saint-Petersburg Institution (Arkhangelskij et al., 2011) and the sq-Globe corpus (1 million words) by the Beijing Foreign Studies University (Ke et al., 2012) annotated with POS tags and lemmas. These two corpora are the largest ones for the Albanian language, but due to them being closed-source, they are not suitable for further research works. Meanwhile, Caka and Caka (2011) have created a closed-source corpus with one million words, which is lemmatized and includes grammatical properties.

Kote et al. (2019) present a corpus containing 118,000 tokens, annotated at the morphological level based on the UD schema, the UniMorph project (Kirov et al., 2018) present a treebank with 33,483-word forms and 589 lemmas and Toska et al. (2020) present the first official UD treebank for

the Albanian language but containing only 60 sentences. Other related works include Kadriu (2013), Kabashi and Proisl (2018), Misini et al. (2020), Ebert et al. (2022) and Mati et al. (2021) that have contributed in different tagsets and small treebank for standard and Gheg Albanian.

## 3 Standard Albanian Language Treebank

This section outlines the development of the SALT treebank. Two expert linguistics conducted manual annotation due to the lack of preprocessing tools and resources for SAL. Tables 1 and 2 show statistics about the SALT treebank.

|                             |        |
|-----------------------------|--------|
| <b>Number of sentences</b>  | 1,400  |
| <b>Number of tokens</b>     | 24,537 |
| <b>Multiword tokens</b>     | 87     |
| <b>Avg. sentence length</b> | 18     |

Table 1: Statistics of the SALT treebank.

| <b>UPOS</b> | <b>frequency</b> | <b>Deprel</b> | <b>frequency</b> |
|-------------|------------------|---------------|------------------|
| NOUN        | 5,353            | punct         | 2,893            |
| PUNCT       | 2,908            | det           | 2,304            |
| DET         | 2,736            | case          | 2,099            |
| VERB        | 2,697            | advmod        | 1,647            |
| ADP         | 2,139            | nsubj         | 1,620            |
| PRON        | 1,738            | nmod          | 1,467            |

Table 2: Frequency of the most used tags.

### 3.1 Data Collection and Selection

The treebank consists of 1,400 sentences containing 24,537 tokens. To prevent potential proprietary rights conflicts, we selected sentences from open corpora. The sentences are extracted from fiction books, a grammar book, and the Leipzig Corpora Collection (Goldhahn et al., 2012). Before annotation, all the sentences are grammatically corrected for any error by the expert linguistics. This step is necessary because texts in the Albanian language available in open-source corpora often exhibit grammatical errors such as missing letters like “ë” or “ç”, typographical mistakes, etc.

### 3.2 The Annotation Process

To facilitate the annotation process the selected sentences are pre-annotated using the model proposed by Kote et al. (2019) for segmentation, lemmatization, part-of-speech, and morphological features. Subsequently, the annotated sentences are reviewed

|                   | POS tag     | Morphological features  |
|-------------------|-------------|---|
| verb              | VERB/AUX    | mood, time, person, number, voice;<br>*verb form only in case of participle |
| noun              | NOUN        | gender, number, case, definiteness  |
| proper noun       | PROPN       | gender, number, case, definiteness;<br>Abbr in case of abbreviation         |
| adjective         | ADJ         | gender, number, case, degree  |
| pronoun           | PRON        | depends on the type<br>(case, number, gender, person, prontype)             |
| adverb            | ADV         | AdvType   |
| numeral           | NUM         | NumType   |
| interjection      | INTJ        |   |
| preposition       | ADP         | case  |
| particle          | PART        |   |
| conjunction       | CCONJ/SCONJ |   |
| articles          | DET         | gender, number, case and prontype   |
| symbols           | SYM         |   |
| punctuation marks | PUNCT       |   |
| others            | X           | Abbr in case of abbreviation  |

Table 3: The list of the POS tags and morphological features.

to ensure accurate sentence and word segmentation, with any errors corrected as needed. Additional scripts are used to identify and correct other errors and add missing morphological features. Following this, expert linguistics manually reviewed and annotated all the sentences. The syntactic annotation is entirely done manually, as there is no available trained model for it. The expert linguistics have reached an agreement among themselves to use a standardized method for annotation.

Two software applications, Conllu Editor (Heinecke, 2019) and Arborator Grew (Guibon et al., 2020), are used for annotation.

The annotation include:

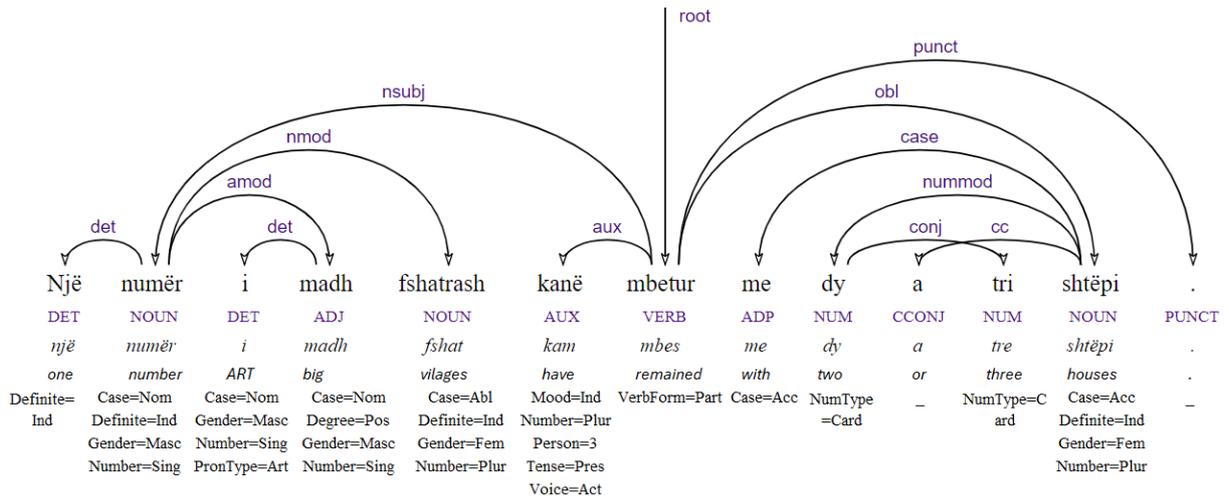
- **Sentence segmentation:** The selected text is segmented in sentences, with titles segmented as a separate sentence.
- **Word segmentation:** Word segmentation was performed using white space and punctuation marks as boundaries, leading to challenges in identifying analytical grammatical forms and various expressions. As the Albanian language is a synthetic-analytical language, with a prevalence of synthetic features but a tendency towards analytic structures (compound verb tenses, the future tense, certain verb forms, nonfinite verbs, the nominative adjective, genitive of nouns, conjunction locu-

tions, prepositions, and phraseological expressions), labeling becomes intricate.

- **Lemmaization:** Expert linguistics used the Albanian National Dictionary, (ASHSH, 1998, 2002, 2006) to determine the lemma of a word. The lemma is assignment based on the context and meaning of the word form in the sentence.
- **Part-of-speech tags:** A total of 17 part-of-speech tags from the UD tagset are utilized.
- **Morphological features:** We applied corresponding morphology features based on the word’s part-of-speech tag.
- **Syntactic annotation:** A total of 32 syntactic tags from the UD tagset are utilized.

#### 4 Annotation Schema

In this section, we discuss the key considerations of the annotation schema used to annotate text data within the UD framework. The grammar of the Albanian language has a complex inflection schema and a rich morphological and syntactic structure, which presents several challenges in annotation due to the presence of unique features specific to the language.



“A large number of villages are left with two or three houses.”

Figure 1: Annotation of a sentence with compound verb tense.

In the annotation process, the grammatical guidelines published by the Albanian Academy of Sciences (Agalliu et al., 2002) are taken into account.

#### 4.1 Part-of-Speech and Morphological Annotation

We have utilized 10 universal part-of-speech tags (verb, noun, adjective, pronoun, adverb, numeral, interjection, preposition, particle, conjunction) from the UD framework, along with their respective grammatical features. Furthermore, we have utilized 4 other tags for various elements present in SAL grammar, such as articles, abbreviations, symbols, and punctuation marks. Table 3 shows the list of used UPOS tags and their corresponding morphological features.

The verb system is one of the most complex aspects of SAL grammar, comprising 6 moods and 14 tenses in total (Agalliu et al., 2002). However, when annotating each word separately, some moods and tenses for compound verb tenses may not be explicitly displayed, such as the future indicative tense, future perfect tense, etc. Each part of the compound verb is separately annotated depending on its form. Another case to discuss is the annotation of the verbs “kam/to have” and “jam/to be,” that are used for three different purposes:

- As copula, it is annotated with VERB tag.
- As auxiliary verb to form the compound tenses, it is annotated with AUX tag.
- As main verb, it is annotated with VERB tag.

Unfortunately, there is no specific morphological tag available to distinguish between the uses of this verb. This necessitates defining these cases as verbs with their full lexical meaning, even when they function as auxiliary verbs. Figure 1 illustrates an example.

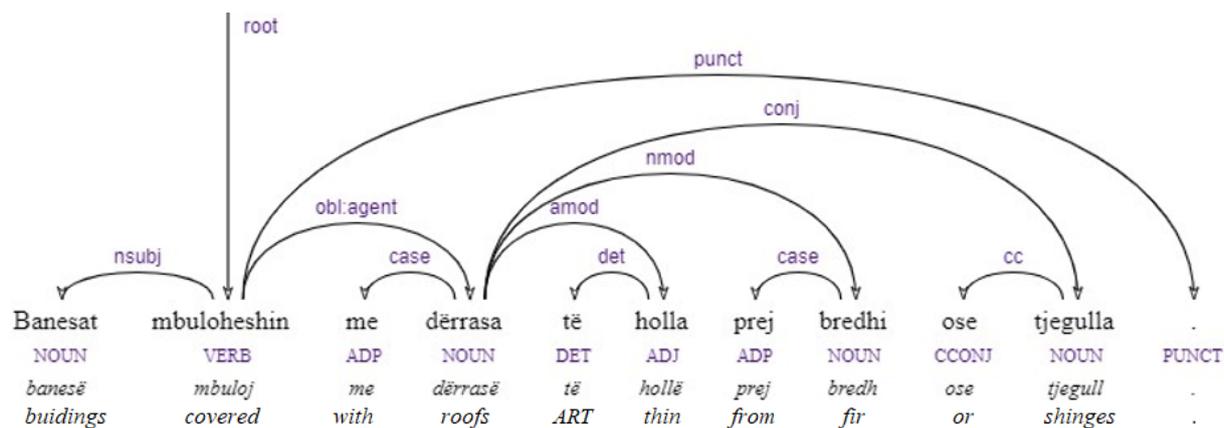
The adjectives agree with the governing nouns in number, gender, and case regardless of the degree. There are three degree categories: positive, comparative, and superlative. The comparative and superlative forms are created as analytic forms utilizing lexical elements, with the positive degree being the main focus of annotation, as it serves as the base for these forms. Another aspect not covered in adjective annotation is the categorization of articulated and non-articulated adjectives. The article of the articulated adjective is annotated as DET, like the adjective “i madh/big” in the sentence presented in Figure 1.

Pronouns are classified into seven distinct classes, each annotated with different morphological features. Some share common attributes such as case, number, and gender, while others lack specific categories, including abbreviations that are also annotated as pronouns.

The conjunctions and adverbs are annotated based on their types, but the annotation for conjunctions doesn’t encompass semantically related subtypes.

Prepositions are annotated with case morphological features to aid in syntactic analysis and to determine the type of syntactic relationships formed with prepositional phrases.

Given the diverse and multifaceted nature of the



“The flats were covered with thin fir boards or tiles.”

Figure 2: Annotated sentence where the root is a verb.

particles, they are not annotated with specific type tags, as they encompass both semantic and grammatical dimensions.

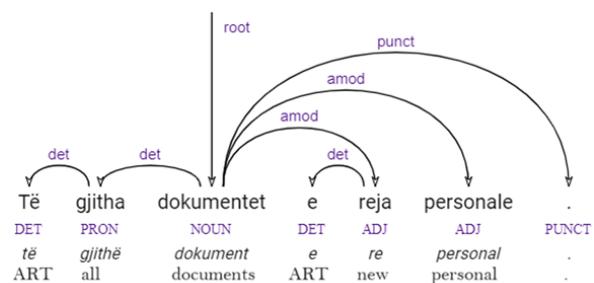
It is important to clarify how we have used the determiner (DET) in our annotation based on the UD framework. In traditional Albanian grammar, articles are not categorized as a separate part of speech "Determiner - DET" but are instead treated as morphological elements associated with nouns, adjectives, and pronouns, inheriting features such as gender, number, and case. Despite this, we have aligned articles with the DET category in UD based on their functional roles, which include indicating definiteness, gender, number, and case. This mapping preserves the grammatical features of articles while ensuring consistency with UD principles. By doing so, we maintain the traditional grammatical structure of Albanian while leveraging UD's universal annotation schema, thereby providing an accurate and comprehensive representation of Albanian articles within the UD framework.

### 4.2 Syntactic Annotation

The root of the sentence is indicated by the root tag, usually by labeling the principal verb, sentence designer, or principal unit verb in compound sentences. When the verb (predicate), which marks thematic roles in a sentence, is absent (due to ellipses), and multiple orphaned subordinates exist, by agreement we decide that one of these subordinates takes on the role of the root while the others relate to it. As a result, a noun is labeled as the root, although in Albanian language, an adjective or another noun can also assume the root role through agreement. Figure 1, Figure 2, and Figure 3 show examples of annotated sentences with a verb as the

root and a noun as the root.

The nsubj (nominal subject) tag is the external argument (the headword) or syntactic subject representing the agent acting, whether expressed through a noun as shown in Figure 1 and Figure 2, pronoun, numeral, or nominal expression.



“All new personal documents.”

Figure 3: Annotated sentence where the root is a noun.

The obl (nominal oblique) tag is used for a noun phrase, specifically when its head is a nominative case preposition+noun, pronoun, or a noun phrase, or any preposition in another case, serving as a non-core or complement argument. Often, this functions like an adverb linked to another verb, adjective, or adverb. To specify the noun subject of a passive verb, the subtype pass (passive) is typically used in the UD framework, but during our labeling, we employed obl: agent for such structures. The obl:agent is used for an oblique noun phrase indicating the agent, and the obl: arg is used for an oblique noun phrase functioning as an argument, as in the example shown in Figure 4.

The obj (direct object) tag is used for direct object of the verb, where the verb action falls, irrespective of whether it's a noun or a pronoun (full

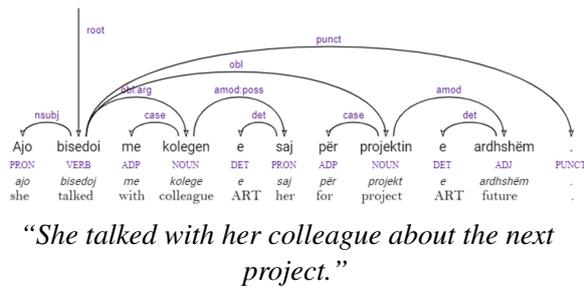


Figure 4: Example using obl:arg tag.

form+clitic/full form/clitic) in the accusative case. It happens in languages where obj is labeled with the morphological case. Figure 5 shows examples.

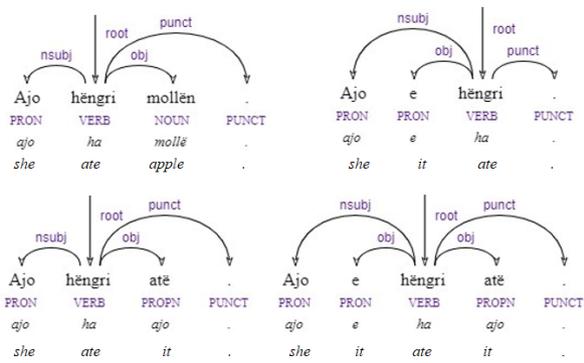


Figure 5: Examples using obj tag.

In the Albanian language, iobj (indirect object) tag is associated with arguments in the dative case, such as in the example shown in Figure 6.

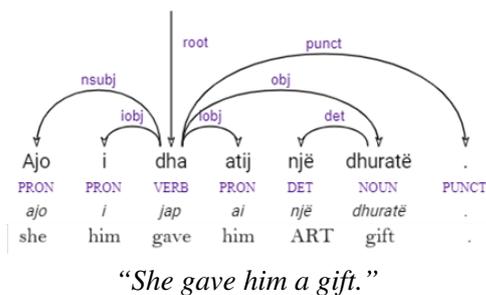


Figure 6: Example using iobj tag.

The nmod (nominal modifier) tag represents a nominal modifier that modifies a noun and we use it for noun dependents of another phrase, such as attributes or complements, associated with the head of the noun phrase.

The amod (adjectival modifier) tag is used not only for adjectival modifiers of a noun or a pronoun but also pronominal modifiers to which the poss subcategory has been added when they are possessive pronouns.

The appos (appositional modifier) tag is used for a noun or noun phrase that explains or defines another noun in the role of an affix. It is also used to link names or noun phrases when providing supplementary information such as email addresses, phone numbers, or residential addresses.

The advmod (adverbial modifier) tag is used for adverb or adverbial phrase emphasizes the modification of another verb, adjective, or adverb, as in the example shown in Figure 7.

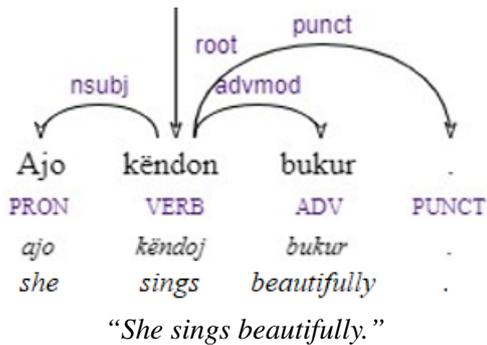


Figure 7: Example using advmod tag.

In Albanian language, a limited set of adverbs can also modify nouns, as in “vetëm të hënë/only on Monday”, where the advmod tag is used, emphasizing the adverb’s role, like advmod:emph.

The term “adverbial modifier” encompasses compounds functioning as adverbs, whether adverbs, non-clausal phrases, or nouns in specific morphological cases, as is the case in the Albanian language. We distinguish modifiers as adverbs (advmod) and others as non-clausal phrases or adverbs (obl). However, we do not differentiate between predicate verb modifiers, so adverbs in a strict sense, and modifiers of other modifying words like adjectives or adverbs, as all these are under the advmod category. The obl tag is used when the circumstantial element is obligatory (an argument), while the ad mod tag is used when the circumstantial element is optional.

The aux (auxiliary) tag indicates a verb that is linked to another verb (predicate), typically serving as an auxiliary verb used to form analytical verb forms as shown in Figure 1. It also includes semi-auxiliary verbs with modal significance, which express the manner of an action and can often have full lexical meanings, “Ajo nisi/mund të kryejë detyrat e shtëpisë / She begin to/can do her homeworks”. The aux tag is utilized to represent passive voice constructions and, in languages with a grammaticalized (periphrastic) passive form, the subtype

aux:pass is encouraged for usage. So, the aux: pass (passive auxiliary) tag is used for auxiliary verbs associated with the past participle, as observed in the sentence shown in Figure 8.

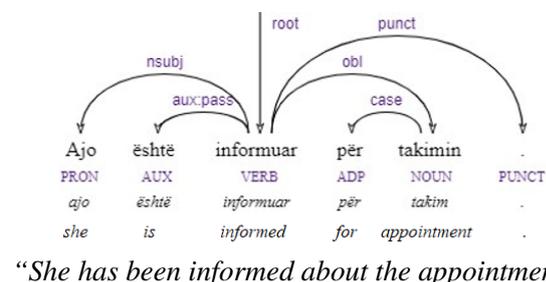


Figure 8: Example using aux: pass tag.

The cop (copula) tag represents the relationship between a subject and a non-verbal predicate, connecting the noun with the subject. In UD, the non-verbal predicate is the root from which all other syntactic connections are created, but referring to the specifics of Albanian, we have annotated as root the verb “jam/to be” which is a copula. Figure 9 shows the difference between the UD annotation of the copula and our annotation.

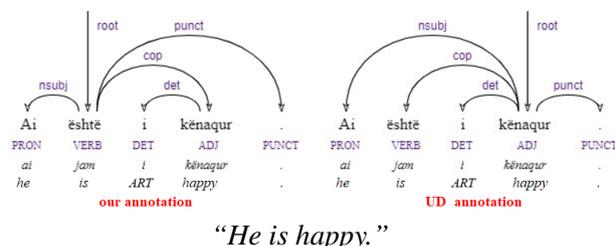


Figure 9: Examples using cop tag.

The nummod (numeric modifier) tag is used for a numerical modifier of a noun, which can be any number that modifies and indicates its quantity. A number that serves as a label for another entity and does not indicate quantity is not labeled as nummod but nmod as in: “Takimi do të jetë në dhomën 4./The meeting will be in room 4.”

The det (determiner) tag is used for a determiner that specifies the noun it modifies. It connects a noun/adjective head, as well as certain types of pronouns with their determiner or modifier. In the context of SAL, this tag applies to all non-significant determiners of the name, all word-forming and shape-forming determine, and so in addition to the det tag, there must be subcategories like det: adj, det: pron, and det: poss.

The case (case marking) tag is used for the analytical case marker, which is treated as a special

syntactic word, such as prepositions in SAL as shown in Figure 1.

The conj (conjunct) tag indicates connections with coordination between members and parts of the compound sentence. The main part of the connection is called the first part, and all subsequent parts are connected to it via the conj tag.

The cc (coordinating conjunction) tag connects a coordinating conjunction to a word-member or compound sentence with coordination.

The csubj (clausal subject) tag is used in the subject sentence.

The xcomp (open clausal complement) tag is used for predicate clauses, and in the case of complementary clauses, the ccomp (clausal complement) tag is used.

The acl (clausal modifier of a noun) and acl:relcl (relative clause modifier) are used for a dependent and relative clause that modifies nouns.

The advcl (adverbial clause modifier) tag is used to indicate a dependent clause that functions as an adverbial modifier. Figure 10 shows an example.

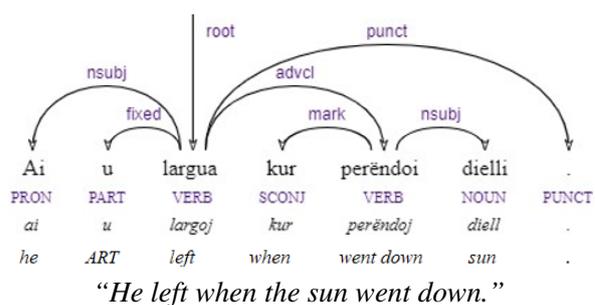


Figure 10: Example using advcl tag.

We should emphasize that tags such as csubj, ccomp, xcomp, acl, and advcl, which are used to connect in minimal structures, the leading verb with a dependent unit or a noun with a dependent unit, referring to the classificatory criterion of which component of the leading unit the subordinate unit modifies, so they are labeled as either complementary, necessary dependents (csubj, ccomp, xcomp) or as optional determiner-relational dependents modifying a noun in the governing unit acl, or as additional adverbial circumstantial dependents advcl.

The discourse tag is used when an element of discourse is in a sentence, such as an exclamation that shows emotional content.

The dislocated tag is used for an element that appears separately from its syntactic position, and also intermediate words or phrases, interlaced, and

sentences that have no grammatical connection with the rest of the sentence.

Adverbial, prepositional, and conjunctive locutions are annotated using the fixed tag that expresses fixed expression.

A fixed phrase composed of two or more words that convey a particular grammatical structure is annotated using the fixed:form tag.

The flat (flat multiword expression) tag is used for phraseological expressions or various nouns.

The list tag is used when parallel elements are listed syntactically. In the case of homogeneous members with coordinating conjunctions, the conj tag is used. Meanwhile, the list tag can be used only when there are no conjunctions.

The parataxis tag refers to phrases or sentences which lack conjunctions or connecting elements.

The mark (marker) tag is used for subordinating conjunctions, particles, etc.

The punct (punctuation) tag is used for punctuation marks. Since they do not follow a typical dependency relationship, several criteria are used to determine their main associated word. Generally, periods, exclamation points, and question marks, which indicate sentence conclusions, are connected to the main verb serving as the root. Commas are connected to coordinating or subordinating conjunctions, or words designated as part of a list. Paired punctuation marks, such as quotation marks, brackets, and sometimes hyphens, are connected to the same word.

The vocative tag indicates an element in the sentence that is a noun used to address a person.

The orphan tag expresses dependencies in the case of a missing head.

### 4.3 Annotation Discussions

Acknowledging the complexity of defining word boundaries in SAL due to its various grammatical structures, including compound verb tenses and phraseological expressions, we use the "fixed:" relation of UD for lexicalized multi-word expressions like the subjunctive mood form "të lexoj". However, we encountered challenges with the particle "të," especially when it merges with the accusative short form "e," forming "ta (të + e) lexoj". This merging complicates the use of the "fixed:" relation, as the resulting form must be accurately represented in the annotation. Therefore, while we have applied the "fixed:" relation where appropriate, addressing the merging of particles requires

additional scripts or manual intervention to ensure precise tagging of these compound structures.

Determining the lemma of adjectives can be challenging because, depending on the noun they modify, they can exist in both feminine and masculine forms. In these cases, we have chosen to use the masculine form specified in the dictionary as the lemma (e.g., "mahnitshëm/amazing" is the lemma used even for the masculine form "i mahnitshëm" and feminine form "e mahnitshme").

Articles, pronouns, and abbreviations can appear in different forms depending on the words they modify. In this case, each distinct variant found in the dictionary should be a lemma to ensure a precise and unambiguous representation with accurate labeling for these linguistic components.

Different elements, including subordinating conjunctions and various participles, are labeled using a mark tag. However, there is no specific label for passive particle and particle "të" that form the relative mood, the future tense, and the affirmative and negative participles.

The nsubj tag is used for all types of headwords (expressed by noun, pronoun, phrase) without the option to use additional tags for subcategorization.

There is no additional syntactic relation label for verbs that lack full lexical meaning and necessitate a complementary predicate. In this case, cop tag is used similar to the verb "jam/to be" in the nominal predicate. Indeed, the presence or absence of this sentence's element is a discussion in SAL syntax, much like the acceptance of the nominal predicate itself. Even for modal verbs (e.g., "mund/can", "duhet/should") and aspectual verbs expressing the initiation, continuation, or completion of actions (e.g., "filloj/start", "vazhdoj/continue"), there are no dedicated labels. In these cases, we use the aux relation used for the auxiliary verbs "jam/to be" and "kam/to have".

The oblique nominal with prepositions remains a subject of controversy among researchers and linguists, even to this day. As UD does not have a distinct label, we used obl tag.

No specific label to distinguish pronouns before or after a noun. For example, to label indefinite pronouns like "asnjë/none", "pak/few", "shumë/many" before a noun, the nummod tag is used because they express indefinite quantities and are treated as quantifiers, but this tag is not used when they follow a noun.

Conjunctions formed by a noun with a posses-

sive pronoun are labeled with *amod*, similar to the conjunctions formed by a noun with an adjective by adding a subcategory *poss*, as *amod:poss*. This tag is used even for indeclinable pronouns, thus broadening the usage of this label. Also, an issue to discuss is the textual conjunctions that appear at the beginning of a sentence, intermediate words, and interjections. The question is: "How should they be integrated into the sentence structure?"

For various types of punctuation marks (period, comma, question mark, exclamation mark, quotation marks, brackets, etc.), there is no specific label or subcategorization to define their functions. Therefore, they are all labeled with *punct* tag.

## 5 Conclusion and Future Work

This paper presents the Standard Albanian Language Treebank (SALT), the first Universal Dependencies (UD) treebank for the Standard Albanian Language (SAL), annotated by expert linguists. SALT includes annotations for syntactic dependencies, part-of-speech tags, morphological features, and lemmas. Adapting the rich and complex grammar of the Albanian language to the UD schema involves significant challenges, such as the absence of direct mappings and the ambiguity in assigning appropriate tags.

We present an overview of the language's grammatical structure, providing a detailed analysis of its key linguistic features. Additionally, we discuss methods for annotating texts in SAL according to the UD framework. Overcoming annotation challenges requires delicately balancing and harmonizing the richness of the language's grammar with adaptation to the UD tagset, while addressing the unique language-specific features for unified annotation. Expert linguists initially mapped the UD tagset to align with the language's grammar and subsequently performed manual annotations on the treebank.

As future work, we aim to use our proposed treebank, composed of 1,400 sentences, as a training and testing dataset for an Albanian language parser. This automated tool will facilitate the annotation of a larger treebank, aiding linguistics, computer scientists and related fields to conduct further research work on the Albanian language.

Our treebank will be available online with open access for research purposes, aiming to foster advancements in NLP research for SAL.

## Acknowledgements

We gratefully acknowledge the support of the National Agency for Scientific Research and Innovation for funding this project under the National Research and Development Program.

## References

- F. Agalliu, E. Angoni, S. Demiraj, and e. al. 2002. *Gramatika e gjuhës shqipe*. Instituti i Gjuhësisë dhe i Letërsisë (Akademia e Shkencave e RSH), Tirana.
- T. Arkhangelskij, M. Daniel, M. Morozova, and Rusakov A. 2011. Korpusi i gjuhës shqipe: drejtimet kryesore të punës. In *Proceedings of Shqipja dhe gjuhët e Ballkanit — Albanian and Balkan Languages.*, page 635–642, Prishtina, Kosovo. ASHAK.
- ASHSH. 1998, 2002, 2006. *Fjalor i gjuhës së sotme shqipe*. Akademia e Shkencave e Shqipëris.
- Nebi Caka and Ali Caka. 2011. Korpusi i gjuhës shqipe – rezultatet e para, problemet dhe detyrat.
- Christian Ebert, Adrian Kuqi, Paul Widmer, and Barbara Sonnenhauser. 2022. *Ud gheg pear stories: An annotated treebank of gheg albanian as spoken in switzerland*. *PREPRINT*, Version 1.
- Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. *Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages*. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).
- Gaël Guibon, Marine Courtin, Kim Gerdes, and Bruno Guillaume. 2020. *When collaborative treebank curation meets graph grammars*. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5291–5300, Marseille, France. European Language Resources Association.
- Eric P. Hamp. 2023. *Albanian Language*.
- Johannes Heinecke. 2019. *ConlluEditor: a fully graphical editor for universal dependencies treebank files*. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 87–93, Paris, France. Association for Computational Linguistics.
- Besim Kabashi and Thomas Proisl. 2018. *Albanian part-of-speech tagging: Gold standard and evaluation*. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Arbana Kadriu. 2013. *Nltk tagger for albanian using iterative approach*. *Proceedings of the ITI 2013 35th International Conference on Information Technology Interfaces*, pages 283–288.

- J. Ke, Q. Jin, S. You, T. Han, Y. Feng, X. Wang, Z. Hu, E. Laçi, E. Allmetaj, T. Chen, W. Zhang, H. Zhang, Y. Lu, and W. Ai. 2012. [The sqglobe corpus \(a balanced corpus of 1m-word contemporary written albanian, lemmatised and pos-tagged\)](#).
- Christo Kirov, Ryan Cotterell, John Syla-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqi, Sabrina J. Mielke, Arya McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. [UniMorph 2.0: Universal Morphology](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Nelda Kote, Marenglen Biba, Jenna Kanerva, Samuel Rönnqvist, and Filip Ginter. 2019. [Morphological tagging and lemmatization of albanian: A manually annotated corpus and neural models](#). *CoRR*, abs/1912.00991.
- Diellza Nagavci Mati, Mentor Hamiti, and Elissa Molakuqe. 2021. [Morphological tagging and lemmatization in the albanian language](#). *SEEU Review*, 16(2):3–16.
- Arta Misini, Ercan Canhasi, and Samedin Krrabaj. 2020. Albanian syntactic parsing. *ICT Innovations 2020, Web Proceedings ISSN null*, pages 135–150.
- 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 Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.
- Marsida Toska, Joakim Nivre, and Daniel Zeman. 2020. [Universal Dependencies for Albanian](#). In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 178–188, Barcelona, Spain (Online). Association for Computational Linguistics.

# Function Multiword Expressions Annotated with Discourse Relations in the Romanian Reference Treebank

**Verginica Barbu Mititelu**

Romanian Academy  
Research Institute for Artificial Intelligence  
Bucharest, Romania  
vergi@racai.ro

**Tudor Voicu**

Tudor Vianu National High School  
of Computer Science  
Bucharest, Romania  
tudor.c.voicu@gmail.com

## Abstract

For the Romanian Reference Treebank, a general language corpus, covering several genres and annotated according to the principles of Universal Dependencies, we present here the annotation of some function words, namely multiword conjunctions, with discourse relations from the Penn Discourse Treebank version 3.0 inventory of such relations. The annotation process was manual, with two annotators for each occurrence of the conjunctions. Lexical-semantic relations of the types synonymy, polysemy can be established between the senses of such conjunctions. The discourse relations are added to the CoNLL-U file in which the treebank is represented.

**Keywords:** function MWE, discourse relation, Romanian Reference Treebank.

## 1 Introduction

One important characteristic of a text is its cohesion, i.e., the presence of linguistic cues to guide the reader into making connections between the ideas expressed therein (Halliday and Hasan, 1976). One such linguistic cue are the connectives, with conjunctions being one subtype thereof and the focus of this paper.

We identify the occurrences of conjunctions in a corpus that is already morpho-syntactically annotated, the Romanian Reference Treebank (Barbu Mititelu, 2018), and annotate them with discourse relations from an inventory already applied at a larger scale (Prasad et al., 2019), namely that from Penn Discourse Treebank (PDTB) version 3.0 annotation manual (Webber et al., 2019). The aim of our endeavour is to identify the possible senses Romanian conjunctions have, as well as to find the possible lexical devices to express these discourse relations in Romanian, particularly in this corpus. As a further step, we aim to use these annotated

occurrences of conjunctions as training material for understanding the structure of a text.

We start by presenting similar initiatives of annotating connectives with PDTB inventory of discourse relations (Section 2) and then outline the system of Romanian conjunctions (Section 3). The annotation method we adopted is described in Section 4 and the obtained results are presented in Section 5 and are then discussed in Section 6, before concluding the paper and envisaging further steps (Section 7).

## 2 Related Work

Prasad et al. (2019)’s work of creating the PDTB corpus annotated with discourse relations has proven seminal to a certain extent: the same inventory of relations was used for annotating a parallel corpus of TED-talks in 6 languages (English, Polish, German, Russian, European Portuguese, and Turkish) (Zeyrek et al., 2020), a corpus for Lithuanian (Oleškevičienė et al., 2023) and for Italian (Feltracco et al., 2017). Our work adds a new language to this landscape, i.e., Romanian. For it, version 2.0 of the PDTB annotation manual (Prasad et al., 2007) was used by Postolea (2018) for annotating adversative conjunctions in a set of 200 sentences extracted from EuroParl corpus (Koehn, 2005). However, this is not made available and no other such endeavour has been reported for Romanian yet.

## 3 Romanian Inventory of Conjunctions

Romanian conjunctions are devices for expressing either coordination or subordination. The former function both at the clause level and at the sentence level, connecting words and, respectively, clauses entering the same syntactic relation with their head: in ex. (1) the conjunction *and* connects two direct objects, and in ex. (2) it connects two clauses

functioning as direct objects. Subordinating conjunctions, however, are only clause linking devices (complementizers), linking a subordinate clause to the clause containing its syntactic head (ex. (3)). One conjunction is either subordinating or coordinating, never both.

(1)  
Cumpăr mere și pere.  
Buy.1SG apple.PL and pear.PL  
'I buy apples and pears.'

(2)  
Cumpăr ce gălesc sau ce îmi  
Buy.1SG what find.1SG or what CL.REFL.1SG.DAT  
permit  
afford.1SG  
'I buy what I find or what I afford.'

(3)  
Știu că mă iubești.  
Know.1SG that me lov.2SG  
'I know you love me.'

In Romanian linguistics, the class of conjunctions is made up of conjunctions and conjunctive locutions. As far as their structure is concerned, conjunctions are simple (e.g., *că* "that") or compound (e.g., *ca să* "so that", *fiindcă* (lit. 'being\_that') "because"). As one can notice, the latter can be written either as distinct words (e.g., *ca să*) or as a single word (e.g., *fiindcă*). Conjunctive locutions are always made up of at least two (separate) words (e.g., *pentru că* 'for that' "because"). Under focus in this paper are only the conjunctive locutions, though further annotation will extend to simple and compound conjunctions as well (see Section 7).

A remark is necessary here with respect to the Romanian conjunction *să*. This is a complementizer that specialized as the subjunctive mood marker (Dindelegan, 2013). It can occur in main clauses, where it is only a subjunctive marker (ex. (4)), but it can also occur in subordinate clauses, where its status varies, depending on the presence of absence of another subordinating device (be it another conjunction or a relative pronoun or adverb): it is either (a) both a subjunctive marker and a complementizer when (i) no other subordinating device is present (ex. (5)), or (b) only a subjunctive marker when the clause is introduced by a relative pronoun or a relative adverb (ex. (6)). There are also cases when *să* is a component (the last one in linear order) of a conjunctive locution: e.g., *fără să* "without SĂ", *pentru ca să* "so as to".

(4)  
Să vină secretara!  
SĂ come.SUBJ secretary.SG.DEF  
'Let the secretary come!'

(5)  
Îmi dai voie să te ajut?  
Me.DAT give.2SG permission SĂ you.ACC.SG help.1SG  
'Do you let me help you?'

(6)  
Nu m-am hotărât când  
Not CL.ACC.1SG have.1SG decided when  
/cu cine să vizitez parcul cel nou.  
/ with who SĂ visit.1SG.SUBJ park.DEF the new  
'I haven't decided when/who to visit the new park with.'

For the analysis of *să* in RRT, the relation `mark`<sup>1</sup> is always used to attach it to the verb in the subjunctive mood, irrespective of whether it is also a complementizer or only a subjunctive marker. Consequently, the relation `fixed`, used for fixed expressions that are function words, cannot be used for attaching *să* as the last component in a conjunctive locution, thus resulting into an inconsistent treatment of conjunctive locutions in Romanian UD treebanks: in those that do not contain *să* the components are linked the relation `fixed` to the first component in linear order (Figure 1 for example (7)), while in those containing *să* this component is not attached to the rest of the expression, but is treated like a separate word (Figure 2 for example (8)).

(7)  
Echipajul său de opt oameni a pierit în timp  
Crew.DEF his of eight people has vanished in time  
ce se zbătea să salveze echipajul  
what CL.REFL.3SG.ACC striving SĂ save.3SG crew.DEF  
Santampa.  
Santampa  
'His crew of eight people also died while striving to save the Santampa crew.'

(8)  
El continuase să meargă fără să  
He had.continued SĂ walk without SĂ  
se oprească  
CL.REFL.3SG.ACC stop  
'He had continued walking, without stopping.'

## 4 Work Methodology

**The Corpus.** We chose to annotate the conjunctions in the Romanian Reference Treebank (RRT) (Barbu Mititelu, 2018). The corpus contains 9,523 sentences and 218,511 tokens, distributed in several genres. It is released and distributed within Universal Dependencies<sup>2</sup> (de Marneffe et al., 2021), thus being tokenized, lemmatized and morpho-syntactically annotated according to the principles thereof.

<sup>1</sup>In UD, `mark` is the relation used for linking a subordinating word to the head of the clause it introduces.

<sup>2</sup><https://universaldependencies.org/>

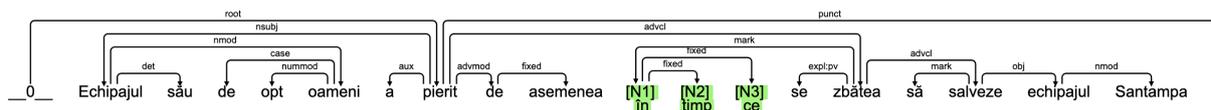


Figure 1: The annotation of a conjunctive locution with the relation *fixed*. This is the Universal Dependencies representation of ex. (7).

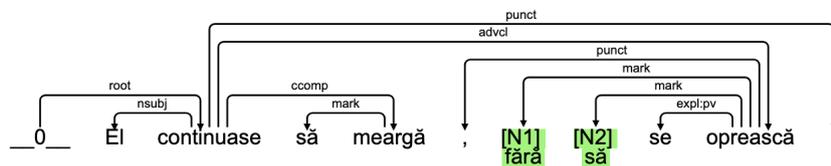


Figure 2: The annotation of conjunctive locutions containing *să*. This is the Universal Dependencies representation of ex. (8).

**The Inventory of Discourse Relations.** For annotation we used the discourse relations defined in the PDTB 3.0 manual. There are three levels on which relations are defined here, from more general to more specific (see Figure 3). The most refined level was always used in the annotation process.

**The Inventory of Conjunctions.** The aim of this work presented here is to specify the discourse relation expressed by each occurrence of conjunctive locutions in RRT. We started from a list of such conjunctions extracted from the Morphological, Orthographic and Orthoepic Romanian Dictionary (DOOM, 3rd edition)<sup>3</sup>, 71 of which were found in RRT with a total number of 479 occurrences.

**Annotators.** The data underwent double annotation, with a linguist and a student contributing to this task. The former was already familiar with the PDTB 3.0 manual, while for the latter this was the first experience of the kind. The degree of agreement between them is presented in Table 1. The accuracy represents the percent of annotations both annotators agreed on and it is calculated by dividing the number of cases both annotators agreed on to the total number of analyzed conjunctions. The values of the accuracy are also relevant for the possibility of automatically assigning such discourse relations, showing that the task is quite challenging for humans, thus problematic for machines.

**The methodology.** Each occurrence of the conjunction was annotated independently by each annotator. Only one relation was assigned, always at the lowest level possible in the PDTB hierarchy

(Figure 3). Only in a couple of cases, when the context was not considered enough to identify the sense or to distinguish between two senses, no annotation was assigned or, respectively, two senses were assigned.

## 5 Results

All occurrences of conjunctive locutions in RRT were annotated. In Table 2 we present the possible discourse relations they express alongside the frequency of each such relation in RRT. Although all 479 occurrences of conjunctions were annotated in the corpus, we selected only conjunctions with a frequency above 5 to show here and left out those with less occurrences.

Table 3 shows the same data, but it is more explicit in rendering the conjunctions that express relations from PDTB.

The annotation is added to the CoNLL-U format<sup>4</sup> of RRT. In Figure 4 we present the current annotation of the sentence in example (9). The information about the discourse relation is added on the last column of the file: the same number is used in this last column to identify the components of the conjunctive locution, while the label of the discourse relation appears only with the first component<sup>5</sup>: see the highlighted lines in Figure 4, where number 1 is added in the last column of the first occurrence of components of the conjunction *pentru că*, number 2 is used for the second occurrence of the same conjunction, while the discourse

<sup>3</sup><https://doom.lingv.ro/>

<sup>4</sup><https://universaldependencies.org/format.html>

<sup>5</sup>It is the same system of encoding the multiword expressions used in the PARSEME treebanks (Ramisch et al., 2018).

| Level-1     | Level-2              | Level-3                              |
|-------------|----------------------|--------------------------------------|
| TEMPORAL    | SYNCHRONOUS          | –                                    |
|             | ASYNCHRONOUS         | PRECEDENCE<br>SUCCESSION             |
| CONTINGENCY | CAUSE                | REASON                               |
|             |                      | RESULT                               |
|             | CAUSE+BELIEF         | NEGRESULT                            |
|             |                      | REASON+BELIEF<br>RESULT+BELIEF       |
|             | CAUSE+SPEECHACT      | REASON+SPEECHACT<br>RESULT+SPEECHACT |
|             |                      | CONDITION                            |
|             | CONDITION+SPEECHACT  | –                                    |
|             | NEGATIVE-CONDITION   | ARG1-AS-NEGCOND<br>ARG2-AS-NEGCOND   |
|             |                      | NEGATIVE-CONDITION+SPEECHACT         |
|             | PURPOSE              | ARG1-AS-GOAL<br>ARG2-AS-GOAL         |
| COMPARISON  |                      | CONCESSION                           |
|             | CONCESSION+SPEECHACT | ARG2-AS-DENIER+SPEECHACT             |
|             | CONTRAST             | –                                    |
|             | SIMILARITY           | –                                    |
| EXPANSION   | CONJUNCTION          | –                                    |
|             | DISJUNCTION          | –                                    |
|             | EQUIVALENCE          | –                                    |
|             | EXCEPTION            | ARG1-AS-EXCPT<br>ARG2-AS-EXCPT       |
|             |                      | INSTANTIATION                        |
|             | LEVEL-OF-DETAIL      | ARG1-AS-DETAIL<br>ARG2-AS-DETAIL     |
|             |                      | MANNER                               |
|             | SUBSTITUTION         |                                      |

Figure 3: The inventory of PDTB 3.0 discourse relations.

relation expressed by each appears only for the first component, in each occurrence.

Articulațiile între șanțurile de plastic  
 Joints.DEF between ditches.DEF of plastic  
 se pot scurge, de asemenea - de  
 CL.REFL.3SG.ACC can.3PL leak of alike - of

(9)

| Conjunction | după ce<br>“after”     | pentru că<br>“because”    | în timp ce<br>“while”        | înainte de<br>“before” | pentru ca să<br>“in order to” | astfel încât<br>“so that” |
|-------------|------------------------|---------------------------|------------------------------|------------------------|-------------------------------|---------------------------|
| No.         | 87                     | 54                        | 51                           | 35                     | 33                            | 27                        |
| Accuracy    | 89                     | 70.4                      | 86.3                         | 100.0                  | 93.9                          | 88.9                      |
| Conjunction | pe măsură ce<br>“as”   | chiar dacă<br>“even if”   | asa încât<br>“so that”       | asa că<br>“so that”    | fără să<br>“without”          | înainte ca<br>“before”    |
| No.         | 16                     | 13                        | 10                           | 9                      | 9                             | 9                         |
| Accuracy    | 75.0                   | 100.0                     | 100.0                        | 100.0                  | 77.8                          | 100.0                     |
| Conjunction | astfel că<br>“so that” | cu toate că<br>“although” | de îndată ce<br>“as soon as” | până când<br>“until”   |                               |                           |
| No.         | 8                      | 8                         | 7                            | 5                      |                               |                           |
| Accuracy    | 87.5                   | 25.0                      | 28.6                         | 100.0                  |                               |                           |

Table 1: The agreement between annotators for each conjunction.

obicei pentru că gunoiul sau pietrișul s-  
custom for that garbage or gravel CL.REFL.3PL.ACC  
au adunat între șanț și sigiliu sau pentru că  
have gathered between ditch and seal or for that  
însuși sigiliul s- a stricat.  
itself seal.DEF CL.REFL.3SG.ACC has broken\_down.  
‘The joints between the plastic ditches may also leak - usually  
because the garbage or the gravel have gathered between the  
ditch and the seal or because the seal itself has broken down.’

## 6 Discussion of Results

We notice in Table 2 that the more frequent conjunctions are also more polysemous, in the sense that they are not specialized for one discourse relation; they tend to have a dominant meaning and also other meanings, more or less frequent: e.g., *pentru că* (54 total occurrences) seems specialised for Contingency.Cause.Reason (35 occurrences), but also expresses Contingency.Cause+Belief.Reason+Belief (10 occurrences), Contingency.Cause+SpeechAct.Reason+SpeechAct (4 occurrences), Contingency.Cause+SpeechAct.Result+SpeechAct (3 occurrences), Contingency.Cause.Result (2 occurrences). The most polysemous seems to be *pentru că* “because”, with 5 senses, though they are rather pragmatically distinguished than semantically: three of the relations are distinguished by the association of epistemic knowledge (belief) or a speech act. However, the more diverse polysemy is displayed by *în timp ce* “while” and *pentru ca să* “in order to”, each expressing four different discourse relations, as shown in the table.

One of the relatively frequent (35 occurrences) conjunctions which is specialized for a relation is *înainte de* “before”, which expresses Tempo-

ral.Asynchronous. Precedence.

Table 3 shows which of the relations defined in the PDTB manual are lexicalized by the analysed conjunctions. We notice again that there are prevailing ways of expressing such a relation: e.g. the relation Temporal.Synchronous tends to be expressed by *în timp ce* “while”, but there are other conjunctive locutions for it as well: *pe măsură ce* “as” *de îndată ce* “as soon as”. A rather tight competition between two conjunctions for expressing a relation is seen between *pentru ca să* “in order to” and *astfel încât* “so that” for rendering Contingency.Purpose:Arg2-as-Goal, though the former has a small advantage in our corpus.

The fact that not all PDTB relations occur in these tables does not mean they are not lexicalized in Romanian or by these conjunctions; it simply means that the analysed conjunctions in this corpus do not express them.

There are cases when one conjunction expresses two senses for the same occurrence: see the case of *după ce* which is assigned two senses for 38 occurrences: Temporal.Asynchronous.Succession and Contingency.Cause.Reason. This is expected not to be a singular case, as this is also reported for PDBT (Webber et al., 2019).

For the moment, we were not able to find any correlations between the text genre and the relation expressed by a conjunction, nor between the order of arguments and such relation.

## 7 Conclusions and Further Work

Our work so far ensured the familiarity of annotators with the PDTB annotation manual and the inventory of discourse relations thereof. This is

| Conjunction  | Total no. | No.                     | Sense  |
|--------------|-----------|-------------------------|--|
| după ce      | 86        | 48<br>38                | TEMPORAL:ASYNCHRONOUS:SUCCESSION<br>TEMPORAL:ASYNCHRONOUS:SUCCESSION—<br>CONTINGENCY:CAUSE:REASON  |
| pentru că    | 54        | 35<br>10<br>4<br>3<br>2 | CONTINGENCY:CAUSE:REASON<br>CONTINGENCY:CAUSE+BELIEF:REASON+BELIEF<br>CONTINGENCY:CAUSE+SPEECHACT:REASON+SPEECHACT<br>CONTINGENCY:CAUSE+SPEECHACT:RESULT+SPEECHACT<br>CONTINGENCY:CAUSE:RESULT |
| în timp ce   | 50        | 29<br>18<br>2<br>1      | TEMPORAL:SYNCHRONOUS<br>COMPARISON:CONTRAST<br>COMPARISON:CONCESSION:ARG2-AS-DENIER<br>COMPARISON:SIMILARITY   |
| înainte de   | 35        | 35                      | TEMPORAL:ASYNCHRONOUS:PRECEDENCE   |
| pentru ca să | 33        | 30<br>1<br>1<br>1       | CONTINGENCY:PURPOSE:ARG2-AS-GOAL<br>CONTINGENCY:CAUSE.NEGRESULT<br>TEMPORAL:ASYNCHRONOUS:SUCCESSION<br>CONTINGENCY:CONDITION:ARG1-AS-COND  |
| astfel încât | 27        | 20<br>7                 | CONTINGENCY:PURPOSE:ARG2-AS-GOAL<br>CONTINGENCY:CAUSE:RESULT   |
| pe măsură ce | 16        | 9<br>4<br>3             | TEMPORAL:SYNCHRONOUS<br>CONTINGENCY:CAUSE:REASON<br>CONTINGENCY:CONDITION:ARG2-AS-COND   |
| chiar dacă   | 13        | 13                      | COMPARISON:CONCESSION:ARG1-AS-DENIER   |
| așa încât    | 9         | 7<br>2                  | CONTINGENCY:CAUSE:RESULT<br>CONTINGENCY:PURPOSE:ARG2-AS-GOAL   |
| așa că       | 9         | 7<br>2                  | CONTINGENCY:CAUSE:RESULT<br>CONTINGENCY:PURPOSE:ARG2-AS-GOAL   |
| fără să      | 9         | 9                       | EXPANSION:MANNER:ARG2-AS-MANNER  |
| înainte ca   | 9         | 9                       | TEMPORAL:ASYNCHRONOUS:PRECEDENCE   |
| astfel că    | 8         | 8                       | CONTINGENCY:CAUSE:RESULT   |
| cu toate că  | 8         | 8                       | COMPARISON:CONCESSION:ARG1-AS-DENIER   |
| de îndată ce | 7         | 5<br>2                  | TEMPORAL:ASYNCHRONOUS:SUCCESSION<br>TEMPORAL:SYNCHRONOUS   |
| până când    | 5         | 5                       | TEMPORAL:ASYNCHRONOUS:PRECEDENCE   |
| TOTAL        | 378       |                         |  |

Table 2: The PDTB discourse relations expressed by the annotated conjunctions in RRT. For the translation of the conjunctions into English see Table 1.

experience that will be further harnessed in annotating the simple and compound conjunctions in the same corpus, thus increasing the size of such data, to serve for linguistic analysis and interpretation, as well as for experiments of automatic identification of such relations in texts.

This annotation can help drawing comparisons between cross-lingually equivalent conjunctive locutions, which is of paramount importance in translation (be it manual or automatic). Corpora anno-

tated with discourse relations are used in training, tuning and testing of systems for discourse parsing, which Romanian lacks at the moment. Understanding such relations between parts of a text is vital in many NLP applications (from question-answering and summarization to automatic reasoning).

As conjunctions are not the only means of expressing discourse relations, we also envisage extending our work to the annotation of adverbs and other textual connectors and, eventually, to identify-

| Sense   | Total No. | No. | Connectives (count)   |
|---|-----------|-----|---|
| TEMPORAL:SYNCHRONOUS  | 40        | 29  | în timp ce<br>9<br>pe măsură ce<br>2<br>de îndată ce                              |
| TEMPORAL:ASYNCHRONOUS:PRECEDENCE                              | 49        | 35  | înainte de<br>9<br>înainte ca<br>5<br>până când                                   |
| TEMPORAL:ASYNCHRONOUS:SUCCESSION                              | 54        | 48  | după ce<br>1<br>pentru ca să<br>5<br>de îndată ce                                 |
| TEMPORAL:ASYNCHRONOUS:SUCCESSION—<br>CONTINGENCY:CAUSE:REASON | 38        | 38  | după ce   |
| CONTINGENCY:CAUSE:REASON                                      | 39        | 35  | pentru că<br>4<br>pe măsură ce  |
| CONTINGENCY:CAUSE:RESULT                                      | 31        | 2   | pentru că<br>7<br>astfel încât<br>7<br>așa încât<br>7<br>așa că<br>8<br>astfel că |
| CONTINGENCY.CAUSE.NEGRESULT                                   | 1         | 1   | pentru ca să  |
| CONTINGENCY:CAUSE+BELIEF:REASON+BELIEF                        | 10        | 10  | pentru că   |
| CONTINGENCY:CAUSE+SPEECHACT:REASON<br>+SPEECHACT              | 4         | 4   | pentru că   |
| CONTINGENCY:CAUSE+SPEECHACT:RESULT<br>+SPEECHACT              | 3         | 3   | pentru că   |
| CONTINGENCY:CONDITION:ARG1-AS-COND                            | 1         | 1   | pentru ca să  |
| CONTINGENCY:CONDITION:ARG2-AS-COND                            | 3         | 3   | pe măsură ce  |
| CONTINGENCY:PURPOSE:ARG2-AS-GOAL                              | 54        | 30  | pentru ca să<br>20<br>astfel încât<br>2<br>așa încât<br>2<br>așa că               |
| COMPARISON:CONCESSION:ARG1-AS-DENIER                          | 21        | 13  | chiar dacă<br>8<br>cu toate că  |
| COMPARISON:CONCESSION:ARG2-AS-DENIER                          | 2         | 2   | în timp ce  |
| COMPARISON:CONTRAST   | 18        | 18  | în timp ce  |
| COMPARISON:SIMILARITY   | 1         | 1   | în timp ce  |
| EXPANSION:MANNER:ARG2-AS-MANNER                               | 9         | 9   | fără să   |

Table 3: The conjunctions that lexicalize various PDTB discourse relations in RRT. For the translation of the conjunctions into English see Table 1.

ing cases of implicit instantiation of such relations, i.e. the relation exists in the absence of a lexicalized connector.

The Romanian inventory of such connecting devices will be added to Connective-Lex (Stede et al., 2019), a multilingual online resource of connectors and the discourse relations expressed by them.

The newly added annotation of RRT will be made freely available with the forthcoming UD

release.

## References

- Verginica Barbu Mititelu. 2018. Modern syntactic analysis of Romanian. In *Clasic și modern în cercetarea filologică românească actuală*, pages 67–78, Iași, Romania.
- Gabriela Pană Dindelegan, editor. 2013. *The Grammar of Romanian*. Oxford University Press.

|    |               |             |       |               |   |    |         |    |                            |
|----|---------------|-------------|-------|---------------|---|----|---------|----|----------------------------|
| 1  | Articulațiile | articulație | NOUN  | Ncfpry        | Case=Acc,Nom Definite=Def Gender=Fem Number=Plur                      | 7  | nsubj   | -- | --                         |
| 2  | între         | între       | ADP   | Spsa          | AdpType=Prep Case=Acc   | 3  | case    | -- | --                         |
| 3  | șanțurile     | șanț        | NOUN  | Ncfpry        | Case=Acc,Nom Definite=Def Gender=Fem Number=Plur                      | 1  | nmod    | -- | --                         |
| 4  | de            | de          | ADP   | Spsa          | AdpType=Prep Case=Acc   | 5  | case    | -- | --                         |
| 5  | plastic       | plastic     | NOUN  | Ncms-n        | Definite=Ind Gender=Masc Number=Sing                                  | 3  | nmod    | -- | --                         |
| 6  | se            | sine        | PRON  | Px3-a-y-----w | Case=Acc Person=3 PronType=Prs Reflex=Yes Strength=Weak               | 8  | expl:pv | -- | --                         |
| 7  | pot           | putea       | VERB  | Vmip3p        | Mood=Ind Number=Plur Person=3 Tense=Pres VerbForm=Fin                 | 0  | root    | -- | --                         |
| 8  | scurge        | scurge      | VERB  | COMMA         | Tense=Pres VerbForm=Inf   | 7  | ccomp   | -- | SpaceAfter=No              |
| 9  | ,             |             | PUNCT | Spsa          |   | 10 | punct   | -- | --                         |
| 10 | de            | de          | ADP   | Spsa          | AdpType=Prep Case=Acc   | 7  | advmod  | -- | --                         |
| 11 | asemenea      | asemenea    | ADJ   | Afp           | Degree=Pos  | 10 | fixed   | -- | --                         |
| 12 | -             |             | PUNCT | DASH          |   | 22 | punct   | -- | --                         |
| 13 | de            | de          | ADP   | Spsa          | AdpType=Prep Case=Acc   | 22 | advmod  | -- | --                         |
| 14 | obicei        | obicei      | NOUN  | Ncms-n        | Definite=Ind Gender=Masc Number=Sing                                  | 13 | fixed   | -- | --                         |
| 15 | pentru        | pentru      | ADP   | Spsa          | AdpType=Prep Case=Acc   | 22 | mark    | -- | 1:CONTINGENCY_CAUSE_REASON |
| 16 | și            | și          | SCONJ | Ccssp         | Polarity=Pos  | 15 | fixed   | -- | --                         |
| 17 | gunoiul       | gunoi       | NOUN  | Ncmsry        | Case=Acc,Nom Definite=Def Gender=Masc Number=Sing                     | 22 | nsubj   | -- | --                         |
| 18 | sau           | sau         | CCONJ | Ccssp         | Polarity=Pos  | 19 | cc      | -- | --                         |
| 19 | pietrișul     | pietriș     | NOUN  | Ncmsry        | Case=Acc,Nom Definite=Def Gender=Masc Number=Sing                     | 17 | conj    | -- | --                         |
| 20 | s-            | sine        | PRON  | Px3-a-y-----w | Case=Acc Person=3 PronType=Prs Reflex=Yes Strength=Weak Variant=Short | 22 | expl:pv | -- | SpaceAfter=No              |
| 21 | au            | avea        | AUX   | Va--3p        | Number=Plur Person=3  | 22 | aux     | -- | --                         |
| 22 | adunat        | aduna       | VERB  | Vmp--sm       | Gender=Masc Number=Sing VerbForm=Part                                 | 7  | advcl   | -- | --                         |
| 23 | între         | între       | ADP   | Spsa          | AdpType=Prep Case=Acc   | 24 | case    | -- | --                         |
| 24 | șanț          | șanț        | NOUN  | Ncms-n        | Definite=Ind Gender=Masc Number=Sing                                  | 22 | obl     | -- | --                         |
| 25 | și            | și          | CCONJ | Ccssp         | Polarity=Pos  | 26 | cc      | -- | --                         |
| 26 | sigiliu       | sigiliu     | NOUN  | Ncms-n        | Definite=Ind Gender=Masc Number=Sing                                  | 24 | conj    | -- | --                         |
| 27 | sau           | sau         | CCONJ | Ccssp         | Polarity=Pos  | 34 | cc      | -- | --                         |
| 28 | pentru        | pentru      | ADP   | Spsa          | AdpType=Prep Case=Acc   | 34 | mark    | -- | 2:CONTINGENCY_CAUSE_REASON |
| 29 | și            | și          | SCONJ | Ccssp         | Polarity=Pos  | 38 | fixed   | -- | --                         |
| 30 | insuși        | insuși      | DET   | Dh3ms         | Gender=Masc Number=Sing Person=3 PronType=Emp                         | 31 | det     | -- | --                         |
| 31 | sigiliul      | sigiliu     | NOUN  | Ncmsry        | Case=Acc,Nom Definite=Def Gender=Masc Number=Sing                     | 34 | nsubj   | -- | --                         |
| 32 | s-            | sine        | PRON  | Px3-a-y-----w | Case=Acc Person=3 PronType=Prs Reflex=Yes Strength=Weak Variant=Short | 34 | expl:pv | -- | SpaceAfter=No              |
| 33 | a             | avea        | AUX   | Va--3s        | Number=Sing Person=3  | 34 | aux     | -- | --                         |
| 34 | STRICAT       | strica      | VERB  | Vmp--sm       | Gender=Masc Number=Sing VerbForm=Part                                 | 22 | conj    | -- | SpaceAfter=No              |
| 35 | .             |             | PUNCT | PERIOD        |   | 7  | punct   | -- | --                         |

Figure 4: Adding discourse relations in the CoNLL-U file of RRT.

Anna Feltracco, Bernardo Magnini, and Elisabetta Jezek. 2017. Contrast-Ita Bank: A corpus for Italian Annotated with Discourse Contrast Relations. In *Proceedings of the Fourth Italian Conference on Computational Linguistics CLiC-it 2017*, pages 159–164, Rome, Italy.

M. A. K. Halliday and R. Hasan. 1976. *Cohesion in English*. Longman.

Philipp Koehn. 2005. *Europarl: A parallel corpus for statistical machine translation*. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. *Universal Dependencies*. *Computational Linguistics*, 47(2):255–308.

Giedrė Valūnaitė Oleškevičienė, Vitalija Karaciejūtė, and Dalia Gulbinskienė. 2023. *Lithuanian discourse markers and their relations in a multilingual corpus*. *Sustainable Multilingualism*, 22(1):258–272.

Sorina Postolea. 2018. A particle filter algorithm for Bayesian wordsegmentation using annotation to identify connective meanings in a multilingual environment. romanian and english contrast markers in a parallel corpus. In *Cross-Linguistic Discourse Annotation: applications and perspectives. TextLink2018*, pages 107–113, Toulouse, France.

Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie Webber. 2007. *The Penn Discourse Treebank 2.0 Annotation Manual*. University of Pennsylvania.

Rashmi Prasad, Bonnie Webber, Alan Lee, and Aravind Joshi. 2019. *Penn Discourse Treebank Version 3.0*. Linguistic Data Consortium.

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.

Manfred Stede, Tatjana Scheffler, and Amalia Mendes. 2019. *Connective-lex: A web-based multilingual lexical resource for connectives*. *Discours. Revue de linguistique, psycholinguistique et informatique*, 24.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The Penn Discourse Treebank 3.0 Annotation Manual. *Philadelphia, University of Pennsylvania*, 35:108.

Deniz Zeyrek, Amália Mendes, Yulia Grishina, Murathan Kurfalı, Samuel Gibbon, and Maciej Ogródniczuk. 2020. TED multilingual discourse bank (TED-MDB): a parallel corpus annotated in the PDTB style. *Language Resources Evaluation*, 54:587–613.

# Dependency parser for Bulgarian

Atanas Atanasov

Faculty of Slavic Studies,  
Sofia University "St. Kliment Ohridski"  
atanasov@slav.uni-sofia.bg

## Abstract

This paper delves into the implementation of a Biaffine Attention Model, a sophisticated neural network architecture employed for dependency parsing tasks. Proposed by Dozat and Manning, this model is applied to Bulgarian language processing. The model's training and evaluation are conducted using the Bulgarian Universal Dependencies dataset. The paper offers a comprehensive explanation of the model's architecture and the data preparation process, aiming to demonstrate that for highly inflected languages, the inclusion of two additional input layers - lemmas and language-specific morphological information - is beneficial. The results of the experiments are subsequently presented and discussed. The paper concludes with a reflection on the model's performance and suggestions for potential future work.

**Keywords:** syntactic parsing, Universal Dependencies, Biaffine Attention, Bulgarian.

## 1 Introduction

The paper presents an implementation of a neural network-based dependency parser using TensorFlow and Keras<sup>1</sup>. The parser is trained and evaluated on the Bulgarian Universal Dependencies dataset.

The article introduces a model in its intermediate stage of development, with the current focus entirely on dependency analysis. At this stage, it does not aim to predict parts of speech, lemmas, etc. (these are expected to be added to the model later). For this reason, only the results for Labeled

and Unlabeled Attachment Scores will be presented.

### 1.1 Linguistic background

From a linguistic perspective, Generative Grammar (GG) and Dependency Grammar (DG) have emerged as the two primary approaches to syntax study over the past few decades<sup>2</sup>. GG is a grammar model that operates on the premise that a sentence's syntactic structure is generated by a set of rules. These rules are applied to a set of terminal nodes, which are the words (or potentially empty functional categories) of the sentence. The rules are recursively applied until the sentence is parsed into its smallest constituents. The result of constituency parsing is a tree, known as a constituency parse tree, which represents the sentence's syntactic structure.

On the other hand, DG is a grammar model that posits that a sentence's syntactic structure is represented by a set of dependencies between the sentence's words. These dependencies are directed links between the words of the sentence and are represented by a tree, known as a dependency parse tree. The nodes of the dependency parse tree are the words of the sentence, and the edges (arcs) represent the dependencies between the words. The dependency parse tree has a root node, typically the main verb (predicate) of the sentence. This root node has no incoming edges but can have multiple outgoing edges. The dependency parse tree is a directed acyclic graph (DAG), meaning there are no cycles in the graph and only one path exists between any two nodes in the tree.

---

<sup>1</sup> The code, written in Python 3.11, has been tested on Ubuntu 22.04.3 LTS running on WSL2. It is accessible at [https://github.com/nassoo/dependency\\_parser](https://github.com/nassoo/dependency_parser). The necessary packages are enumerated in the `requirements.txt` file. Additionally, an `environment.yml` file is provided for convenience. Note that the virtual environment includes packages not directly used in the notebook, such as PyTorch

and Transformers, which are for side experiments. If disk space is limited, consider manually installing only the necessary packages.

<sup>2</sup> In fact, the concept behind the dependency approach dates back several thousand years (Kruijff, 2002: 7-17). However, it wasn't until the 20th century that it was formalized and evolved into a comprehensive theory.

In recent decades, Generative Grammar (GG) has been the preferred approach for representing syntactic structures in most linguistic studies. However, due to advancements in Natural Language Processing (NLP), particularly the Universal Dependencies (UD) project and deep learning models based on it, DG is gaining popularity. The primary advantage of DG lies in its simplicity and intuitiveness, despite GG's greater expressiveness.

## 1.2 Universal Dependencies

Marneffe et al. (2021) point out that "Universal dependencies (UD) is a framework for morphosyntactic annotation of human language, which to date has been used to create treebanks for more than 100 languages" and "the linguistic theory of the UD framework ... draws on a long tradition of typologically oriented grammatical theories. Grammatical relations between words are centrally used to explain how predicate–argument structures are encoded morphosyntactically in different languages while morphological features and part-of-speech classes give the properties of words". The UD project aims to facilitate multilingual parser development, cross-lingual learning, and parsing research from a language typology perspective. It has significantly influenced the field of Natural Language Processing (NLP). The project provides a set of universal guidelines, applicable to all languages, with language-specific extensions. These guidelines cover annotation at both the word level (morphology) and the sentence level (syntax). The UD treebanks, collections of annotated sentences, serve as a valuable resource for training and evaluating models such as part-of-speech taggers and dependency parsers.

## 1.3 Dependency parsing background

Two primary data-driven approaches exist for dependency parsing: transition-based and graph-based methods (Kübler et al., 2009). The transition-based (shift-reduce) approach is a greedy algorithm that builds a dependency tree by applying a sequence of actions to a partially built tree. They maintain a stack and a buffer of words to be processed. The parsing process starts with all words in the buffer and an empty stack. The parser can perform three types of actions: 'SHIFT' (moves a word from the buffer to the stack); 'REDUCE' (removes a word from the stack); 'ARC' (creates a

dependency relation between two words, one from the stack and one from the buffer). The parser makes these actions based on a set of features extracted from the current state of the stack and buffer, and the previously built dependency relations. The process continues until the buffer is empty and all words have been incorporated into the dependency tree. These parsers do not use deep learning algorithms for direct prediction of dependencies, instead they use it to predict the next action to be taken.

Graph-based parsers, on the other hand, work by considering all possible dependency trees for a sentence and choosing the one with the highest score. The score of a tree is typically computed as the sum of the scores of its individual dependencies. The scoring function is learned from a treebank during training. The learning process involves finding weights for the features of the dependencies such that the correct trees in the training data get higher scores than incorrect trees.

Transition-based parsers are generally more efficient, as they parse a sentence in linear time. This makes them suitable for real-time applications or large-scale data processing. However, these parsers suffer from error propagation. The decisions are made greedily, and once a parsing action is taken, it cannot be undone. This means that an error early in the parsing process can affect the rest of the parse. Graph-based parsers often achieve higher accuracy than transition-based parsers. They globally optimize the parse tree of a sentence, considering all possible trees before making a decision. However, this global optimization comes at a cost. Graph-based parsers have a higher time complexity (usually cubic in the length of the sentence), making them slower than transition-based parsers.

The development of dependency parsing has been significantly influenced by the CoNLL Shared Tasks, which have provided standardized datasets and evaluation benchmarks. The CoNLL 2017 (Zeman et al., 2017) and 2018 (Zeman et al., 2018) Shared Tasks, in particular, focused on multilingual dependency parsing, advancing the field through cross-lingual comparisons. While this work concentrates solely on the dependency analysis of Bulgarian, insights gained from these shared tasks have informed the approach to model development and evaluation.

## 1.4 Deep Biaffine Attention

This paper presents an implementation of a graph-based dependency parser, following the algorithm proposed by Dozat and Manning (2017). Their architecture utilizes a bi-LSTM (Bidirectional Long Short-Term Memory) to read the input sentence from both directions, thereby capturing a rich set of syntactic and semantic features. Furthermore, the authors introduce a biaffine attention mechanism, which is a bilinear function supplemented with an affine transformation. Instead of employing shallow bilinear attention that operates directly on recurrent state representations, deep biaffine attention uses a multi-layer perceptron (MLP) to project these representations into a higher-dimensional space prior to applying the bilinear attention function. This approach enables the model to capture more complex relationships between words and their potential dependencies.

The parser proposed in this paper is trained and evaluated using the Bulgarian UD treebank (Osenova and Simov, 2015).

## 2 Data preparation

The UD treebanks come in various formats, including CoNLL-U (a format used for linguistic treebanks in the Conference on Natural Language Learning), TensorFlow Datasets (TFDS), and HuggingFace Datasets. Given that this project uses TensorFlow, TFDS is the most convenient format. The treebanks are always split into training, development, and test sets.

The Bulgarian UD dataset includes 8,907 training sentences, 1,115 development sentences, and 1,116 test sentences. This distribution, which approximates an 80-10-10 split for training, validation, and testing respectively, is typical for UD treebanks. Given its suitability for developing a neural network parser, no additional splitting of the data is required.

In Dozat and Manning's architecture, only specific parts of the data are utilized. The *tokens* column, which contains the tokenized sentence, and the *upos* column, which includes Universal Part-of-Speech tags, are used to train the parser. The *deprel* column, with dependency relations, and the *head* column, indicating word head indices, serve as the parser's targets. Given the focus on using the parser for Bulgarian, as highlighted in the Experiments and Results section, two additional

columns are incorporated into the input layers: *lemmas* (containing the lemmatized forms of the words) and *xpos* (providing more detailed morphological language-specific information, although these tags are not consistent across languages).

After loading the data, the next step is to construct vocabularies. These vocabularies consist of unique words, lemmas, Universal Part-of-Speech tags, language-specific tags, and dependency relations from the training set, all of which are converted into numerical representations for model processing. The vocabularies facilitate the conversion of words and tags into these representations, while the dependency relations are used to construct the target vectors for the parser. The unique values from the dataset are extracted and stored in TensorFlow's hash tables for more efficient tensor handling. Special tokens are added to equalize input sequence lengths, represent the root of the dependency parse tree (the dummy token that governs the main verb of the sentence), and denote unknown words, tags, or relations. These tables are utilized to convert between numerical IDs and their corresponding labels during both training and prediction.

A configuration management module was implemented to streamline experimentation and reproducibility. This module incorporates hyperparameter loading from a JSON file and manages essential data structures, such as hash tables, necessary for both model training and data preprocessing. By centralizing configuration parameters, the development process was optimized, facilitating efficient exploration of the hyperparameter space.

The vectorization process is a critical step in preparing the data for the neural network. This process transforms the dataset into a format where each element is a tuple of inputs and outputs.

The inputs consist of tokenized sentences, Universal Part-of-Speech tags, lemmas, language-specific tags, and sentence lengths. The outputs are word head indices and dependency relations. Word heads are cast to numbers (specifically, `tf.int32` to meet the requirements of `tf.lookup.StaticHashTable`), while the remaining elements are encoded as integers using their respective hash tables. To ensure uniformity, sequences are either padded or truncated to a predetermined length, as specified in the configuration. Each sequence begins with a

dummy value, and an additional token is included in the sentence length to represent the root of the dependency parse tree.

The vectorized inputs and outputs are then used to generate a new dataset.

The data preparation process includes batching and shuffling steps to optimize the training of the neural network. In the batching step, the dataset is divided into smaller groups or *batches*. Each batch contains a certain number of examples that the model will process simultaneously. The size of these batches is a configurable parameter and can be adjusted based on the computational resources available.

The shuffling step randomizes the order of the examples in the dataset. This is done to ensure that the model does not learn any unintended patterns from the order of the examples, which could lead to overfitting. This step is performed before each epoch, i.e., each pass through the entire dataset, to ensure that the model is exposed to a different order of examples in each epoch.

This dataset, vectorized and batched, can be used for model training or prediction.

### 3 The model

This chapter introduces the primary neural network model, designed for dependency parsing.

Built with TensorFlow and the Keras API, the model incorporates several components. These include embedding layers for words, lemmas, POS, and language-specific tags, BiLSTM layers for sentence encoding, and MLP layers for transforming the sentence encodings into a form suitable for predicting arcs and relations between words. The model also features Biaffine layers, which take these transformed outputs and make the

actual predictions of arcs and relations between words.

The model is equipped with a custom loss function and utilizes the Adam optimizer for training. It also includes metrics for monitoring the loss and accuracy during both training and evaluation.

#### 3.1 Model Architecture

As it was already mentioned, the model is based on the architecture proposed in *Deep Biaffine Attention for Neural Dependency Parsing* (Dozat and Manning, 2017), which builds off the work from Kiperwasser and Goldberg (2016) with a few modifications. The graph-based algorithm is illustrated in Figure 1.

Dozat and Manning use a larger but more thoroughly regularized parser, with biaffine classifiers to predict arcs and labels. They use biaffine attention instead of bilinear or traditional MLP-based attention; a biaffine dependency label classifier; and apply dimension-reducing MLPs to each recurrent output vector  $r_i$  before applying the biaffine transformation. The biaffine mechanism is similar to traditional affine classifiers, where the vector of scores  $s_i$  for all classes equals the weight matrix  $W$  multiplied by single LSTM output state  $r_i$  (or other vector input) plus the bias term  $b$ :

$$(1) \quad s_i = W r_i + b$$

(Fixed-class affine classifier)

In the biaffine mechanism, the weight matrix  $W$  in (1) is replaced by a  $(d \times d)$  linear transformation of the stacked LSTM output  $RU^{(1)}$  in (2) and a  $(d \times d)$  transformation  $Ru^{(2)}$  replaces the bias term  $b$ :

$$(2) \quad s_i^{(arc)} = (RU^{(1)})r_i + (Ru^{(2)})$$

(Variable-class biaffine classifier)

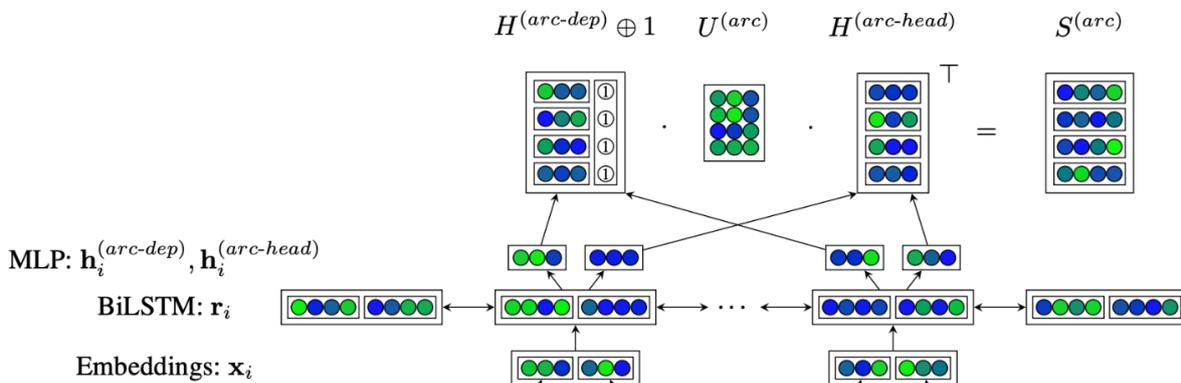


Figure 1: The graph-based architecture, proposed by Dozat and Manning.

Along with being simpler than the MLP-based approach, this has the conceptual advantage of directly modeling both the prior probability of a word  $j$  receiving any dependents in the term  $r_j^{Tu(2)}$  and the likelihood of  $j$  receiving a specific dependent  $i$  in the term  $r_j^r U_i^{(1)r}$ . The authors also use a biaffine classifier to predict dependency labels given the gold or predicted head  $y_i$ :

$$(3) \quad s_i^{(label)} = r_{y_i}^{TU(1)} r_i + (r_{y_i} \otimes r_i)^T U^{(2)} + b$$

(Fixed-class biaffine classifier)

Dozat and Manning point out that applying smaller MLPs to the recurrent output states before the biaffine classifier has the advantage of stripping

away information not relevant to the current decision. They also claim that reducing dimensionality and applying a nonlinearity (4, 5, 6) increases parsing speed and decreases the risk of overfitting.

$$(4) \quad h_i^{(arc-dep)} = \text{MLP}^{(arc-dep)}(r_i)$$

$$(5) \quad h_j^{(arc-head)} = \text{MLP}^{(arc-head)}(r_j)$$

$$(6) \quad s_i^{(arc)} = H^{(arc-head)} U^{(1)} h_i^{(arc-dep)} + H^{(arc-head)} u^{(2)}$$

They call this a **deep** bilinear attention mechanism, as opposed to **shallow** bilinear attention, which uses the recurrent states directly.

MLPs are applied to the recurrent states before using them in the label classifier as well.

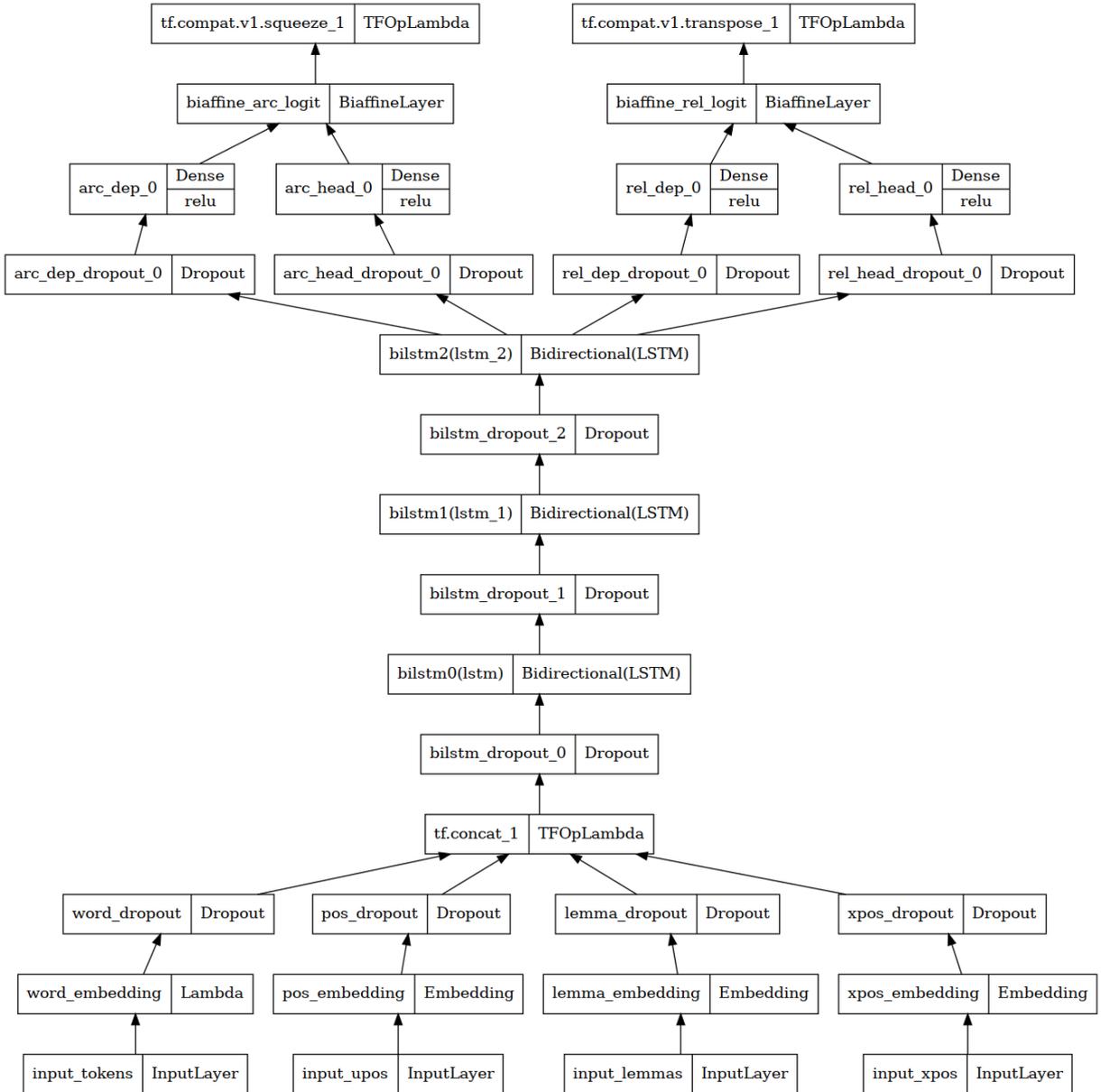


Figure 2: The architecture of the model with additional input layers.

### 3.2 Model Implementation

The parser proposed here is implemented following this architecture, with some modifications. Given that Bulgarian is a morphologically rich language, the model includes optional input layers for lemmas and morphological tags. These additions could potentially enhance its performance and warrant further evaluation. The model is constructed using the Keras API and features several custom components.

The primary component is responsible for constructing the neural network model. It includes methods for building the model's components, such as the embedding, BiLSTM, MLP, and biaffine attention layers. The model is compiled with a custom loss function, based on the sparse categorical cross-entropy loss object. This function computes the loss between the true and predicted values for the arcs and relations, returning a tensor that represents the average loss per example in the batch. This average loss is used during training to update the model's weights.

While the Adam optimizer is a popular choice for training deep learning models due to its adaptive learning rate, this implementation also includes the option to use exponential decay to potentially improve results.

The class tracks three metrics: the mean loss, the unlabeled attachment score (UAS), and the labeled attachment score (LAS).

Since TensorFlow does not provide a built-in biaffine layer, a custom one is implemented to perform the transformations. It follows the formulas for calculating the scores for potential arcs (6) and labels (3), and computes the weighted sum of the input tensors according to the weight matrix.

The model, comprising 31,036,906 parameters, is depicted in [Figure 2](#).

The training process includes an option to log summaries in both TensorBoard and MLflow. This feature allows for the monitoring and comparison of all hyperparameter changes.

## 4 Experiments and Results

The model underwent testing on the Bulgarian UD treebank with varying hyperparameters, and the results were evaluated using the Unlabeled Attachment Score (UAS) and Labeled Attachment Score (LAS). The UAS measures the proportion of

words correctly attached to their head, while the LAS measures the proportion of words correctly attached to their head with the correct dependency relation.

Optimal performance was achieved with a higher dropout rate of 0.5, as opposed to the 0.33 reported by Dozat and Manning. This can be attributed to the smaller size of the Bulgarian UD treebank compared to the treebanks used in their study, necessitating stronger regularization to prevent overfitting and enhance generalization.

Modifications to the learning and decay rates did not yield improved results. The best scores achieved using exponential decay, tested with values between 0.075 and 0.95, were 0.14% for UAS and 0.08% lower than the scores reported in [Table 1](#). Similarly, increasing the batch size (and correspondingly the number of epochs) did not significantly affect performance. Specifically, training with a batch size of 512 and 160 epochs resulted in scores that were 1.04% lower for UAS and 1.02% lower for LAS compared to training with a batch size of 128 and 80 epochs with the same hyperparameters.

A substantial improvement (0.86% for UAS and 1.31% for LAS) was observed upon the inclusion of lemmas and language-specific morphological information as input layers. This enhancement is anticipated given the complexity of the Bulgarian language and the significance of morphological information in parsing it.

Another enhancement involved replacing the traditional embedding layer for input tokens with RoBERTa embeddings. This change leverages the pre-trained model ([Liu et al., 2019](#)) to generate contextualized embeddings, which capture richer semantic and syntactic information. Since RoBERTa is used in the current version of the parser only for token vectorization, the improvement is not particularly large (0.21% for UAS and 0.49% for LAS). However, it still demonstrates the advantage of using large language models.

The model achieved a UAS of 93.32 and a LAS of 89.73 on the test dataset, thereby demonstrating its ability to accurately predict the dependency parse tree of a sentence. [Table 1](#) compares the performance of this model with other models. It surpasses the NLP pipeline for Bulgarian, developed within the spaCy framework ([Popov et al., 2020](#)). The model also yields superior results (with a 1.90% increase on UAS and a 2.72%

increase on LAS) compared to the model by Dozat and Manning, which uses only two input layers and a dropout rate of 0.33. It is also ahead of NLP-Cube (Boros et al., 2018) and UDPipe 2.0 (Straka, 2018). However, it still falls short of UDify (Kondratyuk and Straka, 2019). One reason for this could be that UDify is trained multilingually. Nevertheless, even when trained solely on Bulgarian, UDify's results are closely matched, suggesting that its primary advantage lies in the use of the BERT self-attention model. It is worth noting that the presented model is considerably smaller in size (especially when trained without RoBERTa embeddings) compared to UDify.

| Model               | UAS   | LAS   |
|---------------------|-------|-------|
| spaCy               | 88.95 | 83.03 |
| Biaffine w/o morph  | 91.21 | 86.52 |
| NLP-Cube            | 92.47 | 88.93 |
| UDPipe 2.0          | 92.82 | 89.70 |
| Biaffine with morph | 93.32 | 89.73 |
| UDify               | 95.54 | 92.40 |

Table 1: Results on Bulgarian UD dataset

## 5 Future Work

The model performs well on the Bulgarian UD treebank, with its results approaching those of state-of-the-art parsers. However, further improvements are necessary. Potential areas for enhancement include:

- Hyperparameter tuning: The model's hyperparameters can be further optimized to improve its performance. This includes (but not restricted to) the learning rate, the number of layers in the BiLSTM, the number of units in the MLPs, and the dropout rate.
- Better embeddings: The model can be improved by using better word embeddings, leveraging pre-trained large language models.

Another crucial step involves integrating POS and XPOS annotations. Currently, the model utilizes the CLASSLA library (Ljubešić and Dobrovoljc, 2019; Terčon and Ljubešić, 2023) to perform POS and XPOS tagging, which are then used as inputs for sentence prediction. Therefore, the next development step is to directly incorporate this functionality by training a morphological tagger.

While the focus of this study was specifically on Bulgarian, the model should be evaluated with other highly inflected languages to determine if the inclusion of lemmas and morphological tags improves performance for these languages as well.

## 6 Conclusion

The implementation of the neural network-based dependency parser, utilizing TensorFlow and Keras, gave near state-of-the-art results (UAS: 93.32, LAS: 89.73). The parser underwent training and evaluation on the Bulgarian Universal Dependencies dataset, yielding competitive results and thereby demonstrating the efficacy of the proposed architecture. Although the original model by Dozat and Manning was evaluated on considerably larger datasets and languages with simpler morphological structures, the results are comparable. For instance, the parser's results for Bulgarian outperforms the original model's scores for Chinese and Czech, the latter possessing one of the largest treebanks. The parser can predict the dependency syntax structure of Bulgarian sentences, and the *displacy* module from the *spacy* library can visualize these predictions. The parser's performance can be further enhanced by optimizing its hyperparameters and employing superior word embeddings.

## References

- Tiberiu Boros, Stefan Daniel Dumitrescu, and Ruxandra Burtica. 2018. [NLP-Cube: End-to-End Raw Text Processing With Neural Networks](#). In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 171–179, Brussels, Belgium. Association for Computational Linguistics.
- Timothy Dozat and Christopher D. Manning. 2017. [Deep Biaffine Attention for Neural Dependency Parsing](#). In *International Conference on Learning Representations*.
- Diederik P. Kingma and Jimmy Ba. 2014. [Adam: A method for stochastic optimization](#). *arXiv preprint arXiv:1412.6980*.
- Eliyahu Kiperwasser and Yoav Goldberg. 2016. [Simple and accurate dependency parsing using bidirectional LSTM feature representations](#). *Transactions of the Association for Computational Linguistics* 4: 313-327. MIT Press.

- Dan Kondratyuk and Milan Straka. 2019. *75 Languages, 1 Model: Parsing Universal Dependencies Universally*. 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 2779–2795, Hong Kong, China. Association for Computational Linguistics.
- Geert-Jan Kruijff. 2002. Formal and computational aspects of dependency grammar: History and development of DG. Technical report, ESSLI-2002.
- Sandra Kübler, Ryan McDonald and Joakim Nivre, 2009. Dependency Parsing. *Synthesis lectures on human language technologies 2*. Morgan & Claypool Publishers
- 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*. <https://arxiv.org/abs/1907.11692>
- Nikola Ljubešić and Kaja Dobrovoljc. 2019. *What does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian*. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 29–34, Florence, Italy. Association for Computational Linguistics.
- Marie-Catherine de Marneffe, Christopher Manning, Joakim Nivre, and Daniel Zeman. 2021. *Universal Dependencies*. *Computational Linguistics 2021*; 47 (2): 255–308.
- Petya Osenova and Kiril Simov. 2015. *Universalizing BulTreeBank: a Linguistic Tale about Glocalization*. In *Proceedings of BSNLP 2015*, Hissar, Bulgaria.
- Alexander Popov, Petya Osenova, and Kiril Simov. 2020. *Implementing an End-to-End Treebank-Informed Pipeline for Bulgarian*. In *Proceedings of the 19th International Workshop on Treebanks and Linguistic Theories*, pages 162–167, Düsseldorf, Germany. Association for Computational Linguistics.
- 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.
- Luka Terçon and Nikola Ljubešić. 2023. *CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages*. arXiv:2308.04255.
- Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, et al. 2017. *CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.
- Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. *CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium. Association for Computational Linguistics.

# Towards a Romanian phrasal academic lexicon

Madalina Chitez<sup>1</sup>, Ana-Maria Bucur<sup>2</sup>, Andreea Dinca<sup>1</sup>, Roxana Rogobete<sup>1</sup>

<sup>1</sup>West University of Timisoara, Timisoara, Romania

<sup>2</sup>University of Bucharest, Bucharest, Romania

{madalina.chitez, andreea.dinca, roxana.rogobete}@e-uvt.ro  
ana-maria.bucur@drd.unibuc.ro

## Abstract

The lack of NLP based research studies on academic writing in Romania results in an unbalanced development of automatic support tools in Romanian compared to other languages, such as English. For this study, we use Romanian subsets of two bilingual academic writing corpora: the ROGER corpus, consisting of university student papers, and the EXPRES corpus, composed of expert research articles. Working with the Romanian Academic Word List / Ro-AWL, we present two phrase extraction phases: (i) use Ro-AWL words as node words to extract collocations according to the thresholds of statistical measures and (ii) classify extracted phrases into general versus domain-specific multi-word units. We show how manual rhetorical function annotation of resulting phrases can be combined with automatic function detection. The comparison between academic phrases in ROGER and EXPRES validates the final phrase list. The Romanian phrasal academic lexicon (ROPAL), similar to the Oxford Phrasal Academic Lexicon (OPAL), is a written academic phrase lexicon for Romanian language made available for academic use and further research or applications.

**Keywords:** Romanian academic writing corpora, Romanian phrasal academic lexicon, EXPRES corpus.

## 1 Introduction

We present the first approach to creating the Romanian phrasal academic lexicon (ROPAL)<sup>1</sup> for written Romanian language. Academic writing is a challenge for students and experienced writers alike. Studies (Saber et al., 2020; Chitez and Dinca, 2023) have pointed out that linguistic support is effective at all levels of language use and

acquisition. Among the multitude of support options, e.g. grammatical accuracy, lexical diversity, paraphrasing (Strobl et al., 2019), multi-word units with a rhetorical function, i.e. academic phrases, are the most productive in enhancing the coherence and persuasiveness of academic writing (AlHassan and Wood, 2015; Hinkel, 2003). Thus “opportunities to practice bundle usage so as to trigger acquisition of formulaic language” (Pérez-Llantada, 2014) are associated with higher academic writing proficiency. These phrases ensure that writers comply with specialized academic conventions and further contribute to the development of writers’ language skills, facilitating their ability to articulate complex ideas and arguments with precision and refinement. From a computational linguistics perspective, phrasal academic lexicons can serve as invaluable resources for training large language models to improve their proficiency in language generation and paraphrasing. At the same time, academic phrase lists can contribute to training LLM models to automatically annotate rhetorical functions. As a result, AI tools can be improved, which can help with tasks such as genre classification, stylistic analysis, and discourse segmentation, thereby enhancing the automated processing and understanding of academic texts.

When compiling lists of academic phrases, specialized corpora such as academic writing corpora can be of great assistance in identifying prevalent language patterns and rhetorical strategies across different academic disciplines or cross-disciplinarily. Although most extraction and analysis methods have been performed for phraseology extraction and use in L2 (English) (Section 2), their replicability for academic writing in L1 has been demonstrated (Pérez-Llantada, 2014). In the present paper, we use the EXPRES corpus and the Romanian Academic World List / Ro-AWL (Bucur et al., 2023) (Section 3) to extract phrases

<sup>1</sup><https://github.com/chia-AR/ROPAL-Romanian-phrasal-academic-lexicon>

in ROPAL. Highly frequent phrases (first 150 occurrences in ROPAL) are manually annotated for rhetorical functions (Section 5). We use an adaptation of two rhetoric function systems proposed by Morley (2018) and by OPAL (2019) for the English language. In the end, we verify whether ROPAL can detect learner academic writing phrase patterns by using the ROGER corpus (Chitez et al., 2022a) (Section 5).

## 2 Related Work

Most research on academic multiword units was done in the context of English for Academic Purposes (EAP), where considerable attention was placed on finding a core academic phrasal lexicon that could be used across disciplines. These phraseological resources are informed by both EAP research and EAP practice, as academic writing is widely taught in university settings. Typically, EAP research uses academic writing corpora from which common multiword units are identified using computational methods. After extraction, the phrases are assigned rhetorical and functional categories (see e.g. Hyland (2008)). These findings are then integrated into various learning and teaching resources for academic writing. For instance, they may be presented into standalone phrase lists (e.g. Simpson-Vlach and Ellis (2010)) or phrasebooks (e.g. Morley (2018)) categorized by rhetorical and functional attributes. More recently, phrases have also been integrated into digital learners' dictionaries (see OPAL (2019)), or have become part of AI-powered academic writing platforms, such as Writefull<sup>2</sup>.

In contrast to the abundance of resources available for English, research on academic phraseological units in other languages is scarce. To the best of our knowledge, no study to date has aimed at identifying cross-disciplinary academic phraseological units in a language other than English. Existing investigations predominantly adopt a comparative approach, such as Cortes (2008), which compares English and Spanish academic phraseological units. In the case of Romanian, research is rather characterized by heterogeneity, both in theoretical frameworks, terminologies and methodologies employed (Zafu, 1989). While some researchers refer to 'pragmatic functions' (Roventța-Frumușani, 2012; Ștefănescu, 2007), others may use terms such as 'metalinguistic markers' (Bîtea, 1986), 'special-

ized syntagms' (Pricop, 2014). This lack of standardization in terminology makes it challenging to compile comprehensive online resources, such as lists of various types of phraseological units or academic vocabulary or conventions. Unlike languages with more robust digital resources, such as English, the availability of openly accessible linguistic data for Romanian is limited. Even though the last two decades have seen various initiatives launched (Tamba, 2017; Cristea et al., 2011), efforts have lacked sufficient coordination. While specific micro-studies have been conducted to analyze a restricted number of phraseological units across limited disciplines comparatively (Bocoș, 2018; Stoichițoiu-Ichim, 2001), only one study has focused on multiword units in Romanian academic writing (Muresan et al., 2022). It explored the difference between expert academic writing produced by Romanian scholars and novice academic writing produced by Romanian university students.

## 3 Method

Several language datasets are used in this study. The Corpus of Expert Academic Writing in Romanian and English / EXPRES (Chitez et al., 2022b) and the Romanian Academic Word List / Ro-AWL (Bucur et al., 2023) were used in the creation of ROPAL. Subsequently, The Corpus of Romanian Academic Genres / ROGER (Chitez et al., 2022a) is used to test ROPAL. The datasets and the extraction procedures are described in the following sections.

### 3.1 EXPRES and ROGER

**EXPRES** is a bilingual multidisciplinary corpus representative of expert academic writing in English and in Romanian. This investigation uses a subset of EXPRES containing academic writing in Romanian. As shown in Table 1, the dataset has a size of more than 3 million words, and more than 200 thousand unique words. The disciplines represented in the corpus are Linguistics (LG), Economics (EC), Information Technology (IT) and Political Science (PS). There are 800 research articles, 200 articles for each discipline.

**ROGER** is a bilingual comparable learner corpus containing academic writing by students enrolled in Romanian universities. The full corpus contains various text genres from five disciplines, written in either students' native language, Romanian, or in English as a Foreign Language. The

<sup>2</sup><https://www.writefull.com/>

| Domain       | Tokens           | Types          |
|--------------|------------------|----------------|
| EC           | 1,092,846        | 48,807         |
| LG           | 674,277          | 73,667         |
| IT           | 750,236          | 40,494         |
| PS           | 963,061          | 62,096         |
| <b>Total</b> | <b>3,480,420</b> | <b>225,064</b> |

Table 1: EXPRES Statistics.

| Domain       | Texts      | Tokens           | Types          |
|--------------|------------|------------------|----------------|
| EC           | 97         | 238,770          | 39,605         |
| HUM          | 653        | 476,232          | 75,785         |
| CS           | 42         | 118,507          | 21,131         |
| PS           | 119        | 346,388          | 52,796         |
| <b>Total</b> | <b>911</b> | <b>1,179,897</b> | <b>139,283</b> |

Table 2: ROGER Statistics.

present study uses the Romanian subset of ROGER. As can be seen from Table 2, it contains text genres such as essays, or B.A. theses from the disciplines of Economics (EC), Humanities (HUM), Computer Science (CS) and Political Sciences (PS). The corpus amounts to more than one million words.

### 3.2 Ro-AWL

The Romanian Academic Word List<sup>3</sup> (Bucur et al., 2023) is derived from the EXPRES corpus and consists of a set of academic words that are frequently encountered in academic texts. It is made up of 673 lemmas, distributed among the main part-of-speech categories (nouns, verbs, adverbs, adjectives). The list is freely available, and a detailed description of its creation is available in Bucur et al. (2023).

### 3.3 Extraction and annotation of the phrases

The EXPRES corpus is organized into .txt files. We removed specific tags such as {FIG}, {REF\_LIST}, {JOURNAL\_TITLE}, {AUTHOR\_NAME}, or tags used to indicate the title (<TITLE>, </TITLE>), abstract (<ABS\_INT>, </ABS\_INT>), keywords (<KW\_INT>, </KW\_INT>), etc. We used Stanza (Qi et al., 2020) for lemmatization, and all the lemmas from the texts were converted to lowercase for further analysis. We used Ro-AWL terms as node words and extracted collocations of 2 to 6 words from EXPRES that contain the academic words found in Ro-AWL.

<sup>3</sup><https://github.com/bucuram/Ro-AWL>

We compiled the phrasal academic lexicon using the criteria of frequency and dispersion, following the works of Ädel and Erman (2012) and Ebeling and Hasselgård (2015). Given that we extracted collocations containing up to 6 words, we opted for a lower threshold for frequency – a collocation had to occur at least 10 times per million words. Dispersion was used to complement frequency measures to ensure that a collocation was not idiosyncratically confined to a limited number of texts. This is why the collocations that appeared in less than 5 different texts were excluded from the analysis. Manual filtering was further performed to remove collocations that were part of larger phrases. For example, the collocation “punct de vedere” (En: “point of view”) was manually removed, as it was contained into “din punct de vedere” (En: “from the point of view”). Even if the purpose of the lexicon was to encompass 6-grams, all collocations with 6 words were excluded from the final list either because of their low frequency and dispersion, or because they were removed during manual filtering.

In this study, the frequency of the collocations is presented in two ways: raw frequency, which is the number of times a collocation appears in the dataset, and normalized frequency per 1,000 words (Lancaster, n.d.). Since the datasets differ in size, it is common practice to use normalized frequencies for comparing the results.

## 4 The Romanian phrasal academic lexicon

### 4.1 Statistics

| Collocation         | English Translation    | Frequency Normalized per 1,000 words |
|---------------------|------------------------|--------------------------------------|
| de asemenea         | also                   | 0.67                                 |
| de tip              | of type                | 0.49                                 |
| de exemplu          | e.g.                   | 0.42                                 |
| din punct de vedere | from the point of view | 0.38                                 |
| cu privire la       | regarding              | 0.36                                 |
| în funcție de       | depending on           | 0.36                                 |
| avea în vedere      | considering            | 0.33                                 |
| în timp ce          | while                  | 0.32                                 |
| pe de altă parte    | on the other hand      | 0.25                                 |
| pe termen           | on term                | 0.22                                 |

Table 3: Top 10 collocations in ROPAL.

ROPAL includes a total of 794 collocations, which are comprised of 477 2-grams, 252 3-grams, 52 4-grams and 13 5-grams. Table 3 displays the 10 most frequent collocations in ROPAL. The number

| Category  | Example translated from Romanian               | No of collocations |
|---|--|--------------------|
| Perspective and Viewpoint, Considering or Taking into Account a Specific Perspective      | from the point of view; regarding; considering | 28                 |
| Miscellaneous (e.g. Expressing quantity; Timeframe)                                       | short term; real time                          | 17                 |
| Compare and contrast (discourse)  | while; on the other hand                       | 14                 |
| Defining; Explaining terms  | of type; consists of                           | 14                 |
| Evaluation; Expressing degree   | extremely; quality                             | 13                 |
| Signaling transition; Introduce additional information                                    | also; furthermore                              | 9                  |
| Conclusion; Summary; Hypothesis   | thus; this work                                | 9                  |
| Literature review; Discussing related work; Appeal to Authority; Referring to other texts | in the specialized literature; in discussion   | 9                  |
| Explaining causality  | as a result; having as a goal                  | 7                  |
| Giving examples   | e.g.; for example in                           | 3                  |

Table 4: Classification of academic language functions in ROPAL.

is influenced by the choice of the method extraction, i.e. use of Ro-AWL list (see Sections 3.2 and 3.3).

To verify ROPAL’s reliability, we tested the coverage of the list in two academic writing corpora, EXPRES and ROGER. The coverage of ROPAL in the EXPRES corpus is 3.6%. When it comes to the coverage of ROPAL in writings by university students from the ROGER corpus, the coverage is lower, at 1.6%. ROPAL used EXPRES for validation since the list was based on Ro-AWL, which, in turn, used another reference corpus for list validation (Bucur et al., 2023). At the same time, ROPAL list was manually filtered, thus being quite different from the automatically generated list.

## 4.2 Classification

To classify the most frequent collocations in ROPAL, we used a mixed approach, by adapting established models such as the OPAL framework (OPAL, 2019) and the Manchester Academic Phrasebank (Morley, 2018). The decision to draw from these models was motivated by the lack of a standardized classification system in Romanian (Stefanescu, 2017). These models provided valuable insights into the categorization of academic language functions and served as guiding templates for the refinement process for Romanian phrases in ROPAL. The final 10 categories were developed for the first 129 units/collocations, which occurred in the corpus with a frequency of at least 0.05. We simplified overlapping concepts and reorganized them into broader groups. The expertise of

the team, in both Romanian linguistics and didactics, contributed to the creation of a unique classification model inspired by previous international models. For example, categories like “Being critical” and “Describing trends” were integrated into broader categories such as “Perspective and Viewpoint; Considering or Taking into Account a Specific Perspective”. We also merged categories such as “Making contrast” and “Comparing”. Finally, we developed a more versatile category - “Literature review; Discussing related work; Appeal to Authority/ Referring to other texts”, which encompasses classes such as “Hedging”, “Writing about the past” and “Describing trends”. The final functional areas developed for this study can be seen in Table 4.

Results show that phrases falling under the category of “Perspective and Viewpoint/Considering or Taking into Account a Specific Perspective” are the most common in research articles from EXPRES corpus. This is because academic writing requires authors to discuss or evaluate various viewpoints or theoretical frameworks (particularly within sections like the literature review) to present their own arguments.

Furthermore, the “Compare and contrast” section is well represented, since scholars often choose a comparative analysis or contextualize their research within the existing literature and academic community.

### 4.3 Evaluation

To describe how the phrasal academic units from ROPAL are distributed among disciplines, we conducted a comparative linguistics analysis looking at how these units are used in four disciplines.

The EXPRES platform (<https://expres-corpus.org/>) where the corpus is freely available for use, was used to extract examples. The top three collocations in each of the four disciplines, are presented in Table 5. Most of these collocations serve as linguistic tools for authors to introduce, contextualize, and evaluate different viewpoints, while contributing to the overall coherence and logical flow of the text. By looking at discipline-based phrases in ROPAL, common patterns, phrase overlaps and discipline specific units be highlighted. Research articles in Economics contain specialized phrases such as “at a national level”, “on term”, “on long-term”, “at the global level”. The field of Economics appears to be the most specialized among the disciplines analyzed, given the presence of N-grams that are entirely absent in the other disciplines examined, like “from a statistical point of view”, “had a positive impact on”, “at an average pace”, “annual average of”, “growth trend”, “influencing factors”. When shifting our attention to Information Technology, structures such as “database analysis”, “real time”, “model of”, “in the database”, “of classification” are specific to data management and analysis. An N-gram which occurs only in IT is “allow access”, emphasizing the central role of security, databases, and information processing for the domain. The field of linguistics employs most of the rhetorical strategies and connective phrases, such as “e.g.”, “also”, “by point of view”, “of type”, “depending on”, “considering”, “while”. A structure like “in the paradigm” occurs only in this domain, which appears to have a predilection for expansive explanations and exemplification. The last analyzed discipline, political sciences, is characterized by inserting multiple perspectives, since the most frequent structures are “regarding”, “also”, “by ... point of view”, “on the other hand”, “while”. This field appears to be the most nonspecific in using academic phrases, since there is none used in political sciences and absent in other disciplines (even an N-gram like “in the public space” appears also in other corpora).

Overall, it can be observed that political sciences appear less prone to specific linguistic patterns, per-

| Collocation                   | Examples from the EXPRES   |
|-------------------------------|--|
| <b>Economics</b>              |  |
| regarding                     | “member states will have reports regarding financial aspects”  |
| considering                   | “an analysis considering a causal relationship between macroeconomic variables and [...]”                    |
| from the point of view        | “the most developed regions from the economical point of view”   |
| <b>Information Technology</b> |  |
| depending on                  | “depending on the GPU memory”  |
| from the point of view        | “from the point of view of data management functionalities”  |
| by analysis                   | “The functionality and effectiveness of MOOC projects will be highlighted by traffic and event log analysis” |
| <b>Linguistics</b>            |  |
| from the point of view        | “from a semantic point of view”  |
| depending on                  | “depending on conjugation”   |
| considering                   | “Considering semantic equivalences/ analogies [...]”   |
| <b>Political Sciences</b>     |  |
| regarding                     | “we have identified several approaches regarding foreign affairs”  |
| from the point of view        | “from the point of view of the political route”  |
| believes that                 | “a large part of the population believes that the state should intervene”                                    |

Table 5: Top 3 collocations for each discipline in ROPAL.

haps due to its interdisciplinary nature, while academic writing in linguistics tends to overuse explanations. Although information technology seems to have a more technical focus, it often employs a more complex discourse. The field that exhibits the most distinctive academic phrase patterns is Economics.

### 5 Utilization

In this section, we will test the ROPAL list on the Romanian section of the ROGER corpus (Chitez et al., 2022a), also available online (<https://roger-corpus.org/>). We first selected the most frequent 10 N-grams from ROPAL, and searched them in the whole ROGER corpus, then in similar disciplinary datasets in EXPRES (economics, humanities, political sciences, computer science). The numbers are listed in Table 6 (normalized frequency in the entire ROGER corpus vs normalized frequency in the analyzed disciplines).

The distribution of the ROPAL academic phrases related to discourse cohesion follows, in ROGER,

| Collocation<br>English Translation | ROGER<br>Total | EC   | HUM  | CS   | PS   |
|------------------------------------|----------------|------|------|------|------|
| also                               | 0.32           | 0.26 | 0.45 | 0.22 | 0.21 |
| of type                            | 0.16           | 0.15 | 0.20 | 0.25 | 0.09 |
| e.g.                               | 0.20           | 0.20 | 0.25 | 0.19 | 0.12 |
| from the point of view             | 0.26           | 0.20 | 0.42 | 0.11 | 0.15 |
| regarding                          | 0.13           | 0.12 | 0.15 | 0.11 | 0.12 |
| depending on                       | 0.11           | 0.10 | 0.15 | 0.11 | 0.07 |
| considering                        | 0.14           | 0.12 | 0.19 | 0.12 | 0.10 |
| while                              | 0.12           | 0.17 | 0.11 | 0.04 | 0.11 |
| on the other hand                  | 0.10           | 0.07 | 0.12 | 0.03 | 0.09 |
| on term                            | 0.05           | 0.16 | 0.01 | 0.03 | 0.05 |

Table 6: Frequencies of the top 10 ROPAL discourse cohesion collocations in the ROGER corpus.

an expected pattern, since we identified common features across disciplines. In fact, the novice writing samples from ROGER do not present sufficient discipline-specific characteristics. For instance, the domain-specific units identified in EXPRES are poorly represented in ROGER. Students tend to focus more on elaborating their discourse than on the development of a discipline-specific language. Therefore, greater exposure to specialized articles would enable students to familiarize themselves with the conventions and expectations of their micro-academic community.

## 6 Conclusions

The findings presented in the study highlight several key points regarding the development and application of the Romanian Phrasal Academic Lexicon (ROPAL) in computational linguistics and language education. Firstly, since the main objective of the current study was to verify the capacity of our generated academic phrase lexicon to support the automatic assessment of academic writing proficiency in the native language Romanian (which was demonstrated by the fact that coverage percentages are lower in novice versus expert writing), IAA was not performed.

Secondly, the extraction and annotation of academic phrases from the EXPRES corpus demonstrate the didactic applicability of using corpora to compile resources adapted for the needs of Romanian academic writers. The phrase extraction and classification approaches provide insights into prevalent language patterns and rhetorical strategies across different academic disciplines, but also across disciplines. Thus, a pilot ROPAL list (i.e. ROPAL for teaching), similar to the OPAL list (OPAL, 2019), to be used for teaching and edu-

cation purposes, was compiled<sup>4</sup>. It contains the prominent rhetorical categories in Romanian academic phrases, based on the complete OPAL list generated for computational purposes. Such a list is particularly novel because, until now, no equivalent resource has existed for the Romanian academic setting.

Most items in ROPAL for teaching represent general academic writing multi word units (e.g. [translation from Romanian into English] “also”, “from the point of view”, “regarding”) whereas others are rather discipline-specific or have a higher probability of being used more frequently in a particular disciplinary field. This list will be adapted for various disciplines and validated through interrater agreement (IAA) methods. Further studies using ROPAL and the rhetoric function annotation for each phrase may result in the creation of AI tools for academic text correction, text suggestion and text assessment in the Romanian language. ROPAL enhances students’ ability to engage with and produce academic work that adheres to the rhetorical standards of their field.

## Acknowledgments

This study was supported by the UVT 1000 Develop Fund of the West University of Timisoara.

## References

- Annelie Ädel and Britt Erman. 2012. Recurrent word combinations in academic writing by native and non-native speakers of english: A lexical bundles approach. *English for specific purposes*, 31(2):81–92.
- Lina AlHassan and David Wood. 2015. The effectiveness of focused instruction of formulaic sequences in augmenting 12 learners’ academic writing skills: A quantitative research study. *Journal of English for Academic Purposes*, 17:51–62.
- Ioan N Bîtea. 1986. Conectorii metalingvistici: adverbe ori conjuncții? *Editura Academiei Republicii Socialiste România*.
- Cristina Bocoș. 2018. Unitățile frazeologice cu termeni religioși în română și spaniolă. o analiză comparativ-contrastivă. *Philologica Jassyensia*, 14(2):337–340.
- Ana-Maria Bucur, Andreea Dincă, Mădălina Chitez, and Roxana Rogobete. 2023. Automatic extraction of the romanian academic word list: Data and methods. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 234–241.

<sup>4</sup>ROPAL for teaching is freely accessible online at: <https://codhus.projects.uvt.ro/corpus-tools/>.

- Mădălina Chitez, Loredana Bercuci, Andreea Dincă, Roxana Rogobete, and Karla Csürös. 2022a. *Corpus of romanian academic genres (ROGER)*. West University of Timisoara.
- Madalina Chitez and Andreea Dinca. 2023. On corpora and writing. In *Digital Writing Technologies in Higher Education: Theory, Research, and Practice*, pages 385–403. Springer.
- Mădălina Chitez, Valentina Muresan, Roxana Rogobete, and Andreea Dincă. 2022b. *Corpus of expert writing in romanian and english (EXPRES)*. West University of Timisoara.
- Viviana Cortes. 2008. A comparative analysis of lexical bundles in academic history writing in english and spanish. *Corpora*, 3(1):43–57.
- Dan Cristea, Gabriela Haja, Alex Moruz, Marius Răschip, and Mădălin Pătrașcu. 2011. Statistici parțiale la încheierea proiectului edtlr–dicționarul tezaur al limbii române în format electronic. În *volumul Lucrările conferinței naționale Limba română: ipostaze ale variației lingvistice*, pages 3–4.
- Signe Oksefjell Ebeling and Hilde Hasselgård. 2015. Learners’ and native speakers’ use of recurrent word-combinations across disciplines. *Bergen Language and Linguistics Studies*, 6.
- Eli Hinkel. 2003. *Teaching academic ESL writing: Practical techniques in vocabulary and grammar*. Routledge.
- Ken Hyland. 2008. As can be seen: Lexical bundles and disciplinary variation. *English for specific purposes*, 27(1):4–21.
- University of Lancaster. n.d. *Comparing frequencies for corpora of different sizes*.
- John Morley. 2018. *Academic phrasebank: A compendium of commonly used phrasal elements in academic English in PDF format*. The University of Manchester.
- Valentina Muresan, Roxana Rogobete, Ana-Maria Bucur, Madalina Chitez, and Andreea Dinca. 2022. Phraseology in romanian academic writing: Corpus based explorations into field-specific multiword units. *Recent Advances in Digital Humanities. Romance Language Applications*, Peter Lang, pages 29–48.
- OPAL. 2019. *Oxford phrasal academic lexicon*.
- Carmen Pérez-Llantada. 2014. Formulaic language in L1 and L2 expert academic writing: Convergent and divergent usage. *Journal of English for Academic Purposes*, 14:84–94.
- Alina-Mihaela Pricop. 2014. Tendințe în tratarea sintagmelor specializate. *Lexicografia academică românească*, page 139.
- 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*, pages 101–108.
- Daniela Roventța-Frumușani. 2012. *Analiza discursului: ipoteze și ipostaze*. Tritonic.
- Dariush Saberi, John Lee, and Jonathan Webster. 2020. Automatic assistance for academic word usage. In *The 28th International Conference on Computational Linguistics (COLING’2020)*, pages 2163–2168. International Committee on Computational Linguistics.
- Rita Simpson-Vlach and Nick C Ellis. 2010. An academic formulas list: New methods in phraseology research. *Applied linguistics*, 31(4):487–512.
- Ariadna Ștefănescu. 2007. Conectori pragmatici.
- Maria Ștefănescu. 2017. Some remarks on phraseological units in the thesaurus dictionary of the romanian language. *Dacoromania*, 22(1):51–62.
- Adriana Stoichițoiu-Ichim. 2001. *Semiotica discursului juridic*. Editura Universității din București.
- Carola Strobl, Emilie Ailhaud, Kalliopi Benetos, Ann Devitt, Otto Kruse, Antje Proske, and Christian Rapp. 2019. Digital support for academic writing: A review of technologies and pedagogies. *Computers & education*, 131:33–48.
- Elena Isabelle Tamba. 2017. Clre. corpus lexicografic românesc esențial. 100 de dicționare din bibliografia dlr alinate la nivel de intrare și la nivel de sens. haja, gabriela (éd.). 2017. *Lexicografia academică românească. Studii. Proiecte*, pages 221–234.
- Rodica Zafiu. 1989. Câteva observații asupra conectorilor pragmatici din limba română. *Studii și cercetări lingvistice, Vol. 40, nr. 3,(1989); p. 315-319*.

# Classifying Multi-Word Expressions in the Latvian Monolingual Electronic Dictionary Tēzaurs.lv

Laura Rituma, Gunta Nešpore-Bērzkalne, Agute Klints,  
Ilze Lokmane, Madara Stāde, Pēteris Paikens

Institute of Mathematics and Computer Science

University of Latvia

Raiņa bulvāris 29, Rīga, Latvia

{laura.rituma, gunta.nespore, agute.klints, peteris.paikens}@lumii.lv  
ilze.lokmane@lu.lv, madara.stade@gmail.com

## Abstract

The electronic dictionary Tēzaurs.lv contains more than 400,000 entries from which 73,000 entries are multi-word expressions (MWEs). Over the past two years, there has been an ongoing division of these MWEs into subgroups (proper names, multi-word terms, taxa, phraseological units, collocations). The article describes the classification of MWEs, focusing on phraseological units (approximately 7,250 entries), as well as on borderline cases of phraseological unit types (phrasemes and idioms) and different MWE groups in general. The division of phraseological units depends on semantic divisibility and figurativeness. In a phraseme, at least one of the constituents retains its literal sense, whereas the meaning of an idiom is not dependent on the literal sense of any of its constituents. As a result, 65919 entries of MWE have been manually classified, and now this information of MWE type is available for the users of the electronic dictionary Tēzaurs.lv.

**Keywords:** multi-word expression, phraseological unit, idiom, phraseme, semantics.

## 1 Introduction

Tēzaurs.lv<sup>1</sup> is the largest Latvian electronic explanatory dictionary with more than 400,000 entries. It emerged as a compilation from nearly 300 prior dictionaries and other sources (Grasmanis et al., 2023). Besides entries for single words Tēzaurs.lv also contains approximately 73,000 multi-word expressions (MWEs; dictionary entries that contain more than one orthographic word (Bauer, 2021: 5)) stored as separate entries. Most MWEs are linked to the corresponding word entries or a specific word sense that is included in the MWE. Therefore, dictionary users can either

search a specific expression or find it in the matching word entry.

Up until now, all Latvian studies of MWEs have been carried out to accommodate machine translation. A bilingual MWE dictionary has been created, listing the relevant syntactic patterns both in English and the respective Latvian MWEs; this helps obtain syntactic rules for better machine translation (Deksne et al., 2008). Additionally, there have been studies regarding the methods of obtaining MWE lists to improve the quality of translation (Skadiņa, 2016; Mandravickaitė and Krilavičius, 2017) or to expand the dictionary data (Skadiņa, 2018). However, the aim of these studies was not creating a system of MWE classification based on their function and meaning. The MWE lists do not contain sense descriptions and most of the data is not freely accessible. In contrast, Tēzaurs.lv open-access data contains MWE sense descriptions but lacks information on morphological and syntactic structure.

Over the past two years, functionally diverse expressions have been linguistically analyzed and manually sorted into following categories:

- multi-word place names, e.g. *Juglas ezers* ‘Jugla Lake’, *Egļu ciems* ‘Egļu Village’;
- taxonomic group names, such as species, families or classes, both international, e.g. *Vultur pryphus*, *Tulipa lanata*, and Latvian, e.g. *aklais dundurs* lit. ‘Blind Horse-Fly’, *vilnainā tulpe* ‘Woolly Tulip’;
- complex terms and term candidates, e.g. *centrbēdzes spēks* ‘centrifugal force’; *ciešamā kārta* ‘passive voice’
- phraseological units, e.g. *mest plinti krūmos* lit. ‘to throw the rifle into the bushes’ (to give up);

<sup>1</sup>Available interactively at <https://tezaurs.lv> or as data from <https://repository.clarin.lv/repository/xmlui/handle/20.500.12574/104>

- collocations, e.g. *pieļaut kļūdu* ‘to make a mistake’, *apģērba gabals* ‘piece of clothing’. We have adopted a rather narrow understanding of collocations, which are statistically significant co-occurrences of words outside of all previously mentioned groups. In other words, collocations are fixed word combinations with no semantic reinterpretation (Veisbergs, 2019: 114).

Table 1 shows the total number of MWEs in the dictionary Tēzaur.lv and the number of MWEs in each category and subcategory.

Certain groups were left outside of this classification, such as expressions in foreign languages (excluding taxa), e.g., *de facto*, *per aspera ad astra*, and abbreviations consisting of multiple tokens, e.g., *t. sk.* ‘incl.’, *a. god.* ‘esteemed’, as well as MWEs mentioned in Chapter 5.

This classification provides additional information to the dictionary users regarding MWE functions within the language and promote the study of phraseology in Latvian linguistics. A more formal structure of MWEs is also useful for computational linguistics tasks that involve structured, explicit semantic models, such as semantic parsing and information extraction, controlled natural languages, and structured natural language generation. This is especially relevant in multilingual solutions, as some concepts are expressed as individual word senses in one language and as MWEs in another, necessitating a structured inventory of the applicable MWEs. In addition to the MWE classes, their review improved the overall quality of data, e.g. combining close MWE variants into one dictionary entry. However, we noted that it was often difficult to decide whether two close MWEs are separate and further work is needed to develop objective criteria for this decision.

In this study, we focused on the semantic analysis of phraseological units by separating them into two subgroups depending on the relationship of the words forming the MWEs to the general meaning of the MWE itself (for a more detailed distinction between the two subgroups, phrasemes and idioms, see Chapter 3). The creation of this division is the first step, so that in the future, when the morphosyntactic and lexical variation of these phraseological units, including word order and derivation options (see, e.g. Leseva et al. (2020)), will be analyzed, it would be possible to test the hypothesis that phrasemes are more prone to morphosyntactic

and lexical variation than idioms. Other studies also emphasize that decomposable phraseological units tend to be syntactically flexible to some degree (see, e.g., Sag et al. (2002: 5–7)).

Chapter 2 deals with the borderline cases of phraseological units and other MWE groups mentioned above, namely, collocations, taxa, and terms. Chapter 3 outlines the distinction between idioms and phrasemes. Chapter 4 describes the borderline cases involving idioms and phrasemes to show that semantic transparency is essentially scalar. Chapter 5 describes MWE groups that were not included in any of the defined categories. Finally, the last chapter of the article consists of conclusions and future work for MWE processing.

## 2 Borderline Cases of the Phraseological Unit and Other MWE Classes

Before creating division of phraseological units, we had to establish terms for defining each MWE group. Difficulties arose when borders between two MWE classes were not that clear and fixed.

In this study, a MWE was classified as a collocation if all of the words that form it are used in their literal sense, i.e., the senses can be found in the dictionary entries of the corresponding words. For example, *izdzert līdz dibenam* lit. ‘to drink to the bottom’ is a collocation (and not a phraseological unit), since “dibens” ‘bottom’ has a literal meaning ‘lower part (e.g., of a dish)’.

However, during data processing, difficulties arose in separating collocations and phraseological units as latter possess some degree of figurative, transferred or metaphorical meaning (Veisbergs, 2019: 114). Figurativeness fades over time and it is difficult to decide the point at which the use of a word meaning transitions from figurative to literal, therefore to decide whether a MWE has to be classified as a phraseological unit or a collocation. The words that form a MWE are occasionally used in a sense that could be perceived as figurative, but may already be listed in the dictionary as literal, because most language users no longer note the meaning transfer. In that case the MWE is still sorted as a collocation. For example, in the expression *labas acis* lit. ‘good eyes’, the dictionary entry *acis* ‘eyes’ lists the meaning of vision without the “figurative” tag. Similarly, the expression *celt trauksmi* ‘to raise the alarm’ contains the word *celt* ‘raise’, which has a figurative meaning ‘radīt’ ‘to make’ listed in the dictionary without the “figurative” tag. Thus, both

| Name of Category                          | Name of Subcategory               | Number of MWEs |
|---|-----------------------------------|----------------|
| complex terms                             |                                   | 22,552         |
| multi-word place names                    |                                   | 14,733         |
| taxonomic group names                     | International                     | 10,347         |
|   | Latvian                           | 7,854          |
| phraseological units                      | phrasemes                         | 2,863          |
|   | idioms                            | 4,029          |
|   | unclassified phraseological units | 358            |
| collocations                              |                                   | 3,183          |
| <b>Classified MWEs in total</b>           |                                   | <b>65,919</b>  |
| Unclassified MWEs                         |                                   | 5,385          |
| <b>Total number of MWEs in Tēzaurs.lv</b> |                                   | <b>71,304</b>  |

Table 1: The number of MWEs sorted into each category and subcategory.

of the mentioned MWEs have been classified as collocations, even though they could also easily be seen as phraseological units, since they do display a certain degree of fading figurativeness.

Additionally, over time, certain figurative meanings have been preserved only in one expression. For example, the entry *apaļš* ‘round’ lists a meaning “not having any family”, which nowadays is only used in the expression *apaļš bārenis* lit. ‘a round orphan’. In such cases, it is advisable to delete this meaning of *apaļš* from the dictionary and sort the MWE as a phraseme.

These issues show that, at times, the line between figurative and direct meanings can be vague – the more frequent and varied the use of a figurative meaning is, the more likely it is that the meaning will lose its figurativeness. Therefore, with certain expressions it is more difficult to discern whether they still count as phraseological units or have already become collocations. In this study, it was decided not to delve into the borderline cases of figurativeness, but instead agree on clear criteria for separation based on dictionary data.

Further difficulties arose from the fact that both terms and taxa can be figurative, e.g., term *auss gliemene* lit. ‘ear clam’, taxon *atvērtā pērtiķmutīte* lit. ‘open monkey-mouth’. Although in Latvian linguistics figurative names are traditionally not recognized as phraseological units, they are essentially idioms, which only differ in their naming function (for a more detailed description of idioms, see Chapter 3). One MWE cannot simultaneously belong to several categories (e.g., term and idiom), so it was decided to classify such cases as terms or

taxa despite their figurativeness. In the future, these cases could be re-sorted into further sub-categories.

A distinct group is formed by expressions, that can be used in both literal and figurative sense. Stephen G. Pulman also examines such phraseological units as a special, separate group. He notes that the components of such unit have literal meanings, but that these are not what is involved in their interpretation as a phraseological unit. It is certainly the case that someone unfamiliar with the phraseological unit nevertheless can arrive at an appropriate meaning for it by processing it as a metaphor (Pulman, 1993: 260). For example, expressions *atmest ar roku* lit. ‘throw one’s hand at something’ (to stop, abandon doing something) and *grozīt galvu* lit. ‘turn one’s head around’ (express surprise, concern) can be used in their direct sense to describe a physical action, as well as figuratively. In such cases, the MWE has two meanings: one is direct (categorised as a collocation, given that the expression is also often used in its direct sense) and the other is figurative (categorised as an idiom).

### 3 Semantic Types of Phraseological Units and Representation in Tēzaurs.lv

Phraseological units are usually expected to comply with three fundamental criteria: they are fixed, consist of multiple words and possess some degree of figurative, transferred or metaphorical meaning (Veisbergs, 2019: 114). In Latvian linguistics, the hyperonymic term ‘phraseological unit’ encompasses both phrasemes and idioms (Laua, 1992; Skujiņa, 2007), thus the term ‘idiom’ is used in a narrower sense, as a sub-type of a phraseological

unit.

Semantically, phrasemes are partially compositional and transparent, as one of their components functions in its direct, literal sense, e.g., *domu grauds* ‘a grain of thought’, where *doma* ‘thought’ is used in the sense ‘the result of thinking’, whereas the other component of the phraseme, *grauds* ‘grain’ in itself does not represent the specific meaning realized in the phraseme. A similar example is *caurs miegs* lit. ‘leaky sleep’ (fitful, poor sleep), where *miegs* ‘sleep’ is used in its basic sense, whereas *caurs* ‘leaky’ acquires the meaning of fitful or poor only in this expression and is not used in the same way in any other distribution. It is generally important that one of the components of a phraseme is used in a literal sense (which can be either basic or secondary) while the other component draws its specific semantic value exclusively from the corresponding MWE.

The meanings of idioms, in turn, are non-transparent, e.g., *karāties mata galā* lit. ‘to hang by a thread of hair’ (to be in a precarious situation), *kārt zobus vadzī* lit. ‘to hang one’s teeth on a wedge’ (to starve). This means that idioms cannot be worked out by the usual semantic rules (Pulman, 1993: 260).

This distinction is represented in the Tēzaur.lv entries as well: the phrasemes are linked to the corresponding, literal senses of the used words, e.g. *slinkuma maiss* lit. ‘a bag of laziness’ (a lazy person) is linked to the basic sense of the word *slinkums* ‘laziness’. The same phraseme is also linked to the entry *maiss* ‘bag’ as a whole (and not to any specific sense) as the word *maiss* does not list a meaning of ‘person’.

Unlike phrasemes, idioms should be linked to entries as a whole (and not separate word senses), e.g. the idiom *cieta galva* lit. ‘a hard head’ has two meanings: 1) difficulty learning, remembering, and 2) a stubborn, rebellious character; this idiom is linked to both entries, *galva* ‘head’, and *ciets* ‘hard’.

#### 4 Borderline Cases of Phraseological Unit Classification

To some extent, the separation of phrasemes and idioms is linked to the notion of idiom decomposability mentioned in linguistic literature (Sag et al., 2002: 5) which demonstrates how the overall sense of a given idiom is related to its parts. Although we use a similar approach, it does not easily pro-

vide a simple and indisputable division into categories, since phraseological units are very diverse both formally and semantically. One could agree with the view that MWEs have varying degrees of semantic transparency and should be described with reference to a semantic scale ranging from totally transparent in meaning to completely opaque (Parra Escartín et al., 2013: 346). However, there is no consensus on how many intermediate sections and corresponding types would exist on such a scale.

In this study, problems arose when a phraseological unit is decomposable in principle – each word meaning can be discerned – but some of them are used figuratively. For example, in the expression *aizlaist vējā* lit. ‘to let loose in the wind’, to squander (classified as an idiom), the locative *vējā* ‘wind’ has a listed figurative sense ‘a way in which (something) disappears, ceases to exist’, so it can be used in different distributions, whereas the meaning *aizlaist* ‘let loose’ is used in its literal sense: ‘to let something go by acting passively’. This expression cannot be classified as a phraseme since other components of a phraseme acquire figurative meanings only in that specific combination. In this expression, both components retain their own meanings – literal for one and figurative for the other – therefore it is classified as an idiom. In such cases, there are two potential solutions: to define subtypes for idioms, or to introduce a third group of phraseological units that is neither a phraseme nor an idiom.

Even though the degrees of semantic transparency and semantic types of phraseological units are still under study, from the perspective of data processing, separating phraseological units from other MWE groups and dividing them into at least two subtypes provides significant benefits, since this data will be available for further research as a separate group.

#### 5 MWEs Not Included in The Existing Classification

A small part (7.5%) of the existing MWEs within Tēzaur.lv have not been categorized yet. This is either because they cannot be assigned to any of the existing MWE categories, or because some entries have been listed as MWEs by mistake. The classification of these MWEs will be addressed in future work.

Firstly, there are naming units that are difficult to fit into any of the current categories, such as mytho-

logical entities (*Meža māte* lit. ‘Forest Mother’), names of dances and games (*vistiņu ķeršana* ‘tag, catchers’), old names for months (*lapu mēnesis* lit. ‘leaf month’, May), names for fingers (*garais Ancis* lit. ‘Long Ancis’, middle finger), etc.

Secondly, there are names that contain nomenclature words, for example, *ātrvilciens Eurostar* ‘high-speed train Eurostar’, *operētājsistēma UNIX* ‘operating system UNIX’. Based on Tēzaurs.lv principles, such entries should not count as MWEs and the lemma should only consist of the proper name.

Furthermore, in many cases, the names of food dishes have not been classified at the moment. This thematically and semantically varied group has been set aside for future research and testing of more fine-grained classification, since they often belong to one or more overlapping categories. For instance, certain dish names can be idioms and food technology terms (*viltotais zaķis*, lit. ‘mock rabbit’, meatloaf), idioms but not terms (*ērzeļa pauti* lit. ‘stallion’s testicles’, deep-fried balls of batter), terms and phrasemes (*smilšu mīkla*, lit. ‘sand dough’, shortcrust pastry), phrasemes but not terms (*aklā putra* lit. ‘blind porridge’, porridge with no fat), as well as collocations that can either be terms (*rauga mīkla*, yeast dough) or not (*balta putra* lit. ‘white porridge’, milk porridge).

## 6 Conclusions and Future Work

Firstly, extensive work has been carried out to sort various MWEs into distinct categories, during which it was concluded that the existing system of classification does not cover all types of MWEs in Tēzaurs.lv; there are certain groups (e.g., abbreviations, certain naming units and dish names) that remain unsorted. This, in turn, shows the need for additional MWE categories. The results of this work are integrated in the relevant entries of the dictionary and are accessible to all its users.

Secondly, certain borderline cases between different MWE categories were observed. A part of these cases stems from the fact that figurativeness is also used in term creation, and currently they are sorted in the category of terms. Other borderline cases arise when frequently used figurative senses gradually become literal and thus cause difficulties to distinguish phraseological units from collocations that do not contain figurative meanings.

Future work includes combining MWE variants in one entry and the continued analysis of morphosyntactic and lexical variations of phraseolog-

ical units, e.g., the expression *Kā putns gaisā* lit. ‘like a bird in air’ can vary as *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’. All variants have the same syntactic structure and meaning (to be without obligations, worries or cares). Determining variants is also related to distinguishing between the fixed components of a phraseological unit and its characteristic environment, which is not a part of the unit itself. For example, the phraseological unit *gaisā tās, ka cirvi var pakārt* lit. ‘(one) could hang an axe in this air’ is a phraseme, but in certain environments it can appear simply as *cirvi var pakārt* lit. ‘(one) could hang an axe here’, (a feeling of stuffiness indoors). Thus, a phraseme can be reduced and subsequently become an idiom. After collecting such variants, we will test the hypothesis of whether phrasemes are lexically and syntactically more flexible than idioms.

## Acknowledgments

This work was supported by the Latvian Council of Science project “Advancing Latvian computational lexical resources for natural language understanding and generation” (LZP2022/1-0443) in synergy with the State Research Programme project LATE (VPP-LETONIKA-2021/1-0006). We also thank the anonymous reviewers for their input in improving this paper.

## References

- Laurie Bauer. 2021. *An Introduction to English Lexicology*. Edinburgh University Press, Edinburgh.
- Daiga Deksnē, Raivis Skadiņš, and Inguna Skadiņa. 2008. *Dictionary of multiword expressions for translation into highly inflected languages*. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08)*, Marrakech, Morocco. European Language Resources Association (ELRA).
- 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.
- Alīse Laua. 1992. *Latviešu valodas frazeoloģija*. Zvaigzne, Rīga.
- Svetlozara Leseva, Verginica Barbu Mititelu, and Ivelina Stoyanova. 2020. *It takes two to tango – towards a*

- multilingual MWE resource. In *Proceedings of the 4th International Conference on Computational Linguistics in Bulgaria (CLIB 2020)*, pages 101–111, Sofia, Bulgaria. Department of Computational Linguistics, IBL – BAS.
- Justina Mandravickaitė and Tomas Krilavičius. 2017. Identification of multiword expressions for Latvian and Lithuanian: Hybrid approach. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 97–101, Valencia, Spain. Association for Computational Linguistics.
- Carla Parra Escartín, Gyri Smørdal Losnegaard, Gunn Inger Lyse Samdal, and Pedro Patiño García. 2013. Representing multiword expressions in lexical and terminological resources: An analysis for natural language processing purposes. In *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia.*, pages 338–357, Ljubljana/Tallinn. Institute for Applied Slovene Studies/Eesti Keele Instituut.
- Stephen G. Pulman. 1993. The recognition and interpretation of idioms. In C. Cacciari and P. Tabossi, editors, *Idioms: Processing, Structure, and Interpretation*, Laurence Erlbaum Cognitive Science Monographs, pages 249–270. Lawrence Erlbaum Associates, New Jersey.
- Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing*, pages 1–15, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Inguna Skadiņa. 2016. Multi-word expressions in english-latvian machine translation. *Baltic Journal of Modern Computing*, 4(4).
- Inguna Skadiņa. 2018. Looking for a needle in a haystack: Semi-automatic creation of a latvian multiword dictionary from small monolingual corpora. In *Proceedings of the 18th EURALEX International Congress*, pages 255–265.
- Valentīna Skujiņa, editor. 2007. *Valodniecības pamattermiņu skaidrojošā vārdnīca*. Valsts valodas agentūra, Rīga.
- Andrejs Veisbergs. 2019. The fuzzy concept of idiom and what it might mean for bilingual dictionaries. *Baltic Journal of English Language, Literature and Culture*, 9:111–129.

# Complex Word Identification for Italian Language: a dictionary-based approach

Laura Occhipinti

University of Bologna, Italy

`laura.occhipinti3@unibo.it`

## Abstract

Assessing word complexity in Italian poses significant challenges, particularly due to the absence of a standardized dataset. This study introduces the first automatic model designed to identify word complexity for native Italian speakers. A dictionary of simple and complex words was constructed, and various configurations of linguistic features were explored to find the best statistical classifier based on Random Forest algorithm. Considering the probabilities of a word to belong to a class, a comparison between the models' predictions and human assessments derived from a dataset annotated for complexity perception was made. Finally, the degree of accord between the model predictions and the human inter-annotator agreement was analyzed using Spearman correlation. Our findings indicate that a model incorporating both linguistic features and word embeddings performed better than other simpler models, also showing a value of correlation with the human judgements similar to the inter-annotator agreement. This study demonstrates the feasibility of an automatic system for detecting complexity in the Italian language with good performances and comparable effectiveness to humans in this subjective task.

**Keywords:** complex word identification, Italian language, lexical complexity.

## 1 Introduction

Identifying the complexity of a word is a very challenging process that requires a series of linguistic reflections intertwined with the concept of complexity itself (Pallotti, 2015). While humans can intuitively perceive word simplicity, translating this intuition into quantitative parameters for automatic systems is challenging.

The task of Complex Word Identification (CWI) aims to pinpoint those words that may pose decoding challenges for certain readers due to a variety of linguistic features (Shardlow, 2013). The

concept of linguistic complexity indeed is closely intertwined with the readability and accessibility of texts (Chen and Meurers, 2019). Recognizing complex words is crucial, not only for readers with learning difficulties, such as dyslexia or aphasia (Stajner, 2021; De Hertog and Tack, 2018), but also for native speakers, since understanding word meanings is fundamental for comprehension (Carroll et al., 1998). Studies related to CWI have seen a significant increase in recent years, either as a part of lexical simplification systems (Saggion and Hirst, 2017), or as an independent task, promoted by several shared tasks (Paetzold and Specia, 2016; Yimam et al., 2018; Shardlow et al., 2021). In the latter case, it is very useful for the development of systems aiming at facilitating foreign language acquisition, creating reading tools for individuals with limited linguistic skills, and enhancing accessibility for native speakers (Gooding and Kochmar, 2018, 2019). Despite the importance of CWI, the development of such systems has been limited to a few languages, mainly due to the scarcity of necessary linguistic resources and the high costs associated with their development (Štajner et al., 2022).

To the best of our knowledge, there have been no studies directly addressing the CWI in the Italian language, even though research has focused on text simplification (Brunato et al., 2022). The absence of requisite databases classifies Italian as a 'low-resource language' for this specific task. The main contribution of this article is to propose the first automatic system to identify lexical complexity specifically designed for native Italian speakers, motivated by educational concerns (ISTAT, 2021) and the need to understand perceived complexity under typical conditions. We created a dataset of individual lexical entries, labelled as simple or complex (3.1) and selected various linguistic features (3.2), through which a classifier system could be trained in a supervised setting (3.3). Our approach

is dictionary-based and context agnostic (Billami et al., 2018; Baeza-Yates et al., 2015). We considered the probability of each item belonging to a certain class as the prediction of word complexity. Finally, the system was validated against a dataset containing human judgements regarding the perceived complexity of selected words (3.3).

## 2 Related Work

Recent investigation in CWI focused on the development of statistical classifiers that can accurately assign lexical items to specific complexity classes based on labelled data (Paetzold and Specia, 2016; Yimam et al., 2018). Classification systems typically utilize feature-based approaches or neural networks with word embeddings to enhance prediction accuracy (Aroyehun et al., 2018). Most CWI studies classify word complexity in two primary ways: **binary classification**, labeling words as either simple or complex (0-1), and **continuous classification**, where words receive a complexity score on a continuum from very simple to very complex. In recent years, it has become more common to use Lexical complexity prediction name to refer to the latter (North et al., 2023). Among the statistical classifiers, Support Vector Machines, Decision Trees, Random Forests, Logistic Regression, and Recurrent Neural Networks have been prominently used (Yimam et al., 2018; Shardlow et al., 2021).

For the Italian language, the few studies concerning lexical simplification (Tonelli et al., 2016; Brunato et al., 2015) have overlooked this task deemed crucial for the proper execution of simplification (Shardlow, 2014). The words to be simplified were selected exclusively on the basis of the frequency parameter (Brunato et al., 2022) relying on *Nuovo Vocabolario di base* (De Mauro and Chiari, 2016), which is a fundamental lexicon for the Italian language, comprising approximately 7,000 selected words. This approach poses significant limitations, as words outside this vocabulary are often prematurely considered complex, and potential substitutions are restricted to those within the same lexicon. This approach does not take into account the nuanced and multifaceted nature of linguistic complexity, so relying on a single measure such as frequency can lead to oversimplification (Bott et al., 2012): frequency is strongly linked to the reference corpus used to calculate it.

## 3 Methods

In this study, we developed a binary classification system using a dataset of isolated words, created due to the absence of comprehensive resources for the CWI task. Recognizing that word complexity is intrinsically context-dependent and that complexity itself is a gradient, our approach was constrained by resource limitations. The creation of a large-scale dataset capable of training models with nuanced human judgments in context would require significant time and resources. Consequently, we opted for a more manageable solution by employing a word list, which is computationally 'small and easily tractable' (Kilgarriff et al., 2014: 124). This choice was motivated by its ability to provide representative data for our purpose. The decision to adopt binary classification reflects not only these practical constraints but also avoids the subjectivity inherent in gradual classifications without extensive contextual data. This pragmatic approach aims to establish a foundational methodology that can be expanded as more comprehensive data become available.

We selected various linguistic features to characterize the complexity of our items. From these features, our model learned to predict the complexity classification and the likelihood of a word belonging to a particular class. While it is recognized that the probability of a target word being classified as simple or complex does not directly predict its degree of complexity (North et al., 2023), the continuous probabilistic values generated by our model provide valuable insights into the nuanced nature of word complexity. These probabilistic values reflect the uncertainty inherent in the classification process: higher probabilities indicate a stronger likelihood that a word has intricate linguistic properties, whereas lower probabilities suggest simpler linguistic structures. These predicted values can then be compared with complexity assessments derived from human judgments, which serve as our gold standard. This methodology helps bridge the gap between objective classification and subjective perception, enhancing our understanding of lexical complexity.

### 3.1 Dataset

Recognizing the importance of context in the domain of complexity perception, our approach was limited by the absence of an available dataset for the CWI task. As a result, we opted to build a list

of words for the purpose of training an automatic complexity classification system. The word selection was not based on frequency parameter, but on a series of heuristics aimed at minimising personal bias in the selection of lexical items. Considering the challenges in defining complexity (Miestamo et al., 2008), we decided to classify as simple all words that should be known or learned by Italian L2 learners, as outlined in levels A1-B1 of the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001). The selection of these words was made by exploring various linguistic resources developed and/or used for teaching Italian to non-native speakers.

**Merlin Corpus** (Wisniewski et al., 2013; Boyd et al., 2014) This corpus is a linguistic resource aimed at exploring texts produced by L2 students for Italian, Czech, and German languages. The Italian section includes 813 texts, each associated with specific CEFR levels by professional evaluators and featuring metadata related to various linguistic levels. We specifically selected texts with metadata corresponding to the 'Vocabulary Range' from levels A1 to B1. We extracted all forms and manually corrected any orthographic errors, recognizing that this corpus also serves as a representation of the errors made by the writers. Despite these graphical errors, we opted to include 'wrong' words as they are undoubtedly familiar to the writers, who employ them in their writing.

**Kelly** (Kokkinakis and Volodina, 2011) This resource was developed as part of the European Kelly project, aimed at creating vocabularies for nine languages, including Italian. The Kelly word list reflects modern usage and captures the core vocabulary of each language, selected through an objective process based on corpus analysis and pedagogical criteria. Words were categorized across levels based on daily themes deemed essential by the CEFR. This categorization guided their inclusion in our study due to their alignment with established language proficiency standards.

**ELI: vocabolario illustrato junior** (ELI Publishing Group, 2020) This dictionary is designed for a target audience of young students, presenting basic vocabulary ranging from levels A1 to A2, using graphical representations to link images with words effectively. It organizes 936 words into 45 themes relevant to everyday contexts. We chose this tool because it targets a beginner audience,

suggesting that the included words are widely recognized within the native speaking population.

**Word lists identified by University for Foreigners of Perugia**<sup>1</sup> The University for Foreigners of Perugia offers a range of open-access resources essential for teaching Italian to foreign learners. We focused on the section relating to the lexical lists for each level from A1 to B2, developed through extensive validation by linguistic and pedagogical experts. After downloading these lists, we removed additional details such as word index numbers and grammatical descriptions.

To these resources, purely related to L2 teaching, we added

**Varless** (Burani et al., 2001; Barca et al., 2002) This resource includes a list of simple Italian nouns accompanied by various lexical and sub-lexical variables such as age of acquisition, familiarity, concreteness, and frequency metrics. These variables significantly affect how words are perceived by speakers, with early-acquired words being recognized and named more rapidly and accurately. We included this resource because the words are classified as simple based on their acquisition and familiarity profiles.

The integration of these resources involved the exclusion of common vocabulary, multi-word expressions (which are not within the scope of this paper), and the normalization of word forms to their respective reference lemma. This process yielded a consolidated list comprising 5,382 lemmas.

It was not feasible to apply the same criterion in selecting complex words, as digital resources available for levels B2-C2 are limited and primarily focus on pragmatic aspects of the language. Therefore, for complex words, a dictionary containing words defined as difficult or truly difficult in the Italian language was utilized.

**Dizionario delle parole difficili e difficilissime** (Vallardi, 2016) This dictionary comprises Italian words that are arcane, remote, or enigmatic, and seldom used in colloquial, television, or journalistic contexts. It spans various domains such as literature, science, and technology, serving as a repository of linguistic richness and cultural heritage. From its approximately 13,000 lemmas, we carefully selected about 8,000 terms for our dataset to ensure sample balance and integrity.

<sup>1</sup><https://www.unistrapg.it>.

The final dataset consists of a list of words labelled as simple (0) or complex (1), comprising 13,319 lemmas distributed between the two categories as follows: 5,382 simple lemmas; 7,937 complex lemmas.

### 3.2 Features

Defining the linguistic features for identifying lexical complexity is critical, involving several interrelated aspects (Collins-Thompson, 2014). The selection of features is based on their strong psycholinguistic evidence, which significantly impacts the perception of complexity. These features are calculated using both the word form and its lemma, with lemmatization performed using the Italian SpaCy model<sup>2</sup>. Given their general applicability and the robust psycholinguistic backing, these measures are particularly suited for our target population of native Italian speakers.

**Frequency** Frequency appears to be the predominant and essential parameter in all approaches to CWI, supported by various pieces of psycholinguistic evidence (Segui et al., 1982). For instance, frequency is significant for gauging familiarity with a term. We used two reference corpora to calculate frequency, aiming to reduce bias from corpus composition. The first corpus we considered is the **ItWac corpus** (Baroni et al., 2009), that is a 2 billions word corpus, created from the web. The other is **Subtlex-it** (Crepaldi et al., 2015), a word frequency list based on movie and tv show subtitles for approximately 520,000 Italian word-forms. For both, we calculated the row frequency for each lemma, representing the number of occurrences within the corpus. The two frequencies are treated separately and the values were converted into base 10 logarithmic scales, returning 0 if before normalization frequency value was 0.

#### 3.2.1 Surface Features

We considered some surface linguistic parameters that are crucial from a psycholinguistic standpoint (Perfetti et al., 2001) because they significantly affect reading and decoding times:

**Word length** The number of characters in the word.

**Syllable count** The number of syllables of the word calculated using Pyphen<sup>3</sup>.

<sup>2</sup>[https://spacy.io/models/it#it\\_core\\_news\\_sm](https://spacy.io/models/it#it_core_news_sm).

<sup>3</sup><https://pyphen.org/>.

**Vowels count** The number of vowels presented in the word. This feature was determined by iterating through each character in the word and checking if it corresponds to any vowel, including accented characters. Notably, we included vowels typical of the Italian language in our analysis.

#### 3.2.2 Linguistic features

In addition to the superficial characteristics of words, it is necessary to carry out deeper analyses concerning the types of words and the meanings attached to them.

**Stop words** Recognizing whether a word is a stopword is crucial for determining its complexity. Stopwords, such as articles, prepositions, and conjunctions, are frequently encountered and widely understood by readers. Therefore, identifying whether a word is a stopword provides insights into its familiarity and ease of comprehension. This measure was computed using SpaCy.

**Number of senses** We assessed the number of senses for each lemma using the ItalWordNet (Roventini et al., 2000). This analysis helps clarify the semantic complexity of words by revealing how many different meanings a word can have, indicating its potential to cause decoding ambiguities for readers.

#### 3.2.3 Morphological Features

We selected features related to word morphology, crucial for defining lexical complexity. Most of the morphosyntactic information we have for Italian language from existing corpora or from readability measures concerns the class to which words belong. Beyond this, we incorporated details about internal structure of the word (Baerman et al., 2015).

**POS-tag** We categorized the lemma into predefined POS labels, assessing the presence or absence of each label using a list. The provided method iterates through the lemma, assigning a value of 1 to the corresponding POS label if matched, otherwise 0. Using SpaCy, we predicted the POS labels while consolidating certain subcategories into broader groups to simplify analysis. We merged ‘VERB’ and ‘AUX’ into a category ‘VERB’, ‘NOUN’ and ‘PROPN’ into ‘NOUN’, and ‘CCONJ’ and ‘SCONJ’ into ‘CONJ’.

**Number of morphemes** We calculated the number of morphemes, the smallest units of meaning,

that composed the word. In this way, we can provide indications about the amount of information readers must decode to understand the term they are facing (Brezina and Pallotti, 2019). Italian is an inflected language (Grandi, 2011) that employs inflection, derivation, and composition to modify words. The number and type of morphemes in a word are crucial indicators of its complexity; for instance, a derived word is more complex than a simple one, as it contains more elements to decode (Rastle and Davis, 2003).

**felice** is simpler than **infelice**

The adding of the prefix **in-** to the base form leads us to decode the meaning of **felice** (happy) to which a negation is added. For this reason, we could argue that the word *infelice* (unhappy) is marked compared to *felice* and increases its degree of complexity. To calculate the morphological composition of a word, we used a Convolutional Neural Model<sup>4</sup> trained on an Italian hand-checked dataset<sup>5</sup> to obtain an automatic morphological segmentation.

**Morphological Density** This measure quantifies morphological complexity at the word level (Sandra, 1994; Manova et al., 2020), defined as the ratio of the number of morphemes to word length. It helps analyze a word’s structural complexity, indicating how densely packed it is with meaningful units. A higher morphological density suggests a more complex word, with many meaningful units condensed into a shorter length, possibly making it more challenging to comprehend. Lower density, conversely, implies simpler and potentially easier-to-understand words.

**Frequency of lexical morpheme** We determined the frequency of the lexical morpheme that most conveys the meaning of the word (Amenta and Crepaldi, 2012). Employing our morphological segmentator on the ItWac corpus, enabled us to dissect the word into segments and aggregate the frequencies of individual morphemes. The use of lexical morpheme frequency as a complexity indicator is based on the idea that even if a word is unfamiliar as a whole, its component morphemes may be common in the language and more recognizable

<sup>4</sup><https://github.com/AlexeySorokin/NeuralMorphemeSegmentation/tree/master>.

<sup>5</sup>The details of the implementation of this system and the database used will be discussed in a forthcoming paper currently in preparation.

(Colé et al., 1997). Such words are inherently more relatable to familiar concepts due to the frequent occurrence of their constituent morphemes. Leveraging the familiarity of these morphemes enhances the transparency and interpretability of the word’s meaning. We adopted the longest splitting morpheme as the lexical one, as this heuristic aligns with many cases in Italian, acknowledging that there are exceptions to this rule. Additionally, the frequency values have been logarithmized to facilitate analysis.

**Word Embedding** We utilized pre-trained word embeddings from FastText for Italian (Joulin et al., 2016, 2017), which provides word vector representations with 300 dimensions. The model used in our study was trained on Wikipedia and Common Crawl datasets<sup>6</sup>. These embeddings provided vector representations for each word in our dataset, primarily comprising isolated items, allowing us to incorporate contextual features into our analysis.

### 3.3 Models

We evaluated the performance of a classifier built using Random Forest (Breiman, 2001) implemented with the `scikit-learn` library (Pedregosa et al., 2011). Specifically, we utilized the `RandomForestClassifier` module provided by `scikit-learn`<sup>7</sup>. To assess the classifier’s performance, we employed 12-fold cross-validation over the training data. We selected different configuration of features to understand which model is the best in prediction<sup>8</sup> after training and to compare that only one frequency value is not enough for an efficient prediction:

1. Frequency model, that utilizes the two parameters related to Frequency in 3.2.3.
2. Feature-based Model, that leverages the eleven linguistic features discussed above (frequencies, surface, linguistic and morphological features presented in 3.2.3).
3. Embedding Model, that utilizes only pre-trained word embeddings (Word embedding paragraph in 3.2.3).

<sup>6</sup><https://fasttext.cc/docs/en/crawl-vectors.html>.

<sup>7</sup><https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>.

<sup>8</sup>For further practical details concerning the best performing model, the source code, and the resources used, interested parties are encouraged to directly contact the author via email.

4. Total model, that integrates both feature-based and embedding-based features.

These models were trained and evaluated on the dataset presented in Section 3.1, consisting of 13,319 words. To establish a robust evaluation, we employed the `train_test_split` function from `scikit-learn` to partition the dataset into training and testing sets. The split allocated 70% of the data for training, amounting to approximately 9,000 words, while the remaining 30% (about 4,000 words) was reserved for testing. We shuffled the data before splitting to mitigate any bias, and subsequently instantiated a Random Forest Classifier model with the random state set to 42 for reproducibility. For performance evaluation, each model underwent rigorous assessment on four key parameters: Accuracy, Precision, Recall, and F1 Score on the test set. These metrics are commonly employed to evaluate the performance of classification systems (North et al., 2023) (results in Section 4.1).

### 3.4 External validation

After validating our models on the original test set, we extended our evaluation by testing the models on an external resource. This dataset<sup>9</sup> consists of 600 sentences, in each of which a target word was identified. For each word, we gathered a minimum of 10 human judgements regarding the complexity level of the target for a generic native speaker of Italian. The data were annotated exclusively by native speakers, that had the task of assigning a level of complexity to each target word, using a Likert scale ranging from 1 to 5:

- 1: very easy - Words which are very familiar
- 2: easy - Words which are mostly familiar
- 3: neutral - When the word is neither difficult or easy
- 4: difficult - Words which you are unclear of the meaning, but may be able to infer from the context
- 5: very difficult - Words that are very unclear.

This dataset was built as a resource of lexical complexity prediction; for information on how the

<sup>9</sup>[https://github.com/MLSP2024/MLSP\\_Data/](https://github.com/MLSP2024/MLSP_Data/).

dataset was constructed and annotated, please refer to (Shardlow et al., 2024). This resource represents our gold standard. For each target word, the average score between annotations was used as a single human-derived complexity value that was compared with our models predictions. We transformed the values from a range of 1 to 5 to a scale of 0 to 1 using the min-max normalization (Abdi, 2007). This normalisation aligns the data with our model’s output range (between 0 and 1), facilitating effective analysis and consistent evaluation of model performance. The validation metrics normally used to evaluate lexical prediction system performance (North et al., 2023) are Pearson Correlation, Spearman’s Rank, mean absolute error, and mean squared error (Hastie et al., 2009). We calculated these measures evaluating the relations between our predictive model outputs (we excluded the model with the lowest performance) and the aggregated human judgements (results in Section 4.2). We conducted a further analysis by comparing the predictions of the best model with the level of inter-annotator agreement (Artstein, 2017) observed in our resource. While the initial comparison provided valuable insights into the model’s performance against a consolidated human judgement, assessing its agreement with multiple human annotations offers a more comprehensive understanding of its effectiveness. The choice of agreement measure depends on the data nature and the objectives of the study. Since our data are ordinal, with complexity values ranging from 1 to 5, we chose to use Spearman correlation to calculate agreement and not kappa (Rau and Shih, 2021), which is more suited for nominal or categorical data. The Spearman correlation is suitable for ordinal variables as it accounts for the rank order of values without assuming a linear relationship, offering greater flexibility in measuring agreement. Furthermore, it is particularly adequate in cases where the order of the values is significant, but no specific assumptions can be made about the distribution of the data or the uniform intervals between the categories. Our annotation task involved ordinal ratings, where the magnitude of difference between ratings carries significance, thus making this measure a more appropriate choice for assessing agreement. The Spearman correlation coefficients were calculated using the `spearmanr` function from the `scipy.stats` module by iterating through combinations of annotator pairs. After calculat-

| Model           | Accuracy      | Precision     | Recall        | F1            |
|-----------------|---------------|---------------|---------------|---------------|
| Frequency_based | 0.8826        | 0.8995        | 0.9214        | 0.9103        |
| Feature_based   | 0.9006        | 0.9137        | 0.9346        | 0.9243        |
| Embedding_model | 0.8943        | 0.9055        | 0.9289        | 0.9170        |
| Total_model     | <b>0.9149</b> | <b>0.9237</b> | <b>0.9466</b> | <b>0.9350</b> |

Table 1: Classifier results

ing the correlation coefficients, we computed the overall average correlation coefficient across the entire dataset. We operated in the same way with the results of our best model. We also treated our model’s predictions as an additional annotator to calculate its agreement with all human judgments. The final value is the result of the average of correlation values of our model with all the single value of complexity defined by annotators. The comparison between the two values is reported in Section (4.3).

## 4 Results and discussion

### 4.1 Model results on classification

The calculated performances of the four models on the test set are reported in Table 1. The Total\_Model showed the best performances across all validation metrics, outperforming not only the simplest model based on frequency but also the more complex Feature-based and Embedding-based models. The Frequency-based model, even if inferior to the others, still demonstrated acceptable performance, thus highlighting the key significance of frequency. However, a model exclusively based on frequency has a key limitation: words that are not represented in the corpus considered will be labelled as complex with a probability of around 0.99. Thus it is essential to include more words classification features in the training of the model. The Feature-based Model and the Embedding Model exhibit comparable performance across various evaluation metrics. However, it is crucial to recognize the underlying differences in their methodologies and interpretability. The Feature-Based Model provides a transparent framework enabling granular analysis of the impact of individual features on prediction. This transparency facilitates the identification of specific features that contribute significantly to the model’s predictive performance. In contrast, the Embedding model operates on distributed representations of words in a highly dimensional semantic space, making it inherently more opaque and dif-

ficult to interpret. The Total\_Model, thanks to its adeptness in harnessing the respective strengths of each method, shows superior performance due to its capacity to leverage not only the linguistic features we selected but also the connected semantic representations in word embeddings.

### 4.2 Model results on complexity prediction

In Table 2, we reported the results of our models in comparison with the gold standard dataset, containing the human annotations. These results provide insights into the effectiveness of our models in predicting complexity, with the Total\_Model demonstrating again an overall superior performance compared to the others. Pearson’s Correlation evaluates the linear relationship between predicted and actual complexity values, indicating the strength of this relationship. For the Total\_model, the Pearson Correlation is 0.5503, demonstrating a relatively strong linear relationship between predicted and actual values. On the other hand, Spearman’s Rank assesses the monotonic relationship between predicted and actual complexity values, regardless of linearity. The Total\_model achieved a Spearman’s Rank of 0.5528, indicating a strong monotonic relationship between predicted and actual values. The two correlation coefficients are quite similar, suggesting that the model performs well in capturing both linear and non-linear trends in complexity prediction. Mean Squared Error (MSE) measures the average squared difference between predicted and actual complexity values, with lower values indicating better performance. For this parameter the Embedding\_Model shows a slightly lower value than the Total\_Model. Similarly, Mean Absolute Error (MAE) calculates the average absolute difference between predicted and actual complexity values. The Total\_model achieved a MAE of 0.2393, suggesting that its predictions are closer to the true complexity values. Despite the slightly lower MSE for the Embedding\_model, the Total\_model still demonstrates superior performance overall, as evidenced by its higher correlation coefficients and

| Model           | Pearson Correlation | Spearman's    | Mean squared error | Mean absolute error |
|-----------------|---------------------|---------------|--------------------|---------------------|
| Feature_based   | 0.5331              | 0.5403        | 0.1231             | 0.2718              |
| Embedding_model | 0.4762              | 0.4752        | <b>0.0927</b>      | 0.2482              |
| Total_model     | <b>0.5503</b>       | <b>0.5528</b> | 0.0965             | <b>0.2393</b>       |

Table 2: Results of models on complexity prediction

| System agreement           | Spearman's Rank |
|----------------------------|-----------------|
| Inter-annotation agreement | <b>0.4196</b>   |
| Total_Model agreement      | 0.4145          |

Table 3: Comparison between inter-annotator agreement and model predictions

lower error metrics compared to the other models.

### 4.3 Comparison with inter-annotator agreement

The comparison between the Spearman correlation coefficients obtained from the assessments of human annotators and those derived from the predictions of our best model reveals a notable similarity. The results are reported in Table 3. Both values, falling within the same range, demonstrate a significant degree of agreement between the model's predictions and human evaluations. The close proximity of these figures underscores the model's proficiency in capturing the complexity assessed by humans. These findings imply that there are opportunities for improvement both within our system and in fostering increased inter-agreement among human annotators, thereby potentially refining the model's ability to accurately capture the complexities inherent in the task.

## 5 Conclusions

In this study, we introduced the first system aimed at identifying complex words within the Italian language, marking the initial exploration of this task for this linguistic domain.

The absence of specific datasets prompted us to build a dictionary comprising approximately 13,000 words annotated for simplicity and complexity. An appropriate selection of descriptive features of word complexity made it possible to train a classification model in different configurations. We tested our models on a test set and on an external dataset, containing human judgements on word complexity, our gold standard. From the different validation analyses we saw that the best

model is Total\_Model that integrates the linguistic features with the word embedding.

We conducted a further analysis by comparing the Total\_Model to our gold standard. This dataset was annotated by multiple native Italian speakers and for this reason we decided to calculate the inter-annotation agreement and compared it with model-human correlation. In this way we not only validated the reliability of our dataset and the fidelity of our predictive model but also established the basis for a meaningful comparison between human and machine assessments.

Our analysis revealed that the average correlation of each predicted value from our model with the inter-annotator agreement falls within the same range, suggesting that our system is as effective as human judgment in subjective tasks such as this. To enhance inter-annotator agreement and the robustness of our findings, future efforts will focus on increasing the sample size and the number of annotators. Expanding the sample size will cover a broader lexical domain and provide a diverse set of words and contexts, thereby improving the model's generalizability to unseen data. This broader coverage supports robust statistical testing and validation, minimizing the influence of outliers. Incorporating more annotators is crucial for enriching the diversity of perspectives in the evaluation process, which is particularly important in subjective assessments where personal experiences, linguistic backgrounds, and individual biases might skew judgments. A larger pool of annotators diminishes these biases, fostering a balanced and representative consensus on lexical complexity. Furthermore, this approach allows for more detailed inter-annotator agreement analyses, clearly highlighting areas of consensus and disagreement. Together, these strategies not only enhance the reliability of our annotations but also improve the overall accuracy and applicability of our model.

The main limitation of our approach resides in the characteristics of the dataset we used to train our models. Our dataset is composed by words presented in isolation, thus disregarding crucial

contextual cues essential for understanding word meanings and disambiguation. We acknowledge the critical role of context in complexity analysis and recognize the necessity of incorporating specific contextual information where the target word appears. Moving forward, our aim is to advance in this direction by expanding effective datasets that integrate contextual frameworks for training our word CWI systems. Although challenges persist in the field of CWI, our study lays some groundwork for exploring this task for Italian and underscores the potential of automated systems in this domain.

In the future, collaborative efforts and advancements in building datasets and refining models will be crucial for advancing the field and uncovering new insights into language complexity. This methodology could enhance the precision of readability measures (Dell’Orletta et al., 2011), particularly in terms of lexical range and lexical sophistication. Moreover, such a system can be an essential component in a text simplification pipeline. By identifying words that may pose comprehension challenges, the system not only flags these words for potential replacement but also assists in suggesting simpler alternatives. This functionality ensures that the replacements not only match the original words’ meanings as closely as possible but also contribute to a text that is overall easier to understand.

While this study focuses on the Italian language, the methodologies and models we have developed have the potential to be adapted for other languages, especially those considered low-resource in the context of computational linguistic tools. By leveraging similar linguistic resources and adjusting the feature sets to accommodate language-specific characteristics, researchers can extend this approach to support complex word identification across diverse linguistic domains.

## References

- Herve Abdi. 2007. Multiple correlation coefficient. *Encyclopedia of measurement and statistics*, 648(651):19.
- Simona Amenta and Davide Crepaldi. 2012. Morphological processing as we know it: An analytical review of morphological effects in visual word identification. *Frontiers in psychology*, 3:232.
- Segun Taofeek Aroyehun, Jason Angel, Daniel Alejandro Pérez Alvarez, and Alexander Gelbukh. 2018. Complex word identification: Convolutional neural network vs. feature engineering. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 322–327.
- Ron Artstein. 2017. Inter-annotator agreement. *Handbook of linguistic annotation*, pages 297–313.
- Matthew Baerman, Dunstan Brown, and Greville Corbett. 2015. *Understanding and measuring morphological complexity*. Oxford University Press, USA.
- Ricardo Baeza-Yates, Luz Rello, and Julia Dembowski. 2015. **CASSA: A context-aware synonym simplification algorithm**. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1380–1385, Denver, Colorado. Association for Computational Linguistics.
- Laura Barca, Cristina Burani, and Lisa S Arduino. 2002. Word naming times and psycholinguistic norms for Italian nouns. *Behavior research methods, instruments, & computers*, 34:424–434.
- Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43:209–226.
- Mokhtar Billami, Thomas François, and Núria Gala. 2018. **ReSyf: a French lexicon with ranked synonyms**. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2570–2581, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Stefan Bott, Luz Rello, Biljana Drndarević, and Horacio Saggion. 2012. Can Spanish be simpler? lexis: Lexical simplification for Spanish. In *Proceedings of COLING 2012*, pages 357–374.
- Adriane Boyd, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schöne, Barbora Stindlová, and Chiara Vettori. 2014. The merlin corpus: Learner language and the CEFR. In *LREC*, pages 1281–1288. Reykjavik, Iceland.
- Leo Breiman. 2001. Random forests. *Machine learning*, 45:5–32.
- Vaclav Brezina and Gabriele Pallotti. 2019. Morphological complexity in written L2 texts. *Second language research*, 35(1):99–119.
- Dominique Brunato, Felice Dell’Orletta, and Giulia Venturi. 2022. Linguistically-based comparison of different approaches to building corpora for text simplification: A case study on Italian. *Frontiers in Psychology*, 13:707630.
- Dominique Brunato, Felice Dell’Orletta, Giulia Venturi, and Simonetta Montemagni. 2015. Design and annotation of the first Italian corpus for text simplification. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 31–41.

- Cristina Burani, Laura Barca, and Lisa Saskia Arduino. 2001. Una base di dati sui valori di età di acquisizione, frequenza, familiarità, immaginabilità, concretezza, e altre variabili lessicali e sublessicali per 626 nomi dell'italiano. *Giornale Italiano di Psicologia*, 28(4):839–856.
- John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. 1998. Practical simplification of english newspaper text to assist aphasic readers. In *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, pages 7–10. Association for the Advancement of Artificial Intelligence.
- Xiaobin Chen and Detmar Meurers. 2019. Linking text readability and learner proficiency using linguistic complexity feature vector distance. *Computer Assisted Language Learning*, 32(4):418–447.
- Pascale Colé, Juan Segui, and Marcus Taft. 1997. Words and morphemes as units for lexical access. *Journal of Memory and Language*, 37(3):312–330.
- Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*, 165(2):97–135.
- Council of Europe. 2001. *Common of European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.
- Davide Crepaldi, Simona Amenta, Mandra Pawel, Emmanuel Keuleers, and Marc Brysbaert. 2015. Subtlex-it. subtitle-based word frequency estimates for italian. In *Proceedings of the Annual Meeting of the Italian Association For Experimental Psychology*, pages 10–12.
- Dirk De Hertog and Anaïs Tack. 2018. Deep learning architecture for complex word identification. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 328–334.
- Tullio De Mauro and I Chiari. 2016. Il nuovo vocabolario di base della lingua italiana. *Internazionale*. [28/11/2020]. <https://www.internazionale.it/opinione/tullio-de-mauro/2016/12/23/il-nuovo-vocabolario-di-base-della-lingua-italiana>.
- 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.
- ELI Publishing Group. 2020. *ELI Picture Dictionary Junior: Picture Dictionary Junior - Italian*. ELI Publishing.
- Sian Gooding and Ekaterina Kochmar. 2018. Camb at cwi shared task 2018: Complex word identification with ensemble-based voting. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 184–194.
- Sian Gooding and Ekaterina Kochmar. 2019. Complex word identification as a sequence labelling task. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1148–1153.
- Nicola Grandi. 2011. Evaluative affixes between inflection and derivation: a typological survey. In *Societas Linguistica Europaea—44th Annual Meeting*.
- Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.
- ISTAT. 2021. **Livelli di istruzione e ritorni occupazionali**. Technical report, Istituto nazionale di Statistica.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Herve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. **Bag of tricks for efficient text classification**. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431.
- Adam Kilgarriff, Frieda Charalabopoulou, Maria Gavrilidou, Janne Bondi Johannessen, Saussan Khalil, Sofie Johansson Kokkinakis, Robert Lew, Serge Sharoff, Ravikiran Vadlapudi, and Elena Volodina. 2014. Corpus-based vocabulary lists for language learners for nine languages. *Language resources and evaluation*, 48:121–163.
- Sofie Johansson Kokkinakis and Elena Volodina. 2011. Corpus-based approaches for the creation of a frequency based vocabulary list in the eu project kelly—issues on reliability, validity and coverage. *Proceedings of eLex*, 2011:129–139.
- Stela Manova, Harald Hammarström, Itamar Kastner, and Yining Nie. 2020. What is in a morpheme? theoretical, experimental and computational approaches to the relation of meaning and form in morphology. *Word Structure*, 13(1):1–21.
- Matti Miestamo, Fred Karlsson, and Kaius Sinnemäki. 2008. Language complexity. *Language Complexity*, pages 1–374.
- Kai North, Marcos Zampieri, and Matthew Shardlow. 2023. Lexical complexity prediction: An overview. *ACM Computing Surveys*, 55(9):1–42.
- Gustavo Paetzold and Lucia Specia. 2016. SemEval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569.

- Gabriele Pallotti. 2015. A simple view of linguistic complexity. *Second Language Research*, 31(1):117–134.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.
- Charles Perfetti, Julie Van Dyke, and Lesley Hart. 2001. The psycholinguistics of basic literacy. *Annual review of applied linguistics*, 21:127–149.
- Kathleen Rastle and Matthew Davis. 2003. Reading morphologically complex words. *Masked priming: The state of the art*, pages 279–305.
- Gerald Rau and Yu-Shan Shih. 2021. Evaluation of Cohen’s kappa and other measures of inter-rater agreement for genre analysis and other nominal data. *Journal of english for academic purposes*, 53:101026.
- Adriana Roventini, Antonietta Alonge, Nicoletta Calzolari, Bernardo Magnini, and Francesca Bertagna. 2000. Italwordnet: a large semantic database for italian. In *LREC*.
- Horacio Saggion and Graeme Hirst. 2017. *Automatic text simplification*, volume 32. Springer.
- Dominiek Sandra. 1994. The morphology of the mental lexicon: Internal word structure viewed from a psycholinguistic perspective. *Language and cognitive processes*, 9(3):227–269.
- Juan Segui, Jacques Mehler, Uli Frauenfelder, and John Morton. 1982. The word frequency effect and lexical access. *Neuropsychologia*, 20(6):615–627.
- Matthew Shardlow. 2013. A comparison of techniques to automatically identify complex words. In *51st annual meeting of the association for computational linguistics proceedings of the student research workshop*, pages 103–109.
- Matthew Shardlow. 2014. [Out in the open: Finding and categorising errors in the lexical simplification pipeline](#). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, pages 1583–1590, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Matthew Shardlow, Fernando Alva-Manchego, Riza Theresa Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Huelsing, Laura Occhipinti, et al. 2024. An extensible massively multilingual lexical simplification pipeline dataset using the multils framework. In *Proceedings of the 3rd Workshop on Tools and Resources for People with READING Difficulties (READI)@ LREC-COLING 2024*, pages 38–46.
- Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, and Marcos Zampieri. 2021. [SemEval-2021 task 1: Lexical complexity prediction](#). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1–16, Online. Association for Computational Linguistics.
- Sanja Stajner. 2021. [Automatic text simplification for social good: Progress and challenges](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2637–2652, Online. Association for Computational Linguistics.
- Sanja Štajner, Daniel Ferrés, Matthew Shardlow, Kai North, Marcos Zampieri, and Horacio Saggion. 2022. Lexical simplification benchmarks for english, portuguese, and spanish. *Frontiers in Artificial Intelligence*, 5:991242.
- Sara Tonelli, Alessio Palmero Aprosio, and Francesca Saltori. 2016. Simpitiki: a simplification corpus for italian. *Proceedings of the Third Italian Conference on Computational Linguistics CLiC-it 2016*, pages 291–296.
- Vallardi. 2016. *Dizionario delle parole difficili e difficilissime*. Vallardi Editore.
- Katrin Wisniewski, Karin Schöne, Lionel Nicolas, Chiara Vettori, Adriane Boyd, Detmar Meurers, Andrea Abel, and Jirka Hana. 2013. Merlin: An online trilingual learner corpus empirically grounding the european reference levels in authentic learner data. In *ICT for Language Learning 2013, Conference Proceedings, Florence, Italy. Libreriauniversitaria. it Edizioni*.
- Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo H Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A report on the complex word identification shared task 2018. In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 66–78.

# Verbal Multiword Expressions in the Croatian Verb Lexicon

Ivana Brač

Institute for the Croatian  
language  
ibrac@ihjj.hr

Matea Birtić

Institute for the Croatian  
language  
mbirtic@ihjj.hr

## Abstract

The paper examines the complexities of encoding verbal multiword expressions in the Croatian verb lexicon. The lexicon incorporates a verb's description at the syntactic, morphological, and semantic levels. This study explores the treatment of reflexive verbs, light verb constructions, and verbal idioms across several Croatian and Slavic language resources to find the best solution for the verb lexicon. It addresses the following research questions: 1. How should reflexive verbs, i.e., verbs with the reflexive marker *se*, be treated? Should they be considered as separate lemmas, sublemmas of non-reflexive counterparts, or as one of their senses? 2. What syntactic label and semantic role should be assigned to a predicative noun in light verb constructions? 3. Should verbal idioms be included, and, if so, at which level of a description? Our conclusion is that all reflexive verbs should be treated as separate lemmas since they are distinct lexemes that have undergone semantic and syntactic change. To differentiate between a semantically full verb and a light verb, we have introduced the label LV and decided not to assign a semantic role to a predicative noun. By including verbal idioms and their translation into English, non-native users can benefit from the lexicon. The aim is to enhance the verb lexicon for the more effective description and recognition of verbal multiword expressions.

**Keywords:** reflexive verbs, light verbs constructions, verbal idioms, verb lexicon.

## 1 Introduction

Multiword expressions (MWEs) in running text pose challenges in natural language processing (e.g., Sag et al., 2002; Constant and Nivre, 2016; Savary et al., 2017; Osenova and Simov, 2018), in

lexicographic resources (Koeva et al., 2016), and in theoretical syntactic and semantic research (e.g., Grimshaw and Mester, 1988; Butt, 2010). A verb lexicon containing the description of verbal multiword expressions (VMWEs) (see Kettnerová and Lopatková, 2013), such as the Croatian verb lexicon Verbion, can be useful for improving the accuracy and efficiency of processing and in understanding these expressions in various linguistic applications.

The verb lexicon Verbion contains a verb lemma accompanied by labels about aspect and reflexive uses, a morphological block with inflectional forms, and verb senses. Each sense is accompanied by a semantic class, a semantic frame from FrameNet (Baker et al., 2003), and one or more valency frames. Each valency frame includes an example from corpora which is analyzed at syntactic, morphological, and semantic levels. For the syntactic description, phrase type labels such as NP, PP, etc., are used, and for the semantic level, semantic roles mainly adopted from VerbNet (Kipper Schuler, 2005) and frame elements from FrameNet (Baker et al., 2003) are assigned. In the first phase of the project, a description of the 500 most frequent verbs in Croatian will be provided, and the results will be publicly available on the project's webpage

(<https://semtactic.jezik.hr/>). The questions that need to be answered at this stage of planning are: 1. How should reflexive verbs, i.e., verbs with the reflexive marker *se*, be treated? Should they be considered as separate lemmas, sublemmas of transitive or (rarely) intransitive verbs, or as one of the senses of their non-reflexive counterparts? 2. What syntactic label and semantic role should be assigned to a predicative noun in light verb constructions (LVCs)? 3. Should verbal idioms be included, and, if so, at which level of a description?

In this paper, we present the treatment of reflexive verbs, light verb constructions, and verbal

idioms in Croatian general language online dictionaries (*Hrvatski jezični portal* = Croatian Language Portal: <https://hjp.znanje.hr/>; the online version of *Hrvatski školski rječnik* = Croatian School Dictionary: <https://rjecnik.hr/>; *Hrvatski mrežni rječnik* = Croatian Web Dictionary – Mrežnik: <https://rjecnik.hr/mreznik/>) and online valency lexicons (CROVALLEX: <http://theta.ffzg.hr/crovallex/>; e-Glava: <http://valencije.ihjj.hr/>), as well as other Slavic valency lexicons with rich syntactic and semantic descriptions, such as the Czech VALLEX (Kettnerová and Lopatková, 2013, Kettnerová, 2023). In Section 2, the treatment of inherently reflexive verbs or reflexiva tantum, “proper” reflexive, derived reflexive, and reciprocal reflexive verbs is presented. Section 3 addresses the processing of light verb constructions. Section 4 presents the recording of verbal idioms, which is followed by the Conclusion. Our aim is to present solutions for the description of VMWEs at syntactic and semantic levels in the Croatian verb lexicon Verbion.

## 2 Reflexive verbs

There are many classifications of reflexive verbs made both for the Croatian language and crosslinguistically. As for Croatian, at one end are 1. inherently reflexive verbs or *reflexiva tantum*, which cannot appear without the reflexive marker *se* (e.g., *smijati se* ‘laugh’, *natjecati se* ‘compete’), and at the other end there are 2. “proper” (or true) reflexive verbs, which are basically transitive verbs whose object can be replaced by the reflexive pronoun *sebe* ‘oneself’ or its shorter variant *se* (e.g., *češljati se(be)* ‘comb oneself’).<sup>1</sup> There is a third distinct group: 3. reciprocal reflexive verbs (e.g., *ljubiti se* ‘kiss each other’). We could add a fourth group – 4. derived reflexives – a group that is between inherently reflexive verbs and “proper” reflexive verbs. They have transitive and intransitive counterparts, but the reflexive marker *se* cannot be replaced by the pronoun *sebe* (*igrati* ‘play (tran.)’ – *igrati se* ‘play (refl.)’).

<sup>1</sup> There are many discussions regarding the status of *se* in Croatian. Some authors consider it a pronoun (Barić et al., 1997; Raguž, 2010), others view it as a particle (Babić et al., 2007; Oraić Rabušić, 2018), while some argue that it is a particle with reflexiva tantum and derived reflexive verbs, and a pronoun with “proper” reflexive verbs (Silić and

Since there are many yet unaligned approaches, we investigated the treatment of these four groups of verbs in the resources mentioned in the Introduction. *Mrežnik* (Hudeček and Mihaljević, 2020) and *Hrvatski školski rječnik* (ŠR) (Birtić et al., 2012) list inherently reflexive verbs as separate lemmas (e.g., *čuditi se* ‘wonder’). “Proper” or syntactically reflexive verbs are also listed with the reflexive marker, but it is placed in parentheses (e.g., *kupati (se)* ‘bathe (oneself)’). If the meaning of a transitive verb and its reflexive variant which belongs to “proper” reflexive verbs is the same with a different object reference, the reflexive marker *se* in parentheses is written next to the headword. For example, *kupati (se)* is the main lemma which means ‘to wash somebody in the bathtub or container full of water’. If there is another meaning of the reflexive variant, it is written as an additional sense and the reflexive verb is repeated as a sublemma without parentheses: *kupati se* ‘be in water or swim’. If the meanings of transitive and reflexive verbs are not similar, the reflexive variant is listed only as one of the senses of the transitive verb. For example, in both *Mrežnik* and ŠR the verb *češljati (se)* ‘comb (refl.)’ is recorded only as one sense of the lemma *češljati* ‘comb’, and in this case, *se* is placed in parentheses. The derived reflexive verb *buditi se* ‘wake up’ is treated parallel to the “proper” reflexive verb *kupati se* ‘bathe’: the reflexive marker *se* is placed in parentheses next to the main lemma and is also listed as a separate sublemma under one of the main verb’s senses. Reciprocal reflexive verbs, which are not reflexiva tantum, are treated as a sense of a transitive verb. For example, *dogovarati se* ‘arrange things together’ is a sense of the lemma *dogovarati* ‘arrange, fix’.

*Hrvatski jezični portal* (HJP), like all other Croatian dictionaries, treats inherently reflexive verbs as separate lemmas. However, the treatment of other groups is highly inconsistent. For example, the verbs *kupati* ‘bathe’ and *kupati se* ‘bathe oneself’ are listed as two separate headwords. In contrast, the verb *češljati se* ‘comb oneself’ is treated as a sublemma of the main lemma *češljati* ‘comb’ and is recorded as the fourth sense of the

Pranjković, 2005; Belaj, 2001). In a recent work, Belaj (2024) distinguishes between single-participant and multiparticipant middle verbs, i.e., reciprocal middle verbs. In both cases, Belaj (2004, 98) considers *se* as an integral part of the verb, and consequently a particle. We believe that *se* diachronically originates from the pronoun *sebe*, but, synchronically, it is a particle.

headword *češljati* ‘comb’. The verb *prati se* ‘wash oneself’ is treated differently from both *kupati se* ‘bathe oneself’ and *češljati se* ‘comb oneself’. The reflexive marker *se* is not introduced in the morphological block and there is no separate sense for *prati se* ‘wash oneself’, but the reflexive marker *se* is listed next to the object pronoun within the definition of the transitive verb *prati* ‘wash’, see (1).

(1) *prati* ‘wash’

1. (*koga, što, se*) *ispiranjem u tekućini (ob. u vodi) uklanjati nečistoću*

‘(somebody, something, oneself) remove dirt by rinsing in a liquid (usually in water)’

In the morphological block of the derived reflexive verb *buditi se* ‘wake up’, *se* is recorded next to the object pronoun *što* ‘something’, but both senses (transitive and reflexive) are defined together. The treatment of reciprocal reflexive verbs in HJP does not follow any uniform pattern either. For example, *dogovarati se* ‘arrange things together’ is a separate lemma since the HJP does not contain the verb *dogovarati* ‘arrange, fix’, and the reciprocal usage of the verb *sresti se* ‘meet each other’ is not recorded at all. With the verb *ljubiti se* ‘kiss each other; to love each other’, the reflexive marker *se* is listed in the morphological block, next to the object pronoun *koga, što* ‘somebody, something’. This is very confusing since there is no indication of to which of the listed senses the reflexive marker is connected. If *se* is a marker of the “proper” reflexive verb, it means ‘love oneself’ and if the reflexive marker *se* denotes reciprocity, the verb means ‘kiss one another or to love one another’. It seems that reciprocity is marked in HJP only if the verb is introduced as a separate reflexive verb, and its definition indicates that the meaning is reciprocal.

In CROVALLEX (Mikelić Preradović, 2020), reflexiva tantum, derived, and reciprocal reflexive verbs are recorded as separate lemmas. Therefore, for example, the verbs *penjati se* ‘climb’ (reflexiva tantum), *buditi se* ‘wake up’ (which is considered a derived reflexive verb) and *ljubiti se* ‘kiss each other; love each other’ (as a reciprocal verb) are introduced as separate lemmas. In contrast, *prati se* ‘to wash oneself’ (a “proper” reflexive verb) is treated as one of the senses of the verb *prati* ‘wash’. However, the sublemma is not accompanied by the marker *se* nor is there any label indicating

reflexivity (2). Reflexivity is only visible in the example and in the verb’s definition, which contains the reflexive pronoun *sebe* ‘oneself’. This could pose a problem for non-native users of the dictionary.

(2) 4. *prati (pràti) ≈ uklanjati sa sebe prljavštinu vodom i sapunom*

‘wash ≈ to remove dirt from oneself with water and soap’

-example: *Kupač se prao četkom* ‘The swimmer washed himself with a brush.’

-class: dress

In e-Glava (Birtić, Brač, and Runjaić, 2017), which contains only approximately 50 psychological verbs, the main principle for handling reflexive verbs is to treat reflexiva tantum as a separate lemma exclusively. All other reflexive verbs, including those with transitive and intransitive counterparts, as well as reciprocals, are not listed as separate lemmas but rather as senses of the main lemma. Additionally, they are accompanied by the label *pov. ‘refl.’* to indicate reflexivity. If a verb can have both derived reflexive and reciprocal reflexive variants, these are introduced as separate sublemmas, each with its distinct definition. In (3), the second sense pertains to the derived reflexive verb, whereas the fourth is the reciprocal reflexive verb. Among psychological verbs, there are no examples of “proper” reflexive verbs in e-Glava; however, it is presumed they would be treated similarly to derived reflexive verbs.

(3) 1 *vrijeđati* ‘to insult’ *nanositi uvrede komu, često riječima ili postupcima* ‘to inflict insults on someone, often through words or actions’

2 *vrijeđati se* ‘to take offense’ *povr. ‘refl.’ osjećati se uvrijeđen, često čime; primati uvrede* ‘to feel offended, often by something; to receive insults’

3 *vrijeđati* ‘to irritate’ *pobuđivati bol nadražujući bolno mjesto* ‘to provoke pain by irritating a sore spot’

4 *vrijeđati se* ‘to insult each other’ *povr. ‘refl.’ nanositi uvrede jedan drugomu* ‘to inflict insults upon each other’

The Czech VALLEX is available in several versions. In VALLEX 4.0 and 4.5, which include a data component, i.e., valency frames for active,

non-reflexive and non-reciprocal uses of verbs, and a grammar component, i.e., derived valency frames, such as passive, reflexive and reciprocal uses of verbs (see Kettnerová et al. 2022: 40), a distinction is made between reflexiva tantum and derived reflexive verbs. Each verb classified as reflexiva tantum is assigned the attribute “reflexverb” with the value “tantum”, while derived reflexive verbs are assigned the attribute “reflexverb” with the value “derived”. Additionally, derived reflexive verbs are categorized into seven groups: decausative, autocausative, ‘partitive object’, reciprocal, converse, quasiconverse, and deaccusative. All reflexives are treated as separate lemmas.

In Verbion, all verbs with the reflexive marker *se* are considered as separate lexemes; therefore, they are separate lemmas. The reflexiva tantum is labeled with “REFL tantum” (e.g., *nadati se* ‘hope’) (4).

(4) NADATI SE

Eng. *hope*

1. *očekivati da će se ostvariti željeno*  
‘to expect that the desired will be fulfilled’

**Semantic class:** Psych-Verbs

**Frame:** Desiring

**Example:** *Iskreno, nadam se čudu.*

‘Sincerely, I hope for a miracle.’

|        |             |                |          |
|--------|-------------|----------------|----------|
| EX:    | (ja)        | se<br>nadam    | čudu     |
| SYN:   | NP          |                | NP       |
| MORPH: | NOM         | V              | DAT      |
| SemR:  | Experiencer | REFL<br>tantum | Stimulus |
| FE:    | Experiencer |                | Event    |

Other reflexives are also considered as separate lemmas, as they have undergone semantic and syntactic changes and *se* is not a verb argument but a particle due to desemanticization, i.e., the loss of the semantic function (see Belaj, 2024: 100). The meaning of the reflexive verbs may be predictable from an (in)transitive use, e.g., the aforementioned verb *prati* ‘wash’ and *prati se* ‘wash oneself’. However, some reflexive verbs show only an indirect semantic and syntactic relation to the (in)transitive verbs (see Kettnerová et al. 2022: 45). For example, the transitive verb *praviti* ‘make, create’ means ‘to act with the intention of creating something; produce, create’ and it has a direct

object in the accusative case as a complement, while the reflexive verb *praviti se* means ‘to attribute to oneself qualities that are not real; pretend’ with the predicative complement realized as NP or AP in the nominative or instrumental cases. Additionally, they belong to different semantic classes; the transitive verb belongs to the class of verbs of creation and transformation, while the reflexive verb belongs to the class of verbs with predicative components (see in Levin, 1993).

In Verbion, derived reflexive verbs are labeled with “REFL derived” (5), and reciprocal reflexive verbs with “REFL recipr” (6).

(5) BUDITI SE

Eng. *wake up*

1. *prestajati spavati, dovoditi se u budno stanje*

‘to stop sleeping, to bring oneself into a wakeful state’

**Semantic class:** Verbs of Change of State

**Frame:** Waking up

**Example:** *Budi li se dijete zbog zubića...*

‘If the child is waking up because of teething...’

|       |                             |                 |             |
|-------|-----------------------------|-----------------|-------------|
| EX:   | dijete                      | se<br>budi      | zbog zubića |
| SYN:  | NP                          |                 | PP          |
| MORPH |                             |                 |             |
| :     | NOM                         | V               | zbog + GEN  |
| SemR: | Agent_involu<br>n./Experie. | REFL<br>derived | Cause       |
| FE:   | Sleeper                     |                 | Explanation |

(6) LJUBITI SE

Eng. *kiss each other*

1. *uzajamno izmjenjivati poljupce*  
‘to mutually exchange kisses’

**Semantic class:** Verbs of Contact

**Frame:** Manipulation

**Example:** *Zatim se par strastveno ljubio.*

‘Then, the couple was kissing passionately.’

|        |       |                |            |
|--------|-------|----------------|------------|
| EX:    | par   | se<br>ljubio   | strastveno |
| SYN:   | NP    |                | AdvP       |
| MORPH: | NOM   | V              | adv        |
| SemR:  | Agent | REFL<br>recipr | Manner     |
| FE:    |       |                |            |

**Example:** *Ona se strastveno ljubi s novim dečkom.*

‘She is passionately kissing (with) her new boyfriend.’

|        |       |        |                |
|--------|-------|--------|----------------|
| EX:    | Ona   | se     | s novim dečkom |
|        |       | ljubi  |                |
| SYN:   | NP    | V      | PP             |
|        |       | REFL   | s + INST       |
| MORPH: | NOM   |        |                |
| SemR:  | Agent | recipr | Co_Agent       |
| FE:    |       |        |                |

By treating reflexive verbs as separate lemmas and introducing the aforementioned labels, we believe that other resources dealing with reflexive verbs could benefit from more accessible and precise data. Since this resource is a dynamic database, it offers the flexibility to record reflexive verbs as separate lemmas, and ultimately enhancing the usability and accessibility of the resource for researchers dealing with the topic.

### 3 Light verb constructions

Dealing with light verb constructions is significantly more complex from syntactic, semantic, and technical perspectives. In certain constructions, it can be challenging to distinguish between a semantically full (main) verb and a semantically bleached verb.<sup>2</sup> Consequently, determining whether NP functions as an object or as part of the predicate can be difficult. There is also the question of semantic roles assignment, whether the verb assigns the role itself, the predicative noun, or both (see Grimshaw and Mester, 1988; Butt, 2010; Wittenberg, 2014). Since a light verb and a predicative noun form a single unit, it needs to be decided how to show it in the database.

In order to find the best solution for the database, we consulted other resources that included descriptions of LVCs. In the Czech VALLEX, a light verb determines a syntactic structure, i.e., valency frames, which are identical to those of their full verb counterparts. However, a predicative noun

provides semantic participants since the verb is semantically incomplete (Kettnerová and Lopatková, 2013) via coreference (Alonso Ramos, 2007). A predicative noun in LVCs is labeled as Compound PHrase (CPHR functor) (Kettnerová et al., 2018), in contrast to a full verb valency frame where a noun in that position is labeled as Patient. It is neither an actant nor a free modifier, i.e., it is not a participant, and thus, a semantic role is not assigned to it (Cinková and Kolářová, 2006; Kettnerová and Lopatková, 2013). However, according to Kettnerová and Lopatková (2013), the number of complements in LVCs can be reduced in comparison with the full verbs since only verb complements that are semantically linked to the noun complements can be realized. Noun complements that are not linked to the verb complements remain on the surface structure as noun complements. An exception is causative light verbs with an Instigator or Causator in the subject position (Kettnerová et al., 2018), which is assigned by the verb, while other complements are semantic actants of the noun. In VALLEX 3.5, 4.0 and 4.5, LVCs are listed as senses “complex predicates (light verb)”, accompanied by a frame, light verb constructions that contain nouns belonging to the same or a similar semantic class and that form a complex predicate with an LV, and a map. A verb can have more senses defined as complex predicates, depending on the variety of different frames (see, e.g., *činit, činívat*). Each noun has its own frame, which can be accessed by clicking on the noun.<sup>3</sup>

Regarding Croatian resources, in CROVALLEX (Mikelić Preradović, 2020), there is no detailed description of LVCs since it is based on VALLEX 1.0. LVCs are listed as verb senses along with an equivalent simple verb, but without defining the frame and the semantic class of the verb. They are marked as idioms, without distinguishing between LVCs (7) and phraseological idioms (8).

(7) *donijeti (dònijeti)* ‘bring’ ≈ *odlučiti* ‘decide’ (idiom)  
frame:

<sup>2</sup> Some authors consider light verbs semantically empty, insignificant, or vague (e.g., Jespersen, 1942; Poutsma, 1926; Grimshaw and Mester, 1998; Cattell, 1984), while others argue that they affect a sentence’s meaning since both verb and noun choices are constrained (Wierzbicka, 1982; Butt, 1995). We agree with the later perspective; therefore, we provide sense definitions even for semantically bleached verbs.

<sup>3</sup> PropBank (Hwang et al., 2010) uses the label ARG-M-LVB for a light verb and ARG-PRX (ARGument-Predicating eXpression) for a predicative noun. They treat a light verb and true predicate, as they refer to the predicative noun, as a single predicating unit (REL), which assigns semantic roles by combining the arguments from both the light verb and the noun.

example: *donijeti odluku* ‘bring a decision = make a decision’

(8) *donijeti (dònijeti)* ‘bring’ ≈ *dati kome tko nije uložio nikakav trud* ‘to give to someone who has not made any effort’ (idiom)

frame:

example: *donijeti na tanjuru* ‘to bring on a plate = to hand (something) to someone on a silver platter’

In e-Glava (Birtić, Brač, and Runjaić, 2017), only psych-verbs are available; therefore, it does not currently contain LVCs.

Regarding general language online dictionaries, LVCs are listed as a separate sense of the verbs only in *Mrežnik* (Hudeček and Mihaljević, 2020). This sense includes a generic definition: ‘VERB appears as a light verb with nouns and can often be replaced by a full verb related to the corresponding noun’, along with examples from corpora. The example of the description of the verb *napraviti* ‘make, do’ from *Mrežnik* is given in (9).

(9) *Napraviti se kao nepunoznačni glagol pojavljuje uz imenice i najčešće se može zamijeniti punoznačnim glagolom izrazno povezanim s odgovarajućom imenicom.*

‘To make as a light verb appears with nouns and can most often be replaced with a full verb that is connected to the corresponding noun.’

*napraviti analizu*

do/make analysis.ACC.SG

‘make an analysis’

*Napraviti analizu znači analizirati što, pomno što proučiti.*

‘Do/make (conduct) an analysis means to analyze something, to investigate something thoroughly.’

- *Također sam navela da prije izrade zakona treba napraviti detaljnu analizu i procjenu učinaka propisa.*

‘I also stated that before drafting a law, a detailed analysis and assessment of the regulation’s effects should be conducted.’

However, this is not systematically processed (see, e.g., *donijeti* ‘bring’).

In Verbion, the light verb is categorized as a distinct sense, based on conclusions that light verbs have semantic content beyond a mere functional role (e.g., Butt, 2010; Brugman, 2001; Jackendoff, 2007). It is paired with a predicative noun that gives it full meaning.<sup>4</sup> Verbs with a general meaning or highly schematic verbs, such as the verb *vršiti* ‘do, conduct’, are described with a generic definition “light verb that with a deverbal noun means to perform an activity” (12).

## (12) VRŠITI

Eng. *do, conduct*

1. *nepunoznačni glagol koji s odglagolskom imenicom znači izvoditi ili obavljati kakvu radnju; obavljati*

‘light verb that with a deverbal noun means to perform or carry out some activity; carry out’

**Semantic class:** Verbs of General Activity

**Frame:** Intentionally\_act

**Example:** *Inženjeri su vršili ispitivanja podmorja.*

‘The engineers conducted tests of the seabed.’

|       |                  |           |               |  |              |
|-------|------------------|-----------|---------------|--|--------------|
| EX:   | <b>Inženjeri</b> | <b>su</b> | <b>vršili</b> | ispitivanja                              | podm<br>orja |
| SYN:  | NP               |           |               | NP                                       | NP           |
| MORP  |                  |           |               |  | GEN          |
| H:    | NOM              |           |               | ACC                                      |              |
|       |                  |           | LV            |  | Them<br>e    |
| SemR: | Agent            |           |               | 0  |              |
| FE:   | Agent            |           |               |  |              |
| LU:   |                  |           |               | analizu;<br>čišćenje;<br>dostavu,<br>... |              |

The light verb is annotated with LV to differentiate it from semantically full verbs, which are marked with V. The question is whether a predicative noun bears a semantic role. One possible answer is that direct object NPs in LVCs do not bear a semantic role since the light verb is incapable of assigning one. However, due to the argument transfer, the direct object NPs transfer their argument structure to the argument structure of the light verb, resulting in the semantic role of the Theme being assigned to the object in the genitive case (Grimshaw and Mester, 1988;

<sup>4</sup> Deverbal nouns are listed using Word Sketch from (mainly) the Croatian Web Corpus (Ljubešić and Klubička, 2014).

Karimi-Doostan, 2004). Another answer could be that due to the argument sharing, both the light verb and the noun assign semantic roles (e.g., Culicover and Jackendoff, 2005; Butt 1995, 2010). At the moment, we have decided to label the semantic role of a predicative noun with 0. This choice also prompts the question of whether the NP in the genitive case functions as a noun complement or a verb complement (10). In this case, we treat it as a verb complement as when the LVC is substituted with the full verb *testirati* ‘test’, the NP in the genitive case is in the object positions in the accusative case (11).<sup>5</sup>

- (10) *Inženjeri su vršili testiranje podmorja.*  
 engineers AUX did testing.ACC  
 seabed.GEN  
 ‘The engineers conducted testing of the seabed.’
- (11) *Inženjeri su testirali podmorje.*  
 engineers AUX test seabed.ACC  
 ‘The engineers tested the seabed.’

If a light verb accompanied by different nouns has a different sense, each sense will be recorded (e.g., *donijeti velike brige* ‘bring big worries = cause great worry’ – ‘*izazvati kakvu psihičku promjenu; zabrinuti* = cause a psychological change, to worry’; *donijeti zaradu* ‘bring a profit’ – ‘*doprinijeti čemu, biti koristan, često u materijalnom smislu* = to contribute to something, to be useful, often in a material sense’).

#### 4 Verbal idioms

In the general language dictionaries we analyzed, verbal idioms are a separate category within a lexicographic entry. In *Mrežnik*, an explanation of the verbal idiom and an example from the corpora are provided (12).

- (12) *frazem: prodavati zjake*  
 idiom: ‘twiddle one’s thumbs  
*razg. Prodavati zjake znači dangubiti, besposličariti.*  
 ‘colloq. *Prodavati zjake* means to laze around, to waste time.’

- *Božo i Špela, naime, žive zajedno, no dok Špela radi i financira ih oboje, Božo po cijele dane prodaje zjake.*  
 ‘Božo and Špela live together, but while Špela works and financially supports both of them, Božo spends his days twiddling his thumbs.’

Some online and printed valency dictionaries treat idioms as one of the verb’s senses, while others record them separately from the verb’s sense. VALLEX and CROVALLEX, which follows the methodology of VALLEX 1.0, record idioms as one of the senses of the main lemma. For example, in CROVALLEX, the verb *bacati* ‘throw’ includes as its sixth sense the idiom *bacati drvlje i kamenje (na nekoga)* (lit. throw wood and stones (at somebody) = scold somebody very much). The idiom’s meaning is defined next to the sublemma with the label *idiom* in parentheses. The full idiom is quoted under the example section, but without the actual example, see (8) and (13).

- (13) 6. *bacati (bàcati) ≈ jako grediti (idiom)*  
 ‘throw ≈ to scold severely’  
 -frame:  
 -example: *bacati drvlje i kamenje (na koga)*  
 ‘lit. throw wood and stones (at somebody) = to scold somebody severely’

In VALLEX, idioms are also recorded as one of the verb’s senses, but with a detailed analysis of the idiom’s components (similar to the semantic-syntactic characterization of other valency frames). For example, the ninth sense of the verb *házet<sup>impf</sup>, hodi<sup>pf</sup>* ‘throw’ includes a description of the idiom *hoditi se do gala* (lit. throw (one)self in / at a gala = to dress up) (14).<sup>6</sup>

- (14) frame ACT      obl      obl      obl  
                          1            4            do gala

example: *hodit se do gala*

The variable elements of the idiom are described as other valency elements by virtue of semantic roles (ACT, PAT), the case form expressed by a number and label indicating obligatoriness. The idiom’s fixed element is invariably defined with an abbreviation DPHR (Dependent Phraseme),

<sup>5</sup> There is still no agreement among researchers on the project regarding the status of the NP in the genitive case (*podmorje* ‘seabed’), as we are aware it can be treated as a noun

complement (*testiranje* ‘testing’) due to the case assignment to *podmorje* ‘seabed’.

<sup>6</sup> Special attention to idioms is given in the Polish Walenty (see Przepiórkowski et al., 2014).

followed by the exact form of the fixed element in subscript. In this case, it is the prepositional phrase *do gala* ‘to the gala’.

On the other hand, e-Glava does not include idioms among the verb’s senses but treats them in a separate section, after listing the verb’s senses. There is a special section called *Čvrste sveze* ‘Fixed phrases’ under which all collocations and idioms connected with a certain verb are listed. This section contains the verbal idiom, its explanation, and an example from the corpora, but no detailed description of the syntax of the idiom, following VALBU (Schumacher et al., 2004) (15).

(15) *plašiti se svoje/vlastite sjene - biti vrlo plašljiv, pretjerano oprezan, biti kukavica*  
 ‘lit. to be afraid of one’s/own shadow = to be very timid, overly cautious, to be a coward’  
 ◇ *Za razliku od mnogih, koji se plaše vlastite sjene, Koki otvoreno progovara o svom poslu.*  
 ‘Unlike many who are cowards, Koki openly speaks about his work.’

In Verbion, idioms are recorded in a separate tab, as in e-Glava, but the components of the idioms are accompanied by a more detailed description, as demonstrated in (16). Although these descriptions are less detailed than those of the verb senses, as verbal idioms are not the primary focus of this phase of the project, they still provide valuable insights. Notably, the inclusion of translations of idioms into English is a significant contribution, especially given the current lack of online resources offering Croatian idioms with English translations, as far as we know.

(16) *donijeti na tanjuru = dati što komu tko nije uložio nikakav trud*  
 ‘bring on a silver platter = to give something to someone who hasn’t put in any effort’  
 Example: *On joj je sve u životu donio na tanjuru.*  
 ‘He brought her everything in life on a silver platter.’  
 Agent<sub>NP\_nom</sub> Theme<sub>NP\_acc</sub> Recipient<sub>NP\_dat</sub>  
**donijeti na tanjuru** \_VID

The verb, which can be conjugated and can appear in different tenses and moods, and the fixed part of the idiom are marked with the label VID (verbal idiom).

## 5 Conclusion

Investigating the processing of VMWEs in Croatian general language dictionaries and valency lexicons has highlighted the complexities of their recognition, classification, and description. The inconsistency in their treatment underscores the need for a more harmonized approach to the documentation and analysis of VMWEs. As showed in Section 2, the treatment of reflexive verbs in Croatian general language dictionaries is not unified either within a single dictionary or across dictionaries. The Czech VALLEX offers a compelling solution by treating all reflexive verbs as separate lemmas since they are distinct lexemes. By introducing different labels, we differentiate between reflexiva tantum (REFL tantum), derived reflexive verbs (REFL derived), and reciprocal reflexive verbs (REFL recipr). Currently, we lack resources for advanced classification and for the introduction of syntactic operations as in VALLEX, but this is planned for the future.

Light verb constructions present their own set of challenges, particularly in terms of, firstly, criteria for their recognition (which is not the topic of this paper), and for their semantic and syntactic representation. To distinguish between a semantically full verb and a light verb, we introduced the label V for a semantically full verb and LV for a light verb. By listing the most frequent lexemes in the position of a predicative noun and linking it to a full verb entry, we aim to improve the lexicon’s usefulness.

Verbal idioms are included as a separate section, each accompanied by an explanation, a translation into English, an example from the corpus, and a syntactic and semantic description of its participants.

Other linguistic resources and the processing of VMEWs in Croatian can benefit from a verb lexicon that contains clearly marked and described reflexive verbs, light verb constructions, and, to some extent, verbal idioms.

## Acknowledgments

This work has been supported by the Croatian Science Foundation under the project *Semantic-Syntactic Classification of Croatian Verbs* (SEM-TACTIC) (IP-2022-10-8074) and by the European Union – NextGeneration EU. However, the views and opinions expressed are solely those of the authors and do not necessarily reflect those

of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

## References

- Margarita Alonso Ramos. 2007. Towards the Synthesis of Support Verb Constructions: Distribution of Syntactic Actants between the Verb and the Noun. In Leo Wanner, editor, *Selected Lexical and Grammatical Issues in the Meaning-Text Theory*. John Benjamins Publishing Company. 97–137.
- Stjepan Babić, Dalibor Brozović, Ivo Škarić, and Stjepko Težak. 2007. *Glasovi i oblici hrvatskoga književnoga jezika*. Zagreb: Nakladni zavod Globus.
- Collin F. Baker, Charles J. Fillmore, and Beau Cronin. 2003. The Structure of the FrameNet Database. *International Journal of Lexicography*, 16(3): 281–296.
- Eugenia Barić, Mijo Lončarić, Dragica Malić, Slavko Pavešić, Mirko Peti, Vesna Zečević, and Marija Znika. 1997. *Hrvatska gramatika*. Zagreb: Školska knjiga.
- Branimir Belaj. 2001. Prototipno-kontekstualna analiza povratnih glagola u hrvatskom jeziku. *Suvremena lingvistika*, 51-52(1-2): 1–11.
- Branimir Belaj. 2024. Croatian middle *se*-constructions. In M. Batinić Angster and M. Angster, editors, *The verbal kaleidoscope. Perspectives on the syntax and semantics of verbs*. University of Zadar, Zadar. 93–132.
- Matea Birtić et al. 2012. *Hrvatski školski rječnik*. Zagreb: Školska knjiga.
- Matea Birtić, Ivana Brač, and Siniša Runjaić. 2017. The main features of the e-Glava online valency dictionary. *Electronic lexicography in the 21<sup>st</sup> century. Proceedings of eLex 2017 conference*, pages 43–62, Brno. Lexical Computing CZ s.r.o.
- Claudia Brugman. 2001. Light verbs and polysemy. *Language Sciences*, 23: 551–578.
- Miriam Butt. 1995. *The Structure of Complex Predicates in Urdu*. CSLI Publications, Stanford.
- Miriam Butt. 2010. The Light Verb Jungle: Still Hacking Away. In M. Amberber, M. Harvey and B. Baker, editors, *Complex predicates in cross-linguistic perspective*. Cambridge University Press, Cambridge. 48–78.
- Ray Cattell. 1984. *Composite Predicates in English*. Brill.
- Silvie Cinková and Veronika Kolářová. 2006. Nouns as Components of Support Verb Constructions in the Prague Dependency Treebank. In Mária Šimková, editor, *Insight into Slovak and Czech Corpus Linguistics*. L'udovít Štúr Institute of Li Linguistics of the SAS, Slovak National Corpus, Bratislava. 113–139.
- Matthieu Constant and Joakim Nivre. 2016. A Transition-Based System for Joint Lexical and Syntactic Analysis. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 161–171, Berlin. ACL.
- Peter W. Culicover and Ray Jackendoff. 2005. *Simpler Syntax*. Oxford: Oxford University Press.
- Charles J. Fillmore and Collin F. Baker. 2001. Frame Semantics for Text Understanding. In *WordNet and Other Lexical Resources: Applications, Extensions and Customizations, Workshop*, pages 1–6, Pittsburgh. Association for computational linguistics.
- Jane Grimshaw and Armin Mester. 1988. Light Verbs and  $\Theta$ -Marking. *Linguistic Inquiry*, 19(2): 205–232.
- Lana Hudeček and Milica Mihaljević. 2020. The Croatian Web Dictionary – Mrežnik project – goals and achievements. *Rasprave*, 46(2): 645–667.
- Jena D. Hwang, Archana Bhatia, Clare Bonial, Aous Mansouri, Ashwini Vaidya, Nianwen Xue, and Martha Palmer. 2010. PropBank Annotation of Multilingual Light Verb Constructions. In *Proceedings of the Fourth Linguistickett Annotation Workshop, ACL 2010*, pages 82–90, Uppsala. Association for Computational Linguistics.
- Ray Jackendoff. 2007. A Parallel Architecture perspective on language processing. *Brain research*, 1146: 2–22.
- Otto Jespersen. 1942. *A Modern English Grammar on Historical Principles. Part VI. Morphology*. Ejnar Munksgaard, Copenhagen.
- Gholamhossein Karimi-Doostan. 2005. Light verbs and structural case. *Lingua*, 115: 1737–1756.
- Václava Kettnerová. 2023. Valency structure of complex predicates with Light Verbs. The case of Czech. In Anna Pompei, Lunella Mereu, and Valentina Piunno, editors, *Light Verb Constructions as Complex Verbs*. De Gruyter Mouton, Berlin. 19–43.
- Václava Kettnerová, Markéta Lopatková, and Anna Vernerová. 2022. Reflexives as Part of Verb

- Lexemes in the VALLEX Lexicon. *The Prague Bulletin of Mathematical Linguistics*, 119: 37–66.
- Václava Kettnerová, Markéta Lopatková, Eduard Bejček, and Petra Barančíková. 2018. Enriching VALLEX with Light Verbs: From Theory to Data and Back Again. *The Prague Bulletin of Mathematical Linguistics*, 111: 29–56.
- Václava Kettnerová, Petra Barančíková, and Marketa Lopatková. 2016. Lexicographic Description of Czech Complex Predicates: Between Lexicon and Grammar. In *Proceedings of the XVII EURALEX International Congress*, pages 881–892, Tbilisi. Ivane Javakhishvili Tbilisi State University.
- Václava Kettnerová and Marketa Lopatková. 2013. The Representation of Czech Light Verb Constructions in a Valency Lexicon. In *Proceedings of the Second Conference on Dependency Linguistics, Depling 2013*, pages 147–156, Prague. Charles University in Prague.
- Karin Kipper Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. University of Pennsylvania.
- Svetla Koeva, Ivelina Stoyanova, Maria Todorova, and Svetlozara Leseva. 2016. Semi-automatic Compilation of the Dictionary of Bulgarian Multiword Expressions. In *Proceedings of the GLOBALEX 2016: Lexicographic Resources for Human Language Technology*, pages 86–95, Portorož.
- Beth Levin. 1993. *English Verb Classes and Alternations*. Chicago – London: The University of Chicago Press.
- Nikola Ljubešić and Filip Klubička. 2014. {bs,hr,sr} WaC - Web Corpora of Bosnian, Croatian and Serbian. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 29–35, Gothenburg. Association for Computational Linguistics.
- Nives Mikelić Preradović. 2020. *CROVALLEX: valencijski leksikon glagola hrvatskoga jezika*. Zagreb: Filozofski fakultet.
- Ivana Oraić Rabušić. 2018. *Struktura povratnih glagola i konstrukcije sa se u hrvatskom jeziku*. Zagreb: Institut za hrvatski jezik i jezikoslovlje.
- Petya Osenova and Kiril Simov. 2018. The data-driven Bulgarian Wordnet: BTBWN. *Cognitive studies / Études cognitives*, 18: 1–18.
- Hendrik Poutsma. 1926. *A Grammar of Late Modern English. For the use of continenta, especially Dutch, students. Part II. The Parts of Speech*. Noordhoff, Groningen.
- Adam Przepiórkowski, Elżbieta Hajnicz, Agnieszka Patejuk, Marcin Woliński, Filip Skwarski, and Marek Świdziński. 2014. Walenty: Towards a comprehensive valence dictionary of Polish. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2785–2792, Reykjavik. European Language Resources Association.
- Dragutin Raguž. 2010. *Gramatika hrvatskoga jezika*. Zagreb: Vlastito izdanje.
- Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of Computational Linguistics and Intelligent Text Processing: Third International Conference: CICLing 2002, Lecture Notes in Computer Science, Volume 2276*, pages 1–15, Mexico City. Springer.
- Agata Savary et al. 2017. The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In *Proceedings of the 13th Workshop on Multiword Expressions*, pages 31–47, Valencia. Association for Computational Linguistics.
- Helmut Schumacher, Jacqueline Kubczak, Renate Schmidt, and Vera de Ruyter. 2004. *VALBU – Valenzwörterbuch deutscher Verben*. Tübingen: Gunter Narr Verlag Tübingen.
- Josip Silić and Ivo Pranjković. 2005. *Gramatika hrvatskoga jezika za gimnazije i visoka učilišta*. Zagreb: Školska knjiga.
- Anna Wierzbicka. 1982. Why Can You Have a Drink When You Can't \*Have an Eat?. *Language*, 58(4): 753–799.
- Eva Wittenberg. 2014. *With Light Verb Constructions from Syntax to Concepts*. Potsdam: Potsdam Cognitive Science Series.

# Assessing Reading Literacy of Bulgarian Pupils with Finger-tracking

Alessandro Lento<sup>1</sup>, Andrea Nadalini<sup>2</sup>, Marcello Ferro<sup>2</sup>, Claudia Marzi<sup>2</sup>, Vito Pirrelli<sup>2</sup>,  
Tsvetana Dimitrova<sup>3</sup>, Hristina Kukova<sup>3</sup>, Valentina Stefanova<sup>3</sup>, Maria Todorova<sup>3</sup>, Svetla Koeva<sup>3</sup>

<sup>1</sup> Biomedical Campus University, Rome Italy

<sup>2</sup> Italian National Research Council, Institute for Computational Linguistics

<sup>3</sup> Institute for Bulgarian Language, Bulgarian Academy of Sciences

<sup>1</sup>alessandro.lento@unicampus.it

<sup>2</sup>{andrea.nadalini,marcello.ferro,claudia.marzi,vito.pirrelli}@cnr.it

<sup>3</sup>{cvetana,hristina,valentina,maria,svetla}@dcl.bas.bg

## Abstract

The paper reports on the first steps in developing a time-stamped multimodal dataset of reading data by Bulgarian children. Data are being collected, structured and analysed by means of *ReadLet*, an innovative infrastructure for multimodal language data collection that uses a tablet as a reader's front-end. The overall goal of the project is to quantitatively analyse the reading skills of a sample of early Bulgarian readers collected over a two-year period, and compare them with the reading data of early readers of Italian, collected using the same protocol. We illustrate design issues of the experimental protocol, as well as the data acquisition process and the post-processing phase of data annotation/augmentation. To evaluate the potential and usefulness of the Bulgarian dataset for reading research, we present some preliminary statistical analyses of our recently collected data. They show robust convergence trends between Bulgarian and Italian early reading development stages.

**Keywords:** reading literacy assessment, finger-tracking, multimodal dataset.

## 1 Introduction

Good reading and text comprehension skills are key competences and essential prerequisites for high-quality education (Pikulski and David, 2011). Reading ability can predict performance in all subjects where reading and writing play a role, with reading speed being the most important indicator. In the long term, students with early reading difficulties have serious challenges with general learning, academic performance and social integration (Tichá et al., 2009; Nese et al., 2013). Hence, it is not surprising that educational systems across Europe have put considerable effort into teaching children core reading competencies. Literacy research has been at the forefront of this effort, supporting evidence-based practices for reading and

language classes of schools everywhere. Ideally, education should be supported by continual observation of actual reading behaviour. However, major international organisations such as UNICEF (Chzhen et al., 2018) and OECD (OECD, 2023) have lamented a regrettable shortage of large scale reading data.

So far, two major sources of reading complexity have slowed down progress in collecting longitudinal reading data at scale. First, most recent and influential (eye-tracking) research on reading has typically focused on reading single words or sentences (Rayner, 1998, 2009). However, the need to monitor real reading data in real-life settings raises increasing concerns with the ecological validity of behavioural language data (Brennan, 2016; Demberg and Keller, 2019; Hasson et al., 2018; Willems, 2015), and requires shifting the research focus away from specific, highly controlled phenomena, to real-time processing issues (Jarodzka and Brand-Gruwel, 2017; Kaakinen and Hyönä, 2008; Verhoeven and Perfetti, 2008).

Secondly, the advent of eye-tracking technology at the services of eye movement research started a prolonged period of little interest in the vocal component of reading (with only few exceptions such as De Luca et al. (2013)) and a general neglect of the inherently multi-sensory nature of reading. In fact, a key cognitive insight in developing this ability occurs as learners are able to integrate three emerging sources of information about print and speech: i) the auditorily anchored understanding of syllables, ii) the linguistic-conceptual knowledge of words, and iii) the unfolding visuospatial understanding of printed words built upon the visual and (possibly) tactile exploration of the words' spatial dimension, as it occurs in finger-point reading (Mesmer and Lake, 2010; Mesmer and Williams, 2015). In attaining an efficient synchronisation between word pointing and the onset of word artic-

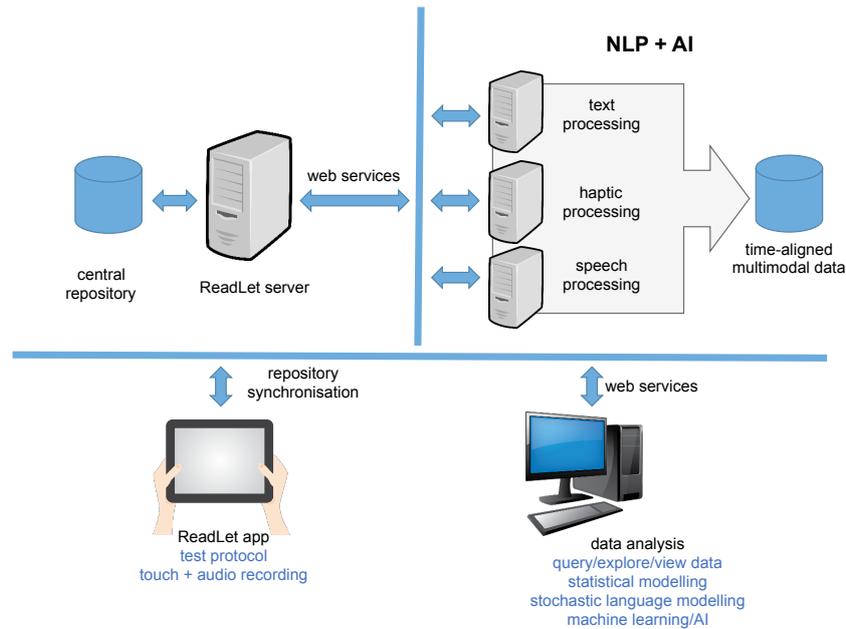


Figure 1: The physical architecture of the ReadLet infrastructure.

ulation, the learner must resolve the competing information between the multiple syllables that (s)he hears and feels and individual words that (s)he sees on a printed page (Mesmer and Lake, 2010; Uhry, 1999, 2002). While some of these reading aspects have been explored and investigated independently, much less work has been conducted so far to study their interaction, also because of the technical difficulty with concurrently recording asynchronous time-series of multimodal signals.

The *ReadLet* infrastructure (Crepaldi et al., 2022; Ferro et al., 2018; Taxitari et al., 2021) was designed and developed to meet most of these methodological and technical *desiderata*, with a view to collecting finely annotated, time-stamped, naturalistic reading sessions of connected texts, in both silent and oral reading modalities, by either child or adult readers. Here, we report on an ongoing project that uses *ReadLet* to collect sessions of reading data of Bulgarian primary school pupils, audio-recorded and “finger-tracked” in Sofia. The project aims to:

- A) design and implement assessment strategies for monitoring and evaluating the reading and word comprehension skills of Bulgarian and Italian early graders;
- B) collect reading and comprehension data from the two populations of children using the same protocols and comparable, rigorously controlled reading texts;

- C) compare the data collected from children of different ages and languages (Bulgarian and Italian) and modelling the results.

Activities (B) and (C) are still underway and only some preliminary results are reported here. In what follows, we provide a broad outline of the *Readlet* architecture (section 2), a technical description of the materials and the experimental protocol adopted for data collection (section 3), an overview of the data collected so far (section 4), and a battery of preliminary analyses (section 5). Some conclusions and prospective directions of the current project are offered in the final section.

## 2 ReadLet

The *ReadLet* infrastructure supports a battery of specialised web services based on a cloud computing architecture (Figure 1). The user endpoint is a native app running on an ordinary tablet with an Internet connection. The app manages the administration of the reading protocol and the concurrent recording of finger touch events and the reader’s voice in oral reading sessions. Upon a reading session being completed, data are encrypted, pseudonymised and transmitted to the central repository through a secure Internet protocol. No session data are locally saved in the tablet’s internal storage after data transmission is successfully carried out.

The core of the architecture is a cloud server that exposes a set of functionalities interfacing

the central repository with the app user. As new data are stored in the repository, cloud services are run to process text, audio and haptic data offline. Since all multi-modal data are eventually time-aligned, processes can cooperate to make the analysis more robust. Results are stored back to the central repository, where they remain available for post-processing.

A simple interface is provided for clinicians and operators to manage customised protocols and user profiles, configure a screening session, personalise the text files (e.g reading texts, questionnaires etc.) required by a specific screening protocol. Recorded data, as well as the output of offline data processing, are eventually made available through a set of web services provided by the cloud server. Web services are accessed through personal authentication by means of username and password, to allow for the rapid prototyping of third-party applications for data visualisation, analysis and modelling, while complying with requirements for personal data protection.

### 3 Data Acquisition

#### 3.1 Participants and protocol

73 children were selected from single classes in a primary school in Sofia, from grade 2 to 5. All classes share the same teachers, and follow the Bulgarian curriculum and state education system. All children whose parents gave their consent took part in the study, with no selection bias for students with better or worse reading skills. The vast majority of participants had normal vision, with a very small percentage of them having their vision corrected to normal. None of them had hearing problems, attention deficit hyperactivity disorder, dyslexia, dysgraphia, dystrophy, aphasia, autism spectrum disorder, cognitive impairment or memory impairment.

During a reading session, participants were sitting at a school desk, in front of a tablet in portrait position. For the present collection campaign, we used a 10.5 inch Samsung Galaxy TAB A8 (2.0 GHz Dual+1.8 GHz Hexa-core, 3 GB RAM, 32 GB eMMC, Android 11), with a 246.8 x 161.9 x 6.9 mm screen and a 1920x1200 pixels resolution and a 1.4 inch Samsung Galaxy TAB S6 Lite (2.3 GHz, 1.8 GHz, 4 GB RAM, 64 GB eMMC, Android 12), with a 244.5 x 154.3 x 7 mm screen and a 2000x1200 pixels resolution. The reading text was displayed in Arial font (21.25pt).

Each child participated in two consecutive reading sessions, one silent and one oral. In each session, the child was asked to read one of four short children's stories originally created for this project, consisting of 5 self-contained episodes organised according to levels of increasing reading complexity (see section 3.2). The length of each episode varied between 120 and 155 words (see Table 1), with all episodes fitting a single tablet page. Second graders read only the first two episodes of each story, third graders read the first three episodes, fourth graders read the first four episodes, and fifth graders read all five episodes. The order of the two reading conditions (oral and silent) was counterbalanced across participants, with no child reading the same story in both conditions.

Before starting a reading session, participants were instructed to use the tip of the index finger of their dominant hand for finger-point reading. A short excerpt from a Bulgarian translation of the *Pinocchio* novel was used as practice session. The session was repeated if the child finger-tracked less than 60% of the practice text. After reading each text episode, the child was asked two reading comprehension questions consisting of a question stem (i.e. the actual question) and four randomly shuffled answers, only one of which was correct. Due to the different number of episodes read by children in different grades, the number of questions ranged from a minimum of 4 (2<sup>nd</sup> graders) to a maximum of 10 questions (5<sup>th</sup> graders).

#### 3.2 Texts for the experiments

Five original Italian texts, created for the specific purposes of the *ReadLet* project (Taxitari et al., 2021), were translated into Bulgarian. In Italian, the readability of each text was automatically controlled according to a data-driven methodology that evaluates the reading difficulty of a text as a machine-learning binary classification problem (Dell'Orletta et al., 2011).

The linguistic features used to predict readability are categorised into four main groups: raw text, lexical, morpho-syntactic and syntactic features. *Raw text features* include sentence length, calculated as the average number of words per sentence, and word length, calculated as the average number of characters per word. *Lexical features* refer to the internal composition of the vocabulary of the text. For Italian, two different features were determined by comparing a text with a reference resource con-

|                             | episode 1 |      | episode 2 |       | episode 3 |      | episode 4 |       | episode 5 |      |
|-----------------------------|-----------|------|-----------|-------|-----------|------|-----------|-------|-----------|------|
|                             | Mean      | SD   | Mean      | SD    | Mean      | SD   | Mean      | SD    | Mean      | SD   |
| BULGARIAN                   |           |      |           |       |           |      |           |       |           |      |
| word length [letters]       | 4.57      | 2.28 | 4.78      | 2.68  | 4.88      | 2.82 | 4.80      | 2.80  | 4.82      | 2.94 |
| text length [words]         | 119.75    | 7.14 | 133.25    | 14.13 | 140.75    | 6.85 | 151.00    | 12.99 | 155.50    | 9.33 |
| type-token ratio            | 0.76      | 0.04 | 0.72      | 0.04  | 0.74      | 0.04 | 0.75      | 0.02  | 0.72      | 0.05 |
| lexical density             | 0.61      | 0.07 | 0.61      | 0.05  | 0.59      | 0.04 | 0.58      | 0.02  | 0.56      | 0.02 |
| PoS type                    | 11.75     | 0.50 | 11.25     | 0.50  | 12.00     | 0.00 | 11.75     | 0.50  | 11.75     | 0.50 |
| IPU length [words]          | 6.30      | 0.88 | 5.93      | 0.58  | 6.75      | 0.72 | 6.78      | 0.82  | 6.38      | 0.47 |
| sentence length [words]     | 8.70      | 0.27 | 11.59     | 1.15  | 14.07     | 0.68 | 18.92     | 1.09  | 20.82     | 1.99 |
| dependency length [words]   | 1.73      | 1.42 | 2.03      | 2.09  | 2.22      | 2.22 | 2.30      | 2.45  | 2.59      | 2.94 |
| word frequency <sup>1</sup> | 5.42      | 1.32 | 5.27      | 1.46  | 5.18      | 1.51 | 5.31      | 1.47  | 5.31      | 1.47 |
| ITALIAN                     |           |      |           |       |           |      |           |       |           |      |
| word length [letters]       | 4.21      | 2.17 | 4.66      | 2.46  | 4.71      | 2.53 | 4.87      | 2.76  | 5.07      | 3.01 |
| text length [words]         | 141.00    | 0.82 | 152.00    | 0.82  | 166.00    | 4.83 | 169.75    | 3.2   | 178.00    | 7.96 |
| type-token ratio            | 0.71      | 0.03 | 0.72      | 0.03  | 0.72      | 0.01 | 0.73      | 0.01  | 0.73      | 0.03 |
| lexical density             | 0.59      | 0.03 | 0.61      | 0.02  | 0.60      | 0.01 | 0.60      | 0.03  | 0.60      | 0.01 |
| PoS type                    | 11.00     | 0.82 | 10.75     | 1.26  | 12.00     | 1.15 | 11.75     | 0.96  | 11.25     | 0.96 |
| IPU length [words]          | 8.24      | 0.83 | 6.70      | 0.36  | 7.74      | 1.87 | 8.95      | 1.68  | 7.51      | 0.85 |
| sentence length [words]     | 10.07     | 0.06 | 13.24     | 0.62  | 16.6      | 0.48 | 21.36     | 1.92  | 23.85     | 2.35 |
| dependency length [words]   | 1.57      | 1.39 | 1.78      | 1.79  | 1.93      | 2.04 | 2.06      | 2.43  | 2.33      | 2.97 |
| word frequency <sup>1</sup> | 5.76      | 1.27 | 5.45      | 1.47  | 5.46      | 1.49 | 5.41      | 1.51  | 5.38      | 1.53 |

Table 1: Descriptive statistics for Bulgarian and Italian reading texts by text episodes (IPU = Implicit Prosodic Unit, SD = standard deviation).

taining a basic vocabulary: a) the percentage of unique words in the text that are also included in the reference list (calculated per lemma); and b) the internal distribution into usage classification classes as very frequent words, frequent words and words with relatively low frequency that refer to everyday objects or actions and are therefore well known to speakers. *Morpho-syntactic features* refer to lexical density, which refers to the ratio of content words (verbs, nouns, adjectives and adverbs) to the total number of lexical tokens in a text. *Syntactic features* are numerous, including the depth of the dependency tree, the relative order of the subordinate clauses in relation to the main clause and the length of the dependency.

Likewise, the Bulgarian translation of the Italian reading texts was preprocessed using the Bulgarian Natural Language Processing pipeline, which orchestrates several natural language processing tools, including the Bulgarian language processing chain (BGLPC) and the Universal Dependencies parser. The Bulgarian language processing chain consists of a sentence splitter, a tokeniser, a Part-of-speech (POS) tagger, a lemmatiser, a noun phrase (NP) extractor, a named-entity recogniser and a stop-word recogniser. All tools are self-contained and designed to work in a pipeline; i.e., the output of the previous component is the input for the

next component, starting with the sentence splitter, and followed by the tokeniser, POS tagger and lemmatiser (Karagiozov et al., 2011).

The current version of the Bulgarian language processing chain uses an improved version of its components enabling simultaneous segmentation of texts into single words and multiword expressions (MWEs) as well as simultaneous POS tagging and lemmatisation of individual words and MWEs (Koeva et al., 2020). Although the accuracy of POS tagging was improved only marginally compared to the accuracy before retraining (0.033%), the most important result is the simultaneous processing of single words and MWEs, which is also reflected in the improvement of the existing lemmatiser. Universal dependency parsing is carried out with the NLP-Cube framework in API mode (Boroş et al., 2018). A Python script was created to enable access to the NLP-Cube functionality, and automate the processing of the Bulgarian texts. For each text, the NLP-Cube annotation and the BGLPC annotation are synchronised token by token, and a correspondence map is created between identical tokens in both documents. Based on this synchronisation, Universal Dependency relations are transferred to the BGLPC CoNLL-U Plus output and the relation index is recalculated.

In order to replicate the methodology used for

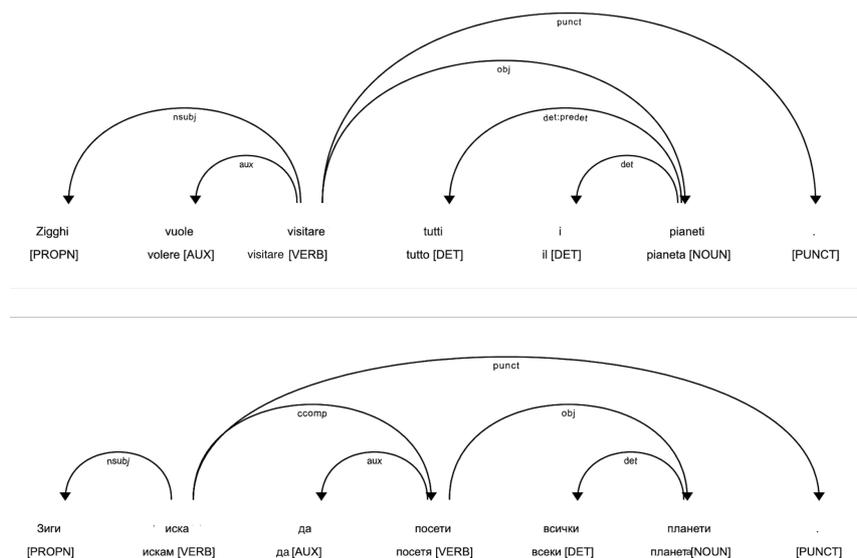


Figure 2: A Universal Dependency analysis of an Italian sentence (top) and its Bulgarian translation (bottom)

the Italian texts, the following principles and steps were applied to Bulgarian translations. First, lexical features in the Bulgarian texts were controlled in two ways: a) by comparing the vocabulary used in each translated text with the general Bulgarian lexis (Koeva and Doychev, 2022); and b) by calculating the *type/token ratio*, i.e. the ratio between the number of lexical *types* (i.e. lemmas) and the number of *tokens* (i.e. lexical forms) that actually occur in the texts. For example, the *type/token ratio* for the first episode of one of the Bulgarian texts is 0.67 (80 unique words and 119 tokens), while in Italian it is 0.61 (87 unique words and 141 tokens). Here, the general grammatical structure of the two languages (as well as the principles of tokenisation and lemmatisation) accounts for the difference between the ratios. Nevertheless, the overall lexical scores in the two languages are comparable.

To ensure a predominant usage of Bulgarian general lexis, we compared the vocabulary of each translated text with the vocabulary found in textbooks and dictionaries for the same age and educational level targeted by the reading texts. Some form of cultural adaptation of the target text was occasionally deemed necessary, as witnessed by some proper names in the Italian texts being replaced by more familiar Bulgarian names in the corresponding translation. For example, *Ivo* was used as a substitute for *Gianni*, and *Violetta* for *Viola*. To the extent possible, both number and type

of sentences and clauses in the original Italian texts were kept in their Bulgarian translations. This is illustrated by the simple example in Figure 2, where two translationally-equivalent sentences are provided, together with their Universal Dependency annotations. The comparison of word and clause lengths in Bulgarian and Italian texts shows that the relatively small number of words in a sentence is maintained in Bulgarian, as is the tendency to use relatively short words with a relatively simple morphological structure. The slightly greater average word length in the Bulgarian texts is due to the morphological structure of nouns, adjectives and some pronouns and numerals, which incorporate definite articles. Likewise, the relatively lower average number of words in Bulgarian can generally be explained by Italian determiners being spelled as independent word tokens (Table 1).

Although the two languages present many morphological and syntactic similarities, there are also significant differences that manifest themselves in the grammatical structure. Overall, the Italian-Bulgarian translation required considerable effort to reproduce the meaning of the original text as faithfully as possible while maintaining the corresponding grammatical structure of the target language (Pirrelli and Koeva, 2024: 35-36).

The descriptive statistics of Table 1 show that all basic parameters of text complexity in the Bulgarian reading texts (letter length, morpho-syntactic

<sup>1</sup>Italian word frequency distributions are extracted from SUBTLEX-IT (Crepaldi et al., 2013). Bulgarian word fre-

quency distributions are extracted from Bulgarian National Corpus, amounting in 1.2 Billion tokens (Koeva et al., 2012).

complexity, lexical density and syntactic complexity) increase with the text episodes. To illustrate: the average word length in letters increases from 4.57 to 4.82 in Bulgarian and from 4.21 to 5.07 in Italian, while the average syntactic dependency length (i.e. the number of tokens that can occur between a syntactic head and its dependent/modifying element in a reading text) increases from 1.73 to 2.59 in Bulgarian and from 1.57 to 2.33 in Italian. At the same time, as expected, the token frequency shows an inverse trend, ranging from a Zipf scale value of 5.42 to 5.31 for the Bulgarian episodes and from 5.76 to 5.38 for the Italian episodes.<sup>2</sup>

#### 4 Data post-processing

The data collected with *ReadLet* include: speech recording, the tracking record of finger movements on the tablet touchscreen, the time taken to answer comprehension questions, and the answers provided by the reader. Data are then post-processed for them to be converted into time-series and then aligned with the text through the following steps.

The position of the text on the tablet touchscreen is encoded with the pixel coordinates ( $x_{min}$ ,  $y_{min}$ ,  $x_{max}$ ,  $y_{max}$ ) of the *bounding box* of each text character on the screen (including blank spaces and punctuation marks).<sup>3</sup> In turn, continuous finger movements are discretized into *touchmove* events on the tablet touchscreen, with each *touchmove* event being associated with its time onset and its pixel coordinates on the screen. Text-coordinates and finger-coordinates are then aligned using a custom convolutional algorithm that finds the largest match between text and finger coordinates (Ferro et al., 2024). Finally, after finger-coordinates are matched with letter-coordinates, we compute the finger-tracking time of each letter in the text as the difference between the last time tick and the first time tick in the time series of touch events falling within a letter's bounding box. The finger-tracking time for a text unit containing more letters is computed as a summation of the tracking times of the letters the unit spans over.

At the moment of writing the paper, the audio-recordings of oral reading sessions are in the pro-

cess of being automatically converted into text using Whisperx (Bain et al., 2023), a free open-source toolkit built on top of Whisper (Radford et al., 2022). For each spoken word, the toolkit outputs an alphabetic transcription and the associated confidence level, together with onset and offset time-points of the word's articulation. After this first processing step, a procedure aligning word transcriptions with the original text is executed using an alignment model (downloadable here), based on a version of Wav2Vec2 XLS-R (Babu et al., 2021) fine-tuned on Bulgarian speech data (downloadable here). This second step is taken to provide more reliable timestamps associated with the actual word in the original text. At the time of writing the paper, Bulgarian children's reading data are being post-processed for speech-to-text conversion. Thus, the present preliminary analyses are exclusively focused on finger-tracking data.

Original audio-recordings of reading sessions will not be made openly available. Nonetheless, we provide open-access information about the onset and offset time-points of a word's articulation, as computed by the speech-to-text conversion tool.

#### 5 Data analysis

The original dataset was trimmed by excluding individual data points (word tokens) whose finger-tracking time was lower than 0.01 seconds or higher than 3.5 seconds. This procedure resulted in 4.4% of the original data being removed, corresponding to 2 subjects of the original set of participants. The resulting dataset was analysed with *R* using Generalised Additive Models (*GAMs*), using the package *gamm4*, version 0.2-6 (Wood, 2017).

To understand the factors affecting the pace of finger-point reading, we entered *token tracking time* (i.e. the time taken by the finger to underline an individual word token) as the dependent variable of two *GAM* models with the independent variables *grade level* (from 2<sup>nd</sup> to 5<sup>th</sup>) and *reading type* (aloud vs. silent) as categorical factors, and *word length* or *word frequency* as numeric predictors. Finally, to take into account the inter-individual variability in our sample and control for effects of lexical variability in our texts, *subjects* and *word tokens* were entered as random effects.

Results are plotted in Figure 3. Here, the box plot in the top panel shows that finger-tracking times are significantly shorter in silent reading than in the aloud reading condition for grades 3, 4 and 5

<sup>2</sup>As frequency measures for the two languages came from corpora of different size, raw counts were transformed using the *Zipf scale*:  $\log_{10}(\text{frequency per million words}) + 3$  (Van Heuven et al., 2014).

<sup>3</sup> $x_{min}$  and  $y_{min}$  are coordinates of the top-left corner of the bounding box;  $x_{max}$  and  $y_{max}$  are coordinates of the bottom-right corner of the bounding box.

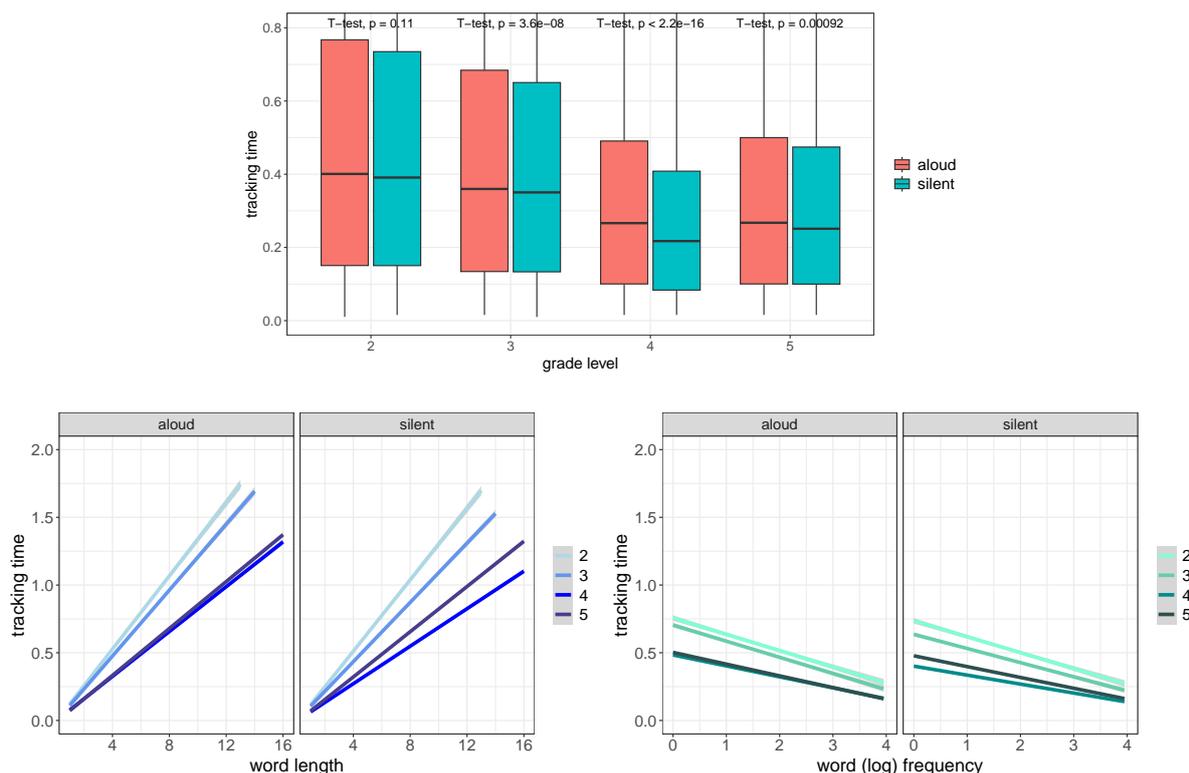


Figure 3: Upper panel: boxplot distributions of tracking time across grades and reading types. Lower panel: linear regression of word length (left) and frequency (right) on tracking time across grades and reading types. Shaded areas refer to 95% confidence intervals.

( $p$ -value  $< 0.001$ ). In addition, one can observe a decreasing trend of finger-tracking times for increasing grade levels ( $p$ -value  $< 2e - 16$ ), with a plateau between grade 4 and 5 for both reading conditions. We believe this levelling effect to be due to the greater complexity of the additional episodes read by 5<sup>th</sup> graders. Effects of word length (left bottom panel) and word frequency (right bottom panel) on tracking times exhibit a significant interaction with the grade level, with early graders (2<sup>nd</sup> and 3<sup>rd</sup>) being more affected by both word length and word frequency than late graders (4<sup>th</sup> and 5<sup>th</sup>) ( $p$ -values  $< 2e - 16$ ). The effect, also observed in Italian child’s reading data collected with the same finger-tracking technology (Marzi et al., 2020; Ferro et al., 2024), shows that the orthographic lexicon of Bulgarian 4 and 5 graders contains not only more word tokens than the orthographic lexicon of 2 and 3 graders, but also distinctively longer and less frequent ones. As a result, the finger-tracking pace of Bulgarian 4 and 5 graders is less prominently slowed down by longer words (left panel) and less frequent words (right panel), than the pace of Bulgarian 2 and 3 graders

is. The evidence confirms that even very young readers of a script such as the Bulgarian Cyrillic alphabet<sup>4</sup> tend to opt for a lexical reading strategy as early as possible for the sake of reading efficiency.

Finally, the two plots also show a significant interaction between word length or frequency and reading conditions, with finger-point reading being significantly more affected by both word length and word frequency in the aloud reading condition than in the silent one ( $p$ -values  $< 2e - 16$ ). Not only does this evidence suggest that longer and rarer words are more difficult to process and access. Also their articulation take longer to be planned and executed. Incidentally, this provides further, indirect evidence of the strong correlation between finger-tracking times and articulation times in finger-point reading. We expect this evidence to be confirmed by text-aligned and time-aligned speech-recognition data.

<sup>4</sup>Although there is no one-to-one correspondence between letters and sounds in either the Bulgarian or Italian alphabet, both scripts are much closer to this ideal condition than – say – in English or French. Accordingly, Bulgarian and Italian are classified as orthographically *transparent* languages, i.e. languages where a word pronunciation can largely be predicted from its spelling.

## 6 Discussion and outlook

In this study, we capitalised on the huge potential of mobile information technology, cloud computing and NLP for behavioural data collection and analysis, to investigate developmental trends in the reading data of Bulgarian early graders. Preliminary results significantly replicate benchmark effects attested in the reading literature on transparent scripts, and are in line with the finger-tracking data collected from Italian children with the same protocol (Ferro et al., 2024; Marzi et al., 2020). Overall, the data confirm that children confronted with a transparent script tend to resort to a lexical reading strategy as early as possible, since direct access to orthographic lexical information allows for a more fluent and efficient reading performance than a sublexical reading strategy.

There are several reasons to recommend an extensive usage of finger-tracking and NLP technologies for literacy research and education. First, the use of a simple tablet supports unobtrusive collection of multimodal reading data in ecological contexts. Following the Italian experience with the *Readlet* infrastructure, we were able to replicate, on a smaller scale, a relatively straightforward and efficient protocol for Bulgarian child’s reading data collection at school. Pupils were extremely responsive to the task, and showed a great familiarity with using the tablet for reading. This made the process of data collection surprisingly quick and most effective, suggesting that finger-tracking can aptly be used for extensive reading assessment in primary schools. In particular, a tablet can be used as a reading book for large-scale studies, paving the way to generalisable results. In addition, the possibility to take repeated single-subject measurements makes finger-tracking evidence suitable not only for group assessment, but also for individual developmental studies.

On a more technical note, the use of a tablet as a reading book allowed us to collect finger-tracking and audio-recording streams concurrently, and take advantage of their being both text-aligned and time-aligned. In spite of recent progress in the accuracy of NLP technologies, they can occasionally be brittle and error-prone, particularly when confronted with real language data, which are collected in inherently noisy, ecological communication contexts.

The bonus of having multiply time-aligned multimodal data streams is that, in processing raw data, noise in one channel can be filtered out by integrat-

ing synchronous information coming from a less noisy channel. For example, the drift of a finger-tracking signal in a particular time window, can be corrected using the voice signal sampled and text-aligned in the same time window (Ferro et al., 2024). In fact, the latter can provide reliable information about which text line the reader is currently processing. This is expected to offer better finger-tracking data but also better transcribed spoken data, which, in turn, can be aligned more reliably both individually with the text being read, and with each other.

In the near future, we consider exploring several lines of research. First, we intend to investigate the complementary role of  $n$ -gram frequencies in affecting Bulgarian finger-point reading, which is likely to attest to a dynamic interaction of sublexical and lexical reading strategies through literacy development (Orsolini et al., 2006). In addition, based on speech-recognition data, we intend to compute the correlation between finger-tracking times and articulation timing in aloud reading, and assess the Implicit Prosody Hypothesis (Breen, 2014), i.e. the idea that, in silent reading, readers activate prosodic representations that are similar to those they would produce when reading the text aloud.

Another aspect of our present study worth exploring in the near future is the interactive dynamic between voice recording and finger-tracking data. A recent analysis of the correlation between voice articulation, finger-tracking patterns and eye-tracking patterns in adult reading (Nadalini et al., 2024) showed that, in finger-point reading, the finger is most often located few characters ahead of the voice, and that the finger’s pace is a rhythmic proxy for a reader’s articulation rate. In particular, the distance of the finger from the voice correlates with the distance of the eye from the voice (or eye-voice span: Inhoff et al., 2011), a measure of the capacity of a reader’s phonological buffer and reading fluency (Laubrock and Kliegl, 2015; Silva et al., 2016).

Thus, one can reasonably expect that monitoring the development of the finger-voice span in early readers can provide evidence of more and less typical developmental patterns of reading skills in the first years of primary school, when reading difficulties are more critical but manifest less clearly. This will offer a suitable benchmark for continual assessment of reading proficiency.

## Acknowledgments

The present study is carried out within the bilateral cooperation project *Assessing reading literacy and comprehension of early graders in Bulgaria and Italy* (2023 – 2025), supported by the National Research Council of Italy (CNR) and the Bulgarian Academy of Sciences. The paper's first author, Alessandro Lento, is a PhD student enrolled in the *National PhD in Artificial Intelligence*, XXXVII cycle, course on Health and Life sciences, organized by Università Campus Bio-Medico of Rome. All other authors contributed equally to the study.

## References

- Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhota, Qiantong Wu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. [Xls-r: Self-supervised cross-lingual speech representation learning at scale](#).
- Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. [WhisperX: Time-Accurate Speech Transcription of Long-Form Audio](#). In *Proc. INTERSPEECH 2023*, pages 4489–4493.
- Tiberiu Boros, Stefan Daniel Dumitrescu, and Ruxandra Burtica. 2018. [NLP-cube: End-to-end raw text processing with neural networks](#). In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 171–179, Brussels, Belgium. Association for Computational Linguistics.
- Mara Breen. 2014. Empirical investigations of the role of implicit prosody in sentence processing. *Language and Linguistics Compass*, 8(2):37–50.
- Jonathan Brennan. 2016. Naturalistic sentence comprehension in the brain. *Language and Linguistics Compass*, 10(7):299–313.
- Yekaterina Chzhen, Anna Gromada, Gwyther Rees, Jose Cuesta, and Zlata Bruckauf. 2018. An unfair start: Inequality in children's education in rich countries. Technical Report 15, UNICEF Office of Research, Innocenti, Florence.
- 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](#). In *Developing Language and Literacy: Studies in Honor of Dorit Diskin Ravid*, pages 443–471. Springer International Publishing.
- Davide Crepaldi, Emmanuel Keuleers, Pavel Mandera, and Michael Brysbaert. 2013. SUBTLEX-IT: A frequency list based on movie subtitles. (unpublished manuscript).
- 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, Edinburgh, Scotland, UK. Association for Computational Linguistics.
- Vera Demberg and Frank Keller. 2019. Cognitive models of syntax and sentence processing. *Human language: From genes and brains to behavior*, pages 293–312.
- Marcello Ferro, Claudia Cappa, Sara Giulivi, Claudia Marzi, Ouafae Nahli, Franco Alberto Cardillo, and Vito Pirrelli. 2018. [Readlet: Reading for understanding](#). In *Proceedings of 5<sup>th</sup> IEEE Congress on Information Science & Technology (IEEE CiST'18)*, Marrakech, Morocco.
- 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*, Turin, Italy.
- Uri Hasson, Giovanna Egidi, Marco Marelli, and Roel M Willems. 2018. Grounding the neurobiology of language in first principles: The necessity of non-language-centric explanations for language comprehension. *Cognition*, 180:135–157.
- 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.
- Halszka Jarodzka and Saskia Brand-Gruwel. 2017. [Tracking the reading eye: Towards a model of real-world reading](#).
- Johanna K Kaakinen and Jukka Hyönä. 2008. [Perspective-driven text comprehension](#). *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 22(3):319–334.
- 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)*, page 223–226, Hamburg, Germany. European Language Resources Association.
- Svetla Koeva and Emil Doychev. 2022. [Ontology supported frame classification](#). In *Proceedings of the 5<sup>th</sup>*

- International Conference on Computational Linguistics in Bulgaria (CLIB 2022)*, pages 203–213, Sofia, Bulgaria. Department of Computational Linguistics, IBL – BAS.
- Svetla Koeva, Nikola Obreshkov, and Martin Yalamov. 2020. [Natural language processing pipeline to annotate Bulgarian legislative documents](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6988–6994, Marseille, France. European Language Resources Association.
- Svetla Koeva, Ivelina Stoyanova, Svetlozara Leseva, Tsvetana Dimitrova, Rositsa Dekova, and Ekaterina Tarpomanova. 2012. The Bulgarian National Corpus: Theory and Practice in Corpus Design. *Journal of Language Modelling*, 0, 1, page 65 — 110.
- Jochen Laubrock and Reinhold Kliegl. 2015. The eye-voice span during reading aloud. *Frontiers in psychology*, 6:1432.
- Claudia Marzi, Anna Rodella, Andrea Nadalini, Loukia Taxitari, and Vito Pirrelli. 2020. Does finger-tracking point to child reading strategies? In *Proceedings of 7<sup>th</sup> Italian Conference on Computational Linguistics*, volume 2769, Bologna.
- 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.
- Heidi Anne E Mesmer and Thomas O Williams. 2015. Examining the role of syllable awareness in a model of concept of word: Findings from preschoolers. *Reading Research Quarterly*, 50(4):483–497.
- Andrea Nadalini, Claudia Marzi, Marcello Ferro, Loukia Taxitari, Alessandro Lento, Davide Crepaldi, and Vito Pirrelli. 2024. [Eye-voice and finger-voice spans in adults’ oral reading of connected texts. Implications for reading research and assessment](#). *The Mental Lexicon*.
- Joseph F. T. Nese, Gina Biancarosa, Kelli D. Cummings, Patrick C. Kennedy, Julie Alonzo, and Gerald A. Tindal. 2013. [In search of average growth: describing within-year oral reading fluency growth across grades 1-8](#). *Journal of school psychology*, 51 5:625–42.
- OECD. 2023. [Equity and inclusion in education: Finding strength through diversity](https://doi.org/10.1787/e9072e21-en). <https://doi.org/10.1787/e9072e21-en>, OECD Publishing, Paris.
- Margherita Orsolini, Rachele Fanari, Valeria Tosi, Barbara De Nigris, and Roberto Carrieri. 2006. From phonological recoding to lexical reading: A longitudinal study on reading development in Italian. *Language and cognitive processes*, 21(5):576–607.
- John J. Pikulski and J. Chard David. 2011. [Fluency: Bridge between decoding and reading comprehension](#). *The Reading Teacher*, 58:510–519.
- 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):29–37.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. [Robust speech recognition via large-scale weak supervision](#).
- Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372.
- Keith Rayner. 2009. The 35th Sir Frederick Bartlett Lecture: Eye movements and attention in reading, scene perception, and visual search. *Quarterly journal of experimental psychology*, 62(8):1457–1506.
- Susana Silva, Alexandra Reis, Luís Casaca, Karl M. Petersson, and Luís Faísca. 2016. [When the Eyes no longer lead: Familiarity and Length Effects on Eye-Voice Span](#). *Frontiers in Psychology*, 7.
- Loukia Taxitari, Claudia Cappa, Marcello Ferro, Claudia Marzi, Andrea Nadalini, and Vito Pirrelli. 2021. Using mobile technology for reading assessment. In *Proceedings of 6<sup>th</sup> IEEE Congress on Information Science & Technology (IEEE CiST’20)*, Agadir, Morocco.
- Renáta Tichá, Christine A. Espin, and Miya Miura Wayman. 2009. [Reading progress monitoring for secondary-school students: Reliability, validity, and sensitivity to growth of reading-aloud and maze-selection measures](#). *Learning Disabilities Research & Practice*, 24(3):132–142.
- Joanna K Uhry. 1999. Invented spelling in kindergarten: The relationship with finger-point reading. *Reading and Writing*, 11:441–464.
- Joanna K. Uhry. 2002. Finger-point reading in kindergarten: The role of phonemic awareness, one-to-one correspondence, and rapid serial naming. *Scientific Studies of Reading*, 6(4):319–342.
- Walter JB Van Heuven, Pawel Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2014. Subtlex-uk: A new and improved word frequency database for British English. *Quarterly journal of experimental psychology*, 67(6):1176–1190.
- Ludo Verhoeven and Charles Perfetti. 2008. Advances in text comprehension: Model, process and development. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 22(3):293–301.
- Roel M Willems, editor. 2015. *Cognitive neuroscience of natural language use*. Cambridge University Press.
- Simon N Wood. 2017. *Generalized additive models: an introduction with R*. Chapman and Hall.

# Educational Horizons: Mapping the Terrain of Artificial Intelligence Integration in Bulgarian Educational Settings

Denitza Kurshumova

University of Plovdiv

“Paisii Hilendarski”

[dcharkova@uni-plovdiv.bg](mailto:dcharkova@uni-plovdiv.bg)

## Abstract

The role of artificial intelligence in education (AIED) has recently become a major topic of discussion and future planning. This article presents data from a large-scale survey involving 1463 Bulgarian educators in primary, secondary, and high schools. The results revealed that 70.30% of the teachers were familiar with or somewhat familiar with the existence of AI applications. Chatbots were the most popular among the surveyed teachers, with ChatGPT ranking as the most familiar. The teachers were almost equally split between those who reported use and those who declared nonuse of AI technology for instructional purposes. A significant association was found between the teachers' familiarity with and use of AI technology and their age-related generational traits. The younger educators (up to 40 years of age) were associated with higher use of AI technology as a support tool for creating lesson plans, lesson content, tests, and exams. The outlined tendencies can be used to inform policy, professional development, and future research in the realm of AI-driven education.

**Keywords:** Artificial Intelligence; Bulgarian schools; Classroom practices; Educational innovation; Education technology; Current trends in education; Survey study

## 1 Introduction

The concept of artificial intelligence in education (AIED) can be traced back to the previous century and the establishment of the International AIED Society (IAIED) in 1997 (Zawacki-Richter et al., 2019). Since then, AIED

has been discussed in scientific forums and journals, but the actual application of AI in education has only recently emerged as a major topic in both the educational community and the general public. The present interest in AI has been ignited by the fast development of various AI apps capable of producing human-like content, audio, and video materials from written prompts (Zhang et al., 2022). It is now evident that AI-based educational technology opens up limitless new opportunities for educators and learners, simultaneously posing numerous challenges stemming from the unknowns (Hsieh and Tsai, 2017; Murugesan and Cherukuri, 2023).

AIED research is currently shifting its focus towards the role of AI in teaching practice, encompassing various aspects of its implementation, including the attitudes and readiness of educators and students for the successful integration of AI-based technology, suitable methodological approaches to the organization of AI-based instruction and assessment, setting ethical standards, and the actual observance of those standards (Baidoo-Anu and Owusu, 2023; Bozkurt, 2023; Grassini, 2023; Zawacki-Richter et al., 2019).

### 1.1 Review of related studies

Although teachers have always played a crucial role in the implementation of teaching innovations, the advent of AI has elevated their role to a whole new level (Ayanwale et al., 2022). The wide range of AI apps and the rapid release of newer ones make it difficult for teachers to explore all options, choose appropriately, and keep up with the learning demand (Sanusi et al., 2022). Moreover, teacher education programs for K-12 grade teachers provide none to minimal knowledge and skills for them to handle the

demands of AI-based technology as a teaching tool and teaching medium (Sanusi et al., 2022; Vazhayil, 2019).

Consequently, insufficient or lacking AI literacy may adversely affect teachers' confidence in using AI-based technology in their instruction. Furthermore, a number of studies have shown a strong association between instructors' preparedness to utilize technology and their degree of confidence (Al-Furaih and Al-Awidi, 2020; Ayanwale et al., 2022; Nikolopoulou, 2021). According to Ayanwale et al.'s research (2022), which included 368 in-service teachers from elementary through high school in Nigeria, teachers' willingness to utilize AI-based technology was also impacted by whether or not they thought its application relevant to the goals of teaching and learning. Another factor impacting teachers' views about AI was their level of trust in the efficacy and accuracy of AI-based technologies (Nazaretsky et al., 2021).

Differences in attitudes towards technology have also been associated with the generational traits of the teacher populations (Hernandez-de-Mendoza et al., 2020; Puiu, 2017). For instance, compared to their older counterparts, teachers born after 1980 (also known as Generation Y or Millennials) and those born in or after 1995 (also known as Generation Z or technoholics) are more likely to hold positive attitudes towards AI technology (Chan and Lee, 2023).

Despite the growing body of research about the relation of teachers to AI technology, most of the studies have focused on teachers' perceptions and hypothetical use of AI (Nazaretsky et al., 2021; Nikolopoulou, 2021; Ayanwale et al., 2022; Sing et al., 2021). With the exception of ChatGPT, whose benefits and drawbacks have been discussed in a number of studies (Alkaiss and McFarlane, 2023; Bozkur et al., 2023; Mathew, 2023), there is little research about teachers' familiarity with current AI applications and their specific functions, including various other chatbots (e.g., Google Bard/Gemini), video and audio generators (e.g., DALL-E), virtual human generators (e.g., PlayHT), and others. Another missing element is the actual purpose of using these AI apps in teaching (in-class, homework assignments, assessments, creative projects, etc.).

This article adds to the existing AIED research insights from a large-scale survey with Bulgarian primary, secondary, and high school teachers

regarding their familiarity with and actual use of trending AI apps in their teaching practice. Connections with the instructors' age-based generational traits are also highlighted.

## 2 Methodology

The present paper is based on a survey study about the role of artificial intelligence (AI) in the Bulgarian school system, including primary school, secondary school and high school. The survey was administered online through Google Forms to school principals from all over Bulgaria. They were asked by the researcher to share the survey with the teachers in their schools and to encourage them to respond. The study was approved by the committee of scientific ethics in the faculty of mathematics and informatics at Plovdiv University "Paisii Hilendarski" under protocol №1252/31.01.2024. Prior to answering any survey questions, the teachers provided an informed consent regarding their participation in the study and the use of the data in academic publications. They were also assured that their responses would remain anonymous.

Apart from the demographic questions, the survey included 26 target items, organized in four sections: 1) teachers' familiarity with AI; 2) teachers' opinions on the application of AI in education, 3) the influence of AI on the teaching practice; 4) the place of AI in teacher training programs. This paper is based on part of the survey questions that were considered the most relevant to its purpose as stated below.

### 2.1 Purpose

The current study examined the familiarity with and actual use of trending AI apps by Bulgarian primary, secondary, and high school educators. The following research questions were addressed:

1. How familiar are teachers with AI? Is there a significant association between the level of familiarity and the teachers' age classified into generations as given in Chan and Lee (2023)?

2. Which are the most popular AI applications among Bulgarian school teachers?

To address question 2, one of the survey items asked teachers to check all AI applications that they were familiar with. The list of AI applications was created following the guidelines about the use of artificial intelligence issued by the Bulgarian ministry of education and science in

January 2024  
[https://www.mon.bg/nfs/2024/02/nasoki-izpolzvane-ii\\_190224.pdf](https://www.mon.bg/nfs/2024/02/nasoki-izpolzvane-ii_190224.pdf)). In this document the current AI applications are conditionally categorized into three main groups (p. 11):

**Chatbots** (e.g. ChatGPT, Google Bard/Gemini, Microsoft Bing AI/ GitHub CoPilot, Perplexity). They generate human-like text based on textual prompts and directions.

**Video, audio, and image generators:** Creation of images and design, music, audio, video, presentations (e.g. DALL-E, Stable Diffusion, Midjourney, Bing Image Creator).

**Virtual human generators:** These systems generate verbalize/speak language and generate audio or video from written text (e.g. Synthesia, Play HT).

3. Do teachers use AI applications in their teaching practice and for what purposes?

### 2.2 Statistical Analysis

The statistical package for the Social Sciences (SPSS) Version 27 (2020) was used to analyze the data. The results are summarized by frequencies and percentages. The association between target variables (e.g., generation category, school level) and the teachers' responses was examined through the Chi-square test and the Kruskal-Wallis test, followed by Bonferroni paired comparisons. All statistical tests were two-tailed and performed at a Type I error ( $\alpha$ ) of 0.05.

## 3 Results

### 3.1 Demographic data

The survey was completed by 1463 teachers, of whom 83.00% were women, 16.60% were men, and 0.40% identified themselves as *other*. The distribution of the teachers by school level showed 21.50% teaching in primary schools, 26.10% in secondary schools, and 52.40% in high schools. The teachers came from 24 out of the 28 regions of Bulgaria. The majority of them (68.10%) worked in schools located in regional cities, 18.50% worked in small town schools, and 13.40% worked in village schools. According to their age, the teachers represented four generation groups, following the classification given in Chan and Lee (2023). The majority of them (68.80%) were of Gen X (Table 1).

| Variables                                    | N    | Percentage |
|--|------|------------|
| <b>Gender</b>                                |      |            |
| Men  | 243  | 16.60%     |
| Women  | 1214 | 83.00%     |
| Other  | 6    | 0.40%      |
| <b>School level</b>                          |      |            |
| Primary                                      | 314  | 21.50%     |
| Secondary                                    | 383  | 26.10%     |
| High   | 766  | 52.40%     |
| <b>School location</b>                       |      |            |
| Regional city                                | 966  | 68.10%     |
| Small town                                   | 271  | 18.50%     |
| Village                                      | 196  | 13.40%     |
| <b>Generation</b>                            |      |            |
| <b>Gen Z</b><br>(20 to 29 years)             | 73   | 5.0%       |
| <b>Gen Y/Millennials</b><br>(30 to 43 years) | 254  | 17.4%      |
| <b>Gen X</b><br>(44 to 60 years)             | 1006 | 68.8%      |
| <b>Baby Boomers</b><br>(over 60 years)       | 130  | 8.9%       |

Table 1: Distribution of the participants by demographic variables

### 3.2 Teachers' familiarity with AI

The distribution of the teachers' responses regarding their overall familiarity with AI was as follows: very familiar (n = 160, 10.90%); familiar (n = 305, 20.80%); somewhat familiar (n = 564, 38.60%); rather unfamiliar (n = 355, 24.30%); and unfamiliar (n = 79, 5.40%). Collapsing the top two and bottom two categories yielded the following distribution: familiar at 31.70%, somewhat familiar at 38.60%, and unfamiliar at 29.70% (Fig. 1).

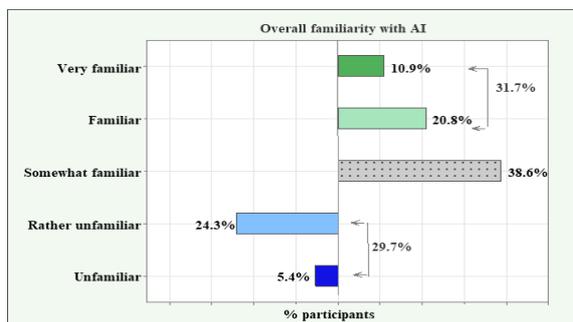


Figure 1: Distribution of the teachers' responses regarding their familiarity with AI

The generation categories, reflecting the age of the participants, showed significant associations with their familiarity with AI according to the Kruskal-Wallis test (test statistic = 41.719, df = 3,  $p < 0.001$ ). The Bonferroni paired comparisons

revealed significant differences between all generations. The proportion of Gen Z teachers who were familiar with AI was significantly higher compared to that of the older generations (Gen Z ↔ Gen Y,  $p = 0.044$ ; Gen Z ↔ Gen X,  $p < 0.001$ ; Gen Z ↔ Baby Boomers,  $p < 0.001$ ). The Gen Y teachers reported significantly higher familiarity than Gen X ( $p = 0.010$ ) and the Baby Boomers ( $p < 0.001$ ). A significant difference was also found between Gen X and Baby Boomers, with a higher percentage of familiarity associated with Gen X ( $p = 0.018$ ).

The negative responses, combining the *rather unfamiliar* and *unfamiliar* options, revealed an opposite trend to the positive ones. Gen Z teachers showed the lowest percentage of unfamiliarity that was significantly different from Gen X ( $p = 0.001$ ) and from the Baby Boomers ( $p < 0.001$ ). The rate of negative responses in the Gen Y age group was significantly lower than that in the Gen X group ( $p < 0.001$ ) and the Baby Boomers ( $p < 0.001$ ). A significant difference was also found between Gen X and Gen Baby Boomers, with a lower percentage of negative responses in the Gen X group ( $p = 0.003$ ).

The middle option *somewhat familiar* did not show substantial variations associated with the generational categories of the participants (Fig. 2).

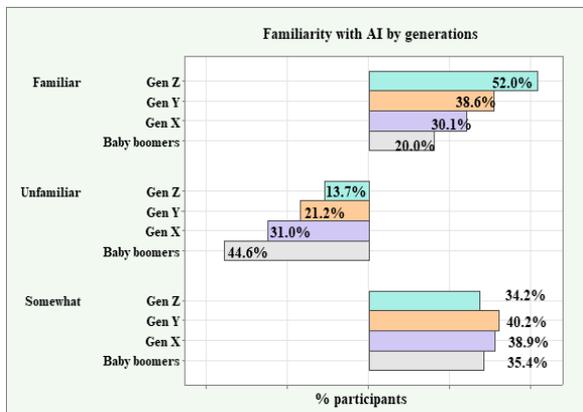


Figure 2: Familiarity with AI by generations

As is seen in figure 3, Chatbots were the most popular among the surveyed teachers. They were indicated as familiar in 121.30% of the responses. The percentage exceeds 100% because the question allowed as many answers as were known to the teachers. ChatGPT was the most popular among them, with 80% of the instructors being familiar with it. Video, audio, and image generators constituted 21.0% of the responses as Bing Image Creator had the highest share of 9.6%. Virtual human generators were the least

popular, constituting only 4.80% of the responses.

The option “*other*” was present in 1.10% of the responses and included Open AI Playground (0.40%), YouChat (0.20%), Jasper AI (0.20%), Claude (0.20%) and Amazon Codewhisperer (0.10%).

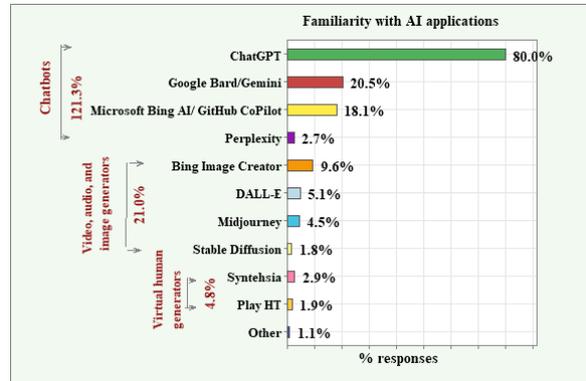


Figure 3: AI applications familiar to the participants in the survey

### 3.3 Teachers’ use of AI in their teaching practices

Of the 1463 teachers who completed the survey, 51.10% ( $n = 747$ ) did not use AI applications in their teaching practice. Frequent use, including options *very often* and *often*, was indicated by 8.70% of the participants, and occasional use (options *from time to time* and *rarely*) was reported by 40.20% of the teachers (Fig. 4). A significant association was found between the reported use of AI and the age-based generation category of the teachers according to the Kruskal-Wallis test (test statistic = 15.851,  $df = 3$ ,  $p = 0.001$ ). The highest percentage of teachers who reported using AI were of the younger generations (Gen X and Gen Y) and the lowest percentage was in the Baby Boomers’ age group.

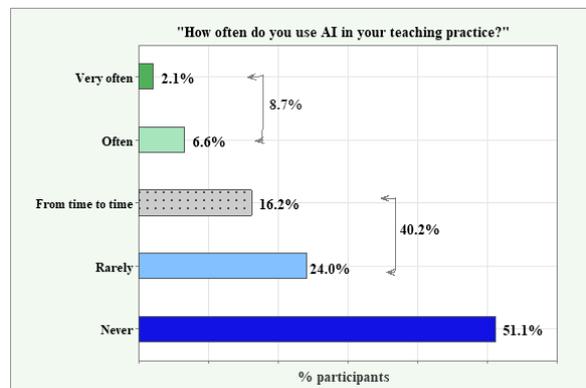


Figure 4: Teachers’ use of AI in the teaching practice

The survey item about the purpose of using AI allowed multiple responses, which is the reason for the cumulative percentage to exceed 100%. In 86.50% of the responses, seven purposes for using AI were indicated. Figure 5 displays these purposes in descending order of use. Among them, *preparation of teaching materials* was the most frequent (24.60%), and the least frequent was *grading tests and exams* (4.10%).

Nonuse of AI was indicated in 61.20% of the responses, as some of the teachers selected both options “*I have not used AI in my teaching.*” and “*I do not want to use AI in my teaching.*”

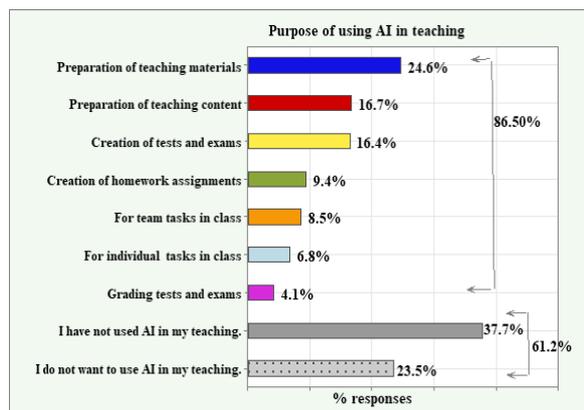


Figure 5: Purpose for using AI in the teaching practice

#### 4 Discussion

The survey data on which this paper is based was collected at the beginning of 2024, when a variety of new AI applications appeared on the Internet and caught the attention of the general public, the media, and the educational community. The Bulgarian Ministry of Education was also involved in discussions and the preparation of guidelines for the use of AI in the education system. Thus, 2024 can be identified as the unofficial starting point of a long-term process of integrating AI in the educational context.

At this onset, the use of AI in teaching practice was not regulated by methodological directions and was mostly driven by the personal motivation of the teachers themselves. Our results showed that of the 1463 teachers who completed the survey, 70.30% were familiar with or somewhat familiar with the existence of AI applications. Given the very recent release of AI apps for general use, this percentage appears high and promising for future teacher training initiatives. The rate of familiarity was the highest among the younger generation of teachers (Gen Z

and Gen Y) and the lowest among the teachers over 60 years of age (Baby boomers). The finding is not surprising given the trends reported in other studies regarding generational differences in attitudes to technology in general and to AI in particular (Chan and Lee, 2023; Hernandez-de-Menendez et al., 2020; Puiu, 2017).

Chatbots were the most popular among the surveyed teachers, with ChatGPT ranking as the most familiar. Much less familiar were video, audio, and image generators and the least familiar were virtual human generators. Given that ChatGPT has been the subject of several research studies discussing both the benefits and drawbacks of the app for teaching and student assessment, its popularity among the surveyed teachers was predicted (Alkaiss and McFarlane, 2023; Bozkur et al., 2023; Mathew, 2023).

Regarding the application of AI in teaching practice, the teachers were almost equally split between those who reported frequent to occasional use and those who declared nonuse at the time of the survey. The highest percentage of teachers who reported using AI were of the younger generations (Gen X and Gen Y), and the lowest percentage was in the Baby Boomers’ age group. Our findings can indirectly be linked to the postulation that insufficient or lacking AI literacy may adversely affect teachers' confidence in using AI-based technology (Al-Furaih and Al-Awidi, 2020; Ayanwale et al., 2022; Nikolopoulou, 2021). Presumably, younger educators—who are naturally more technologically literate than their older counterparts—have the self-confidence to experiment with AI technology as a support tool for creating lesson plans, lesson content, tests, and exams. Although, to a lesser degree, AI was reportedly being utilized in homework assignments and individual and team tasks.

##### 4.1 Limitations

The results reported in this article are time-sensitive due to the rapid release of new AI apps and the expanding field of AIEd, which is trying to catch up with technological advances. In a relatively short time span, teachers’ familiarity with and use of AI apps will grow manifold. Nevertheless, the value of the present research is that it provides a reference point for comparing future developments. Because this paper was based on a limited number of survey questions, it leaves out important issues related to the drawbacks of using AI technology, such as

plagiarism, equitable assessment, the need for technical and methodological assistance in AI use, and others. Additionally, the findings regarding the teachers' familiarity with and utilization of AI technology are predicated on self-reports and may be impacted by an overestimation or underestimation of their actual practices.

## 5 Conclusion

The overarching conclusion that can be drawn on the basis of the reported results is that at the onset of introducing AI applications in the Bulgarian education system, the majority of the school teachers were familiar with current AI apps and were to a certain extent using AI technology in their teaching. This fact indicates a favorable inclination towards AI technology and the existence of some, although unsystematized, baseline knowledge and competence that can serve to inform policy, professional development, and future research in the realm of AI-driven education.

## Acknowledgments

The author would like to thank all of the school principals and administrators in the regional educational departments for their assistance with the survey's dissemination among school teachers from all 28 regions of Bulgaria. A special recognition goes to the teachers who so willingly provided their responses to the survey. Your cooperation is highly appreciated.

## References

Suad A.A. Al-Furaih and Hamed M. Al-Awidi. 2020. Teachers' change readiness for the adoption of smartphone technology: Personal concerns and technological competency. *Technology, Knowledge and Learning*, 25, pages 409–432. <https://doi.org/10.1007/s10758-018-9396-6>.

Hussam Alkaissi and Sammy I. McFarlane. 2023. Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus*, 15(2). <https://doi.org/10.7759/cureus.35179>

Musa Ayanwale, Ismaila Sanusi, Adelana Owolabi, Kehinde Aruleba, and Solomon Oyelere. 2022. Teachers' readiness and intention to teach artificial intelligence in schools. *Computers and Education: Artificial Intelligence*,

3(100099).

<https://doi.org/10.1016/j.caeai.2022.100099>

David Baidoo-anu and Leticia Owusu Ansah. 2023. Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI*, 7(1): 52-62. <https://doi.org/10.61969/jai.1337500>

Aras Bozkurt. 2023. Generative artificial intelligence (AI) powered conversational educational agents: The inevitable paradigm shift. *Journal of Educational Technology*, 18(1). <https://doi.org/10.5281/zenodo.7716416>

Aras Bozkurt, Junhong Xiao, Sarah Lambert, Angelica Pazurek, Helen Crompton, Suzan Koseoglu, Robert Farrow, Melissa Bond, Chrissi Nerantzi, Sarah Honeychurch, et al. 2023. Speculative Futures on ChatGPT and Generative Artificial Intelligence (AI): A collective reflection from the educational landscape. *Asian Journal of Distance Education*, 18(1): 53–130.

Cecilia Ka Yuk Chan and Katherine K. W. Lee. 2023. The AI generation gap: Are Gen Z students more interested in adopting generative AI such as ChatGPT in teaching and learning than their Gen X and millennial generation teachers? *Smart Learning Environments*, 10(60). <https://doi.org/10.1186/s40561-023-00269-3>

Simone Grassini. 2023. Shaping the Future of Education: Exploring the Potential and Consequences of AI and ChatGPT in Educational Settings. *Education Sciences*, 13(7), page 692. <https://doi.org/10.3390/educsci13070692>

Marcela Hernandez-de-Menendez, Carlos A. Escobar and Ruben Morales-Menendez. 2020. Educational experiences with Generation Z. *International Journal on Interactive Design and Manufacturing (IJIDeM)*, 14(3): 847-859. <https://doi.org/10.1007/s12008-020-00674-9>

Wen-Min Hsieh and Chin-Chung Tsai. 2017. Taiwanese high school teachers' conceptions of mobile learning. *Computers and Education*, 115(1): 82–95. <https://doi.org/10.1016/j.compedu.2017.07.013>

Alex Mathew. 2023. Is Artificial Intelligence a World Changer? A Case Study of OpenAI's Chat GPT. *Recent Progress in Science and Technology*, 5, pages 35–42. <https://doi.org/10.9734/bpi/rpst/v5/18240D>

Ministry of Education and Science (2024). Nasoki za izpolzvane na izkustven intelekt v obrazovatelna sistema (Guidelines for utilizing

artificial intelligence in the educational system). Available at:

[https://www.mon.bg/nfs/2024/02/nasoki-izpolzvane-ii\\_190224.pdf](https://www.mon.bg/nfs/2024/02/nasoki-izpolzvane-ii_190224.pdf)

San Murugesan and Aswani Kumar Cherukuri. 2023. The Rise of Generative Artificial Intelligence and Its Impact on Education: The Promises and Perils. *Computer*, 56(5): 116-121. <https://doi.org/10.1109/MC.2023.3253292>

Tanya Nazaretsky, Mutlu Cukurova, Moriah Ariely and Giora Alexandron. 2021. Confirmation Bias and Trust: Human Factors That Influence Teachers' Attitudes Towards AI-Based Educational Technology. *AI for Blended Learning: Empowering Teachers in real Classrooms, Workshop 16<sup>th</sup> European Conference on Technology Enhanced Learning (EC-TEL'2)*. <https://doi.org/10.35542/osf.io/dzqju>

Kleopatra Nikolopoulou, Vasilis Gialamas, Konstantinos Lavidas and Vassilis Komis. 2021. Teachers' readiness to adopt mobile learning in classrooms: A study in Greece. *Technology, Knowledge and Learning*, 26(1): 53-77. <https://doi.org/10.1007/s10758-020-09453-7>

Silvia Puiu. 2017. Generation Z—an educational and managerial perspective. *Revista tinerilor economişti*, (29): 62-72.

Ismaila Temitayo Sanusi, Solomon Sunday Oyelere, Joseph Olamide Omidiora. 2022. Exploring teachers' preconceptions of teaching machine learning in high school: A preliminary insight from Africa. *Computers and Education Open*, 3(4): 100072. <https://doi.org/10.1016/j.caeo.2021.100072>

Anu Vazhayil, Radhika Shetty, Rao R. Bhavani and Nagarajan Akshay. 2019. Focusing on teacher education to introduce AI in schools: perspectives and illustrative findings. In: *Proceedings of the IEEE Tenth international conference on technology for education (T4E)*, pages 71–77. <https://doi.org/10.1109/T4E.2019.00021>

Olaf Zawacki-Richter, Victoria I. Marín, Melissa Bond and Franziska Gouverneur. 2019. Systematic review of research on artificial intelligence applications in higher education – where are the educators? *Int J Educ Technol High Educ* 16, page 39. <https://doi.org/10.1186/s41239-019-0171-0>

Chaoning Zhang, Chenshuang Zhang, Sheng Zheng, Yu Qiao, Chenghao Li, Mengchun Zhang, Sumit Kumar Dam, Chu Myaet Thwal, Ye Lin Tun, Le Luang Huy, Donguk kim, Sung-Ho Bae,

Lik-Hang Lee, Yang Yang, Heng Tao Shen, In So Kweon, and Choong Seon Hong. 2023. A Complete Survey on Generative AI (AIGC): Is ChatGPT from GPT-4 to GPT-5 All You Need?. *Artificial Intelligence*, pages 1-57. <https://doi.org/10.48550/arXiv.2303.11717>

# Evidential auxiliaries as non-reliability markers in Bulgarian parliamentary speech

Ekaterina Tarpomanova

Sofia University Saint Kliment Ohridski

katya@slav.uni-sofia.bg

## Abstract

In the evidentiality system of Bulgarian, there are three evidential auxiliaries that form complex verbal forms. The paper analyzes their potential to mark non-reliability in political discourse by using the ParlaMint-BG corpus of parliamentary debates. The method of the study includes detection, categorisation and context analysis of the evidentials formed with auxiliaries. The results prove that the evidential auxiliaries function as markers of non-reliability, especially in argumentative text type such as political discourse.

**Keywords:** evidentiality, non-reliability, political discourse

## 1 Introduction

Evidentiality in Bulgarian is a grammaticalized category coding the source of information of an utterance and its classification according to the speaker's cognitive states (Nitsolova, 2008: 332). The category is based on the opposition between direct and indirect source of information: the indicative mood is marked for direct (firsthand) evidence whereas a paradigm of three indirect (non-firsthand) evidentials, i.e. reportative, inferential, and dubitative, emerged from the perfect tense. According to Plungian (2001: 354), Balkan systems are an example of modalized evidential systems. He suggests that the evidential system in Bulgarian is based on the opposition personal vs. non-personal information rather than on directly observable vs. non-observable evidence, and furthermore the speaker's mental states are no less personal (and reliable) than their visual observations. Such interpretations of the Bulgarian evidential system rely on the fact that observable evidence is considered more reliable and mediated information less reliable, respectively. On the other hand, the dubitative is the only indirect evidential that includes the value

of reliability in its core semantics, as it expresses the personal stance of the speaker to a non-firsthand information which is evaluated as non-reliable, while the reportative and the inferential are rather neutral to reliability (Tarpomanova, 2016). However, in certain contexts, such as media texts, all indirect evidentials can signal different levels of (non-)reliability (Guentcheva, 1996; Tarpomanova, 2015: 32-33) or even disinformation (Margova, 2022). Political speech can also favour the use of indirect evidentials to mark certain levels of reliability of the information.

## 2 Objectives and methodology

This is a corpus-based study whose objective is to explore three evidential auxiliaries as markers of non-reliability of the information: *бил* 'be-evidential', *щял* 'will-evidential' and *нямал(о)* 'will not-evidential'. As a grammatical form, *бил* is the past active participle of the verb 'be' and can be incorporated in the (present) perfect tense as a full verb – *съм бил* 'have been' but in the evidential paradigm it is an auxiliary that forms different tenses. Similarly, *щял* can be either past active participle of the colloquial verb *ща* 'want' or future evidential auxiliary, and *нямал* – either past active participle of the verb *нямал* 'don't have' or negative future evidential auxiliary. In fact, *нямал* should function as a particle and should be used only in neuter as invariable form – *нямало*, but in the colloquial speech it often agrees in gender and number with the subject. All three evidential auxiliaries are found both in the paradigm of the reliability-neutral evidentials, i.e. the inferential and the reportative, and in the reliability-marked dubitative. The main hypothesis of the study is that in political discourse, which can be viewed as argumentative text type, indirect evidentials are used to suggest non-reliability of the information

and the evidential auxiliaries *бил*, *щял* and *нямал(о)* have mostly dubitative value and to a lower extent reliability-neutral interpretation. To test this assumption, the occurrences of the evidential auxiliaries in the ParlaMint-BG corpus of Bulgarian parliamentary debates have been automatically extracted, manually classified, and analyzed in context.

### 3 Evidentiality recognition and disambiguation

In corpora of Bulgarian as a rule evidentiality is not annotated, but different schemata can be applied for (automatic) recognition of evidential forms. The use of the aorist active participle and especially of the imperfect active participle (the latter occurs only in indirect evidentials, but not in the perfect tenses of the indicative) with omission of the auxiliary *съм* ‘be’ is a reliable marker although it applies for the 3<sup>rd</sup> person only. However, the 3<sup>rd</sup> person is the most frequent grammatical form as compared with the 1<sup>st</sup> and the 2<sup>nd</sup> person. The omission of the auxiliary *съм* ‘be’ in the indicative perfect is possible but rare (Nitsolova, 2008: 296). Additionally, the evidential auxiliaries *бил* ‘be-evidential’, *щял* ‘will-evidential’ and *нямал(о)* ‘will not-evidential’ are unambiguous markers of evidentiality.

The main difficulty in evidentiality recognition is the disambiguation of two types of grammatical homonymy: between the inferential aorist and the indicative perfect (cf. Aleksova, 2003), and in the paradigm of the indirect evidentials. The first type is not relevant for the study as neither the inferential aorist, nor the indicative perfect contain the analyzed auxiliaries, and what is more, both forms are neutral to reliability. As for the homonymies of the second type, several of them should be considered: i. the inferential and the reportative coincide in the 1<sup>st</sup> and the 2<sup>nd</sup> person and differ only in the 3<sup>rd</sup> person by the presence and the omission of the auxiliary *съм* ‘be’, respectively; ii. in all the forms of the dubitative the auxiliary *бил* ‘be-evidential’ can be omitted and in that case, they coincide with the reportative; this means that formally the reportative can express doubt, i.e. non-reliability; iii. the reportative perfect/pluperfect and the dubitative aorist are homonymous (*бил чел* ‘he reportedly has/had read’ or ‘he allegedly read’), i.e. the auxiliary *бил* can function either as a

temporal marker of the reportative or as a non-reliability marker in the structure of the dubitative.

### 4 The corpus

The ParlaMint is a project contributing at the creation of comparable multilingual corpora of parliamentary debates with uniform annotation (Osenova, 2023: 62). The corpora contain transcriptions of the sessions of the 17 European national parliaments with half a billion words (Erjavec et al., 2023: 415). The Bulgarian part used for this study comprises 20 million words attributed to 606 speakers and the political parties and organizations they belong to, collected from October 2014 to July 2020 (Erjavec et al., 2023: 431). The linguistic annotation of the corpus was performed automatically by the CLASSLA pipeline (Ljubešić and Dobrovoljc, 2019). The ParlaMint corpora are available through Sketch Engine, a corpus manager and text analysis software. There are several query types: by word, lemma, character, phrase, and regular expressions based on the Corpus Query Language (CQL).

### 5 Results

To obtain quantitative data for the use of the evidential auxiliaries *бил* ‘be-evidential’, *щял* ‘will-evidential’ and *нямал(о)* ‘will not-evidential’ several regular expressions have been applied. As one of the purposes is to differentiate between evidentials that contain and do not contain the auxiliary *съм* ‘be’, seven regular expressions have been set to search the following strings:

- (1) 1<sup>st</sup>, 2<sup>nd</sup> and 3<sup>rd</sup> person, singular and plural of the auxiliary *съм* ‘be’ + masculine, feminine, neuter and plural of the auxiliary *бил* ‘be-evidential’ + words ending in *-л*, *-ла*, *-ло*, *-ли* (to cover all grammatical genders and numbers of the past active participles);
- (2) no lemma of the auxiliary *съм* ‘be’ + masculine, feminine, neuter and plural of the auxiliary *бил* + words ending in *-л*, *-ла*, *-ло*, *ли*;
- (3) 1<sup>st</sup>, 2<sup>nd</sup> and 3<sup>rd</sup> person, singular and plural of the auxiliary *съм* ‘be’ + masculine, feminine, neuter and plural of the auxiliary *щял* ‘will-evidential’ + particle *да*;
- (4) no lemma of the auxiliary *съм* ‘be’ + masculine, feminine, neuter and plural of the auxiliary *щял* ‘will-evidential’ + particle *да*;
- (5) 1<sup>st</sup>, 2<sup>nd</sup> and 3<sup>rd</sup> person, singular and plural of the auxiliary *съм* ‘be’ + masculine, feminine,

neuter and plural of the auxiliary *нямал* ‘will not-evidential’ + particle *да*;

(6) no lemma of the auxiliary *сѣм* ‘be’+ masculine, feminine, neuter and plural of the auxiliary *нямал* ‘will not evidential’ + particle *да*;

(7) masculine, feminine, neuter and plural of the auxiliary *цѣл* ‘will-evidential’ + masculine, feminine, neuter and plural of the auxiliary *бил* ‘be-evidential’.

Strings (1) and (2) present evidentials containing the auxiliary *бил* ‘be-evidential’ which are ambiguous. Provided that the results are not so numerous, the disambiguation between non-dubitatives and dubitatives has been manually performed. Besides that, as the third element of the string can be any word ending in *-л*, *-ла*, *-ло*, *-ли*, i.e. not necessarily a past active participle, there were many irrelevant results, usually collocations of the perfect tense and a masculine noun ending in the suffix *-тел*, the most frequent among them being *е бил свидетел* ‘he has been a witness’. The irrelevant results were significant in string (1) – 215 out of 298, and minor in string (2) – only 20 out of 365. Of course, they have been excluded from the study. The inverted structure, with the auxiliary *бил* in the first position and the auxiliary *сѣм* in the second one, has been checked too but only 6 occurrences have been found, 5 of which irrelevant, that is why it is not taken into account for the study.

Strings (3) and (4) present positive future evidentials, and strings (5) and (6) exhibit negative future evidentials marked by the respective auxiliaries – *цѣл* and *нямал*. Adding to the string the modal subordinator *да* which functions as a particle linking the future evidential auxiliary and the full verb in a complex verbal form limits the output to the future evidentials. The disambiguation between dubitatives and non-dubitatives has been manual, considering the context.

The output of string (7) should be unambiguously a future dubitative, but it appeared to be a very rare structure.

The results of the search are presented in the tables, which contain the evidentials extracted after disambiguation and elimination of the wrongly recognized outputs (column 1), further divided into reliability-neutral evidentials labeled as non-dubitatives (column 2) and evidentials that express non-reliability labeled as dubitatives (column 3).

Table 1 displays the evidentials containing the auxiliary *сѣм* ‘be’, which allows for person and number differentiation, and the auxiliary *бил* ‘be-evidential’, which has separate forms for masculine, feminine, neuter and plural. The most frequent form is the 3<sup>rd</sup> person, both singular and plural, and it is expected to be reliability neutral as it marks the inferential. However, there are a few occurrences that have dubitative reading and express non-reliable information. The 1<sup>st</sup> and the 2<sup>nd</sup> person, on the other hand, are grammatically ambiguous given that the reported and the inferential may be differentiated in the 3<sup>rd</sup> person only. The results show that the 1<sup>st</sup> and the 2<sup>nd</sup> person tend to be interpreted as dubitatives, with one exception in the 2<sup>nd</sup> person plural.

|        | Evidentials | Non-dubitatives | Dubitatives |
|--------|-------------|-----------------|-------------|
| 1 sg m | 7           | 0               | 7           |
| 1 sg f | 0           | -               | -           |
| 2 sg m | 0           | -               | -           |
| 2 sg f | 0           | -               | -           |
| 3 sg m | 7           | 7               | 0           |
| 3 sg f | 13          | 9               | 4           |
| 3 sg n | 9           | 8               | 1           |
| 1 pl   | 19          | 0               | 19          |
| 2 pl m | 0           | -               | -           |
| 2 pl f | 0           | -               | -           |
| 2 pl   | 5           | 1               | 4           |
| 3 pl   | 22          | 19              | 3           |
| Total  | 82          | 44              | 38          |

Table 1: Auxiliary *сѣм* ‘be’ + auxiliary *бил* ‘be-evidential’ + *-l*-participle.

|       | Evidentials | Non-dubitatives | Dubitatives |
|-------|-------------|-----------------|-------------|
| M sg  | 168         | 1               | 167         |
| F sg  | 48          | 1               | 47          |
| N sg  | 25          | 0               | 25          |
| Pl    | 100         | 1               | 99          |
| Total | 341         | 3               | 338         |

Table 2: Auxiliary *бил* ‘be-evidential’ + *-l*-participle.

Table 2 presents the occurrences of the strings containing the evidential auxiliary *бил* and the past active participle, and excluding the lemma of the auxiliary *сѣм* ‘be’. With those parameters, the query can give as an output only 3<sup>rd</sup> person evidentials differentiated by gender and number. There is a clear preference for dubitative interpretation of that structure, with very few exceptions that are neutral to reliability and have a temporal value of perfect or pluperfect.

In Table 3, future evidentials with the auxiliary *сѣм* ‘be’ are presented. The total is almost equally distributed between non-dubitative and dubitative values, but the preferences for a given value differ among grammatical persons: the 1<sup>st</sup> and the 2<sup>nd</sup> person correlate with dubitative reading, and the 3<sup>rd</sup> person with non-dubitative reading. Similarly to the structure with *бил* ‘be-evidential’, those correlations can be explained with the grammatical ambiguity of the 1<sup>st</sup> and the 2<sup>nd</sup> persons: they can be interpreted either as inferentials or as reportatives and the letter may have dubitative value. On the other hand, the 3<sup>rd</sup> person is expected to have only inferential reading, but still, some of the contexts are clearly dubitative. There were no negative future evidentials containing *сѣм* ‘be’ in the corpus.

Three occurrences of negative future evidentials with the auxiliary *сѣм* ‘be’ have been found, all of them in neuter, one in the 1<sup>st</sup> person singular with dubitative reading, and two in the 3<sup>rd</sup> person singular which have inferential value.

Table 4 and Table 5 present future evidentials with omission of the auxiliary *сѣм* ‘be’ (which occurs in the 3<sup>rd</sup> person only), with the positive and negative future evidential auxiliaries respectively. In both cases, there is a clear preference for dubitative interpretation. The negative future evidential marker is expected to function as a particle, i.e. to be invariable (in neuter), but in fact it is often used inflected for gender and number in the colloquial speech.

In the corpus, there is only one cooccurrence of the two auxiliaries *цѣл* ‘will-evidential’ and *бил* ‘be-evidential’ in neuter – *цѣло било*, and none of the negative future auxiliary *нямал* ‘will not-evidential’ and *бил* ‘be-evidential’.

## 6 Discussion

The results extracted from the ParlaMint-BG corpus justify the hypothesis that political speech is a text type that supports the use of the evidential

|        | Evidentials | Non-dubitatives | Dubitatives |
|--------|-------------|-----------------|-------------|
| 1 sg m | 3           | 0               | 3           |
| 1 sg f | 0           | -               | -           |
| 2 sg m | 1           | 1               | 0           |
| 2 sg f | 0           | -               | -           |
| 3 sg m | 11          | 11              | 0           |
| 3 sg f | 5           | 3               | 2           |
| 3 sg n | 6           | 5               | 1           |
| 1 pl   | 12          | 1               | 11          |
| 2 pl m | 0           | -               | -           |
| 2 pl f | 0           | -               | -           |
| 2 pl   | 8           | 3               | 5           |
| 3 pl   | 8           | 4               | 4           |
| Total  | 54          | 26              | 26          |

Table 3: Auxiliary *цѣл* ‘will-evidential’ + auxiliary *сѣм* ‘be’ + particle *да*.

|       | Evidentials | Non-dubitatives | Dubitatives |
|-------|-------------|-----------------|-------------|
| M sg  | 107         | 5               | 102         |
| F sg  | 56          | 3               | 53          |
| N sg  | 112         | 0               | 112         |
| Pl    | 140         | 1               | 139         |
| Total | 415         | 9               | 406         |

Table 4: Future evidential without auxiliary *сѣм* ‘be’.

|       | Evidentials | Non-dubitatives | Dubitatives |
|-------|-------------|-----------------|-------------|
| M sg  | 0           | -               | -           |
| F sg  | 0           | -               | -           |
| N sg  | 59          | 2               | 57          |
| Pl    | 1           | 0               | 1           |
| Total | 60          | 2               | 58          |

Table 5: Negative future evidential without auxiliary *сѣм* ‘be’.

auxiliaries as markers of non-reliability but depending on the cooccurrence with the auxiliary *сѣм* ‘be’ and its omission. For all evidential auxiliaries analyzed here, the structures containing the auxiliary *сѣм* ‘be’ in the 1<sup>st</sup> and the 2<sup>nd</sup> person are grammatically ambiguous but tend to be interpreted as dubitatives. Constructions with the auxiliary *сѣм* ‘be’ in the 3<sup>rd</sup> person should be interpreted as inferentials, i.e. reliability-neutral, as the inferential cannot express dubitativity. They have temporal value, as in (8).

(8) *Когато ГЕРБ е дошла на власт, е бил влязъл [be-3SgPrInd be-EvMSg enter-AAPMSg]<sup>1</sup> в сила. Така ли е? Така е.* ‘When GERB came to power, it had (supposedly) been put into effect. Is that correct? That’s correct.’

Still, 8 out of 42 cooccurrences of *съм* ‘be’ and *бил* ‘be-evidential’ and 7 out of 30 of *съм* ‘be-evidential’ and *щял* ‘will-evidential’ in the 3<sup>rd</sup> person express non-reliability. As compared with the forms with omission of the auxiliary *съм* ‘be’, they are rather stylistically neutral, i.e. do not imply irony or negative stance.

(9) *Седя и слушам колегите от ляво, голяма част от тях бяха депутати и през 2009 г. – със същите аргументи заставаха пред нас, за да защитят колко добър бюджет са били сътворили [be-3PlPrInd be-EvPl create-AAPPl].* ‘I am listening to my colleagues from the left, a large number of them were MPs in 2009 as well – they stood before us with the same arguments to defend what a good budget they had allegedly created.’

On the contrary, constructions with omitted auxiliary *съм* ‘be’ in the 3<sup>rd</sup> person are most often interpreted as dubitatives, as in (10), with very few exceptions: in contexts (11) and (12), for example, the evidentials with omitted auxiliary *съм* ‘be’ designate tenses of the reportative – perfect in (11) and future in (12). There were also a few contexts, in which the evidentials, although with omitted auxiliary, have inferential reading and do not express non-reliability – (13) and (14).

(10) *Какъв е доводът? Другите били приели [be-EvPl accept-AAPPl]. Ние на състезание ли се намираме?* ‘What is the argument? The others (allegedly) accepted. Are we in a competition?’

(11) *Колко хора в тази зала не са поставили маските си правилно? Тоест маската знаем, че трябва да покрива и носа. Един колега каза, че синът му бил ходил [be-EvMSg go-AAPMSg] една седмица на училище тази година. Уважаеми колеги, от отговорността на всеки един от нас зависи колко време децата няма да бъдат в класните стаи.* ‘How many people in this room didn’t put their masks on correctly? That is, we know that the mask must also cover the nose. A colleague said his son has gone to school for a week this year. Dear colleagues, it is the

responsibility of each of us how long the children will not be in the classrooms.’

(12) *Нещо повече, смятам, че така нареченото дистанционно или електронно обучение трябва много да се внимава и в сферата на висшето образование. Онзи ден синът ми получил в университета съобщение, че изпитът им по физическо щял да се проведе [will-EvMSg Part be held-3SgPresPerf] с тест. Предполагам, че изпитът им по търговско право ще се проведе с лицеви опори.* ‘Moreover, I think that we should be very careful with the so-called distance or electronic learning in the field of higher education as well. The other day, my son received a message at the university that their PE exam will be held with a test. I guess their commercial law exam will be held with push-ups.’

(13) *Били са назначавани и служители на Агенцията срещу подкуп. [...] Престъпната схема била функционирала [be-EvFSg function-AAPFSg], според разследващите органи, чрез добре изградена мрежа от посредници. По време на претърсванията разследващите органи са открили равностойността на 175 хил. лв. в различна валута в дома на Цветанов, както и значителна сума пари в дома на неговия тъст.* ‘Agency employees were also appointed for bribes. [...] The criminal scheme functioned, according to the investigating authorities, through a well-established network of intermediaries. During the searches, the investigating authorities found the equivalent of 175,000 leva in different currencies in Tsvetanov’s home, as well as a significant amount of money in his father-in-law’s home.’

(14) *Невероятни боеве – описани са. Третия ден – решаващия, когато може би Добрич щял да падне [will-EvMSg Part fall-3SgPrPerf], генерал Колев спасява града със своята конна дивизия, като разбива сръбската дивизия.* ‘Amazing fights – they are described. On the third day – the decisive one, when perhaps Dobrich was about to fall, General Kolev saved the town with his cavalry division, defeating the Serbian division.’

As already noted, the cooccurrence of the future evidential auxiliary *щял* (or the negative *нямал*) and the evidential auxiliary *бил* is unambiguously interpreted as future dubitative but the whole verbal

<sup>1</sup> Glosses of the evidentials are provided in square brackets with the following abbreviations: 1, 2, 3 – 1<sup>st</sup>, 2<sup>nd</sup>, 3<sup>rd</sup> person, Sg – singular, Pl – plural, M – masculine, F – feminine, N –

neuter, Pr – present, Ind – indicative, Ev – evidential, AAP – aorist active participle, Part – particle, Perf – perfective, Imperf – imperfective, Refl – reflexive

form becomes overmarked and somehow heavy and subsequently very rare – there is only one such case found in the corpus.

(15) ... **щяло било да попречи** [will-EvNSg be-EvNSg Part hinder-3SgPresPerf] на проверката на сделките и на ревизията на политиката в последните 4 или 10 години. Абсурдни твърдения! Оправдания! ‘... it would have hindered the scrutiny of transactions and the revision of policy in the last 4 or 10 years. Absurd claims! Excuses!’

A survey of the evidentials expressing non-reliability reveals that they occur in specific contexts that support the dubitative reading. In her study on the dubitative in Bulgarian, Krasimira Aleksova argues that as a grammatical form the dubitative is the core of the dubitativity subfield of the functional and semantic field of reliability, whose periphery comprises different lexical modifiers, such as expressions of general evaluation of non-reliability, the particle *уж* ‘allegedly’, words expressing negative subjective stance of the speaker towards the reliability of a non-firsthand information (*lie, bullshit, speculations*, etc.), emotive expressions denoting disapproval (Aleksova 2023: 273-290). The analysis of the occurrences in the ParlaMint-BG corpus confirms Aleksova’s observations and adds some more contexts that are not described by her.

The particle *уж* ‘allegedly’ is generally accepted as one of the most frequent lexical markers of dubitativity as its main semantics is to explicitly express doubt towards non-firsthand source. However, its function in dubitative contexts should not be exaggerated because the evidentials are most often sufficient to express dubitativity, even if the dubitative marker *бил* ‘be-evidential’ is omitted. Indeed, there are 27 cooccurrences of *уж* ‘allegedly’ and *бил* ‘be-evidential’, and only 3 of *уж* ‘allegedly’ and *щял* ‘will-evidential’ (in a distance of 0 to 10 tokens). Another marker of dubitativity and non-reliability with similar semantics is the expression *едва ли не* ‘scarcely, hardly’ which cooccurs 37 times with *бил* ‘be-evidential’, 7 times with *щял* ‘will-evidential’, and 2 times with *нямал(о)* ‘will not-evidential’.

(16) Тук *преждеговоривши* показваха един *междинен доклад*, който *уж* не си **бил свършил** [be-EvMSg do-AAPMSg] работата, но който бил *доказателство*. ‘Here, previous speakers showed an interim report, which allegedly didn’t work, but which was proof.’

(17) *Проблемът, уважаеми господин Председател, е, че вчера министър Радев просто излъга цялата страна, като заяви, че едва ли не той бил получил* [be-EvMSg get-AAPMSg] подкрепа от всички политически сили – нещо, което не е вярно. ‘The problem, Mr. President, is that yesterday Minister Radev simply lied to the whole country by stating that he almost had the support of all political forces – something that is not true.’

Emotional expressions that convey contradiction and irony like *видите ли* ‘you see’, *разбирате ли* ‘you know’, *моля Ви (се)* ‘please’, *извинявайте* ‘excuse me’ may also cooccur with evidentials in a context of non-reliability.

(18) *Господин Димитров си позволи да обяснява, че не се било направило* [Ref] [be-EvNSg do-AAPNSg] нещо *някако от предишните управляващи. Извинявайте, Вие управлявате шест месеца, преди това с малки изключения този Министерски съвет управлява още половин година*. ‘Mr. Dimitrov dares to explain that once upon a time something was not done by the previous government. Excuse me, you govern for six months, before that, with few exceptions, this Council of Ministers governed for another half a year.’

(19) *Никой няма да бъде изправен на съд за това, че бил замислял* [be-EvMSg plan-AAPMSg]. *Много Ви моля, това е несериозно*. ‘No one will be prosecuted for planning. Please, this is not serious.’

(20) *Сега чувам как, видите ли, господин Кирил Ананиев, безспорно с добра биография и авторитет, между другото се чудя как се е съгласил да приеме тази позиция, той щял да бъде* [will-EvMSg Part be-3SgFut] *спасителят*. ‘Now I hear that, you see, Mr. Kiril Ananiev, undoubtedly someone with a good biography and reputation, by the way, I wonder how he agreed to accept this position, he will be the savior.’

Dubitatives are often supported by lexical units denoting false statement – nouns, adjectives or verbs with the meaning of ‘lie’ – *лъжа* ‘a lie’, *спекулация* ‘speculation’, *инсинуации* ‘insinuations’, *неверен* ‘untrue’, *фалшив* ‘false’, *лъжа* ‘to lie’, *мамя* ‘to lie’, and also “softer” synonyms that suggest non-reliability, such as *мит* ‘myth’, *приказки* ‘tales’.

(21) *Каза сума ти лъжи по адрес на ДПС, че, видите ли, Рашков бил нанесъл* [be-EvMSg

*cause-AAPMSg]* щети на ДПС заради купен вот! Пълни лъжи, долни лъжи, господин Председател, от една политическа пеперудка! ‘He told you a lot of lies about the DPS, that, you see, Rashkov had caused damage to the DPS because of a bought vote! Complete lies, vile lies, Mr. Chairman, from a political butterfly!’

(22) Друго невярно твърдение е, че новият акцизен календар щял да облекчи [*will-EvMSg Part relieve-3SgPresPerf*] най-евтините. ‘Another false claim is that the new excise calendar would relieve the cheapest ones.’

(23) Четвърти мит – санирането щяло да подкрепи [*will-EvNSg Part support-3SgPresPerf*] местния бизнес. ‘Fourth myth – remediation will support local businesses.’

The expressions discussed so far are lexical markers of non-reliability. The context analysis shows that there are other mechanisms to suggest or even to intensify the non-reliability of a given information that function rather on textual level. When using an evidential with dubitative value to allude that the reported information is false, the speaker very often offers their version of the situation by making the opposite statement. One of the means to give the opposite information is to use a negation, usually in the next sentence, as disclaimer of the previous statement presented as false (contrariwise, negative evidentials are followed by confirmative statements).

(24) Да, ама те ни били останали [*be-EvPl left-AAPPl*] нула. Не, останали са и изброяваме колко са останали. ‘Yes, but we had zero left. No, there are some left and we are counting how many are left.’

(25) Но това е икономическата философия, която Вие проповядвате: да не вдигаме пенсиите, защото щяла да се вдигне [*will-EvFSg Part Refl raise-3SgPresPerf*] инфлацията. Не, не сме казали това. ‘But this is the economic philosophy you preach: let’s not raise pensions because inflation would rise. No, we didn’t say that.’

(26) Този, който се учудва, че не го бил видял [*Neg be-EvMSg see-AAPMSg*], че го нямало разпечатан. Има го разпечатан – за първо четене е разпечатано, няма промени. ‘The one who is surprised that he had not seen it, that it had not been printed. It’s printed out – for the first reading it’s printed out, there are no changes.’

Another frequent mechanism to introduce a counterstatement is by using adversative

conjunctions or discourse markers: *но* ‘but’, *само че* ‘but, although’, *всъщност* ‘in fact’, *ама* ‘but’, *ами* ‘well’, *обаче* ‘but’, *а* ‘and, but’, etc.

(27) Световната банка се била забавила [*Refl be-EvFSg be late-AAPFSg*]. Ами нали тя Ви беше основният гуру, на който вярвахте? ‘The World Bank was too slow. Well, it was the main guru you trusted, wasn’t it?’

(28) Вече няколко часа от тази трибуна не знам колко човека разпространяват една полуистина – как в Стара Загора някой ги бил блокирал [*be-EvMSg block-AAPMSg*]. Всъщност системата е следната. ‘For several hours now, I don’t know how many people have been spreading a half-truth from this stand – that someone blocked them in Stara Zagora. In fact, the system is as follows.’

(29) Има един общ отговор, как този ток щял да бъде изнасян [*will-EvMSg Part be-3SgFut exported*] някъде. Но на конкретния въпрос къде точно ще го изнасяте, на този въпрос няма отговор. ‘There is a general answer, how this current would be exported somewhere. But there is no answer to the specific question of where exactly you will export it.’

(30) ...защото Брюксел бил казал [*be-EvMSg say-AAPMSg*] така. Точно обратното, ... ‘... because Brussels said so. Just the opposite, ...’

In the near context of evidentials with dubitative value, questions can be seen as a rhetoric device that raises objections to the reliability of the previous statement.

(31) ...днес чувам, че някой се бил подвел [*Refl be-EvMSg mislead-AAPMSg*]. Кой се е подвел? ‘...today I hear that someone was allegedly misled. Who has been misled?’

(32) Как да кажа, има разказ за това как изглеждал основният и допълнителният пакет и това щяло да увреди [*will-EvNSg Part damage-3SgPresPerf*], някой щял за нещо да плаща [*will-EvMSg Part pay-3SgPresImperf*]. Как го разбрахте? ‘How can I say, there is a story about what the basic and additional package looked like, and that would cause damage, someone would pay for something. How did you find out?’

In addition to introducing a question, *wh*-words can be used as intensifiers. In this function, they frequently co-occur with evidentials with dubitative value in repetitive syntactic constructions to suggest non-reliability.

(33) Вие ми говорите тук за окабеляване, под какъв ъгъл щяла да бъде [*will-EvFSg Part be-*

3SgFut] камерата, какво щяло да се заснеме [will-EvNSg Part Refl shoot-3SgPresPerf]? ‘You’re talking to me here about wiring, what angle would the camera be at, what would be shot?’

As a sign of reaction to the false statement, the speaker may address the audience thus imitating a rhetoric dispute. In such context, vocatives or discourse markers for urging reaction of the addressee(s) are used introducing the sentence that follows the evidentials with dubitative meaning.

(34) *Моето предложение към Вас е да не връщаме нещата там, където не бива – служебното правителство какво било казало* [be-EvNSg say-AAPNSg], *какво не било казало* [Neg be-EvNSg say-AAPNSg]. *Хайде да видим този доклад, той действително отразява състоянието на армията.* ‘My proposal to you is that we should not put things back where they should not be – the caretaker government, whatever it said, whatever it did not say. Come on, let’s see this report, it really reflects the state of the army.’

(35) *Да не говорим за това, че беше споменато за Варна как щяла да стане* [will-EvFSg Part become-3SgPresPerf] *логистичен център за износ на украинско зърно. А бе хора, в средите на българските зърнопроизводители ако кажете нещо такова, те ще умрат от смях...* ‘Not to mention that it was said about Varna that it would become a logistics center for the export of Ukrainian grain. Come on, people, in the circles of Bulgarian grain producers, if you say something like that, they will laugh out loud...’

As it can be seen in the contexts discussed so far, different markers supporting the non-reliability interpretation can occur in the same text segment. On the other hand, there are contexts in which evidentials with dubitative value are the only marker of non-reliability.

(36) *Кирил Петков бил излъгал* [be-EvMSg lie-AAPMSg] *пред всички в залата – се каза по-рано днес.* ‘Kiril Petkov had allegedly lied to everyone in the hall – it was said earlier today.’

## 7 Conclusions and applications

The auxiliaries *бил* ‘be-evidential’, *щял* ‘will-evidential’ and *нямал(о)* ‘will not-evidential’ are used to form complex evidentials, which in political discourse are most often interpreted as dubitatives. The research based on the ParlaMint-BG corpus reveals that 867 out of 955 occurrences, or 91% of the complex evidentials have dubitative

value. The percentage is higher if only the evidentials with omission of the auxiliary *съм* ‘be’ are considered: 803 out of 819, or 98%. The analyzed auxiliaries may be seen as a grammatical means to express non-reliability. As compared with some lexical means that suggest non-reliability, they are not less frequent in parliamentary speech: the evidential auxiliaries altogether have 955 occurrences, the lemma of the noun *лъжа* ‘a lie’ 1304, the lemma of the noun *манипулация* ‘manipulation’ 688, the verb *излъга* ‘he/she lied’ 88.

The study has twofold application: for recognition, disambiguation and annotation of evidential forms in corpora, on one hand, and for sentiment analysis, on the other.

## Acknowledgments

The research is supported by the project КП 06-H-80/11 / 15.12.2024 *Bulgarian in state institutions and in the public sphere: crises of communication and communication in crises* financed by the Scientific Research Fund at the Ministry of Education and Science.

## References

- Aleksova, K. 2003. Udostoveritelen perfekt ili umozaklyuchitelten aortist – ot teoretichnite osnovi kam prepodavaneto na chuzhdentsi. In *Publikatsii ot Yubileynata nauchnoprakticheska sesiya “40 godini ICHS”*, pages 60 – 66. Sofia.
- Aleksova, K. 2023. *Dubitativat v savremenniya balgarski ezik*. Sofia: UI “Sv. Kliment Ohridski”.
- Erjavec, T., Ogrodniczuk, M., Osenova, P. et al. 2023. The ParlaMint corpora of parliamentary proceedings. *Lang Resources & Evaluation*, 57, 415–448.
- Guentchéva, Z. 1996. Le médiatif en bulgare. In Guentchéva, Z., editor, *L’énonciation médiatisée*, pages 45 – 70. Louvain – Paris.
- Ljubešić, N., & Dobrovoljc, K. (2019). What does Neural Bring? Analysing improvements in morphosyntactic annotation and lemmatisation of Slovenian, Croatian and Serbian. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 29 – 34. Association for Computational Linguistics.
- Margova, R. 2022. Problemi na upotrebata na renarativa i perfekta za konstatatsiya v presata i onlayn zhurnalistikata. In Koeva, S. & M. Stamenov, editors, *Proceedings of the International Annual Conference of the Institute for*

*Bulgarian Language*, pages 157 – 164. Sofia: Prof. Marin Drinov Publishing house of the Bulgarian Academy of Sciences.

Nitsolova, R. 2008. *Balgarska gramatika. Morfologiya*. Sofia: UI “Sv. Kliment Ohridski”.

Osenova, P. 2023. The Role of Language Technologies in Digital Humanities (The Case of Parliamentary Debates). In *Digital Presentation and Preservation of Cultural and Scientific Heritage*, 13, 61 – 68.

Plungian, V. A. 2001. The place of Evidentiality within the Universal Grammatical Space. *Journal of pragmatics*, 33.3, 349 – 357.

Tarpomanova, E. 2015. *Evidentsialnost v balkanskite ezitsi: balgarski i albanski*. Sofia: Ni plus.

Tarpomanova, E. 2016. Za semantichното sadarzhanie na kategoriyata evidentsialnost v balgarski i albanski. In *Aktualni problemi na lingvistikata. Yubileen sbornik v chest na prof. d.f.n. d-r hon. kauza Stefana Dimitrova*, pages 210 – 216. Sofia: Izdatelstvo na BAN “Prof. Marin Drinov”.

# Extended Context at the Introduction of Complex Vocabulary in Abridged Literary Texts

Iglika Nikolova-Stoupak

Eva Schaeffer-Lacroix

Gaël Lejeune

Sens Texte Informatique Histoire, Sorbonne Université, Paris, France

iglika.nikolova-stoupak@etu.sorbonne-universite.fr,

{eva.lacroix, gael.lejeune}@sorbonne-universite.fr

## Abstract

Psycholinguistics speaks of a fine-tuning process used by parents as they address children, in which complex vocabulary is introduced with additional context (Leung et al., 2021). This somewhat counterintuitive lengthening of text in order to aid one's interlocutor in the process of language acquisition also comes in accord with Harris (1988)'s notion that for every complex sentence, there is an equivalent longer (non-contracted) yet simpler one that contains the same amount of information. Within the proposed work, a corpus of eight renowned literary works (e.g. *Alice's Adventures in Wonderland*, *The Adventures of Tom Sawyer*, *Les Misérables*) in four distinct languages (English, French, Russian and Spanish) is gathered: both the original (or translated) versions and up to four abridged versions for various audiences (e.g. children of a defined age or foreign language learners of a defined level) are present. The contexts of the first appearance of complex words (as determined based on word frequency) in pairs of original and abridged works are compared, and the cases in which the abridged texts offer longer context are investigated. The discovered transformations are consequently classified into three separate categories: addition of vocabulary items from the same lexical field as the complex word, simplification of grammar and insertion of a definition. Context extensions are then statistically analysed as associated with different languages and reader audiences.

**Keywords:** corpus linguistics, literary abridgment, language acquisition, introduction of complex vocabulary

## 1 Introduction

The Cambridge Dictionary defines thus the verb 'to abridge': "to make a book, play, or piece

of writing shorter by removing details and information that is not important". Similarly, the Collins dictionary claims that 'to abridge' means "to reduce the length of (a written work) by condensing or rewriting". Despite this focus on textual length, however, it is logical to assume, in light of linguistic and psycholinguistic theory, that a mere reduction in size would be insufficient for said transformation to act as a 'bridge' (as its name implies) for an audience that would otherwise have difficulty in accessing a given text. This work, which builds upon Nikolova-Stoupak (2023)'s related findings, deliberately seeks and systematises instances where abridged literary works provide longer context than their original counterparts; specifically, when complex vocabulary is introduced. Particular contributions in the process include the composition of a larger corpus of full and abridged versions of famous literary works (including a significant variety of abridged versions in terms of audience, such as readers with comprehension difficulties and foreign language learners of several languages and levels), the exploration of both nouns and verbs as complex words on the basis of frequency lists, the inclusion of contexts that do not necessarily relate to identical parts of the examined texts and an elaborate pipeline that allows for the extraction and analysis of relevant textual samples.

The two main research questions investigated within the present work are: 1. Does complex vocabulary appear in a longer context (sentence) within an abridged rather than full literary text more frequently than random vocabulary? 2. When present, can the mentioned lengthening of context in an abridged text be further systematised with the help of specific examples and relevant linguistic theory?

## 2 Background

### 2.1 Literary Abridgement: Motivation

In a detailed study of 2006, the United Kingdom's *National Literacy Trust* notes that the practice of reading for pleasure has been diminishing over time, much to the detriment of children and adolescents' academic and even social performance (National Literacy Trust, 2006). A direct link between the reading of novels and academic results is established by Moje et al. (2008), who go on to note that motivation to read is triggered by the reading process itself. This vicious circle has also been referred to as the 'Matthew effect', according to which the gap between 'good' and 'bad' readers becomes larger as the latter's difficulty to comprehend reading materials leads to decreased interest in reading (National Literacy Trust, 2006). This is where literary abridgement comes into play and, although it has been claimed that the process compromises an original text's aesthetic or cultural value (Charyulu, 2018), there is significant evidence that the increased comprehensibility of abridged texts helps boost readers' confidence, motivation and ability to enjoy the process of reading (Arias Rodríguez and Roberto Flórez, 2018).

### 2.2 Linguistic Theory

There is evidence in support of the need and practice of expansion as a result of language simplification. The concept of 'baby talk' as well as the related one of 'foreigner talk' have an established presence in linguistic theory. Ferguson (1975) notes that these two modes of address to non-proficient interlocutors are characterised with an array of similarities and differences and that, having their own sets of 'rules', they are more complex than an intuitive imitation of the way the implied interlocutors (young children or foreign language learners) speak. After establishing that the mentioned linguistic varieties are language-specific, he goes on to gather a corpus of English foreigner talk as conceived by a group of students. The issuing language is derived through English based on a series of alterations, including replacements, omissions and, notably, expansion. The last includes, for instance, multiple negation and the inclusion of the subject 'you' in imperatives. Some reported lexical substitutions, such as 'yesterday'

becoming 'the day before', also lead to textual expansion.

A recent study by Leung et al. (2021) specifically addresses the process of introduction of new, complex vocabulary (such as by parents to their young children) as involving fine-tuning, in which the target words are presented along with complementary information that provides cues about their meaning (e.g. 'the dotted leopard' or 'the leopard ran like a cat' rather than the word 'leopard' in isolation). de Bot et al. (1997) seek to define a model for second language (L2) vocabulary acquisition. In particular, the ways in which learners make out the meaning of unknown words during the process of reading are discussed. In first place, the reader establishes the relative importance of the word. Then, they may directly use a dictionary to determine its meaning or rely on inference, which is most often based on grammatical knowledge, word morphology and world knowledge. The understanding and use of vocabulary are established to be inseparable from these of grammar. Ibrahim (2013) points out that lexical complexity may be voluntary or inherent to a language, whilst the complexity of accompanying syntax needs to be deliberately selected accordingly. Christophe et al. (2010) also imply the importance of balance between lexical and syntactic complexity for language acquisition at stating quasi-paradoxically that knowledge of different elements of language is both mutually helpful and inherently mandatory in order for additional knowledge to be achieved.

Harris (1988)'s linguistic theory provides a detailed overview of the way language is composed, with a particular focus on its goal of transmitting information. Discussing grammar, he points out that every complex construction (typically, sentence) can also be expressed in simple terms whilst carrying the same amount of information. The two implied sentences are referred to as having 'descriptive equivalence'. Contractions, i.e. the skipping of inferrable parts of a sentence, cause ambiguity, whilst extension and redundancy facilitate comprehension. Harris (1988) assumes that different languages are associated with different types and frequencies of grammatical contractions.

### 3 Methods

#### 3.1 Corpus Generation

The gathered corpus consists of full and abridged versions of eight famous literary texts with high involvement in school curricula: *Alice's Adventures in Wonderland* (Lewis Carroll), *A Christmas Carol* (Charles Dickens), *Don Quixote* (Miguel de Cervantes), *Les Misérables* (Victor Hugo), *Madame Bovary* (Gustave Flaubert), *Romeo and Juliet* (William Shakespeare), *Oliver Twist* (Charles Dickens) and *The Adventures of Tom Sawyer* (Mark Twain). The languages covered are English, French, Russian and Spanish. Up to four abridged versions per work and per language are included (based on budget and availability) as well as the original work or up to three translations in the cases of languages different from the original one of composition. Almost all texts are present in every language and in at least one full and one abridged version<sup>1</sup>. The main source for full versions is Project Gutenberg<sup>2</sup>, whilst abridged works come from a large variety of sources, mostly as distributed for free. Please refer to Appendix A for more detailed corpus characteristics.

All works were converted into plain text format. In cases where character recognition was necessary, it was run through the Adobe Acrobat software. Texts underwent minor pre-processing, including the removal of metatext (including prefaces), of non-alphanumeric symbols and of readily fixible mistakes resulting from character recognition (such as the presence of '\_' instead of space between words).

The corpus of source texts gathered and used for the purpose of this study has been made publicly available by the authors.<sup>3</sup>

#### 3.2 Context Extraction

The texts in the corpus were grouped into 72 full-abridged pairs, where each pair consisted of one full and one abridged version of the same text in the same language.

Due to the Russian language's strong morphology, Russian texts were firstly lemmatised.

<sup>1</sup>exceptions: *A Christmas Carol*, *Les Misérables* and *Don Quixote* are not present in Russian; *Madame Bovary* is not present in English

<sup>2</sup><https://www.gutenberg.org/>

<sup>3</sup>[https://github.com/iglika88/corpus\\_original\\_and\\_abridged\\_texts/](https://github.com/iglika88/corpus_original_and_abridged_texts/)

All texts underwent part-of-speech (POS) tagging via Python's *spacy* library, following which lists of the nouns and verbs (selected due to carrying the key meaning within a sentence unit) in each text were composed. Then, lists of the intersecting nouns and verbs were extracted for each full-abridged textual pair and they were compared against frequency lists in the respective languages, each of which consisting of 5000 words. The English<sup>4</sup> frequency list used is fittingly based on Project Gutenberg. Based on availability and size as well as the absence of a narrow domain focus, the French<sup>5</sup> and Spanish<sup>6</sup> frequency lists are based on Open Subtitles, and the Russian one<sup>7</sup> – on the Russian National Corpus. All corpora were taken from the *Wiktionary* website. The Russian frequency list contains lemmas, whilst all others contain words as verbatim found in the text.

At this point, the lists of nouns and verbs associated with each full-abridged textual pair contained solely non-frequent (which for the purpose of the experiment is taken to be closely identical to 'complex') words found in both texts. The next step was to extract sentences in which the complex words appear for the first time in each of the two texts and to compare their length. Samples of ten parallel sentences per full-abridged pair were extracted and compared, and the pairs of sentences for which the one associated with the abridged text was longer were counted as well as set apart for further, qualitative analysis.

An additional experiment was carried out in order to test the hypothesis that complex words are more likely to appear within longer context in abridged texts than random words. For the purpose, before the derived lists of nouns and verbs that appear in each textual pair were further filtered based on frequency lists, an additional sample of parallel sentences was extracted from both texts, and the numbers of instances in which an abridged text provided a longer sentence were once again counted.

<sup>4</sup>[https://en.wiktionary.org/wiki/Wiktionary:Frequency\\_lists/English/Project\\_Gutenberg](https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists/English/Project_Gutenberg)

<sup>5</sup>[https://en.wiktionary.org/wiki/Wiktionary:Frequency\\_lists/French\\_wordlist\\_opensubtitles\\_5000\\_](https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists/French_wordlist_opensubtitles_5000_)

<sup>6</sup>[https://en.wiktionary.org/wiki/Wiktionary:Frequency\\_lists/Spanish/Subtitles10K](https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists/Spanish/Subtitles10K)

<sup>7</sup>O. N. Lyashevskaya and S. A. Sharov

## 4 Results

A total of 720 sentence pairs per scenario (i.e. in relation to random versus complex nouns and verbs) were extracted from the corpus, ten per each full-abridged textual pair of the same literary work in the same language. There were 221 cases of longer sentences pertaining to the abridged text within the random sample (i.e. 30.69%) and 222 cases when only complex words were considered (30.83%). Due to the closeness of results, it is concluded that the null hypothesis, namely that abridged versions are as likely to involve longer context when they contain complex words, could not be rejected at this point.

When all sentence pairs related to complex words where the abridged sentence was longer were manually investigated, only 64 of them (28.83%) were discovered to factually conform to the presence of additional context associated with the vocabulary item in question. In the rest of the cases, the added length was attributable to textual processing (e.g. a punctuation sign being erroneously marked as a noun or verb during POS tagging or different punctuation leading to different textual units being extracted as sentences) or to the presence of additional information that, however, has no link to the complex word. In addition, some examples were disregarded where the examined words were not definable as complex despite their absence in the frequency lists (e.g. 'cake', 'pique-nique' ('picnic', Fr.)). Conversely, the present words that were conjugated forms of a common verb, a frequent phenomenon in French and Spanish text (e.g. 'pronunciaba' - 'to pronounce', Sp., 1st/3rd person singular, imperfect tense), were taken to correctly represent complex vocabulary due to their difficult comprehension and use by children and non-native speakers.

### 4.1 Types of Extended Context

In accord with the experiment carried out by Nikolova-Stoupak (2023), three main categories of extended context as used in abridged literary texts were detected: additional vocabulary items pertaining to the complex word's lexical field (59.38% of the discovered instances), grammatical transformations that render the sentence simpler (34.38%) and definitions of

the complex word (6.25%).

#### 4.1.1 Additional vocabulary from the same lexical field

Full: "'Ha!' roared Charley Bates."<sup>a</sup>

Abridged: "Oliver opened his mouth and roared *with all the force of his baby lungs*."<sup>b</sup>

Full: "Nous devrions l'explorer encore."<sup>c</sup>  
("We should explore it further."; Fr.)

Abridged: "Ils découvrent une grotte et *pénètrent à l'intérieur pour explorer les galeries*."<sup>d</sup> ("They discover a cave and enter inside it in order to explore the galleries."; Fr.)

<sup>a</sup>Dickens, Charles. *Oliver Twist*. Project Gutenberg, 1838.

<sup>b</sup>Dickens, Charles. *Oliver Twist*. In *Illustrated Stories by Dickens*, edited by M. Sebag-Montefiore, Abridged edition, Usborne, 2009.

<sup>c</sup>Twain, Mark. *Les Aventures de Tom Sawyer*. Translated by P. F. Caillé and Y. Dubois-Mauvais, Ebooks libres et gratuits, 1917.

<sup>d</sup>Twain, Mark. *Les Aventures de Tom Sawyer*. Abridged edition, edited by A. Culleton, Broché, 2020.

Figure 1: Additional related vocabulary: examples

In the examples in Figure 1, the added context in relation to the complex word aids the reader to both understand its meaning and associate it with common collocations (one roars with a force coming from the lungs; in order to explore a closed area, one needs to enter it).

#### 4.1.2 Transformation into simpler grammar

A variety of grammatical transformations are encountered within the abridged texts, and they neatly point to Harris' claim that language is rendered more complex as a result of contraction. For instance, a clause such as "to ask that someone give you something" is typically contracted as "to ask for something"; anaphora (e.g. "like *this*", a single word in many languages) replaces a longer referent; and personal pronouns are often skipped in languages where they can be inferred from the verb's declension. Within the examined abridged works, one can notice a tendency for sentences to be

|  |
|--|
| <p>Full: "Ningún mendigo le pedía limosna."<sup>a</sup> ("No beggar asked him for alms."; Sp.)</p> <p>Abridged: "Ningún mendigo le suplicó que le diera una limosna."<sup>b</sup> (lit. "No beggar asked him that he give him alms."; Sp.)</p> <p>Full: "так {реветь}"<sup>c</sup> ("to howl like this"; Ru.)</p> <p>Abridged: "{реветь} в тпу ручей"<sup>d</sup> ("to howl like a stream"; Ru.)</p> <p>Full: "corrí"<sup>e</sup> ("ran"; 1st person singular, Sp.)</p> <p>Abridged: "yo corrí"<sup>f</sup> ("I ran"; Sp.)</p> <hr/> <p><sup>a</sup>Dickens, Charles. <i>Canción de Navidad</i>. Ediciones del Sur, 2004.</p> <p><sup>b</sup>Dickens, Charles. <i>Canción de Navidad</i>. Abridged edition, Ediciones la Cueva, 1905.</p> <p><sup>c</sup>Carroll, Lewis. <i>Alisa v Strane chudes</i>. Translated by N. M. Demurova, Nauka, 1978.</p> <p><sup>d</sup>Carroll, Lewis. <i>Alisa v Strane chudes</i>. Abridged edition, edited by L. Yahnin, Eksmo, 1991.</p> <p><sup>e</sup>Shakespeare, William. <i>Romeo y Julieta</i>. Biblioteca digital universal, 2003.</p> <p><sup>f</sup>Shakespeare, William. <i>Romeo y Julieta</i>. Abridged edition, translated by A. L. Pujalte, educ.ar.</p> |
|--|

Figure 2: Grammatical transformations: examples

re-extended to an associated fuller form (see Figure 2).

#### 4.1.3 Definition

As illustrated in Figure 3, definitions within the abridged texts come in two distinct types: metatextual (akin to a glossary) and in-text (naturally embedded within the literary text).

#### 4.2 Results by Language

As seen in Figure 4, the addition of vocabulary items of the same lexical field is the most common scenario of context extension within the abridged texts found in the corpus, and its prevalence is strongest in relation to English and French texts. Grammatical transformations are closely as common within Spanish and Russian abridged works and significantly less prominent in English and French. Def-

|   |
|---|
| <p>Full: "délabré"<sup>a</sup> ("run down", Fr.)</p> <p>Abridged: (in the text's margin) "délabré : en mauvais état."<sup>b</sup> ("run down: in a bad state"; Fr.)</p> <p>Full: "Frère Jean, religieux du même ordre."<sup>c</sup> ("Brother Jean, a monk of the same order"; Fr.)</p> <p>Abridged: "c'est un religieux, un homme d'église."<sup>d</sup> ("He is a monk, a man of the church."; Fr.)</p> <hr/> <p><sup>a</sup>Hugo, Victor. <i>Les Misérables</i>. Project Gutenberg, 1862.</p> <p><sup>b</sup>Hugo, Victor. <i>Les Misérables</i>. Abridged edition, edited by B. Faucard-Martinez, CLE International, 2016.</p> <p><sup>c</sup>Shakespeare, William. <i>Roméo et Juliette</i>. Translated by F.-V. Hugo, Kiléma, 1868.</p> <p><sup>d</sup>Shakespeare, William. <i>Roméo et Juliette</i>. Translated by A. Hubert, edited by F.-V. Hugo, Kiléma.</p> |
|---|

Figure 3: Definition of complex vocabulary: examples

initions are globally rarest, ranging from no instances in Russian to three in French texts.

#### 4.3 Results by Target Audience

Figure 5 shows the presence of extended context in abridged texts by target audience. When it comes to an audience of (native) children, involvement of additional vocabulary is by far most commonly found. In contrast, although this type of extended context is also most common in texts meant for foreign language learners, the other two types follow closely, notably definitions taking the second place. No instances of the investigated context extension are found within the texts marked as being for readers with comprehension difficulties.

The current sample of sentence pairs is insufficient for tendencies to be established in relation to subgroups of readers, such as children of a particular age or language learners with a particular proficiency level.

#### 5 Discussion

Going back to research question 1 as stated in Section 1, the null hypothesis was not rejected in an initial experiment that compared

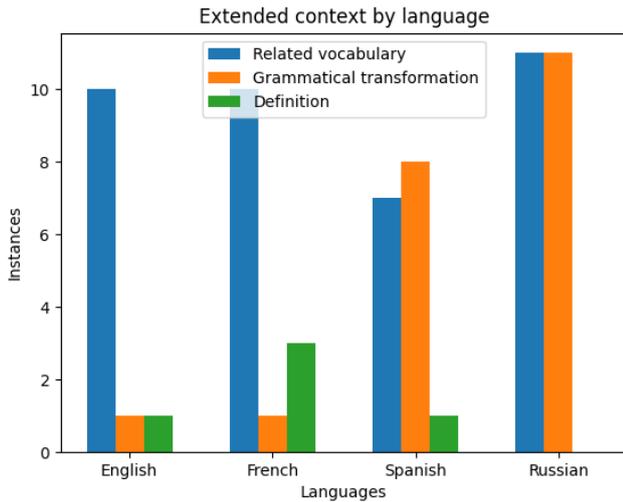


Figure 4: Instances of extended context by language

the number of instances where abridged texts contained longer sentences than their full counterparts in a random sample and in a sample associated with complex vocabulary. However, it was discovered that, in fact, less than half of the cases where the abridged sentences were longer were actually attributable to extended context related to the examined word (rather than to specificities of textual processing, for example). Therefore, results would be more accurate if the experiment is repeated with a larger sample and/or includes manual filtering of the sentences. Moreover, additional factors may exert an influence on the length of sentences within a literary text. For instance, translators have an established tendency to 'simplify' and/or 'explicate' a work's original text, thus respectively reducing and extending the size of context (Volansky et al., 2015).

As regards research question 2, when present, extended context within abridged texts was indeed systematisable as belonging to three main types: addition of vocabulary items of the same lexical field, transformations into simpler grammar (such as anaphora resolution and addition of implied personal pronouns) and definitions of the investigated complex vocabulary item. The first and third types are relatable to Harris (1988)'s notion of 'report' (description, definition), while the second type resembles the linguist's notion of 'paraphrase' (morphological and syntactic modification).

Instances of added vocabulary were most nu-

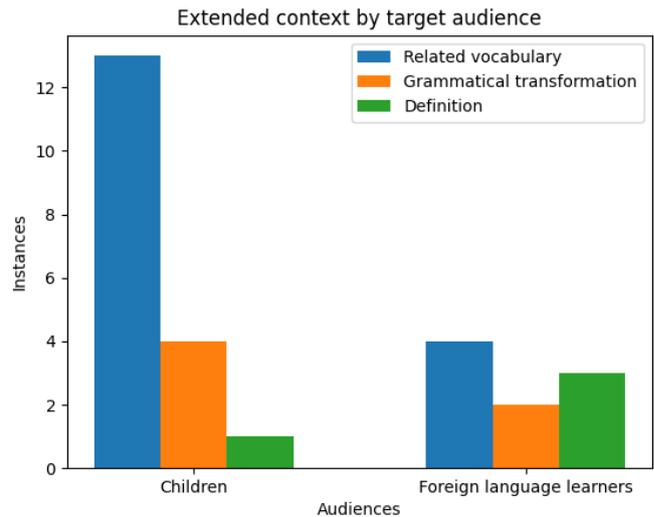


Figure 5: Instances of extended context by target audience

merous, potentially speaking of the importance for learners to acquire knowledge of a new word along with its common collocations and in a common and natural usage scenario (rather than through an explicit definition, for example). In addition, lexical fields are of key importance in the understanding of discourse. The words belonging to one field are used in a systematised pattern throughout a text and may as much as appear synonymously as anaphora referents (e.g. the phrase "I climbed (...)", as used without a direct object being followed by the definitive "The peak(...) in the next sentence) (Christophe et al., 2010). Also, as demonstrated in de Bot et al. (1997)'s experiment, readers make especially high use of discourse content in order to determine the meaning of an unknown lexical item.

Evidence was found that grammatical transformations are significantly more common in morphologically rich languages like Russian and Spanish while, as de Bot et al. (1997) point out, languages like English and French, which have fewer word categories and argument structures, are associated with easier syntactic inference of a new word's meaning. It also appears that a larger variety of extended context is offered to foreign language learners. When it comes to the lack of detected instances of the investigated feature in relation to texts targeted for readers with comprehension difficulties, it is at least partly explainable through the fact that

most of these texts were by a single publisher, Almadraba (Spain) and that they contained designated glossaries, which were not processed as part of the texts.

Finally, the study revealed additional cases of simplified vocabulary and grammar in sentences containing complex vocabulary that are worthy of notice despite not directly leading to an increase in length, as they demonstrate a curious tendency for balance to be sought within the overall complexity present in an abridged textual unit. For instance, an abridged text thus uses a simpler verb as associated with the complex noun 'fragancia': "cuya fragancia hacía recordar" <sup>8</sup> ("whose fragrance made remember"; Sp.) in place of "fragancia que evocaba" <sup>9</sup> ("a fragrance that evoked"; Sp.). Likewise, a negative sentence that contains a relatively complex conjugated verb, like "vous ne me reconnaissez pas ?" <sup>10</sup> ("Don't you recognise me?"; Fr.) is made positive in an abridged text: "vous me reconnaissez ?" <sup>11</sup> ("Do you recognise me?"; Fr.).

## 6 Conclusion and Future Directions

The current study shows that there are cases when abridged literary works make use of extended rather than reduced context, in particular when it comes to the introduction of complex vocabulary. Several distinct types of context extension are detected, and there is evidence that their application and frequency are language-specific. However, further work is required in order to determine the particular cases in which the practice is applied as well as its efficiency. The authors are planning for the latter to be evaluated through a reading comprehension survey, in which pairs of short texts will be presented, one of which will introduce complex vocabulary through extended context in scenarios resembling the ones found within the examined corpus.

To go further, a tendency has been detected for complex vocabulary to influence its context

<sup>8</sup>Dickens, Charles. Canción de Navidad. Abridged edition, edited by S. R. Santerbás, Anaya, 1986.

<sup>9</sup>Dickens, Charles. Canción de Navidad. Ediciones del Sur, 2004.

<sup>10</sup>Hugo, Victor. Les Misérables. Project Gutenberg, 1862.

<sup>11</sup>Hugo, Victor. Les Misérables. Abridged edition, edited by B. Faucard-Martinez, CLE International, 2016.

in ways that are not related to length but to overall complexity in terms of syntax and vocabulary. This claim deserves to be further investigated in order for optimal abridgement practices to be defined that coincidentally render a text simpler and encourage vocabulary acquisition.

## 7 Limitations

Although the utilised corpus has been specifically compiled in a way as to provide a variety of abridged versions of famous literary works, limitations in terms of budget and availability have influenced its extent. In addition, some texts (particularly, Russian-language ones) are not clearly denoted as abridged or full. Finally, the quality of texts is imperfect due to the processes of character recognition and conversion between textual formats; a limitation that has, however, not proven significant.

## References

- Gladis Leonor Arias Rodríguez and Eliana Edith Roberto Flórez. 2018. Using the abridged version of some novels as a way to encourage students' written and oral production. *GiST Education and Learning Research Journal*, (16):6–32.
- Kees de Bot, T. Sima Paribakht, and Marjorie Bingham Wesche. 1997. Toward a lexical processing model for the study of second language vocabulary acquisition: Evidence from esl reading. *Studies in Second Language Acquisition*, 19(3):309–329.
- Gomatam Mohana Charyulu. 2018. Complications in reading abridged texts: a study on cultural destruction by ellipses in meaning-making process. *Review of Research*, 7(9):1–5.
- Anne Christophe, Séverine Millotte, Perrine Brusini, and Elodie Cauvet. 2010. Early bootstrapping of syntactic acquisition. In *Language Acquisition Across Linguistic and Cognitive Systems*, pages 53–66.
- Charles A. Ferguson. 1975. Toward a characterization of english foreigner talk. *Anthropological Linguistics*, 17(1):1–14.
- Zellig Harris. 1988. *Language and Information*. Columbia University Press.
- Amr Helmy Ibrahim. 2013. Une mesure unifiée de la complexité linguistique : l'analyse matricielle définitoire. *Nouvelles perspectives en sciences sociales*, 9:17–80.

- A. Leung, A. Tunkel, and D. Yurovsky. 2021. Parents fine-tune their speech to children’s vocabulary knowledge. *Psychological Science*, 32:975–984.
- Elizabeth Birr Moje, Melanie Overby, Nicole Tysvaer, and Karen Morris. 2008. The complex world of adolescent literacy: Myths, motivations, and mysteries. *Harvard Educational Review*, 78(1):107–154.
- National Literacy Trust. 2006. [Reading for pleasure: A research overview](#). Accessed April 6, 2023.
- Iglika Nikolova-Stoupak. 2023. Introduction of complex vocabulary in literature through fine-tuning: A corpus-based study. *Innovation in Language Learning*, 16.
- Vered Volansky, Noam Ordan, and Shuly Wintner. 2015. [On the features of translationese](#). *Digital Scholarship in the Humanities*, 30(1):98–118.

## Appendix A Corpus Characteristics

| Language | Text                                     | Version    | Length (char.) | Audience                          |
|----------|--|------------|----------------|-----------------------------------|
| English  | <i>Alice in Wonderland</i> <sup>12</sup> | Full       | 144018         |                                   |
| English  | <i>Alice in Wonderland</i> <sup>13</sup> | Abridged 1 | 34836          | FL <sup>14</sup> A2 <sup>15</sup> |
| English  | <i>Alice in Wonderland</i> <sup>16</sup> | Abridged 2 | 53003          | (not indicated)                   |
| English  | <i>A Christmas Carol</i> <sup>17</sup>   | Full       | 158394         |                                   |
| English  | <i>A Christmas Carol</i> <sup>18</sup>   | Abridged   | 39314          | FL B1 <sup>19</sup>               |
| English  | <i>Don Quixote</i> <sup>20</sup>         | Full       | 2167742        |                                   |
| English  | <i>Don Quixote</i> <sup>21</sup>         | Abridged 1 | 130587         | children                          |
| English  | <i>Don Quixote</i> <sup>22</sup>         | Abridged 2 | 81001          | age 8+ <sup>23</sup>              |
| English  | <i>Les Miserables</i> <sup>24</sup>      | Full       | 3199531        |                                   |
| English  | <i>Les Miserables</i> <sup>25</sup>      | Abridged 1 | 1493465        | (not indicated)                   |
| English  | <i>Les Miserables</i> <sup>26</sup>      | Abridged 2 | 166826         | FL C1                             |
| English  | <i>Les Miserables</i> <sup>27</sup>      | Abridged 3 | 1790853        | (not indicated)                   |
| English  | <i>Oliver Twist</i> <sup>28</sup>        | Full       | 887739         |                                   |
| English  | <i>Oliver Twist</i> <sup>29</sup>        | Abridged 1 | 21041          | age 9+                            |
| English  | <i>Oliver Twist</i> <sup>30</sup>        | Abridged 2 | 147207         | FL B2/C1                          |
| English  | <i>Oliver Twist</i> <sup>31</sup>        | Abridged 3 | 22948          | age 7+                            |
| English  | <i>Romeo and Juliet</i> <sup>32</sup>    | Full       | 140481         |                                   |
| English  | <i>Romeo and Juliet</i> <sup>33</sup>    | Abridged 1 | 32375          | children                          |
| English  | <i>Romeo and Juliet</i> <sup>34</sup>    | Abridged 2 | 316247         | (not indicated)                   |
| English  | <i>Romeo and Juliet</i> <sup>35</sup>    | Abridged 3 | 74939          | (not indicated)                   |
| English  | <i>Romeo and Juliet</i> <sup>36</sup>    | Abridged 4 | 19758          | children                          |

<sup>12</sup>Carroll, Lewis. *Alice's Adventures in Wonderland*. Project Gutenberg, 1865.

<sup>13</sup>Carroll, Lewis. *Alice's Adventures in Wonderland*. Abridged edition, edited by J. Bassett, Oxford University Press, 2000.

<sup>14</sup>foreign language

<sup>15</sup>indicated as '700 headwords'

<sup>16</sup>Carroll, Lewis. *Alice's Adventures in Wonderland*. Abridged edition, Sam'l Gabriel Sons and Company, 1916.

<sup>17</sup>Dickens, Charles. *A Christmas Carol*. The Baker & Taylor Company, 1843.

<sup>18</sup>Dickens, Charles. *A Christmas Carol*. Abridged edition, edited by P. Lagendijk, Mediasat Poland Bis, 2004.

<sup>19</sup>1000 headwords

<sup>20</sup>Cervantes, Miguel de. *Don Quixote*. Translated by J. Ormsby, Project Gutenberg, 2004. (Original work published 1605)

<sup>21</sup>Cervantes, Miguel de. *Don Quixote*. Abridged edition, edited by E. Underdown, Thomas Nelson and Sons, 1910.

<sup>22</sup>Cervantes, Miguel de. *Don Quixote*. Abridged edition, edited by K. Casey, Dover Publications, 1979.

<sup>23</sup>where relevant, it is the lowest age for which a text is recommended that is considered

<sup>24</sup>Hugo, Victor. *Les Miserables*. Translated by I. F. Hapgood, Project Gutenberg, 1887.

<sup>25</sup>Hugo, Victor. *Les Misérables*. <https://lesmiserablesabridged.blogspot.com/>.

<sup>26</sup>Hugo, Victor. *Les Miserables*. Abridged edition, easybusyomsk.ru, 2017.

<sup>27</sup>Hugo, Victor. *Les Miserables*. Translated by C. E. Wilbour, edited by F. M. Cooper, Abridged edition, Pranava Books, 2022.

<sup>28</sup>Dickens, Charles. *Oliver Twist*. Project Gutenberg, 1838.

<sup>29</sup>Dickens, Charles. *Oliver Twist*. Abridged edition, translated by G. Tavner, Real Reads, 2009.

<sup>30</sup>Dickens, Charles. *Oliver Twist*. Abridged edition, edited by R. Rogers, Oxford Bookworms, 2009.

<sup>31</sup>Dickens, Charles. *Oliver Twist*. In *Illustrated Stories by Dickens*, edited by M. Sebag-Montefiore, Abridged edition, Usborne, 2009.

<sup>32</sup>Shakespeare, William. *Romeo and Juliet*. Project Gutenberg, 1597.

<sup>33</sup>Shakespeare, William. *Romeo and Juliet*. Abridged edition, edited by M. Lamb and P. Middelboe, Shakespeare Schools Festival.

<sup>34</sup>Shakespeare, William. *Romeo and Juliet*. Abridged edition, Spark Publishing, 2003.

<sup>35</sup>Shakespeare, William. *Romeo and Juliet*. Abridged edition, edited by J. Tanner, The Wichita Shakespeare Co., 2017.

<sup>36</sup>Shakespeare, William. *Romeo and Juliet*. Abridged edition, edited by M. Heyes, Eleanor Palmer Primary

|         |   |            |         |                     |
|---------|---|------------|---------|---------------------|
| English | <i>The Adventures of Tom Sawyer</i> <sup>37</sup> | Full       | 386043  |                     |
| English | <i>The Adventures of Tom Sawyer</i> <sup>38</sup> | Abridged 1 | 141205  | FL                  |
| English | <i>The Adventures of Tom Sawyer</i> <sup>39</sup> | Abridged 2 | 21864   | FL A1 <sup>40</sup> |
| English | <i>The Adventures of Tom Sawyer</i> <sup>41</sup> | Abridged 3 | 32220   | FL A1 <sup>42</sup> |
| French  | <i>Alice in Wonderland</i> <sup>43</sup>          | Full 1     | 183660  |                     |
| French  | <i>Alice in Wonderland</i> <sup>44</sup>          | Full 2     | 187474  |                     |
| French  | <i>Alice in Wonderland</i> <sup>45</sup>          | Abridged 1 | 12235   | age 5+              |
| French  | <i>Alice in Wonderland</i> <sup>46</sup>          | Abridged 2 | 73638   | children            |
| French  | <i>Alice in Wonderland</i> <sup>47</sup>          | Abridged 3 | 64692   | age 9+              |
| French  | <i>A Christmas Carol</i> <sup>48</sup>            | Full 1     | 194592  |                     |
| French  | <i>A Christmas Carol</i> <sup>49</sup>            | Full 2     | 188831  |                     |
| French  | <i>A Christmas Carol</i> <sup>50</sup>            | Abridged   | 6550    | children            |
| French  | <i>Don Quixote</i> <sup>51</sup>                  | Full       | 2391260 |                     |
| French  | <i>Don Quixote</i> <sup>52</sup>                  | Abridged 1 | 93066   | (not indicated)     |
| French  | <i>Don Quixote</i> <sup>53</sup>                  | Abridged 2 | 606051  | (not indicated)     |
| French  | <i>Les Misérables</i> <sup>54</sup>               | Full       | 3674712 |                     |
| French  | <i>Les Misérables</i> <sup>55</sup>               | Abridged 1 | 59606   | FL A2               |
| French  | <i>Les Misérables</i> <sup>56</sup>               | Abridged 2 | 670357  | (not indicated)     |
| French  | <i>Madame Bovary</i> <sup>57</sup>                | Full       | 683958  |                     |
| French  | <i>Madame Bovary</i> <sup>58</sup>                | Abridged   | 60448   | FL B2               |
| French  | <i>Oliver Twist</i> <sup>59</sup>                 | Full 1     | 467498  |                     |
| French  | <i>Oliver Twist</i> <sup>60</sup>                 | Full 2     | 660534  |                     |
| French  | <i>Oliver Twist</i> <sup>61</sup>                 | Full 3     | 932256  |                     |

School, London, 2016.

<sup>37</sup>Twain, Mark. *The Adventures of Tom Sawyer*. Project Gutenberg, 1876.

<sup>38</sup>Twain, Mark. *The Adventures of Tom Sawyer*. Abridged edition, Global Publishing Solutions.

<sup>39</sup>Twain, Mark. *The Adventures of Tom Sawyer*. Translated by J. Kehl, Pearson Education, 2000.

<sup>40</sup>300 headwords

<sup>41</sup>Twain, Mark. *The Adventures of Tom Sawyer*. Abridged edition, edited by N. Bullard, Oxford University Press, 2000.

<sup>42</sup>400 headwords

<sup>43</sup>Carroll, Lewis. *Les Aventures d’Alice au pays des merveilles*. Translated by J. Papy, Gallimard Jeunesse, 2001.

<sup>44</sup>Carroll, Lewis. *Les Aventures d’Alice au pays des merveilles*. Translated by H. Bué, Hachette, 1908.

<sup>45</sup>Carroll, Lewis. *Les Aventures d’Alice au pays des merveilles*. Abridged edition, edited by H. Parisot, Editions Corantin, 1975.

<sup>46</sup>Carroll, Lewis. *Les Aventures d’Alice au pays des merveilles*. Abridged edition, edited by P. Protet, Auzou, 2012.

<sup>47</sup>Carroll, Lewis. *Les Aventures d’Alice au pays des merveilles*. Abridged edition, edited by P. Rouard, Bayard Jeunesse, 1992.

<sup>48</sup>Dickens, Charles. *Conte de Noël*. Translated by A. De Goy and De Saint-Romain, La Bibliothèque électronique du Québec, 1890.

<sup>49</sup>Dickens, Charles. *Conte de Noël*. pitbook.com, 2002.

<sup>50</sup>Dickens, Charles. *Un Chant de Noël*. Abridged edition, edited by L. Papineau, Héritage, 2004.

<sup>51</sup>Cervantes, Miguel de. *Don Quichotte*. Translated by L. Viardot, La Bibliothèque électronique du Québec, 1936.

<sup>52</sup>Cervantes, Miguel de. *Don Quichotte*. Abridged edition, edited by D. Guerin de Bouscal, Theatre classique, 2015.

<sup>53</sup>Cervantes, Miguel de. *Don Quichotte*. Translated by J.-P. Florian, Hachette, 1882.

<sup>54</sup>Hugo, Victor. *Les Misérables*. Project Gutenberg, 1862.

<sup>55</sup>Hugo, Victor. *Les Misérables*. Abridged edition, edited by B. Faucard-Martinez, CLE International, 2016.

<sup>56</sup>Hugo, Victor. *Les Misérables*. La Bibliothèque électronique du Québec.

<sup>57</sup>Flaubert, Gustave. *Madame Bovary: Mœurs de province*. Project Gutenberg, 1857.

<sup>58</sup>Flaubert, Gustave. *Madame Bovary*. Abridged edition, edited by B. Faucard-Martinez, CLE International, 2019.

<sup>59</sup>Dickens, Charles. *Olivier Twist*. Translated by A. Gérardin, Librairie Hachette et Cie., 1893.

<sup>60</sup>Dickens, Charles. *Oliver Twist*. Translated by E. De La Bédolère, Culturea, 2022.

<sup>61</sup>Dickens, Charles. *Oliver Twist*. Translated by A. Girardin, Ink Books, 2017.

|         |   |            |        |                  |
|---------|---|------------|--------|------------------|
| French  | <i>Oliver Twist</i> <sup>62</sup>                 | Abridged   | 22047  | age 10+          |
| French  | <i>Romeo and Juliet</i> <sup>63</sup>             | Full 1     | 166742 |                  |
| French  | <i>Romeo and Juliet</i> <sup>64</sup>             | Abridged 1 | 71651  | (not indicated)  |
| French  | <i>Romeo and Juliet</i> <sup>65</sup>             | Abridged 2 | 19449  | CD <sup>66</sup> |
| French  | <i>The Adventures of Tom Sawyer</i> <sup>67</sup> | Full 1     | 404458 |                  |
| French  | <i>The Adventures of Tom Sawyer</i> <sup>68</sup> | Full 2     | 417108 |                  |
| French  | <i>The Adventures of Tom Sawyer</i> <sup>69</sup> | Abridged   | 7191   | age 7+           |
| Russian | <i>Alice in Wonderland</i> <sup>70</sup>          | Full 1     | 173205 |                  |
| Russian | <i>Alice in Wonderland</i> <sup>71</sup>          | Full 2     | 140446 |                  |
| Russian | <i>Alice in Wonderland</i> <sup>72</sup>          | Abridged 1 | 127433 | (not indicated)  |
| Russian | <i>Alice in Wonderland</i> <sup>73</sup>          | Abridged 2 | 31018  | age 6+           |
| Russian | <i>Madame Bovary</i> <sup>74</sup>                | Full       | 619263 |                  |
| Russian | <i>Madame Bovary</i> <sup>75</sup>                | Abridged   | 610777 | (not indicated)  |
| Russian | <i>Oliver Twist</i> <sup>76</sup>                 | Full       | 854364 |                  |
| Russian | <i>Oliver Twist</i> <sup>77</sup>                 | Abridged   | 74449  | children         |
| Russian | <i>Romeo and Juliet</i> <sup>78</sup>             | Full 1     | 154285 |                  |
| Russian | <i>Romeo and Juliet</i> <sup>79</sup>             | Full 2     | 120633 |                  |
| Russian | <i>Romeo and Juliet</i> <sup>80</sup>             | Abridged   | 33182  | (not indicated)  |
| Russian | <i>The Adventures of Tom Sawyer</i> <sup>81</sup> | Full       | 954134 |                  |
| Russian | <i>The Adventures of Tom Sawyer</i> <sup>82</sup> | Abridged 1 | 218344 | (not indicated)  |
| Russian | <i>The Adventures of Tom Sawyer</i> <sup>83</sup> | Abridged 2 | 374766 | age 11+          |
| Spanish | <i>Alice in Wonderland</i> <sup>84</sup>          | Full 1     | 156666 |                  |
| Spanish | <i>Alice in Wonderland</i> <sup>85</sup>          | Full 2     | 142262 |                  |
| Spanish | <i>Alice in Wonderland</i> <sup>86</sup>          | Abridged   | 66023  | age 10+          |

<sup>62</sup>Dickens, Charles. *Oliver Twist*. Abridged edition, translated by F. Ledoux, edited by P. Arrou-Vignod, Gallimard Jeunesse, 2018.

<sup>63</sup>Shakespeare, William. *Roméo et Juliette*. Translated by F.-V. Hugo, Kiléma, 1868.

<sup>64</sup>Shakespeare, William. *Roméo et Juliette*. Abridged edition, edited by E. Gwénola and P. Fièvre, Theatre classique, 2014.

<sup>65</sup>Shakespeare, William. *Roméo et Juliette*. Translated by A. Hubert, edited by F.-V. Hugo, Kiléma.

<sup>66</sup>readers with comprehension difficulties

<sup>67</sup>Twain, Mark. *Les Aventures de Tom Sawyer*. Translated by P. F. Caillé and Y. Dubois-Mauvais, Ebooks libres et gratuits, 1917. (Original work published 1876)

<sup>68</sup>Twain, Mark. *Les Aventures de Tom Sawyer*. Translated by F. De Gail, Flammarion, 1996.

<sup>69</sup>Twain, Mark. *Les Aventures de Tom Sawyer*. Abridged edition, edited by A. Culleton, Broché, 2020.

<sup>70</sup>Carroll, Lewis. *Alisa v Strane chudes*. Translated by N. M. Demurova, Nauka, 1978.

<sup>71</sup>Carroll, Lewis. *Alisa v Strane chudes*. Translated by Y. Nesterenko, Biblioteka Maksima Moshkova, 2000.

<sup>72</sup>Carroll, Lewis. *Alisa v Strane chudes*. Abridged edition, edited by L. Yahnin, Eksmo, 1991.

<sup>73</sup>Carroll, Lewis. *Alisa v Strane chudes*. Abridged edition, Eksmo, 2018.

<sup>74</sup>Flaubert, Gustave. *Gospozha Bovari*. Translated by A. Chebotarevski, Klub semeynovo dosuga, 2011.

<sup>75</sup>Flaubert, Gustave. *Gospozha Bovari*. Translated by N. Lyubimov. <https://www.100bestbooks.ru/>.

<sup>76</sup>Dickens, Charles. *Priklyucheniya Olivera Tvista*. Translated by A. Krivtsov, Eksmo, 2023.

<sup>77</sup>Dickens, Charles. *Priklyucheniya Olivera Tvista*. Translated by I. O. Rodin, Biblioteka shkol'nika.

<sup>78</sup>Shakespeare, William. *Romeo i Zhuletta*. Translated by D. Mihailovski, educ.ar.

<sup>79</sup>Shakespeare, William. *Romeo i Zhuletta*. Translated by B. Pasternak, Biblioteka dramaturgii agenstva.

<sup>80</sup>Shakespeare, William. *Romeo i Zhuletta*. Abridged edition, edited by Z. Barbe and M. Karre, Orchestre et Choeurs du Théâtre National de l'Opéra de Paris, 1968.

<sup>81</sup>Twain, Mark. *Priklyucheniya Toma Soyera*. Translated by K. Chukovskiy, Kaliningradskoe Knizhnoe Izdatelstvo, 1972.

<sup>82</sup>Twain, Mark. *Priklyucheniya Toma Soyera*. Abridged edition, edited by I. O. Rodin, Biblioteka Shkolnika, 2014.

<sup>83</sup>Twain, Mark. *Priklyucheniya Toma Soyera*. Abridged edition, edited by N. L. Daruzes, Vita Nova, 2014.

<sup>84</sup>Carroll, Lewis. *Las aventuras de Alicia en el país de las maravillas*. Translated by M. Aguirre, Ediciones del Sur, 2003.

<sup>85</sup>Carroll, Lewis. *Las aventuras de Alicia en el país de las maravillas*. Translated by L. Maristany, Titivillus, 1996.

<sup>86</sup>Carroll, Lewis. *Las aventuras de Alicia en el país de las maravillas*. Abridged edition, edited by F. Díez de

|         |  |            |         |                     |
|---------|--|------------|---------|---------------------|
| Spanish | <i>A Christmas Carol</i> <sup>87</sup>             | Full       | 170634  |                     |
| Spanish | <i>A Christmas Carol</i> <sup>88</sup>             | Abridged 1 | 164266  | (not indicated)     |
| Spanish | <i>A Christmas Carol</i> <sup>89</sup>             | Abridged 2 | 40993   | (not indicated)     |
| Spanish | <i>Don Quixote</i> <sup>90</sup>                   | Full       | 2093496 |                     |
| Spanish | <i>Don Quixote</i> <sup>91</sup>                   | Abridged 1 | 212235  | FL B1 <sup>92</sup> |
| Spanish | <i>Don Quixote</i> <sup>93</sup>                   | Abridged 2 | 302115  | (not indicated)     |
| Spanish | <i>Don Quixote</i> <sup>94</sup>                   | Abridged 3 | 1022357 | children            |
| Spanish | <i>Les Miserables</i> <sup>95</sup>                | Full       | 851969  |                     |
| Spanish | <i>Les Miserables</i> <sup>96</sup>                | Abridged   | 15518   | age 12+             |
| Spanish | <i>Madame Bovary</i> <sup>97</sup>                 | Full 1     | 688237  |                     |
| Spanish | <i>Madame Bovary</i> <sup>98</sup>                 | Full 2     | 695263  |                     |
| Spanish | <i>Madame Bovary</i> <sup>99</sup>                 | Abridged 1 | 89352   | CD                  |
| Spanish | <i>Madame Bovary</i> <sup>100</sup>                | Abridged 2 | 661392  | (not indicated)     |
| Spanish | <i>Oliver Twist</i> <sup>101</sup>                 | Full       | 994225  |                     |
| Spanish | <i>Oliver Twist</i> <sup>102</sup>                 | Abridged 1 | 112781  | CD                  |
| Spanish | <i>Oliver Twist</i> <sup>103</sup>                 | Abridged 2 | 95452   | (not indicated)     |
| Spanish | <i>Romeo and Juliet</i> <sup>104</sup>             | Full       | 240967  |                     |
| Spanish | <i>Romeo and Juliet</i> <sup>105</sup>             | Abridged 1 | 78402   | CD                  |
| Spanish | <i>Romeo and Juliet</i> <sup>106</sup>             | Abridged 2 | 127332  | (not indicated)     |
| Spanish | <i>The Adventures of Tom Sawyer</i> <sup>107</sup> | Full       | 415424  |                     |
| Spanish | <i>The Adventures of Tom Sawyer</i> <sup>108</sup> | Abridged 1 | 90334   | (not indicated)     |
| Spanish | <i>The Adventures of Tom Sawyer</i> <sup>109</sup> | Abridged 2 | 135697  | age 11+             |
| Spanish | <i>The Adventures of Tom Sawyer</i> <sup>110</sup> | Abridged 3 | 36930   | CD                  |
| Spanish | <i>The Adventures of Tom Sawyer</i> <sup>111</sup> | Abridged 4 | 95491   | CD                  |

Miranda, Zig Zag, 2018.

<sup>87</sup>Dickens, Charles. Canción de Navidad. Ediciones del Sur, 2004.

<sup>88</sup>Dickens, Charles. Canción de Navidad. Abridged edition, Ediciones la Cueva, 1905.

<sup>89</sup>Dickens, Charles. Canción de Navidad. Abridged edition, edited by S. R. Santerbás, Anaya, 1986.

<sup>90</sup>Cervantes, Miguel de. Don Quijote de la Mancha. Editorial Porrúa S.A. de C.V., 1605.

<sup>91</sup>Cervantes, Miguel de. Don Quijote de la Mancha. Abridged edition, edited by U. Malmlose, Egmont, 1972.

<sup>92</sup>2500 headwords

<sup>93</sup>Cervantes, Miguel de. Don Quijote de la Mancha. Abridged edition, edited by D. Rieu, Librairie des Éditions Espagnoles, 1955.

<sup>94</sup>Cervantes, Miguel de. Don Quijote de la Mancha. IES Maese Rodrigo, 2012.

<sup>95</sup>Hugo, Victor. Les Misérables. Livros grátis, 2013.

<sup>96</sup>Hugo, Victor. Los Miserables. Translated by M. Pujadó, Anaya, 2020.

<sup>97</sup>Flaubert, Gustave. Madame Bovary. Translated by M. Armiño, Siruela, 2014.

<sup>98</sup>Flaubert, Gustave. Madame Bovary. Translated by M. T. G. Urrutia, Hofmiller, 2012.

<sup>99</sup>Flaubert, Gustave. Madame Bovary. Abridged edition, edited by C. S. Baudet, La mar de fácil, 2018.

<sup>100</sup>Flaubert, Gustave. Madame Bovary. Abridged edition, Culturea, 2010.

<sup>101</sup>Dickens, Charles. Las Aventuras de Oliver Twist. Culturea, 2003.

<sup>102</sup>Dickens, Charles. Oliver Twist. Abridged edition, edited by N. Martí Constans, Almadraba, 2016.

<sup>103</sup>Dickens, Charles. Oliver Twist. Abridged edition, Guagua Luna, 2020.

<sup>104</sup>Shakespeare, William. Romeo y Julieta. Biblioteca digital universal, 2003.

<sup>105</sup>Shakespeare, William. Romeo y Julieta. Abridged edition, edited by E. O'Callaghan, Almadraba, 2009.

<sup>106</sup>Shakespeare, William. Romeo y Julieta. Abridged edition, translated by A. L. Pujalte, educ.ar.

<sup>107</sup>Twain, Mark. Las Aventuras de Tom Sawyer. Translated by J. Torroba, Biblioteca Virtual Universal, 2003.

<sup>108</sup>Twain, Mark. Las Aventuras de Tom Sawyer. Abridged edition, edited by B. Palacios, Dirección General de Bibliotecas, 2010.

<sup>109</sup>Twain, Mark. Las Aventuras de Tom Sawyer. Abridged edition, edited by L. I. Barrena, Anaya, 2007.

<sup>110</sup>Twain, Mark. Las Aventuras de Tom Sawyer. Abridged edition, edited by M. A. Rubio and M. Sintés, Almadraba, 2006.

<sup>111</sup>Twain, Mark. Las Aventuras de Tom Sawyer. Abridged edition, edited by M. A. Rubio, Almadraba, 2005.

# Corpus-based Research into Derivational Morphology: A Comparative Study of Japanese and English Verbalization

Junya Morita

Kinjo Gakuin University  
morita@kinjo-u.ac.jp

## Abstract

As part of elucidating the syntax-morphology interaction, this study investigates where and how complex verbs are formed in Japanese and English. Focusing on the Japanese verb-forming suffix *-ka-suru* (e.g. *toshi-o gendai-ka-suru* ‘modernize city’), relevant verbs are extracted from a large-scale corpus and they receive an in-depth analysis from semantic, morphosyntactic, and functional viewpoints. The properties of *-ka-suru* and those of its English counterpart are then compared and contrasted. The result reveals three main points: (i) *-ka-suru* verbs are constantly created in syntactic settings to fulfill the functions of brevity and conceptualization, (ii) while denominal *-ize* derivatives have several submeanings such as ‘result,’ ‘ornative,’ and ‘agentive,’ *-ka-suru* equivalents retain the meaning ‘result,’ and (iii) *-ka-suru* can be combined with compound nouns, but *-ize* cannot. We will demonstrate that the above features originate in the underlying syntactic structure related to each suffix and their difference, thus supporting the thesis of syntactic word formation.

**Keywords:** verb derivation, corpus research, generalization, generative morphology, Japanese, English

## 1 Introduction

It is widely accepted that derived words are divided into two types: non-compositional/ fixed (e.g. *communicate*) and compositional/ productive (e.g. *Hegelianize*). Derived words of the latter type show not only word attributes but also the phrasal attribute of “free composition.” This naturally leads to lively debate on where their formation takes place: lexicalists hold that complex words are made in the lexicon to be inserted into the terminal nodes of a syntactic representation (Di Sciullo and Williams, 1987), while antilexicalists

advocate their direct syntactic derivation (Embick, 2010). As part of the investigation of the syntax-morphology relationship, the present study attempts to show where and how *-ka-suru* derivatives are produced in Japanese, exemplified in (1) and (2). They are picked out from Balanced Corpus of Contemporary Written Japanese (BCCWJ)<sup>1</sup>.

- (1) *ji-kokumin-o moomai-ka-suru*  
one’s-people-ACC ignorant-change-do  
‘make one’s people ignorant’
- (2) *shinikaketa momiji-o bonsai-ka-suru*  
dying maple-ACC bonsai-change-do  
‘turn a dying maple into a bonsai’

The complex suffix *-ka-suru* consists of two bound forms, *-ka* and *-suru*. The verb *moomai-ka-suru* in (1) indicates that the complex suffix *-ka-suru* attaches to an adjective (*moomai*) to derive a verb which expresses a causative meaning. The example in (2) illustrates the same point: the causative verb *bonsai-ka-suru* implies that the object’s referent is affected in some way and becomes the resulting state indicated by the base noun. This causative construction is lexical in that the relevant verbal expression is derived by the addition of the bound morpheme *-ka-suru* to an adjective. Moreover, the derivatives *moomai-ka-suru* and *bonsai-ka-suru* are hapax legomena—token frequency 1—of the 105-million-word corpus, suggesting that they are constructed online in working memory. Verbalization by *-ka-sru* is therefore judged to be a compositional/productive type of word formation process.

Although there are a handful of representatives of descriptive studies like Takubo (1986), there has been no systematic analysis of *-ka-suru* verbs. The aim of this study is to make a comprehensive analysis of them. In section 2, we expose the

<sup>1</sup> All the well-formed *-ka-suru* derivatives given below are detected in BCCWJ, although their original sentences are often simplified for convenience. Additionally, the case

markers *Accusative* and *Nominative* are respectively abbreviated as *ACC* and *NOM*.

communicative functions performed by *-ka-suru* verbs. Section 3 deals with the productivity of *-ka-suru* verbalization and its English counterpart. Section 4 illuminates the semantic properties of verbs in *-ka-suru* and compares them with the properties of their English equivalents. After elucidating the differences in their base structures in section 5, section 6 explains the origin of the semantic and formal differences between these derived verbs from an antilexical standpoint.

## 2 Functional Properties

This section deals with the communicative functions fulfilled by *-ka-suru* verbalization. What motivates the use of complex words rather than corresponding phrases? Two functions are particularly important: brevity and conceptualization. The first function of *-ka-suru* verbalization is to serve as a device for indicating brevity: concise and sensible use of words (cf. Clark and Clark, 1979). In example (3), the action inferred from the preceding context is concisely expressed by the temporally constructed verbs *shihei-ka-suru* and *shitauke-ka-suru*. A “brevity effect” thus results from this verbalization, which is not obtained from the related verbose paraphrase *kokuzeikyoku-o shihei-no (yoona) jootaini suru* ‘make NTB (like) private army.’

- (3) *kokuzeikyoku-o shihei-ka-shi,*  
 NTB-ACC private.army-change-do  
*gyokai-o shitauke-ka-suru*  
 industry-ACC subcontract-change-do  
 ‘make National Tax Bureau private army,  
 make the industry subcontracted’

We turn next to the second function of *-ka-suru* verbalization. A verb in *-ka-suru* has a lexical attribute in that it conceptualizes a causative action with a resulting state change for the patient. Conceptualization is achieved through naming. In example (4), a special conception is produced by compressing a predicate content into the verb *rittaikoosa-ka-suru*; the act of turning a road into a grade-separated intersection is labeled by the innovated verb. In contrast, the act of making one’s hands crossed is not lexicalized by a *-ka-suru* verb, as seen in (5). This is because the action at issue is not worthy of the name.

- (4) *dooro-o rittaiakoosa-ka-suru*

*road-ACC grade.separation-change-do*  
 ‘turn a road into a grade-separated  
 intersection’

- (5) *ryoote-o koosa-ka-suru*  
 both.hands-ACC cross-change-do  
 ‘make one’s hands crossed’

To sum up, we have revealed two important communicative functions of *-ka-suru* verbalization: brevity and conceptualization. These functions demonstrate the lexical attributes of *-ka-suru* expressions.

## 3 Productivity

Productivity is defined as the extent to which a word formation device can give rise to new words (Lieber, 2010: 59). There are several approaches to quantifying productivity, the most reliable of which is one putting great importance to hapax legomena of a large-scale corpus (Baayen and Renouf, 1996). This rests on the view that complex forms that have been observed only once in a large corpus are highly likely to be lexical innovations and hence the capacity of a word formation rule to create neologisms crucially involves the degree to which the rule produces words with extremely low frequency (Hay, 2003). Baayen and Renouf (1996: 73) propose a productivity measure:  $Productivity (P) = n_1/N$ , where  $n_1$  is the number of hapaxes and  $N$  is the total number of tokens. Here we revise it so as to place the total number of types (but not tokens) in the denominator; thus,  $P = n_1/V$  ( $V$ : the number of word types). Our measure rests on the view that the productivity of a particular process is reflected in the type frequency of the process (Goldberg, 1995: 134-139).

According to the proposed measure, we calculate the productivity value of *-ka-suru* derivation, based on Balanced Corpus of Contemporary Written Japanese, a 105-million-word corpus. Our BCCWJ research has attested 130 hapaxes and 516 word types of *-ka-suru* derivatives, and so its productivity value is 0.252<sup>2</sup>. In our view, the productivity of *-ka-suru* is defined as the potentiality of creating one new word when four types of *-ka-suru* verbs are used. It then turns out that *-ka-suru* verbalization is very productive; when we derive four types of verbs by adding *-ka-*

<sup>2</sup> For collecting *-ka-suru* verbs, 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\\_c](https://chunagon.ninjal.ac.jp/2Fj_spring_cas_security_c) heck.

*suru* to an adjective or noun, one of them is newly created.

In addition, words may be coined on the spot in the syntactic contexts of comparison and contrast. Specifically, a *-ka-suru* coinage may be constructed when similar activities are compared. In example (6), the verb *soo-heishi-ka-suru* ‘make themselves all soldiers’ is created in conjunction with the prior comparable phrase *buki-o tori* ‘take up arms.’ A similar observation holds for sentence (7): a series of related activities are conceptualized and lexicalized by four verbs in *-ka-suru*. Among them *kooshinrai-ka-suru* is a hapax, i.e. an innovated verb. As laid out in section 2, this *-ka-suru* verb is coined to carry out the functions of “brevity” and “conceptualization.” The fact that *-ka-suru* final verbs may be coined whenever these functions are required indicates the high creativity of the verb derivation.

(6) *shimin-ga buki-o tori, soo-heishi-ka-shita*  
 citizens-NOM arms-ACC hold all-soldier-  
 change-do.past  
 ‘citizens took up arms and made themselves all soldiers’

(7) *tanmatsukiki-o takinoo-ka, koosoku-ka, kooshinrai-ka, kooseinoo-ka-suru*  
 terminal.equipment-ACC multifunctional-  
 change very.fast-change highly.releable-  
 change highly.efficient-change-do  
 ‘make terminal equipment more multi-  
 functional, very fast, highly reliable, and highly efficient’

Following the calculation method just presented, the productivity values of English verb-forming processes were calculated, based on the British National Corpus (BNC), a 100-million-word corpus (Morita, 2022: 93). As a result, it turned out that the productivity values of *-ize*, *-ify*, *-en*, and *-ate* affixations are 0.323, 0.074, 0.035, and 0.058, respectively. The result shows that *-ize* verbalization is quite productive while the other three verbalizations are not. It is thus confirmed that the English equivalent of *-ka-suru* for productivity is *-ize*, and that creative verbalizers such as *-ka-suru* and *-ize* exist in Japanese and English<sup>3</sup>.

<sup>3</sup> Although not discussed in this article, a group of unproductive verb-forming suffixes exist in Japanese: *-meru*, *-maru*, and *-mu* (cf. *shizu-meru* (calm-make ‘calm down’),

## 4 Semantic Properties

### 4.1 Adjective/Noun bases

The suffix *-ka-suru* attaches to adjectives, nouns, and VNs (verbal nouns). This section deals with some semantic aspects of *-ka-suru* derivatives, focusing on adjective and noun bases. The bound morpheme *-ka* has the meaning of ‘change’ or ‘conversion’ and bound morpheme *-suru* has no semantic content and simply functions as a verb-forming suffix. Hence, the complex suffix *-ka-suru* derives a causative verb that stands for ‘change the state or quality of an entity so that it becomes something different from what it was before.’

Let us first consider the case where *-ka-suru* attaches to adjectives. An example is *kaihatsukyoka-o danryoku-ka-suru* (development.permission-ACC flexible-change-do), where *-ka-suru* combines with the adjective *danryoku(-no)* ‘flexible’ to form the verb *danryoku-ka-suru* ‘change (development permission) into being flexible’ or ‘make (development permission) flexible.’ Thus, deadjectival *-ka-suru* verbs essentially convey transition to a resulting state by causation. We refer to this meaning as ‘result.’

This prototypical meaning of *-ka-suru* verbs leads to a semantic condition of the base; *-ka-suru* cannot attach to adjectives that signify an ‘unchangeable quality or state.’ Among them is a class of adjectives called *jootaigo* ‘emotional word’ such as *onwa(-na)* ‘gentle’ and *rippa(-na)* ‘respectable,’ which designates the condition of the mind (Morioka, 1986: 12). Thus, we cannot say *\*Taro-o shinsetsu-ka-suru* (Taro-ACC kind-change-do ‘make Taro kind’). The adjective *shinsetsu(-na)* ‘kind’ denotes an inherent quality of human beings and it generally cannot be changed, and hence it contradicts the causative meaning of *-ka-suru*.

Turning now to noun-based *-ka-suru* derivatives, they basically have the same meaning as those based on adjectives. For example, *namagomi-o taihi-ka-suru* (kitchen.waste-ACC compost-change-do) is interpreted as ‘change kitchen waste into compost.’ We can see here that the object’s referent is affected to become the resulting state expressed by the base noun.

In this context, it is interesting to compare Japanese verb formations with those in English. English has a set of suffixes that derive verbs from adjectives and nouns (*-ize*, *-ify*, *-en*, *-ate*). Among

*atata-maru* (warm-become ‘warm up’), and *yuru-mu* (loose-become ‘loosen’).

them, *-ize* is the most productive suffix (cf. §3) and hence we will compare *-ize* and *-ka-suru*. One notable semantic difference is that denominal *-ize* derivatives have several submeanings, whereas their Japanese counterparts essentially retain the original meaning. To confirm this point, look at Table 1<sup>4</sup>.

| base meaning of derivative | <i>-ka-suru</i> | <i>-ize</i> |
|----------------------------|-----------------|-------------|
| Adj result                 | 129 (25.0%)     | 215 (58.7%) |
| N result                   | 325 (63.0%)     | 54 (14.7%)  |
| ornative                   | 2 (0.4%)        | 35 (9.6%)   |
| agentive                   | 0 (0%)          | 7 (1.9%)    |
| instrumental               | 0 (0%)          | 5 (1.4%)    |
| similitive                 | 6 (1.2%)        | 15 (4.1%)   |
| VN result                  | 54 (10.4%)      | n/a         |
| acN performative           | n/a             | 35 (9.6%)   |
| total num. of types        | 516 (100 %)     | 366 (100%)  |

Table 1: The submeanings of *-ka-suru* and *-ize*

Table 1 presents the semantic classification of the two suffixes attaching to three kinds of base forms: adjectives, (non-action) nouns, and VNs/action Ns. For instance, our BCCWJ survey has detected 129 word types of deadjectival *-ka-suru* verbs and a BNC survey has discerned 215 word types of deadjectival *-ize* verbs. Both derivatives refer to the ‘result’ meaning of ‘cause sth/sb to become X,’ as in *sutajio-o gendai-ka-suru/modernize studio* and this sense is the most commonly used<sup>5</sup>. The case where the suffixes in question are added to Japanese VNs or their English counterparts (acNs) will be discussed in the next section.

As can be seen from Table 1, the reading of ‘result’ is central to noun-based verbs in both languages. Now, what is noteworthy about the two kinds of denominal verbs is that while there are a certain number of *-ize* derivatives that express meanings other than ‘result,’ their Japanese equivalents are almost never attested. Firstly, *-ka-suru* verbs rarely express an ‘ornative’ meaning: ‘change Y so that X is given.’ While an *-ize* verb

expresses an ‘ornative’ meaning, as in *accessorize the dress* ‘change the dress so that accessory is given,’ its Japanese counterpart cannot, as in *\*doresu-o akusesari-ka-suru* (dress-ACC accessory-change-do). Secondly, *-ka-suru* verbs do not bear an ‘agentive’ sense: ‘change Y into a state where something is being done by X.’ Although *-ize* derivatives can bear an ‘agentive’ sense, as in *patronize the shop* ‘change the shop into a state where support is being done by patron,’ the comparable Japanese derivatives (*\*mise-o patoron-ka-suru*) are not acceptable. Finally, *-ka-suru* verbs do not indicate an ‘instrumental’ reading: ‘change Y into a state where something is being done with X.’ While verbs in *-ize* bear an ‘instrumental’ reading as in *cauterize the wounds* ‘change the wounds into a state where treatment is being given with cauter,’ verbs in *-ka-suru* cannot have this meaning as in *\*kizu-o yakigote-ka-suru* (wounds-ACC cauter-change-do)<sup>6</sup>.

Why is it that pertinent semantic extensions tend to take place in English but not in Japanese? The suffix *-ize* forms a causative verb that means ‘make Y (be) X,’ which entails that the surface object Y is made to have some relation to the base noun X. Here, several relationships such as ‘ornative’ and ‘agentive’ are possible, although the ‘result’ relationship is the most common and natural one. Regarding the comparable Japanese suffix, however, the situation is different. The Japanese causative suffix *-ka-suru* has the comparable content word *baker(u)*. They share the ideographic (Chinese) character 化, meaning ‘transform.’ This written form functions as a suffix when pronounced in a pseudo-Chinese manner (called *onyomi*), [ka], while it functions as an independent verb when pronounced in a Japanese manner (called *kunyomi*), [baker(u)]. The suffix under consideration can be understood to have developed from the cognate synonym *baker(u)* and the bound form itself signifies the meaning ‘transform.’ The verb *transform* entails a resulting state: ‘affect Y and change it to the state

<sup>4</sup> The data on *-ize* suffixation rests on the BNC research done in a previous study (Morita, 2022: 90-91). Additionally, the data on *-ka-suru* suffixation is based on our BCCWJ survey.

<sup>5</sup> Part of the resultative causatives can be intransitive verbs that mean ‘become X’ (inchoative), as in *seiji-o kotei-ka-suru* ‘make the politics fixed’ and *kengen-no sumiwake-ga koteika-suru* ‘the division of authority becomes fixed.’ This shift has been discussed as an alternation between transitive and inchoative verbs (Levin and Rappaport Hovav, 1995). We will not, though, discuss how to connect the two classes of verbs.

<sup>6</sup> As exhibited in Table 1, two exceptions are discerned in BCCWJ, where verbs in *-ka-suru* are extended to denote the meaning of ‘ornative’ (cf. *dairekutomeeru-o raberu-ka-suru* (direct.mail-ACC label-change-do ‘give a label to direct mail’). Additionally, six ‘similitive’ *-ka-suru* verbs are detected, as in *Tsubasa-ga ooji-ka-suru* (Tsubasa-NOM prince-change-do), meaning ‘Tsubasa becomes like a prince.’ The base noun seems involved in a kind of metaphorical extension: ‘change Y into someone with prince properties.’ Since the verbs in question express a ‘result’ reading, what they express is consistent with the fundamental meaning of *-ka-suru* verbs.

X, X representing an outcome state’ (cf. the movie *transformed* her overnight from an unknown schoolgirl into a megastar).

From the above observations, we are justified in asserting that the lack of the relevant semantic extension in Japanese stems from the fact that the suffix *-ka*, which originates in a Sino-Japanese ideographic free form, continues to retain its primal meaning.

#### 4.2 VN bases

Let us go on to the case where the base of *-ka-suru* is VN (which stands for verbal noun). This Japanese-specific category is defined as a stem that has a predicate function and designates the meaning ‘action or process.’ The cardinal meaning of *VN-ka-suru* verbs has three component parts: ‘(a) make Y (b) be in the result of (c) being VN-ed.’ The *VN-ka-suru* construction then denotes causation, as derivable from the semantic element of (a) ‘make Y’: ‘cause Y to become.’ The semantic component of (b) ‘be in the result of’ implies that the patient Y becomes a certain state. And the semantic element of (c) ‘being VN-ed’ shows a passive sense in that the patient Y undergoes the action or process expressed by the base VN.

The meaning of the *VN-ka-suru* construction becomes clearer when contrasted with the meaning of the light verb construction, i.e. *VN-suru* construction. The light verb construction of (9) has a non-causative eventive reading: the object *yasei-shokubutsu* simply takes on the action expressed by the verb *saibai-suru*. In comparison, the *-ka-suru* construction of (8) has a causative eventive reading: the object (*yasei-shokubutsu*) receives the action of the verb (*saibai-ka-suru*) and changes to a certain state (*being grown in a field*). It focuses on the result of the process (Takubo, 1986: 82).

(8) *yasei-shokubutsu-o saibai-ka-suru*  
 wild-plant-ACC grow-change-do  
 ‘domesticate wild plants’

(9) *yasei-shokubutsu-o saibai-suru*  
 wild-plant-ACC grow-do  
 ‘grow wild plants’

Regarding the distinction of *-ka-suru* from *-suru*, there is a notable semantic correlation between Japanese and English verbalization. When

the base noun represents ‘action or process,’ an *-ize* verb has a ‘performative’ meaning (cf. Table 1), but not a causative meaning. (The base form *action noun* is abbreviated as *acN* in Table 1.) In the verb phrase *monopolize the soft drink market*, for instance, the verb *monopolize* has the reading of simple action ‘do monopoly’ and does not have the causative reading ‘make the market be in monopoly.’ Similarly, *anatomize (corpses)* stands for ‘do anatomy’ and *plagiarize (a work)* means ‘do plagiary,’ not ‘make (corpses) be in anatomy’ or ‘make (a work) be in plagiary.’ It can thus be seen that the *-ize* verbs in discussion correspond to the Japanese *-suru* light verb in (9), but not to the *-ka-suru* causative verb in (8).

To summarize, in deriving verbs from action nouns, Japanese distinguishes between causative (*VN-ka-suru*) and non-causative (*VN-suru*) forms. In English, on the other hand, there is no such distinction and the derived verbs are always non-causative verbs. Consequently, while *-ize* verbs formed from non-action nouns become causative verbs (‘make Y (be) X’ e.g. *atomize*), *-ize* verbs derived from action nouns express non-causative simple action (‘do X’ e.g. *anatomize*).

#### 5 The Internal Structure of Bases

This section considers *-ka-suru*’s base from a structural viewpoint. Differences can be found between Japanese and English regarding the base structures of derived verbs. Let us first examine the Japanese case. The verbalizer *-ka -suru* can be attached to compound nouns. A good example of this is the verb phrase *NHK-o kyodai-seisaku-gaisha-ka-suru* (NHK-ACC huge-production-company-change-do ‘turn NHK into a huge production company’), where *-ka-suru* is added to the compound noun *kyodai-seisaku-gaisha*. *-Ka-suru* can also be affixed to a compound VN (*kooso-shuuyaku*), as in *tatemono-o koosoo-shuuyaku-ka-suru* (building-ACC high.rise-concentration-change-do ‘make buildings taller and more concentrated’)<sup>7</sup>.

Furthermore, a base with which *-ka-suru* combines may be a prefixed adjective or noun. In the example of *osenbushitsu-o mu-gai-ka-suru* (pollutant-ACC un-harmful-change-do ‘make pollutants un-harmful’), *-ka-suru* is suffixed to the adjective *mu-gai(-no)* which includes the negative prefix *mu-*. By the same token, *-ka-suru* adjoins to the noun *hi-hanzai* with the negative prefix *hi-*, as

consequently they are difficult to appear inside *-ka-suru* verbs.

<sup>7</sup> There are very few Sino-Japanese compound adjectives in Japanese (cf. *\*ganseki-kengo(-na)* ‘rock-solid’), and

in *tanjun-tobaku-o hi-hanzai-ka-suru* (simple-gambling-ACC non-crime-change-do ‘make simple gambling a non-crime’).

Turning now to the case of English, unlike Japanese, the bases of verb-forming suffixes are limited to smaller sizes. It has been pointed out that English verbalizers do not combine with compounds, and this is confirmed by a BNC survey (cf. Morita, 2022: 92). For example, the combination of *novel* and *-ize* makes *novelize*, whereas *-ize* cannot be associated with the compound *detective novel* to yield *\*detective novelize*. Likewise, with an *-ize* verb whose base is an action noun, the *-ize* construction does not allow the base noun to become a compound (cf. *\*foot-anatomize*). Additionally, a BNC survey indicates that English verb-forming suffixes generally do not attach to prefixed bases. For instance, a verbalizer cannot connect to *atypical*, *transcontinental*, or *ultratrendy* to produce *\*atypicalize*, *\*transcontinentalize*, or *\*ultratrendify*, respectively (Morita, 2022: 92).

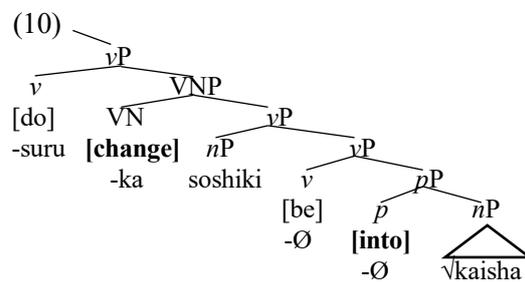
In sum, the corpus-based surveys show that Japanese and English verb-forming suffixes differ in a morphological respect: *-ka-suru* can incorporate compound nouns and prefixed adjectives/nouns, but English verbalizers cannot. The origin of this difference will be accounted for in the next section.

## 6 Theoretical Implications

In the generative theories ranging from the Standard model to the Government-Binding theory, an excessive role was given to the lexicon. It stores not only underived words and stems but also word formation rules for deriving complex words from them. Each simple word and derived word is inserted into a relevant node of a phrase structure to create a D-structure. The point is that word formation only takes place within the lexicon, hence this approach is called lexicalism. However, theoretical and empirical evidence contrary to lexicalism was pointed out (Marantz, 1997; Harley and Noyer, 2000), leading to the rise of antilexicalism. In antilexical Distributed Morphology (DM), a major part of word formation is located in syntax. Specifically, DM attributes the core characteristics of a complex word to its syntactic structure while entrusting the role of its

formal processing to the morphology module. In what follows, we will make an argument for supporting the DM model: after illustrating how the *-ka-suru* construction is realized syntactically, we will show that the properties of the construction observed above follow straightforwardly from the related underlying structure.

Given the DM theoretic viewpoint that derivational suffixes are the heads of phrases, the underlying configuration of *karera-ga soshiki-o kaisha-ka-suru* (they-Nom organization-ACC company-change-do ‘they turn an organization into a company’) are constructed by the merging of a root ( $\sqrt{\text{kaisha}}$ ) and category-defining heads like *v*, *n*, and *p*, as depicted in (10)<sup>8</sup>.



Tree diagram (10) exhibits a causative construction, where the surface object *soshiki* has a predicative relationship with the underlying nominal root  $\sqrt{\text{kaisha}}$  and the “small clause” is dominated by the causative element *-ka*. Thus, the core meaning of the *-ka-suru* construction, ‘change Y into X,’ is derived from its underlying structure (cf. §4.1).

In section 2, we discussed that *-ka-suru* expressions are generated in the comparison/contrast contexts in order to obtain the “brevity” effect (cf. (3)). This context-dependent word formation harmonizes with an approach of deriving *-ka-suru* expressions from their syntactic structure rather than forming them in the lexicon. It was also observed in section 2 that verbs in *-ka-suru* are produced to conceptualize and name causative actions with resulting states. This lexical function distinguishes between noteworthy expressions such as (4) and non-noteworthy expressions like (5). The structural features of the *-ka-suru* construction presented above (cf. (10)) are again the source of its lexical property—conceptualization by naming. Because the affix *-ka* inserted into VN is a category-

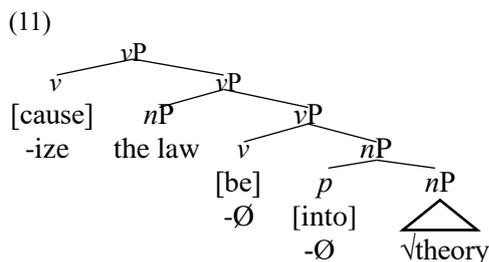
<sup>8</sup> As evidence that the complex suffix *-ka-suru* is separated in underlying structure, we can point out a *-ka-suru* expression in which *-suru* incorporates a coordinate structure: [*koodo-ka•fukuzatsu-ka*]-*suru* ([advanced-change•complex-

change]-do ‘make (sth) advanced and complex’) (see also example (7)).

changing derivational affix, it plays the role of forming a verbal noun.

Furthermore, section 3 demonstrated the creativity of *-ka-suru* affixation: it has the potential to constantly create new expressions on the spot (cf. (6) and (7)). This is a corollary of the idea that syntactic computation is entirely productive and that the core structure of *-ka-suru* expressions is constructed at the syntactic level.

Section 4 clarified the semantic differences between Japanese and English verbal suffixes. A notable difference is that denominal *-ize* derivatives have several submeanings while their *-ka-suru* counterparts retain their prime meaning (cf. Table 1). For example, *-ize* can express the meaning ‘ornative,’ whereas its Japanese equivalent can hardly express it (cf. *accessorize the dress* and *\*doresu-o akusesarii-ka-suru*). This difference can be traced back to a structural distinction between the two languages. Let us now compare the related underlying structures. The underlying configuration of the *-ize* construction *they theorize the law* is presented in (11). From this representation, the *-ize* expression can be interpreted as ‘to cause the law to become a theory.’



What creates the semantic contrast between Japanese and English verbal suffixes is their structural difference, especially the difference in the features of the verbalizers. English verbalization structure (11) implies ‘to cause the object to be in a certain state’; the verbalizer *-ize* with the feature [cause] allows the referent of the object to be in several states. Such states are represented by the spatial and functional relationships expressed by specific prepositions in under-lying structures. Therefore, the reading of ‘ornative’ can be obtained from essentially the same configuration as (11), with the exception that the feature of *p* is changed from [into] to [with]. In Japanese, on the contrary, the

nominalizer *-ka* in (10) has the feature [change], and accordingly the feature of the associated preposition is determined to be [into]. Consequently, denominal *-ka-suru* verbs can only be interpreted as ‘result.’

Section 5 identified a difference in the size of the bases in Japanese and English verbal-ization: *-ka-suru* can be attached to compound nouns and prefixed adjectives/nouns, but its English counterpart cannot. We will show that this contrast is also attributable to the structural difference between *-ka-suru* and *-ize* expressions, focusing on the incorporation of compound nouns. As evidenced in (12), English compound nouns cannot be incorporated to form derived verbs. In contrast, the Japanese verb-forming suffix *-ka-suru* can incorporate compound nouns, as in (13).

- (12) \*They information-theorize the law.
- (13) B-sha-o tokushu-gaisha-ka-suru  
       B-Company-ACC special-company-change-  
       do  
       ‘turn B Company into a special company’

Verbalization in English is subject to stronger restrictions than other categorizations. As is commonly known, English verbs cannot generally be combined with other lexical categories to form compound verbs (cf. *\*rock-throw/\*fast-walk*). This point is similar in Japanese; it is generally not possible to directly form a compound verb by combining a noun or adjective with a verb. Therefore, just like in English, you cannot say *\*ishi-nageru* (rock-throw) or *\*haya-aruku* (fast-walk)<sup>9</sup>.

Interestingly, Marchand (1969: 100-101) states that English compound verbs cannot be directly constructed, but can be built from compound nouns through the processes of zero-derivation and backformation. For example, we can derive the compound verb *spotlight* by adding a zero-morpheme to the compound noun *spotlight* and we can also form the compound verb *window-shop* by removing the ending *-ing* from the compound noun *window-shopping*. Note that zero-morpheme and *-ing* are pure category changers, which have no lexicosemantic content and only serve to transform one category into another.

Verbs ending in *-ka-suru* are derived in a

(see-lose ‘fail to notice’), where the second verb is limited to a specific verb and the entire compound is not necessarily compositional.

<sup>9</sup> In English, two verbs are sometimes combined to form a compound verb, as in *push-walk her* (J. Rossner, *Looking for Mr. Goodbar*, p. 71). Japanese also has V-V type compound verbs, as in *oki-wasureru* (put-forget ‘mislays’) and *mi-otosu*

similar way to this. It is worth noting here that the verbalizer *-ka-suru* is a complex suffix. In structure (10) above, the verb is not directly formed by adding the complex suffix to a complex noun, but rather the nominal suffix *-ka* is used first to form a (verbal) nominal and then it is verbalized by the suffix *-suru*. Example (14) illustrates a case in point: the complex verb *kanzen-kogaisha-ka-suru* is formed by adding the semantically empty category changer *-suru* to the compound noun *kanzen-kogaisha-ka*.

- (14) B-sha-o kanzen-kogaisha-ka-suru  
 B-Company-ACC complete-subsiary-  
 change-do  
 ‘make B Company a wholly owned company’

To summarize this section, Japanese and English have a common feature in that verbs are built from compound nouns by suffixing semantically empty category-changers (zero-morpheme/*-suru*) to them. Since a pure category changer does not cause a semantic change in the base, it functions to change only the category while leaving the base in its original form. Contrastively, the suffix *-ize* in (11) has a causative meaning and is not a simple category changer. Accordingly, *-ize* suffixation is restricted to smaller units.

## 7 Conclusion

Based on an in-depth analysis of Japanese *-ka-suru* derivatives extracted from a large-scale corpus, we have uncovered their semantic, morphological, and functional properties. Specifically, it is revealed that (i) *-ka-suru* derivation gives rise to the causative meaning ‘to bring the object’s referent into a certain resulting state,’ (ii) it is creative enough to coin a variety of neologisms depending on the context, and (iii) it has the functions of conceptualization and naming.

Moreover, comparing them with those of *-ize* verbs, their semantic and morphological differences have been identified. Semantically, *-ka-suru* verbs retain the basic meaning of ‘result,’ while verbs in *-ize* have several submeanings such as ‘ornative’ and ‘agentive’ besides the basic meaning. Morphologically, *-ka-suru* can be attached to larger-sized bases such as compounds or prefixed forms, whereas *-ize* cannot be attached to them.

Then we have demonstrated that the characteristics of *-ka-suru* verbs and the differences of both verbalizations are derived in a unified manner from their underlying syntactic

structures. The core meaning, creativity, and naming function of *-ka-suru* derivation come respectively from its underlying structure, syntactic formation, and the suffix (*-ka-suru*) that governs a verb phrase. Turning to the differences in verbalization between Japanese and English, the fact that only *-ka-suru* verbs retain the meaning of ‘result’ has its origin in the difference in the underlying representations of the two verbalizations; only the Japanese suffix involves the feature [change]. The second difference regarding the size of the base is deduced from whether the verbalizers are pure category-changers or not, that is, from the difference in features between *-suru* ([do]) and *-ize* ([cause]).

This study thus provides strong support for the antilexicalist position, which claims that major properties of complex words can be traced back to the syntactic level, with the power to generate infinite set of sentences. How a syntactic output is constructed into a word form awaits further investigation; there is a need for elucidation of the lexical entries of the verbalizer *-ka-suru* and morphological operations for word make-up.

## Acknowledgments

I would like to express my gratitude to two anonymous reviewers for their valuable comments and suggestions on an earlier draft of this paper. This work is partly supported by a Grant-in-Aid for Scientific Research (C) (No. 22K00562) from the Japan Society for the Promotion of Science.

## References

- Harald R. Baayen and Antoinette Renouf. 1996. Chronicling *the Times*: productive lexical innovations in an English newspaper. *Language*, 72:69-96.
- Eve V. Clark and Herbert H. Clark. 1979. When nouns surface as verbs. *Language*, 55:767-811.
- Anna-Maria Di Sciullo and Edwin Williams. 1987. *On the Definition of Word*. MIT Press, Cambridge, MA.
- David Embick. 2010. *Localism versus Globalism in Morphology and Phonology*. MIT Press, Cambridge, MA.
- Adele E. Goldberg. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press, Chicago.
- 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.
- Jennifer Hay. 2003. *Causes and Consequences of Word Structure*. Routledge, New York.

- Beth Levin and Malka Rappaport Hovav. 1995. *Unaccusativity: At the Syntax-Lexical Semantics Interface*. MIT Press, Cambridge, MA.
- Rochelle Lieber. 2010. *Introducing Morphology*. Cambridge University Press, Cambridge.
- 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*, 2<sup>nd</sup> ed. C. H. Beck, München.
- Kenji Morioka. 1986. *Setsuji to joji* (Affixes and prepositional words). *Nihongogaku* (Japanese Linguistics), 5.3:11-18.
- Junya Morita. 2022. Corpus-based research into verb-forming suffixes in English: its empirical and theoretical consequences. In *Proceedings of the Fifth International Conference, Computational Linguistics in Bulgaria*, pages 89-97.
- Yukinori Takubo. 1986. *-Ka*. *Nihongogaku* (Japanese Linguistics), 5.3:81-84.

# The Verbal Category of Conditionality in Bulgarian and Its Ukrainian Correspondences

Ivan Derzhanski, Olena Siruk  
Institute of Mathematics and Informatics  
Bulgarian Academy of Sciences  
{iad58g,olebosi}@gmail.com

## Abstract

Modern Bulgarian shares a conditional mood with the other Slavic languages, but it also has developed a future-in-the-past tense which is structurally analogous to many Western European languages' category traditionally called a conditional mood in their grammars. The distinction between these two forms is sometimes elusive and can be difficult for native speakers of Slavic languages who are learning Bulgarian. In this paper we consider the uses of the Bulgarian conditional mood and future-in-the-past tense in a parallel corpus of Bulgarian and Ukrainian text, examining the corresponding wording in Ukrainian, where the conditional mood is supplemented by modal verbs, and discuss the breadth of choices open to translators when working in each direction.

**Keywords:** conditional mood, future in the past, tense, parallel corpus, Bulgarian language, Ukrainian language, translation.

## 1 Introduction

Modern Bulgarian shares with the other Slavic languages a conditional mood whose forms are composed of a specific auxiliary verb (originally one of the aorist series of the copula, inflecting for person and number in some languages and reduced to an invariable particle in others) and the aorist participle of the main verb (inflected for gender and number): *бих казал(а)* 'I (m./f.) would say', *биха казали* 'they would say', cf. Ukrainian *я сказав би / сказала б, вони б сказали*. But it also has in its tense system, which is an atypically large one for a Slavic language, a future in the past and a future perfect in the past, analogous to the future and the future perfect, except that they use an imperfect form of the future auxiliary instead of the present (reduced to a particle in the affirmative): future in the past *щях да кажа* 'I would say, was going to say', cf. future *ще кажа* 'I will say' < *ща (да)*

*кажа* (originally 'I want to say'). By virtue of combining preterite and future grammar these tenses bear a structural similarity to verb forms of English and the Romance languages which are sometimes included into the tense system (Колпакчи, 1978: 128ff) but are more often termed a conditional mood, cf. French cond. *je dirais* 'I would say' < *dir(e)* 'to say' + (*av*)*ais* '(I) had', fut. *je dirai* 'I will say' < *dir(e)* + *ai* '(I) have'.

The other Slavic languages have simplified the tense system to various degrees. Modern Ukrainian counts four tenses (pluperfect, past, present, future). Thus the conditional mood (including a past tense form which is to the regular conditional as the pluperfect is to the past: *нішов би був* 'would have gone (sg. m.)', *казала б була* 'would have said (sg. f.)'; Білодід (1969: 378–9, 392) labels it variably as a conditional past tense of the indicative and a past tense of the conditional mood) and constructions with modal verbs, mostly in the past tense, are the chief grammatical ways of expressing conditionality.

## 2 Verbal Conditionality in Bulgarian

The conditional mood and the future-in-the-past tense overlap semantically to a high extent:

- (1) *Ако трябваше да му правя компания в гроздовата, вероятно щях тѣй да се натряскам, та едва ли бих запомнил нещо от уроците.* 'If I had to keep him company in the grape brandy, I would probably get so plastered that I would hardly remember anything from the lessons.'<sup>1</sup> (B. Raynov, *Don't Make Me Laugh*)

Occasionally in translated texts the two forms appear, for no visible reason, side by side in conjoined

<sup>1</sup>The English glosses of corpus examples are the authors', except where published translations could be found.

clauses where the original has identical grammatical forms:

- (2) *Ако не беше синът над шахматната дъска [...], ужасът, който отдавна бе залегнал в душата ѝ, би се надигнал отново [...] и щеше да изсуши безкръвното лице на фрау Райхер.* ‘If it weren’t for her son at the chessboard [...], the terror that had long lain dormant in her soul would rise again [...] and would dry up Frau Reicher’s bloodless face.’ || [ru] *Если бы не сын за шахматной доской [...], ужас, давно прилегший в душе, поднялся бы опять [...] и высушил бескровное личико фрау Рейхер.* (A. Tolstoy, ‘The Hyperboloid of Engineer Garin’)

Sometimes, however, such apparent inconsistency may be grounded in the conditional mood’s tendency to express an inclination, to present the event as a possible rather than a necessary consequence of a (perhaps unexpressed) condition (Андрейчин, 1942: 310):

- (3) *Защото без мене де щеше да си ти сега? Какво би правил?* ‘For without me where would you be? What would you be doing?’ || [fr] *Car, sans moi, où serais-tu ? que ferais-tu ?* (G. Flaubert, *Madame Bovary*)
- (4) *А аз на твоє място щях да я намразя; и не бих ѝ се подчинявала* || [en] *And if I were in your place I should dislike her; I should resist her.* (C. Brontë, *Jane Eyre*)
- (5) *Ако тоя негодник Ликон се беше нахвърлил върху мен, щях да го разбера, дори бих му простил...* ‘If that vile Lykon had attempted my life I could understand, even forgive him.’ || [pl] *Gdyby ten podły Lykon targnął się na moje życie, jeszcze rozumiałbym, nawet przebaczył...* (B. Prus, *Pharaoh*)

In these examples one might sense a slight difference between the two verbs, in that being somewhere, dislike and understanding are less controllable and volitional than doing something, resistance and forgiveness, correlating with the use of the first verb in the future in the past and the second in the conditional mood.

Another difference which can also come up when the forms are coordinated is the future in the past referring to a time following what the conditional refers to:

- (6) *Много неща, много дребни случки, които иначе биха останали незабелязани или щяха да се забравят, сега се припомнаха и се разказваха не за първи и не за втори път.* ‘Many things, many minor events that would otherwise have remained unnoticed or would have been forgotten, were now recalled and retold more than once or twice.’ (Y. Yovkov, ‘Albena’)
- (7) *Наистина с Георгий Васильович бих се чувствувал по-спокоен, но при него дълго щях да си остана момче за всичко.* ‘Of course, I would feel more at ease with Georgy Vasyliovych, but under him I would have remained an errand boy for a long time.’ || [uk] *Звичайно, з Георгієм Васильовичем спокійніше, але під ним я довго був би на побіженьках.* (V. Drozd, *A Lonely Wolf*)

The temporal meaning of the future in the past is, in the well-known scheme of (Reichenbach, 1947: 288–298),  $E > R, R < S$ : the time of the event  $E$  follows the time of reference  $R$ , which precedes the time of speech  $S$  (Figure 1). Reichenbach points out (p. 297f) that this configuration does not correspond to a tense in the traditional grammar of English, although in *I did not expect that he would win the race* what is classified as a form of the conditional mood functions as a tense in the indicative, as do the ‘transcriptions’ in *I did not expect that he was going to win the race* and *the king lavished his favor on the man who was to kill him*; he gives this tense the ‘new name’ **first, second or third posterior past** depending on whether  $E$  precedes, matches or follows  $S$ , these three belonging to the same fundamental form. (Ницолова, 2008: 269, 311f) offers the formula  $R < S, R < E < S$ , which implies that the Bulgarian future in the past is Reichenbach’s first posterior past, although this is neither universal nor essential to its being a future ( $E > R$ ) in the past ( $R < S$ ).

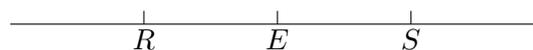


Figure 1: Future in the past: typical temporal meaning

Some examples of its use with this meaning in translated fiction, along with the originals, where the expressions vary from the Italian conditional through the French near future to a German modal verb:

- (8) *По-късно Никита щеше да се смяе, като научи, че този мъж е на повече от шейсет години.* ‘Niketas would soon be even more amazed to learn that this man was over sixty years old.’ || [it] *Niceta si sarebbe poi stupito, apprendendo che quell'uomo aveva più di sessantanni.* (U. Eco, *Vaudolino*)
- (9) *Щеше да се бие на дуел. Не можеше да го избегне.* ‘He was going to fight a duel. He could no longer avoid it.’ || [fr] *Il allait se battre en duel. Il ne pouvait plus éviter cela.* (G. de Maupassant, *Bel-Ami*)
- (10) *Но още на другия ден щеше да бъде изготвен и редовният договор.* ‘But already on the following day the regular contract would be produced also.’ || [de] *Am folgenden Tag schon sollte ein gültiger Vertrag ausgefertigt werden.* (J. Krüss, *Timm Thaler, or the Traded Laughter*)
- (11) *Беше месец май, следобед нямаше да има учебни занимания.* ‘It was May, there would be no classes in the afternoon.’ || [en] *It was May, a half-holiday.* (W. Thackeray, *Vanity Fair*)

The modal (counterfactual) meaning assumes that the future branches at or after the reference time, and the event is located on an alternative (imaginary) timeline, which turns it into a conditional (Figure 2). Here, too, its position in time relative to the utterance is not essential.

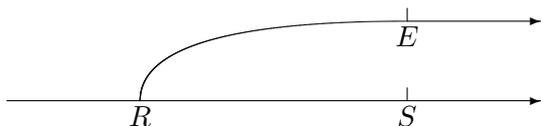


Figure 2: Future in the past: modal meaning

The conditional mood mentions conditionality in its name, but Андрейчин (1942: 306) tellingly only brings it up in the second sentence of the definition of the category, elevating possibility to the status of its principal meaning:

The conditional mood means an action that is presented as conceivably possible in the future. Usually the possibility of performing a certain action is placed in relation to the fulfilment of some condition, therefore this mood is usually found in conditional sentences.

The conditional often replaces the indicative as a means of softening the utterance, with the curious effect that it is more likely to express an unconditional event than the future in the past of the indicative. Compare:

- (12) *А аз бих ви помолил да останете тук, да ми помогнете.* ‘And I would ask you to stay here, to help me.’ || [ru] *А я вас попросил бы остаться здесь, помочь мне.* (V. Shishkov, *Gloomy River*) (In fact he is asking, the conditional merely signalling politeness; an attempt to interpret it as a *bona fide* conditional – ‘I would ask if it were not too much trouble for you’ – breaks against the addressee’s unenthusiastic reply, which proves that he would rather go, but takes the request as made.)
- (13) *Бих ви помолил да ми дадете заповедта — тя лежи на масата ви — каза кротко Авакум.* ‘“I would ask you to give me the order—it is on your desk,” Avakum said softly.’ (A. Gulyashki, ‘The Momchilovo Affair’) (Again, in fact he is asking.)
- (14) *След миг, рече той, щял да помоли присъстващите да вдигнат тост.* (G. Orwell, *Animal Farm*) || [en] *In a moment, he said, he would ask the present company to drink a toast.* (And in fact at the end of his speech he does ask them to.)
- (15) *Искате ли да вземем колите? — Тъкмо това щях да ви помоля.* (E. Hemingway, *A Farewell to Arms*) || [en] *“[...] Would you like us to take the cars?” “That’s what I wanted to ask you.”* (In the event he won’t ask because the other offered it himself, so there was no longer any point.)
- (16) *Щях да помоля да видим рубините, когато телефонът иззвъня и Гетсби взе слушалката.* (F. S. Fitzgerald, *The Great Gatsby*) || [en] *I was going to ask to see the collection of rubies when the phone rang and Gatsby took up the receiver.* (And he never does ask because the events take a different course.)

The effect of all this is that the relation between these two forms can be hard to understand for speakers of other Slavic languages, and is worth studying in contrast.

### 3 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 correction 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 1.52 mln words on the Bulgarian and 1.34 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. (See (Derzhanski and Siruk, 2022) and (Держански and Сирук, 2023) for recent examples of uses of the corpus in comparative research of selected grammatical categories in Bulgarian and Ukrainian.)

In the corpus we sought out all occurrences of the conditional mood and the future (perfect) in the past, regardless of evidentiality, on the Bulgarian side. It would have been beneficial to exclude the future (perfect) conclusive and renarrative (*щял съм да разбера* ‘[apparently or reportedly] I would understand’), but in most forms they coincide with the future (perfect) in the past and are very difficult to differentiate. Non-declarative (conclusive and renarrative) forms constitute 5,76% of the lot. There are 38 occasions on which the mood or tense auxiliary is present but the main verb is absent (elided, or the utterance is cut short). They could have been left out, but we decided not to; that would have had little effect, as they account for just over 0.1% of the data.

A total of 35861 uses of the conditional mood and future-in-the-past tense forms in the Bulgarian texts in the corpus were found. Table 1 presents, in decreasing order of the total number of hits, their distribution by sectors and the number of corresponding  $\bar{o}(u)$  (conditional mood markers) in the Ukrainian texts.

The sectors are conspicuously divided into those where the language of the original is a non-Slavic one and those where it is Slavic: in the former there are  $\frac{3}{4}$  times more occurrences of the forms we are

| Sect | Hits | % $\bar{o}(u)$ | Sect | Hits | % $\bar{o}(u)$ |
|------|------|----------------|------|------|----------------|
| E1   | 5120 | 48.07          | Pl   | 3524 | 67.14          |
| De   | 4729 | 48.76          | Bg   | 2888 | 64.30          |
| E2   | 4618 | 42.38          | Uk   | 2278 | 73.84          |
| Fr   | 4305 | 56.45          | R2   | 2258 | 75.73          |
| It   | 4127 | 43.59          | R1   | 2014 | 78.55          |

Table 1: Number of occurrences of the Bulgarian conditional and future in the past by sector of the corpus.

considering, but in the latter the ratio of corresponding  $\bar{o}(u)$  is higher by  $\frac{1}{4}$ .

In this study we concentrate on the examples from the Bulgarian and the Ukrainian sector, that is, the pairs which consist of an original and a translated sentence.

### 4 The Conditional Mood

As Table 2 demonstrates, the conditional mood is used in Bulgarian original texts (‘Bg orig’) about 1.5 times more often than in Bulgarian translations of Ukrainian literature (‘Bg←Uk’); the columns labelled ‘+  $\bar{o}(u)$ ’ and ‘−  $\bar{o}(u)$ ’ say respectively how many times a corresponding  $\bar{o}(u)$  is present or absent in the parallel Ukrainian text. The verb which

|         | + $\bar{o}(u)$ | − $\bar{o}(u)$ | Total |
|---------|----------------|----------------|-------|
| Bg orig | 1351           | 480            | 1831  |
| Bg←Uk   | 1047           | 192            | 1239  |
| total   | 2398           | 672            | 3070  |

Table 2: Distribution of the conditional mood

appears by far most often in this mood is *мога* ‘can, may’, followed by *трябвам* ‘be needed’, *искам* ‘want’, *съм* ‘be’ and *кажа* ‘say’ (Table 3). (The next

| Sector         | Bg orig | Bg←Uk | Total |
|----------------|---------|-------|-------|
| <i>мога</i>    | 493     | 263   | 756   |
| <i>трябвам</i> | 132     | 103   | 235   |
| <i>искам</i>   | 64      | 122   | 186   |
| <i>съм</i>     | 88      | 73    | 161   |
| <i>кажа</i>    | 101     | 49    | 150   |

Table 3: Verbs most frequently used in the conditional

one on the list, *направя* ‘do’, has 42 occurrences.) The distribution over the sectors is uneven: the ratio of occurrences in the Bulgarian sector to the Ukrainian one is 2:1 for *мога* and *кажа*, 5:4 for *трябвам* and *съм*, but only 1:2 for *искам*. Even more dramatically, *би следвало* ‘it would follow, would be right’

occurs 38 times in original Bulgarian texts and not at all in translations from Ukrainian; its close counterpart and cognate *слід було б* tends to be translated as *би трябвало*.

With its 2398 occurrences the conditional mood is the most frequent Ukrainian correspondence to the cognate Bulgarian category by a wide marge:

- (17) *О, това би било ужасно*. ‘Oh, this would be terrible.’ || *О, то було б жахливо*. (O. Nonchar, *Your Star*)

It is followed by the present, past or future indicative form of the verb (181, 150 and 121 occurrences, respectively):

- (18) *Сега помагам на баща си, но бих искал да си намеря работа*. ‘I help my father for now, but I would like to find a job.’ || *Поки що просто батькові допомагаю, але хочу знайти десь роботу*. (E. Pelin, *Yan Bibiyan on the Moon*)
- (19) *Нищо по-сладко не би могла да ми каже Олена*. ‘There was no sweeter thing that Olena could tell me.’ || *Нічого солодшого для мене не могла сказати Олена*. (V. Drozd, *A Lonely Wolf*)
- (20) *Какво бихте казали за един малък бридж?* ‘What would you say of a little bridge?’ || *Що ви скажете з приводу невеличкого бриджу?* (B. Raynov, *Turhoons with Gentle Names*)

The use of the conditional mood in a conditional subordinate clause, which is very common in other Slavic languages, is largely obsolete in Bulgarian. There are only 18 instances of it in each of the two sectors. In exactly half of these the main verb is in the conditional mood as well. Only on two occasions it is in the future in the past. There are also eight cases when there is no main clause, the sentence expressing longing. In 29 of the Ukrainian sentences there is *б(и)* as well.

- (21) *Да би могъл, би отстранил и унищожил и нея*. ‘If he could, he would depose and destroy her as well.’ || *Коли б сила та воля, він і її усунув би й знищив*. (P. Zahrebelnyi, *Roksolana*)
- (22) *Ако два дни по-рано някой би ми казал, че това стройно и безупречно изваяно тяло принадлежи на Грейс, щях да му предложая да иде да се удави*. ‘If someone had told

me two days earlier that this slender and impeccably sculpted body belonged to Grace, I would have suggested that he go and drown himself.’ || *Коли б два дни тому хто-небудь сказав мені, що це струнке й бездоганне тіло належить Грейс, я запропонував би йому ніти й утопитися*. (B. Raynov, *The Great Boredom*)

- (23) *О, ако биха били пренесени на Земята!* ‘Oh, if they would be transferred to Earth!’ || *О, якби взяти все те на Землю!*... (E. Pelin, *Yan Bibiyan on the Moon*)

## 5 The Future-in-the-Past Tense

We include here the future perfect in the past, but in fact it only occurs twice in these two sectors (out of 34 times in the entire corpus), both times in the same book and with almost the same verb in the same meaning (though translated in different ways):

- (24) *Предполагам, че ако познавах по-добре Майнц, и тримата отдавна щяха да са ме загубили* като съпътник ‘I suppose if I knew Mainz better, all three would have lost me as a companion long ago’ || *Мабуть, коли б я трохи краще знав Майнц, то ця трійця вже тільки б мене й бачила* [...] (B. Raynov, *You Cannot Tell the Day by the Morning*)
- (25) *Ако я виждах, вероятно вече щях да съм ви изгубил* като събеседник. ‘If I saw it, I probably would have already lost you as a conversation partner.’ (ibid.) || *Коли б бачив, то, напевне, уже б змушений був втратити вас* як співрозмовника. (lit. ‘... I would have been obliged to lose you’.)

Table 4 attests that the future in the past appears with approximately equal frequency in Bulgarian translations of Ukrainian literature and in original Bulgarian texts. When Bulgarian authors use it, Ukrainian translators use the conditional somewhat less than half of the time. When Ukrainian authors use the conditional, Bulgarian translators are much more likely to use the future in the past than not to.

The verb which appears most often in this tense is *бъда* ‘be (perf.)’, followed by *съм* ‘be (imperf.)’, *имам* ‘have’ and *стана* ‘become; happen’ (Table 5). (The next one is *видя* ‘see’, with 32 occurrences.) For some reason, although *бъда* and *съм* in the future in the past are in free variation, the first is a

|         |               |               |              |
|---------|---------------|---------------|--------------|
|         | + <i>б(u)</i> | - <i>б(u)</i> | <b>Total</b> |
| Bg orig | 506           | 551           | 1057         |
| Bg←Uk   | 635           | 404           | 1039         |
| total   | 1141          | 955           | 2096         |

Table 4: Distribution of the future in the past

| Sector       | Bg orig | Bg←Uk | Total |
|--------------|---------|-------|-------|
| <i>бъда</i>  | 84      | 54    | 138   |
| <i>съм</i>   | 24      | 46    | 70    |
| <i>имам</i>  | 36      | 28    | 64    |
| <i>стана</i> | 32      | 30    | 62    |

Table 5: Verbs most often used in the future in the past

good deal more frequent in the Bulgarian sector (84) and the second in the Ukrainian one (46).

A reservation may be made concerning the passive voice, whose building is one of the functions of these two verbs. Although they are ‘generally synonymous’ there as well (Ницолова, 2008: 239), there is a tendency for the passive with *съм* to describe the result of the event, which happens only once in the Bulgarian and twice in the Ukrainian sector,

- (26) *Ако имахте късмет, задачата ви щеше вече да е приключена и в тоя час навярно щяхте да се движите нейде между Залцбург и Виена [...]* ‘If you were lucky, your mission would have been completed already and at this moment you would probably be traveling somewhere between Salzburg and Vienna’ || *Коли б вам щастило, ваше завдання вже закінчилося б і в цей час ви вже їхали б десь між Зальцбургом і Віднем [...]* (B. Raynov, *You Cannot Tell the Day by the Morning*)
- (27) *За приказки със стареца повече подхождаше друго време, когато нямаше да е погълнат от риболовното си занимание [...]* ‘Another time was better suited for talking to the old man, when he would not have been engrossed in his fishing activity’ || *Для балачок із Катратим вибирай зручніший час, коли він не буде поглинутий оцим своїм ловецьким заняттям [...]* (O. Nonchar, *The Cathedral*)
- (28) *Да не бе тя [...], сигурно душата му нямаше да е претъпкана с долни грехове.* ‘If it were not for her [...], probably his soul would not be packed with vile sins.’ || *Коли б не вона*

*[...], певне, його нутро не було б до краю забите усяким грехоплутством.* (M. Stelmakh, *The Four Fords*)

and for the passive with *бъда* to denote the event itself, which is much more common (14 and 15 examples, respectively, in the two sectors).

- (29) *Неочакваното излизане на двамата генерали от приема [...] скоро щеше да бъде забравено, ако Фред Барнаби не беше казал на мисис Джаксън [...], че се наложило пашите веднага да заминат за фронта.* ‘The unexpected exit of the two generals from the reception [...] would soon have been forgotten, if Fred Barnaby had not told Mrs Jackson [...] that the pashas had to leave at once for the front.’ || *Раптовий від'їзд двох генералів [...] швидко б забули, якби Фред Барнабі не сказав місіс Джексон [...], що пашам наказано негайно виїхати на фронт.* (S. Dichev, *The Way to Sofia*)
- (30) *Паднеш ли тази най-близка до земята на Османовци твърдина на Хабсбургите, всичко щеше да бъде забравено: и позорът от неуспехите, и бездарността на сераскера — великия везир [...]* ‘If this fortress of the Habsburgs, the closest one to the Ottomans’ land, were to fall, all would be forgotten: the shame of the failures, the ineptitude of the serasker—the grand vizier’ || *Узяти цю твердиню Габсбургів, найближчу до земель Османів, — і забудеться ганьба невдач, нездарність сераскера — великого візира [...].* (P. Zahrebelnyi, *Roksolana*)
- The verb *мога* ‘can, may’, which is the most frequent one in the conditional, with 756 occurrences (Table 3), occurs a total of four times in the future in the past, with different degrees of factuality:
- (31) *Щеше ли да може да ги види още веднаж? ‘Would he be able to see them once more?’ || Чи доведеться йому побачити їх ще хоч раз?* (E. Pelin, *The Gerak Family*)
- (32) *Тогава „гърмящото копие“ нямаше да може да пуска невидими стрели и ние щяхме да нападнем хаубау! ‘Then the “thundering spear” would not have been able to shoot invisible arrows and we would have attacked the Haubau!’ || Тоді б вогненний лук не зміг пускати свої невидимі стріли, й ми вчинили б напад на гаубау...* (G. Ugarov, *In*

*the Exile's Footsteps*) (context: if the magician had doped the white man)

- (33) *Още малко — и нямаше да можем да стреляме.* ‘A while more—and we would not have been able to shoot.’ || *Ще трохи — і не могли б стріляти.* (Y. Yovkov, ‘A Bulgarian Woman’) (context: if the protagonist had not brought water to cool the machine-guns)
- (34) *Хома го стисна толкова силно, че ако пожелалаше да разтвори сега пръсти, едва ли щеше да може.* ‘Khoma squeezed [the bayonet] so hard that if he now wanted to open his fingers, he would hardly have been able to do so.’ || *Хома стис його так міцно, що коли б він і захотів розігнути зараз власні пальці, то вже не розігнув би.* (O. Honchar, *Guide-on Bearers*)

The second in line *трябвам* ‘be needed’ and the fourth *искам* ‘want’ do not occur in this tense in the Bulgarian or Ukrainian sector at all (in the rest of the corpus they do so 6 and 26 times, respectively).

The most common counterpart of the Bulgarian future in the past is the Ukrainian conditional mood again (1141 occurrences as per Table 4, including a few where the Ukrainian form is a past conditional):

- (35) — *Ако ме беше изслушал, щеше всичко да разбереш!* ‘If you had listened to me, you would have understood everything!’ || *Коли б ти вислухав мене, ти б усе зрозумів!* (P. Vezhinov, *Traces Remain*)
- (36) *А още тогава можеше да се намерят и той щеше да остане да живее в колибата, нямаше да се връща със старите си крака през зимата до Унгария.* ‘They would have found one another then, and he, the white-haired one, would have lived next to her, he would not be returning to Hungary now, in the winter.’ || *А вже тоді були б віднайшли себе, був би він, білоголовий, коло неї зажив, нараз тяжкою зимою не вертав в Угорщину.* (O. Kobylanska, *On Sunday Morning She Gathered Herbs*)

It is followed by the future indicative of the verb (528):

- (37) *Войната щеше да има край, а този край можеше да бъде само един — победа.* ‘The war would have an end, and this end could

only be one—victory.’ || *Війна матиме кінець, а цей кінець може бути лише один — перемога.* (M. Marchevski, *Tambuku Island*)

- (38) *Крал Франсоа основа в Париж кралска библиотека, която след време щеше да стане Национална.* ‘King François founded in Paris a royal library which would become National in time.’ || *Король Франціск заснував у Парижі королівську бібліотеку, яка згодом стане Національною.* (P. Zahrebelnyi, *Roksolana*)

This example features an interesting interplay of tenses and moods in the original and the translation:

- (39) *Аз те знам, ти имаш високо мнение за себе си, но Юлия е по-добра от тебе и ако беше на твоите години, сигурно щеше да бъде и по-умна...* ‘I know you, you think highly of yourself, but Yulia is better than you and if she were of your age, surely she would be smarter as well...’ || *Я тебе знаю, ти високої думки про себе, але Юлія краща за тебе, і коли їй буде стільки років, скільки тобі, вона напевно буде розумнішою...* (P. Vezhinov, *Traces Remain*) ‘... and when she is as old as you are [now], she will surely be smarter’.

Yulia is younger than Pesho, and less smart ( $y < p$ ) only because of this; in the original the speaker considers an imaginary situation in which she is as old as Pesho is (Figure 3), and in the translation the

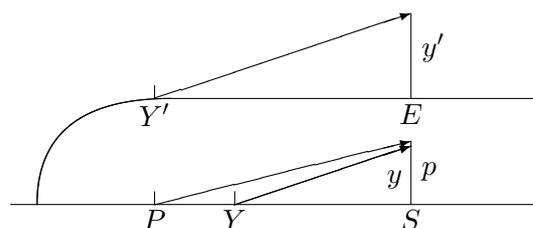


Figure 3: Yulia and Pesho in the Bulgarian original

prospective one in which she will be so old (Figure 4); in other words, the real present Pesho is outsmarted ( $p < y'$ ) by Yulia from the alternative universe where they are of the same age in the original, and by the future Yulia in the translation. Thus the meaning for which Bulgarian uses the future in the past is rendered into Ukrainian without any reference to the past.

The third most frequent choice for rendering the Bulgarian future in the past into Ukrainian is a past

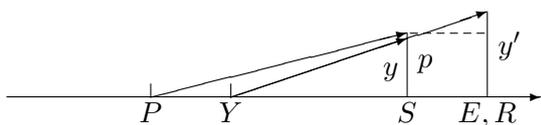


Figure 4: Yulia and Pesho in the Ukrainian translation

tense form of a modal verb plus the infinitive of the main verb (191). The modal verb in the Ukrainian sector is overwhelmingly (47 of 53 times) *мати* ‘have to, need to’, a verb that in the Bulgarian sector is also the most common one, but accounts for under half of the examples (57 of 138), with *хотіти* ‘want’ a significant second, followed by *збиратися* ‘intend’, *могти* ‘can, may’ and several others:

- (40) *Въпреки това вечерта се затътрах към училището където щеше да бъде театърът.* ‘And yet, in the evening I wandered off to the school, where the show was going to be.’ || *И все ж таки увечері я поплентав ся до школи, де мав бути театр.* (M. Stelmakh, *The Geese and Swans Are Flying*)
- (41) *Щеше да каже нещо, но погледът му се спря на Митка:* ‘He was going to say something, but his eyes stopped on Mitko.’ || *Хотів щось сказати, але тут його погляд зупинив ся на Миткові: [...]* (Y. Yovkov, ‘A Night-time Guest’)
- (42) *Нямаше да го направя, разбира се, но защо трябваше да ме постави в таква неудобно положение.* ‘I wasn’t going to do it, of course, but why did he have to put me into such an uncomfortable situation.’ || *Звичайно, я не збирав ся цього робити, але чому він ставить мене в таке незручне становище?* (P. Vezhinov, *When You Are in the Boat*)

In other words, Bulgarian translators reach for the future in the past when the Ukrainian modal verb is *мати*, whereas Ukrainian translators consider a wider variety of modal verbs.

A few examples of two conjoined future-in-the-past tenses to which different forms correspond in Ukrainian:

- (43) *Колѳбаніето беше между зрелия фасул, което щеше да бъде по-евтино, и готвеното с месо, което щеше да бъде по-приятно.* ‘The hesitation was between kidney beans, which would be cheaper, and the meat stew, which would be nicer.’ ||

*Вагалася між квасолею, яка була б дешева, і м’ясною стравою, яка мала бути смачніша.* (P. Vezhinov, *An Incident in the Quiet Street*)

- (44) *Климент си мислеше изтръпнал за ужасите на войната, но не за ужасите, които щеше да види след час, а за ония, които нямаше да види* ‘Kliment was thinking with a shudder about the horrors of the war, but not the horrors he would see within an hour, rather the ones he would not see’ || *Климент невідступно думав про жахи війни, але не про ті жахи, які мав бачити за годину, а про ті, яких йому ніколи не судилось побачити* (S. Dichev, *The Way to Sofia*)
- (45) *Живееше в Топкапъ като в мраморна гробница. Тук щеше и да умре, тук щяха да замлъкнат и нейните песни, и песните за нея.* ‘She lived in Topkapı as in a marble tomb. She would die here too, her songs and the songs for her would fall silent here.’ || *Жила в Топкапи, як у мармуровій корсті, тут мала й умерти, тут умовкнутъ її пісні і пісні для неї.* (P. Zahrebelnyi, *Roksolana*)

What makes the relatively frequent occurrence of a modal verb in the past tense in Ukrainian as a counterpart of the Bulgarian future in the past is that from the example of other languages (and by analogy with the Ukrainian synthetic future in *робитиму* ‘I will do’ < *робити* ‘to do’ + *маю* ‘I have’) it is known that this is how the future-in-the-past tense (or ‘conditional mood’ of Western European languages) is born.

## 6 Conclusions

There are some interesting asymmetries in the translation of tenses and moods.

The Ukrainian present tense serves for rendering the Bulgarian conditional mood with greater frequency than the Bulgarian conditional mood is used for rendering Ukrainian present tense forms. Mostly this concerns *би могло* ‘it could be’, to a lesser extent *би трябвало* ‘it should be’ and some others, which abound in original Bulgarian texts but lose their stylistic palliation in the hands of Ukrainian translators. In the opposite direction the palliation neither exists in the original nor appears in the translation.

The Ukrainian conditional mood is used for rendering the Bulgarian future in the past with lesser

frequency than the Bulgarian future in the past is used for rendering the Ukrainian conditional mood. Here it is harder to name the reasons; the stylistics of the texts in the corpus may be among them, but the issue deserves further study. And certainly so do the Ukrainian conditional mood and its correspondences in Bulgarian in their own right.

## References

- Ivan Derzhanski and Olena Siruk. 2022. Some notes on p(e)re-reduplication in bulgarian and ukrainian: A corpus-based study. In *Proceedings of the 5th International Conference on Computational Linguistics in Bulgaria (CLIB 2022)*, page 98–104. Sofia, Bulgaria. Department of Computational Linguistics, IBL – BAS.
- Hans Reichenbach. 1947. *Elements of Symbolic Logic*. London: Macmillan.
- Л. Д. Андрейчин. 1942. *Основна българска граматика*. София: Хемус. [L. Andreychin. *Osnovna balgarska gramatika*. Sofia: Hemus. 1942.].
- І. К. Білодід, editor. 1969. *Сучасна українська літературна мова. Морфологія*. Київ: Наукова думка. [I. K. Bilodid, editor. *Suchasna ukrainska literaturna mova. Morfologiya*. Kiiv, Naukova dumka. 1969.].
- Иван Держански and Олена Сирук. 2023. Българските неопределителни местоимения с елемента си и съответствията им в украински език. In *Доклади от Международната годишна конференция на Института за български език „Проф. Любомир Андрейчин“ (София, 2023 година)*, page 285–294. [Ivan Derzhanski and Olena Siruk. *Balgarskite neopredelitelni mestoimeniya s elementa si i saotvetstviyata im v ukrainski ezik*. In: *Dokladi ot Mezhdunarodnata godishna konferentsiya na Instituta za balgarski ezik*, Sofia, Bulgaria, 2023.].
- М. А. Колпакчи. 1978. *Дружеские встречи с английским языком*. Ленинград: Издательство Ленинградского университета. [M. A. Kolpakchi. *Druzheskie vstrechi s angliyskim yazykom*. Leningrad: Izdatelstvo leningrtadskogo universiteta, 1978.].
- Руселина Ницолова. 2008. *Българска граматика. Морфология*. София: Университетско издателство „Св. Климент Охридски“. [Ruselina Nitsolova. *Balgarska gramatika. Morfologiya*. Sofia: Universitetsko izdatelstvo Sv. Kliment Ohridski, 2008.].

# Lexical richness of French and Quebec journalistic texts

Natalia Dankova

Université du Québec en Outaouais

natalia.dankova@uqo.ca

## Abstract

This paper presents some results of a quantitative study that focuses on the variety and word frequency in texts from a comparative perspective. The study aims to analyze and compare French and Quebec journalistic texts on political and cultural topics written in French and recently published in major newspapers such as *Le Monde*, *le Figaro*, *Le Devoir*, etc.

The statistical analysis concerns the number of different words in the text, the number of different adjectives, the number of different verbs (and also passive structures, participles and gerunds which contribute to syntactic and stylistic sophistication), and the number of hapaxes.

French texts from France exhibit greater lexical richness and sophistication: they contain more adjectives, a greater variety of adjectives, as well as more participles and gerunds compared to French texts from Quebec.

The originality of the study lies in the fact that it analyzes variation in French using a lexicometric approach.

**Keywords:** lexical statistics, comparative analysis, journalistic texts, variation in French

## 1 Introduction

This quantitative ongoing study focuses on the variety and frequency of words in texts from a comparative perspective. This exploratory study aims to test the hypothesis regarding differences in the lexicon of verbs and adjectives between the two varieties of French.

The study analyzes and compares French and Quebec journalistic texts, written in French and recently published in major newspapers such as *Le Monde*, *le Figaro*, *France-Soir* (France), and *Le Devoir* and *La Presse* (Quebec).

Quebec French is often perceived as being less rich than standard French (i.e., French from

France). The feeling of linguistic insecurity remains strong among Quebec French speakers. Boudreau (2023) mentions that linguistic insecurity, common in the French-speaking world, stems from the way in which the French language has developed, from the idea of a single standard and a unitary, uniform vision of French

Written Quebec French, especially in texts intended for publication, adheres to a model known as international French, which shows minimal signs of regional variation. These differences are mainly lexical; syntax and grammar conform to the standard shared by all French-speaking countries. Consequently, one would reasonably expect few notable differences between French and Quebec texts.

However, our previous studies on variation in French (Dankova, 2002, 2009, and 2017) highlighted the choices of verbs that differ in oral narratives in French from France and in French from Quebec. These differences concern movement verbs and psychological verbs. Thus, we distinguished the following types: verbs of perception such as *regarder* ‘to look’, *voir* ‘to see’, *entendre* ‘to hear’, and verbs of cognition that describe mental representation and activity such as *comprendre* ‘to understand’, *penser* ‘to think’, *décider* ‘to decide’, and *se rendre compte* ‘to realize’. Verbs of visual or auditory perception are more frequent in narratives in French from France compared to those in Quebec French. This provides another avenue for investigating variations in journalistic texts.

Another study that we conducted on the use of adjectives in two types of narratives shows that French informants use a significantly greater variety and quantity of adjectives compared to Quebec speakers (Dankova, 2024). The linguistic means used are more diversified in the French corpus: we noted a greater number of different adjectives, a greater diversity of verbal forms and temporal adverbs.

In the present study, we aim to verify if the observed results hold true in journalistic texts.

## 2 Data and method

There are many lexical richness measures (Lebart and Salem, 1994, Brunet, 2003). The most popular one is the type/token ratio:  $V/N$  where  $N$  = number of tokens and  $V$  = number of types (Muller, 1979). This is the formula we use in our study. Roeland van Hout and Vermeer (2007) note that “lexical richness measures aim to reveal the richness of lexis used in a text, with the intention also of discovering the richness of the lexicon which produced the text. But all kinds of measures turn out to have reliability and/or validity problems” (Roeland van Hout and Vermeer, 2007: 136).

Although in linguistics there is a distinction between *lexicon* and *vocabulary* (de Saussure, (1995/1916), the term *lexical richness* is traditionally used in lexical statistics to measure the vocabulary rather than the lexicon of a language (see also Thoiron and Arnaud, 2008).

At the exploratory stage of the study, the corpus consists of 90 newspaper articles covering current political and cultural topics (45 articles written in each variety of French). This number will be increased in the next phase.

The corpus in French from Quebec (FQ) contains 35975 words and that in French from France (FF) 36999 words; the average length of the texts is respectively 799 words (FQ) and 822 words (FF).

Although it was not possible to select a large number of articles on a single subject, we opted for articles with common themes (post-Covid-19, major political and social events).

The statistical analysis concerns the number of different words in the text, the number of adjectives, the number of different verbs, and the number of hapaxes.

There is no consensus regarding the delimitation of the category of adjectives, the criteria for adjectivity or the classification of adjectives. The literature devoted to adjectives is abundant and reveals several paradoxes (Pottier, 1985; Goes, 2015; Van de Velde, 2009; Baylon and Mignot, 1995; Noailly, 1999; Riegel, 1985; Picabia, 1978; Wilmet, 1997).

When talking about adjectives, we mainly consider their functions rather than their forms, because only some of the adjectives constitute a class (or part of speech) called adjectives. For the

purposes of this study, we therefore take an inclusive stance. The main function of adjectives is to qualify nouns. There are qualifying adjectives (*une robe noire* ‘a black dress’), relational adjectives which can be replaced by a complement of the noun (*une décision gouvernementale* ‘governmental decision’ vs *une décision du gouvernement* (literally ‘a decision of the government’) ‘a government decision’), and referential or third type adjectives (*une carte bleue* (literally ‘a blue card’) ‘a credit card’).

We consider as adjectives nouns that are used as adjectives, for example, in *situation d’urgence* (literally ‘situation of emergency’), *d’urgence* ‘emergency’ plays an adjective role. On the other hand, adjectives that form part of compound nouns which designate, for example, institutions, political parties, or newspapers are excluded from the calculations: *la Cour Suprême* ‘the Supreme Court’, *la République Française* ‘the French Republic’, *le parti vert* ‘the green party’, *Le Monde Diplomatique* (literally ‘Diplomatic World’, name of a newspaper). Possessive adjectives which are considered determiners in the same way as articles are not taken into account either: *leur décision* ‘their decision’.

When analyzing verbs, we examine the number of different verbs and the most frequently used verbs, passive structures, participles, and gerunds which contribute to syntactic and stylistic sophistication; and types of verbs: psychological verbs which describe perception (*to see, to look*) and cognition (*to think, to realize*), and movement verbs (*to go, to come*).

A long indexing work preceded the analyses. We manually annotated adjectives and verbal forms because of the specificity of morphology. In French, verbal morphology is characterized by a large number of temporal forms, simple and compound, marked by endings and including alternations in verb roots. Some homonym forms can be verbs or nouns, adverbs, or adjectives: *pouvoir* ‘to be able / can’ vs *pouvoir* ‘a power’; *note* ‘(I, he, she) ‘note’ vs *note* ‘a note/a mark’, etc.

Adjectives agree in both number (singular or plural) and gender (masculine or feminine) with the noun or pronoun they modify. A limited number of adjectives have several masculine forms for historical reasons: *vieux (tableau)* ‘old (painting) vs *vieil (homme)* ‘old (man)’.

### 3 Discussion and results

#### 3.1 Adjectives

Adjectives are not always as essential as verbs or nouns. Qualifying adjectives appear when the speaker needs to express an inherent quality of the referent (Baylon and Mignot, 1995) and selects one of the dimensions which constitute the essential characteristics (van de Velde, 2009). Goes (2015) emphasizes that the argumentative context also requires the use of adjectives. Chengyu Fang and Cao (2009: 138) state on empirical basis “that adjective density is significantly correlated to degrees of formality of different text categories”.

The use of adjectives varies in the two corpora: in total 2765 adjectives in French from France and 1860 in French from Quebec, i.e. respectively 61.44 and 41.33 adjectives per text. Adjectives ratio is 0.0747 in FF and 0.0517 in FQ (Table 1).

|             | FF     | FQ     |
|-------------|--------|--------|
| Tokens (N)  | 36999  | 35975  |
| Ratio (V/N) | 0.0747 | 0.0517 |

Table 1: Ratio of adjectives

There is an example of the use of adjectives in an article published in France<sup>1</sup>:

- (1) *Fourbes et lâches* qu'ils sont, *tous*, ces «contrôleurs de la bien-pensance», se régalent à appliquer des «circulaires», jusqu'à l'idiotie *manifeste*, qui n'est pas sans rappeler le zèle *sadique* du pervers *narcissique* persécutant sa victime.

*Deceitful* and *coward* as they are, these “controllers of right-thinking” delight in applying “circulars”, to the point of *manifest* idiocy, which is reminiscent of the *sadistic* zeal of the *narcissistic* pervert persecuting their victim.

Among the most frequent adjectives common to both corpora are *social* ‘social’, *public* ‘public’, *d’urgence* ‘emergency’, *bon* ‘good’, *nouveau* ‘new’. The frequency of the adjective *d’urgence* ‘emergency’ is linked to the theme of Covid-19 in newspapers.

The corpus in French from France contains 1.55 times more different adjectives than the corpus in French from Quebec.

Our results confirm those of a previous study on the use of adjectives in two types of oral narratives which also shows that French informants use a greater number and variety of adjectives compared to Quebec speakers (Dankova, 2024).

#### 3.2 Verbs

According to the study known as Français Fondamental (‘Fundamental French’), the verbs *être* ‘to be’, *avoir* ‘to have’, followed by *faire* ‘to do/to make’, *dire* ‘to say’, *pouvoir* ‘to be able/can’, *aller* ‘to go’, *voir* ‘to see’, *savoir* ‘to know’, and *vouloir* ‘to want’ are among the most frequent words in literary French (Niklas-Salminen, 2015: 39).

Regarding the number of verbs and the number of different verbs, we observe no differences between the two corpora. The same observation applies to passive structures.

Among the most frequently used verbs in the French from France corpus, there are *être* ‘to be’, *avoir* ‘to have’, *faire* ‘to do/to make’, *pouvoir* ‘to be able/can’, *mettre* ‘to put’, *exprimer* ‘to express’, *tenir* ‘to hold’, *dire* ‘to say’, *devoir* ‘to have to/must’, *rendre* ‘to realize/to give back’, *considérer* ‘to consider’. The translation of the verbs into English does not reflect all the meanings present in French.



Figure 1: Ten most frequent verbs (FF) (designed with *Voyant Tools*)

In FQ corpus, the most frequently used verbs are *être* ‘to be’, *avoir* ‘to have’, *faire* ‘to do/to make’, *devoir* ‘to have to/must’, *dire* ‘to say’, *pouvoir* ‘to be able/can’, *mettre* ‘to put’; *voir* ‘to

<sup>1</sup> Azalbert, Xavier (2023). Dégénérescence programmée. *France-Soir*, 2023, March, 27.

see', *prendre* 'to take', *penser* 'to think', *opposer* 'to oppose'.



Figure 2: Ten most frequent verbs (FQ) (designed with *Voyant Tools*)

Verbs of perception (such as *to see*, *to look*, *to hear*) and cognition (*to think*, *to realize*, *to understand*) are 2.5 times more frequent in the French from France corpus (360 occurrences) than in the Quebec French corpus (143 occurrences).

While this does not strictly concern the lexicon, it is noteworthy to mention the use of participles and gerunds. Participles that are not part of compound tenses and do not function as adjectives, along with gerunds, predominantly appear in written texts, enhancing stylistic complexity, as illustrated in the following example<sup>2</sup>:

- (2) Certains parents dépassent les limites, *en exposant* leurs enfants à des applications de lecture.

'Some parents exceed boundaries *by exposing* their children to reading apps'.

The corpus in French from France contains 1.71 times more participles and gerunds compared to that in French from Quebec (Table 2):

|  | FF    | FQ    |
|--|-------|-------|
| Tokens (N)                               | 36999 | 35975 |
| Participles/gerunds (occurrences)        | 615   | 360   |
| Number of occurrences per text (average) | 13.67 | 8     |

Table 2: Participles and gerunds

Moreover, our previous studies on oral narratives in French from France and in French from Quebec highlighted differences in verbal morphology and the types of verbs used: verbs of perception and cognition are more frequent in French from France (Dankova, 2017 and 2009), and the verbal morphology is considerably richer compared to the Quebec corpus. (Dankova, 2002).

#### 4 Conclusion

This ongoing research enabled the identification of avenues for analyzing journalistic texts in two varieties of French. While the genre imposes a certain standard and it is the same language, our results indicate that French texts from France exhibit greater lexical richness and sophistication. They contain more adjectives, a wider variety of adjectives, as well as more participles and gerunds compared to French texts from Quebec.

Increasing the size of the corpora will enable us to validate the results obtained at the exploratory stage. Further analyses will be conducted in the future.

#### References

Christian Baylon and Xavier Mignot. 1995. *Sémantique du langage*. Nathan, Paris.

Annette Boudreau. 2023. *Insécurité linguistique dans la francophonie*. Les Presses de l'Université d'Ottawa, Ottawa.

Étienne Brunet. 2003. Peut-on mesurer la distance entre deux textes? *Corpus*, 2. <https://doi.org/10.4000/corpus.30>

Alex Chengyu Fang and Jing Cao. 2009. Adjective Density as a Text Formality Characteristic for Automatic Text Classification: A Study Based on the British National Corpus, *23rd Pacific Asia Conference on Language, Information and Computation*, pages 130-139 <https://aclanthology.org/Y09-1015.pdf>

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*

<sup>2</sup> *France-Soir* (2022). Les assistants vocaux nuisent au développement social et émotionnel des enfants. *France-Soir*, 2022, October 18.

*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. 2024. Variation en français. Un cas des adjectifs. *Moderne Sprachen*, 66: 79-95.

Jan Goes. 2015. Types d'adjectifs et fonctions adjectivales: quelques réflexions. *Studii de lingvistică* 5: 293-322.

Lucien Lebart and André Salem (editors). 1994. *Statistique textuelle*. Dunot, Paris.

Charles Muller. 1977. *Principes et méthodes de statistique lexicale*. Hachette/Champion, Paris.

Aïno Niklas-Salminen. 2015. *La lexicologie*. Arman Colin, Paris.

Michèle Noailly. 1999. *L'adjectif en français*. Ophrys, Gap/Paris.

Lélia Picabia. 1978. *Les constructions adjectivales en français, systématique transformationnelle*. Librairie Droz, Genève/Paris.

Bernard Pottier. 1985. De l'adjectif. *Travaux de linguistique et de littérature*, 23/1: 301-305.

Martin Riegel. 1985. *L'adjectif attribut*. PUF, Paris.

Ferdinand de Saussure. 1995/1916. *Cours de linguistique générale*. Payot, Paris.

Philippe Thoiron and Pierre Arnaud. 2008. Quelques aspects de la perception de la richesse lexicale, *Cycnos*, 8.

<https://epi-revel.univ-cotedazur.fr/cycnos/392.pdf>

Danièle van de Velde. 2009. *Comment, manières d'être et manières de faire* ». *Travaux de linguistique*, 58/1: 39-61.

Roeland van Hout and Anne Vermeer. 2007. Comparing measures of lexical richness. In: H. Daller, J. Milton & J. Treffers-Daller (eds.), *Modelling and assessing vocabulary knowledge*, pages 93-115. CUP, Cambridge.

Marc Wilmet. 1997. *Grammaire critique du français*. Hachette-Duculot, Paris/Louvain-la-Neuve.

*VoyantTools*

<https://voyant-tools.org/>

# A Corpus of Liturgical Texts in German: Towards Multilevel Text Annotation

**Maria Khokhlova**

St Petersburg State University  
m.khokhlova@spbu.ru

**Mikhail Koryshev**

St Petersburg State University  
m.koryshev@spbu.ru

## Abstract

The aim of the study is to create a “documented” literary and theological history of German Catholic hymnography. The paper focuses on the creation of a corpus of liturgical texts in German and describes the first stage of annotation dealing with the metatextual markup of Catholic hymns. The authors dwell in detail on the parameters of the multi-level classification of hymn texts they developed, which allows them to differentiate hymns on different grounds. The parameters include not only characteristics that represent hymns (the period and the source of their origin, rubrics, musical accompaniment), but also ones that are inherent for strophes. Based on the created markup, it is possible to trace general trends in texts divided according to certain meta-features. The developed scheme of annotation is given on the example of the hymnbook Gotteslob (1975). The results present statistics on different parameters used for hymn description.

**Keywords:** corpus of hymns, liturgy, Catholicism, German language.

## 1 Introduction

It would not be an exaggeration to say that modern works related to the analysis of language and literature deal with corpora (being either corpus-based or corpus-driven), which have become a necessary condition for such studies. However, most corpora, regardless of the language they are aimed at, are focused on modern language (most often literary), since automatic processing of texts from other periods can still be difficult. Additional difficulties in building corpora of historical texts are associated with the fact that there is no access to the texts in

electronic form, so the task of searching, scanning and recognizing them still arises.

In the case of church texts, there is a centuries-old tradition of their collection; however, there are no traditions of their systematization and presentation for research purposes. There are various indexes and reference books, archival materials, but they are fragmented and difficult to access. For example, archival materials are available in the archive of the German Liturgical Institute (Deutsches Liturgisches Institut) in Germany. Therefore, our study makes the first attempt to create a corpus of texts that would include such material.

In the article, we will focus on issues related to the creation of a corpus of liturgical texts using texts from Gotteslob (1975), namely, their metatextual markup. In our case, the created corpus can be classified as both historical and poetic one; therefore metatextual parameters combine the features of these two types. We rely on the principles that were elaborated for the development of similar corpora and TEI project and introduce new ones that take into account the features of liturgical texts and have not previously been presented in online systems. In our work, we limit ourselves to the material from Gotteslob (1975), but we also consistently take into account Gotteslob (2013) and the sources of hymn material used by the editorial boards of the respective editions. In further research, we would also like to consider the regional parts of these editions.

The users of the corpus will be historians and theorists of literature, verse scholars, liturgists, ethnographers, musicologists, and cultural researchers. It is this wide range of different specialists and their scientific interests that influenced the selection of features for meta-annotation.

## 2 Related work

The scheme of annotation for corpus data varies depending on the types of texts themselves, as well as on the goals that researchers set for themselves. To the best of our knowledge, there are not many projects on presenting poetic texts in electronic form.

The “Deutsch Diachron Digital” project brought together researchers from 15 universities and research institutes in Germany to create a number of diachronic corpora of the German language that cover different periods of its development: from Old German (Altdeutsch) to Early Modern High German (Frühneuhochdeutsch) (Deutsch Diachron Digital).

The Deutsches Textarchiv (DTA) consists of two parts: the main corpus called DTA-Kernkorpus, which includes texts from 1598 to 1913, and the expanded corpus called DTA-Erweiterungen, which covers the period from 1465 to 1969. The first one is balanced and can serve as a reference corpus for studying New High German. The total volume is about 400 million tokens.

The PO-EMO corpus (Haider et al., 2020) contains German and English poetic texts with annotation for esthetic emotions. The mentioned interdisciplinary project is carried out in line with computational poetry research using methods that are implemented in sentiment analysis. The XML markup was done line by line and includes the following parameters: meter, caesura\_rhythm, main\_accents, caesuras. The corpus of German poetry in New High German (Deutsches Lyrik Korpus, DLK) is described in (Haider and Eger, 2019; Haider, 2021). It includes poetic texts from publicly available German language corpora: the German Text Archive (DTA) and the Digital Library of Textgrid.

The Berliner Repertorium is an interdisciplinary database containing medieval German translations of Latin liturgical songs from the 9th to early 16th centuries (Berliner Repertorium). It comprises 471 Latin hymns, sequences and antiphons with their 3066 Middle High and Low German prose and poetic translations, paraphrases and glosses. Descriptions of the texts are supplemented by digitized copies, which can be useful for further research.

There are no specialized corpora of German-language liturgical poetry, which can be partly

explained by the relatively satisfactory documentation of these texts for literary purposes in the early stages of the formation of the genre thanks to the efforts of 19th-century scholars. At the same time, the collection of the Mainz hymnological archive “Mainzer Gesangbucharchiv”, counting more than 8,000 units, is at the initial stage of digitization (less than a hundred collections of chants, mainly from the 18th century, have been digitized); the creation of a corpus is not currently planned.

Among the poetic corpora for other languages, one can name the poetic corpus of the Russian National Corpus. It is one of the few that contains poetic texts with a total volume of more than 13 million words. The project “Music and Language in Danish Reformation Hymns” (Svendesen, Sørensen, Troelsgård, 2020) focuses on presenting the Reformation hymns in Danish and developing dictionaries of the corresponding period, which present vocabulary from the hymnbooks.

## 3 Corpus of liturgical texts in German

The hymns from Gotteslob (1975) cover the period from the Middle Ages to 1973; in the future, the corpus will be supplemented with the results of the analysis of Gotteslob (2013). Text processing involves five stages: preparation, meta-marking, tokenization, morphological and syntactic analysis. The preparatory stage involved determining the list of texts that should be included in the corpus. We have selected hymns that belong to the genre of Kirchenlied (church chants, which are the core part of the hymns). Outside the scope of our attention were the official translations of original Latin prayers, psalms, as well as litanies, texts of formulas, sacramentals and sacraments set to music.

Text mark-up was performed manually using the close reading method, since not all the information about the hymns that should have been presented in the corpus was indicated explicitly in the texts themselves, so they needed to be additionally read by an expert.

## 4 Text annotation

Text annotation was described in detail in (Sinclair, Ball, 1996), in which the authors identified internal and external markup parameters. The paper by Haider et al. (2020) discusses units that should be used for annotation

of poetic texts. Thus, the authors propose to follow the logical structure of a poetic work, distinguishing between lines, strophes and the text itself (poems).

In our case, we follow the same approach, however, understanding the hymn as a separate text. The markup that is introduced in our corpus

| Unit      | Number of Units |
|-----------|-----------------|
| tokens    | 33,266          |
| words     | 27,191          |
| sentences | 2,011           |
| hymns     | 232             |
| strophes  | 1,149           |

Table 1: Number of units.

includes the following levels: characteristics of the hymn and characteristics of the strophe (see Table 1 for statistics). The former include id, ancestor, year, category, category 2, rubric, while the latter are represented by id, original, year. Below we will dwell on each of them in more detail.

#### 4.1 Hymn id

The “*hymn id*” is understood as its number in the book, which in the traditional annotation scheme in a corpus corresponds to a text title. In total, we described 232 hymns.

#### 4.2 Hymn year

The parameter means the year of creation of the hymn as a whole, that is, the latest date (in case the hymn contain several parts, then the text is dated to an earlier period, i.e. its earliest part). For some hymns two dates may be given, indicating that the strophes were composed at different times and therefore belong to different periods (1940/1970). The most “productive” (in terms of tokens) period includes the 16th century, 17th century and the first half of the 20th century. The average length of a hymn is 143 words. The longest hymns, containing more than 300 words, were written in different periods: the second half of the 20th century (1959/1972), the 16th century (1537 and 1599), the 17th century (1656) and the 18th century (1771). The shortest hymns date mainly from the 16th century. (1522, 1528-1529, 1531), although short hymns from the 20th century are also found (1947 and 1965). Their length varies from 32 up to 45 words.

#### 4.3 Hymn ancestor

Since, as already indicated, the corpus is conceived as a source of data for studying the genesis of a church hymn (its origins, development and transformation), we paid attention to whether the original hymn exists. The “*hymn\_ancestor*” parameter indicates whether the ancestor text exists, and can be “yes” or “no”. The majority of hymns (91.4%) have no preceding text.

#### 4.4 Hymn category

The hymn category denotes the century of hymn creation: before the 16th century, 16th century, 17th century, 18th century, 19th century, first or second half of the 20th century.

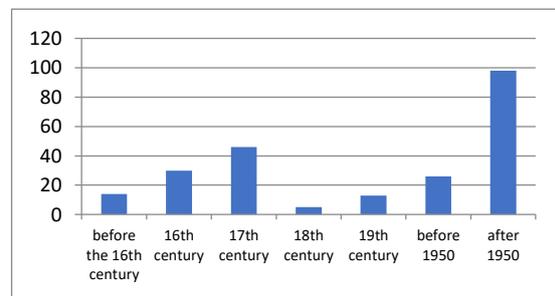


Figure 1: Distribution of hymns (time).

In Fig. 1, bar chart shows the distribution of hymns by time category.

#### 4.5 Hymn category 2

For a more accurate differentiation between hymns by time periods and in order to take into account their musical accompaniment, the texts were divided depending on whether the musical arrangement corresponded to the time of creation of the text (category T+M, namely text and music) or whether the text received a completely different musical arrangement, usually from a later period (category T, namely text only). The texts were marked using appropriate numerical tags (Table 2). Thus, *hymn\_category\_2*="4.1" for the above-mentioned hymn 615, means that it belongs to the 18th century and has musical accompaniment corresponding to the time of its creation.

In general, more than half of the hymns (about 60%) have musical accompaniment created during the period when the text was written.

|     | before the 16th cent. | 16th cent. | 17th cent. | 18th cent. | 19th cent. | before 1950 | after 1950 |
|-----|-----------------------|------------|------------|------------|------------|-------------|------------|
|     | 1                     | 2          | 3          | 4          | 5          | 6           | 7          |
| T   | 1.0                   | 2.0        | 3.0        | 4.0        | 5.0        | 6.0         | 7.0        |
| T+M | 1.1.                  | 2.1        | 3.1        | 4.1        | 5.1        | 6.1         | 7.1        |
|     | MA                    | R          | B          | AE         | H          | LR          | A          |

Table 2: Musical accompaniment in hymns.

We assume that the division into time categories may also correspond to the correlation of hymn texts by significant cultural and historical periods:

- MA – Middle Ages;
- R – Reformation;
- B – Baroque;
- AE – the Age of Enlightenment;
- H – historicism;
- LR – liturgical Renaissance in Germany;
- A – the era of K. Adenauer and the Federal Republic of Germany.

#### 4.6 Hymn rubric

In this case, the labels were applied from the perspective of pastoral-theological categorization, meaning the liturgical rubric. The parameter is important in the study of liturgical texts. But the division that exists in the hymnbook does not always correspond to the time of the church year, hence we observe a number of discrepancies. For example, in the church year, the period of Lent and most of Holy Week are combined together, while the eve of Easter already refers to Easter time. In Gotteslob (1975), the chants of Lent and the chants of Holy Week are separated. At the same time, the hymns for the eve of Easter are included in the circle of hymns for Holy Week.

Additional expert classification took into account the division of hymns corresponding to liturgical time, but also retained the division in the hymnbook that reflects the traditions of German liturgical practice (Table 3 shows the example).

| hymns ID | Gotteslob (1975)    | hymns ID | Expert classification |
|----------|---------------------|----------|-----------------------|
| 195-211  | Karwoche            | 207-256  | Osterzeit             |
| 288-311  | Vertrauen und Bitte | 541-547  | Fronleichnam          |
| 425-540  | Messgesänge         | 548-568  | Jesus Christus        |

Table 3: An example of hymn rubrics.

In total, 18 rubrics were marked in the hymnbook (marked with the “*Hymn\_rubric\_gl*” parameter), while the expert identified 14 items (marked with the “*Hymn\_rubric\_expert*” parameter).

#### 4.7 Strophe id

The parameter marks the strophe number within the hymn. A hymn can have from 1 to 15 strophes. For example, hymn 518, consisting of the maximum number of strophes (15 strophes), dates back to 1962, while hymns 130 and 156, consisting of 14 strophes, date back to 1962 and 1959/1972, respectively.

#### 4.8 Strophe original

A hymn may consist of strophes from different time periods, so for each strophe it is important to indicate whether it is original or not. This logical parameter can take the value “yes” or “no”. Most of the strophes (95.7%) are original.

#### 4.9 Strophe year

The parameter indicates the year when the strophe was written. As noted above, in a number of hymns it is the same for all strophes; in other cases, the first strophe is written earlier, while the rest are later. For the top 10, the classification of strophes by year generally repeats the classification of hymns by year, however, the 20th century prevails, because the strophes were written and added to the rest of hymns during this period.

### 5 Conclusion

This article presents the results of creating a corpus of Catholic hymns, which is the first attempt at developing such a corpus. This material has not been considered in existing text collections, databases, and corpora until now.

The paper described the initial results obtained for metatextual markup of texts. Currently, we plan to take into account such parameters of the verse as meter and rhyme. The next stage will include the morphological annotation of hymns, which will also require additional preparation, because automatic processing of diachronic texts can be tricky and laborious. Future work will include topic modeling of hymns and their clustering in order to identify interrelations between texts.

## References

- Berliner Repertorium (BR),  
<https://repertorium.sprachen.hu-berlin.de>
- Deutsch Diachron Digital (DDL),  
<https://www.deutschdiachrondigital.de/>
- Deutsches Lyrik Korpus (DLK),  
<https://github.com/tnhaider/DLK>
- Deutsches Textarchiv (DTA),  
<https://www.deutschestextarchiv.de/>
- 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.
- John McHardy Sinclair and J. Ball. 1996. EAGLES. Preliminary Recommendations on Text Typology. [online]  
<https://www.ilc.cnr.it/EAGLES96/texttyp/texttyp.html>
- Mette-Marie Møller Svendsen, Nicolai Hartvig Sørensen, and Thomas Troelsgård. 2020. An automatically generated Danish Renaissance Dictionary. In *Proceedings of the 2020 Globalex Workshop on Linked Lexicography*, pages 29–32, Marseille, France. European Language Resources Association.
- PO-EMO Corpus,  
<https://github.com/tnhaider/poetry-emotion>
- Russian National Corpus. <http://ruscorpora.ru>
- Text Encoding Initiative. <https://tei-c.org>
- Thomas Haider and Steffen Eger. 2019. Semantic Change and Emerging Tropes in a Large Corpus of New High German Poetry. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 216–222.
- Thomas Haider, Steffen Eger, Evgeny Kim, Roman Klinger and Winfried Menninghaus. 2020. PO-EMO: Conceptualization, Annotation, and Modeling of Aesthetic Emotions in German and English Poetry. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1652–1663, Marseille, France. European Language Resources Association.
- Thomas Haider. 2021. Metrical Tagging in the Wild: Building and Annotating Poetry Corpora with Rhythmic Features. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3715–3725, Online. Association for Computational Linguistics.

# EurLexSummarization - a new text summarization dataset on EU legislation in 24 languages with GPT evaluation

**Valentin Zmiycharov**

FMI, Sofia University

”St. Kliment Ohridski”

Sofia, Bulgaria

valentin.zmiycharov@gmail.com

**Todor Tsonkov**

FMI, Sofia University

”St. Kliment Ohridski”

Sofia, Bulgaria

ttsonkov@gmail.com

**Ivan Koychev**

FMI, Sofia University

”St. Kliment Ohridski”

Sofia, Bulgaria

koychev@fmi.uni-sofia.bg

## Abstract

Legal documents are notorious for their length and complexity, making it challenging to extract crucial information efficiently. In this paper, we introduce a new dataset for legal text summarization, covering 24 languages. We not only present and analyze the dataset but also conduct experiments using various extractive techniques. We provide a comparison between these techniques and summaries generated by the state-of-the-art GPT models. The abstractive GPT approach outperforms the extractive TextRank approach in 8 languages, but produces slightly lower results in the remaining 16 languages. This research aims to advance the field of legal document summarization by addressing the need for accessible and comprehensive information retrieval from lengthy legal texts.

**Keywords:** Legal texts summarization, Long texts summarization, New dataset.

## 1 Introduction

The task of automatically summarizing legislative documents poses a formidable challenge, primarily due to the extensive nature of these texts, making them intricate to comprehend and process. This complexity is further compounded when dealing with 24 languages, as there is a relative scarcity of pre-trained models for summarization in comparison to more widely-used languages such as English.

While existing text summarization models have shown promise, they are generally trained on shorter texts, such as social media posts or news articles, which significantly differ in complexity and

length from legislative documents. In the forthcoming section, there is an overview of text summarization approaches tailored for longer documents.

This paper presents a new dataset comprising EU legislative documents available in 24 languages. This dataset has undergone cleaning and preprocessing to ensure its utility and accessibility for research and development in the field of legal document summarization.

The contributions of this paper can be summarized as follows:

- Introduction of a multilingual legislative document summarization dataset, cleaned and pre-processed for immediate usage.
- Evaluation of the quality of summaries generated using the GPT model, coupled with a comprehensive comparative analysis against three distinct extractive summarization methods. This research aims to shed light on the efficacy of these techniques in the context of legislative documents and ultimately advance the state of the art in this critical domain.

## 2 Related work

Text summarization is a fundamental task in natural language processing, with a wide range of applications, from news article summarization to document summarization. In this section, we discuss related work in two key areas: text summarization datasets and the summarization of long and legislative documents.

## 2.1 Datasets

One critical aspect of text summarization research is the availability of diverse datasets for training and evaluation. The Document Understanding Conference (DUC) and Text Analysis Conference (TAC) have played a pivotal role in advancing text summarization research by providing benchmark datasets. Notable examples include DUC 2003 (National Institute of Standards and Technology, 2003) and TAC 2008 (National Institute of Standards and Technology, 2008), which have spurred innovation in extractive and abstractive summarization tasks.

Moreover, the CNN/Daily Mail dataset introduced by Hermann et al. has been influential in abstractive summarization. This dataset comprises news articles and corresponding human-generated summaries, serving as a valuable resource for training and evaluating abstractive summarization models (Hermann et al., 2015).

In addition to general text summarization datasets, there is a growing interest in domain-specific datasets, particularly in the field of legal text summarization. Legal documents, characterized by their complexity and extensive use of legal terminology, present unique challenges. Recent efforts, such as the creation of the "Multi-LexSum" dataset (Shen et al., 2022), focus on facilitating summarization specifically for legal texts, thus advancing the state-of-the-art in this domain.

## 2.2 Summarization of long and legal texts

Summarization of long documents, such as legislative texts, requires specialized techniques. Transformer-based models like BERTSUM (Liu and Lapata, 2019) and PEGASUS (Zhang et al., 2020) have demonstrated state-of-the-art performance in handling lengthy documents. These models leverage the ability to capture context over larger text spans, making them particularly well-suited for summarizing extensive legislative documents.

Efforts to improve legal text summarization also extend to the development of domain-specific pre-trained models. Models like LegalBERT (Chalkidis et al., 2020), fine-tuned on legal corpora, show promise in accurately summarizing legal documents, offering valuable resources for legal professionals and researchers.

The challenge of summarizing complex legal case judgments is addressed by conducting the first systematic comparison of various summarization

algorithms (Bhattacharya et al., 2019). Focusing on Indian Supreme Court judgments, the study evaluates both general and legally specialized algorithms, providing assessments against gold standard summaries. The research not only contributes to the advancement of summarization techniques for legal documents but also offers insights from a legal expert's perspective.

Introducing a novel approach to abstractive summarization of lengthy legal opinions, another method prioritizes the document's argument structure (Elaraby et al., 2023). By incorporating argument role information, it generates multiple candidate summaries and reranks them based on alignment with the document's argument structure. Demonstrating superior performance over robust baselines, this approach proves effective in summarizing complex and nuanced legal opinions.

Finally, the challenge of producing abstractive summaries for long texts is an ongoing focus of research. Techniques involving reinforcement learning and advanced decoding mechanisms have enhanced the quality and coherency of abstractive summaries, addressing the unique challenges presented by lengthy documents (Paulus et al., 2017).

In summary, text summarization has made significant strides, especially in the context of legislative documents. Dedicated datasets and advanced models have paved the way for more effective summarization of lengthy and complex texts. Domain-specific challenges in legal text summarization remain a prominent research agenda, with the potential to benefit legal professionals and society at large.

## 3 Dataset collection and preprocessing

The dataset contains legislative documents and their summaries in 24 languages from the European Union<sup>1</sup>. In addition to producing a new dataset to serve other researchers, these were used for various experiments.

The data is downloaded from the official website of the European Union<sup>2</sup>. It contains legal texts on various topics, including laws, acts and others. Each paper has a summary generated by an expert in the field. All documents and summaries, apart from English, have been translated into up to 23 other languages. Some summaries summarize more than 1 document. At the time of data collection, there are 1,816 summaries and their

<sup>1</sup>[https://huggingface.co/datasets/FMISummarization/FML\\_Summarization](https://huggingface.co/datasets/FMISummarization/FML_Summarization)

corresponding full documents in English. A comparison between the data in different languages is made in the following sections.

Many characteristics of the dataset make it valuable to the research field. It contains subject-specific text that is challenging for overtrained generic text models. The texts vary in length, with some documents being extremely long. Having them in different languages provides an opportunity to compare the results and validate whether the experiments work only in a particular case. The small number of documents further complicates the task.

### 3.1 Download and preprocess the data

For each full document and summary, the full HTML content is downloaded. All documents are crawled by taking the search results page by page and for each result the unique identifiers of the full documents and summaries are stored. In cases where there is no translation of the searched language for the relevant pair of (summary, full document), the information is not saved.

The header (Figure 1) and the list of references from the bottom part (Figure 2) are removed. There are cases where several documents correspond to one link. With them, the specified actions are repeated for each document and the content is concatenated.

Summaries are divided into sections. Figure 3 shows the number of documents with a certain number of sections. The most are documents with 6, 7 and 8 sections. After a detailed analysis of the sections, it turns out that a large number of them do not carry essential information for the summary. 408 different section titles have been identified. Table 1 shows the most popular 10 section names and the number of documents in which they occur for English.

As a summary we consider the text contained in one of the following sections: KEY POINTS, SUMMARY. Some summaries do not contain any of these sections. After manual review, it turned out that they were indeed invalid (example Figure 4). All essential information is contained in these sections, while the rest contain ancillary information that is not key to the summary or is already described in the main section.

Other documents that were removed from the dataset were those for which there was no transla-

tion in the respective language. Additionally, there are documents where the summary text is longer than the full document. Following a manual review, additional invalid documents were identified and subsequently removed. A total of 5118 out of 45983 documents (11.1%) were removed.

### 3.2 Data insights for English texts

The final result is 1816 summaries with their corresponding documents. The average length of full documents is 20716 words and of summaries 509 words. The word count ratio between full papers and summaries averaged 44.4. It is important to mention that the length of the texts is unevenly distributed with the presence of many deviations. For example, there are summaries with over 5,000 words (10 times the average) and full texts with over 800,000 words (40 times the average) (Figure 5).

To better visualize and gain a better idea of the distribution of the data, outliers were removed. For charting purposes only, the following have been removed:

- Complete documents with more than 100,000 words.
- Summaries of more than 2000 words.

As can be seen in Figure 6 the ratio of word counts between full papers and summaries is not fixed and varies greatly between examples. This makes the task of automatic summarization even more challenging.

### 3.3 Comparing data on different languages

Using the same methodology, data was downloaded in all 24 available languages. The number of documents in different languages varies, with English expectedly the most (1816) and Irish the fewest (511). The word count ratio between full documents and summaries is fairly constant across all languages with slight variations due to the specifics of the language and the documents being translated.

For each language, the key sections were identified, originating from the names in the language and analyzing the translations in the respective languages and the number of documents with the sections.

<sup>2</sup><https://eur-lex.europa.eu/homepage.html>

Figure 1: Header of the document.

- (<sup>1</sup>) Position of the European Parliament of 20 October 2009 (not yet published in the Official Journal) and Council Decision of 15 February 2010.
- (<sup>2</sup>) OJ L 298, 17.10.1989, p. 23. The original title of the act was 'Council Directive 89/552/EEC of 3 October 1989 on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the pursuit of television broadcasting activities'.
- (<sup>3</sup>) See Annex I, Part A.
- (<sup>4</sup>) OJ C 285 E, 22.11.2006, p. 126.

Figure 2: References of the document.

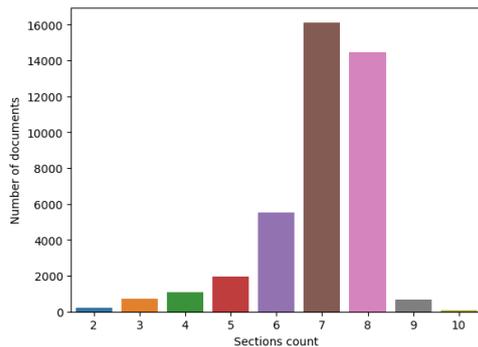


Figure 3: Number of sections in the summaries in the dataset. Most summaries have 6,7 or 8 sections. Minimum number of sections is 2 and maximum is 10.

## 4 Experimented approaches

### 4.1 Baseline

We assumed that the most important information is at the beginning of the document/section. Therefore, our baseline approach collects the first  $k$  consecutive sentences from the full texts. The number  $k$  of sentences is decided based on the number of words in the original summary. We continue adding sentences until we reach the number of words. Abstractive approaches like Pegasus (Zhang et al., 2020) also use the first part of the text. The results from the experiment reported below provide evidence that this assumption is good enough for a baseline.

### 4.2 TF-IDF Summarizer

We have also tested basic extractive summarization (Malik, 2019) based entirely on TF-IDF. The first step of the algorithm is to split the full text into a list of sentences. After that all special characters and stop words are removed. Stop words for various languages were sourced from the Stopwords ISO collection<sup>3</sup>. Then all sentences are tokenized. Next, the weighted frequency of occurrences of all

words must be calculated. The weighted frequency of each word can be found by dividing its frequency by the frequency of the most occurring word. After that, the words in the original sentences are replaced by their respective weighted frequency. The weighted frequency for the words removed during preprocessing is zero. For each sentence, the sum of weighted frequencies is calculated. Only sentences with more than three words are evaluated to avoid the ones that do not contain enough information. Finally, the sentences are sorted in descending order by the sum of the weighted frequencies. The summary contains the sentences at the beginning of the ordered list. The number of sentences to be selected is based on the ratio between the number of sentences in the training dataset. The algorithm does not require training and is entirely based on the content of the full document.

### 4.3 TextRank

For each sentence, we produce a vector of embeddings using Word2Vec (Mikolov et al., 2013). Word2Vec is an algorithm for generating a fixed-length distributed vector representation of all words in a huge corpus. The efficiency of Word2Vec is due to two reasons — one is the use of fixed-size vectors, which means that the size of the vector does not depend on the number of unique words in the corpus. Second, incorporating semantic information into vector representations. Word2Vec vectors are very effective at grouping similar words together. The algorithm can make strong judgments based on the position of the word in the corpus. For example, "handsome" and "nice" are similar, and therefore their vector representation will be very similar. The resulting vectors allow us to represent each sentence as a set of vectors for each word in them. To obtain vectors of the same size for

<sup>3</sup><https://github.com/stopwords-iso/stopwords-iso>

| Section name                         | Number of documents |
|--------------------------------------|---------------------|
| KEY POINTS                           | 1726                |
| SUMMARY OF:                          | 1608                |
| BACKGROUND:                          | 1601                |
| RELATED DOCUMENTS                    | 1172                |
| MAIN DOCUMENT                        | 1087                |
| KEY TERMS                            | 704                 |
| FROM WHEN DOES THE REGULATION APPLY? | 509                 |
| WHAT IS THE AIM OF THE REGULATION?   | 465                 |
| MAIN DOCUMENTS                       | 377                 |
| SUMMARY                              | 341                 |

Table 1: Section names and number of documents in which they appear (10 most popular).

**ACT**

Council Directive [2009/71/Euratom](#) of 25 June 2009 establishing a Community framework for the nuclear safety of nuclear installations.

**SUMMARY**

**WHAT IS THE AIM OF THIS DIRECTIVE?**

Figure 4: Summary with missing text.

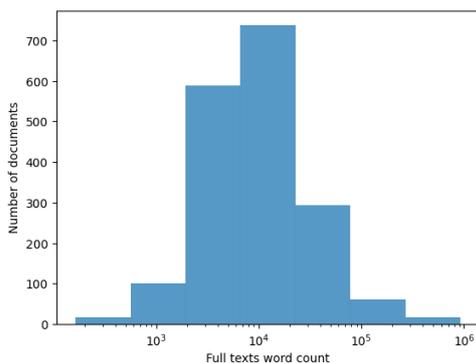


Figure 5: Number of words in full documents plotted on a logarithmic scale (x-axis). Examples with above 100000 words skew the average word count statistics.

all sentences, we zero-padded all but the longest sentences.

Once we have vectors that represent each sentence, we prepare a similarity matrix, which is square with the number of columns and rows equal to the number of sentences in the text. On it we apply the PageRank algorithm, which gives us a score for each sentence.

Once we have the scores, similar to the previous algorithm, we sort the sentences by the results obtained in descending order and concatenate the sentences in order until we reach the length of the original summary.

**4.4 GPT summarization**

We used the latest advancements in the machine learning field to generate summaries. We used OpenAI’s GPT API<sup>4</sup>.

We experimented with different prompts. The system prompt we ended up with is “*You are a lawyer.*” The aim is to order the agent to speak as if it is a lawyer. This matches the profile of the people who created the summaries. The document prompt is “*You will correctly answer the questions about the following text.*” concatenated with the full text. The prompt for the final task for the GPT API is “*Summarize the text without introductory words.*”. The reason we added the ending “*without introductory words*” is the following: The GPT API has maximum token limits, which are not enough for most of the texts. Therefore, we had to break the full text into chunks. Without this addition to the prompt, all the generated texts started with introductory words like “The text outlines” or “The most important parts of the text are”. These are repeated for each chunk and such introductions are not present in the real summaries. Prompts are translated into the 24 different languages using an automatic translator.

The price of GPT API usage is based on the

<sup>4</sup><https://platform.openai.com/docs/guides/gpt>

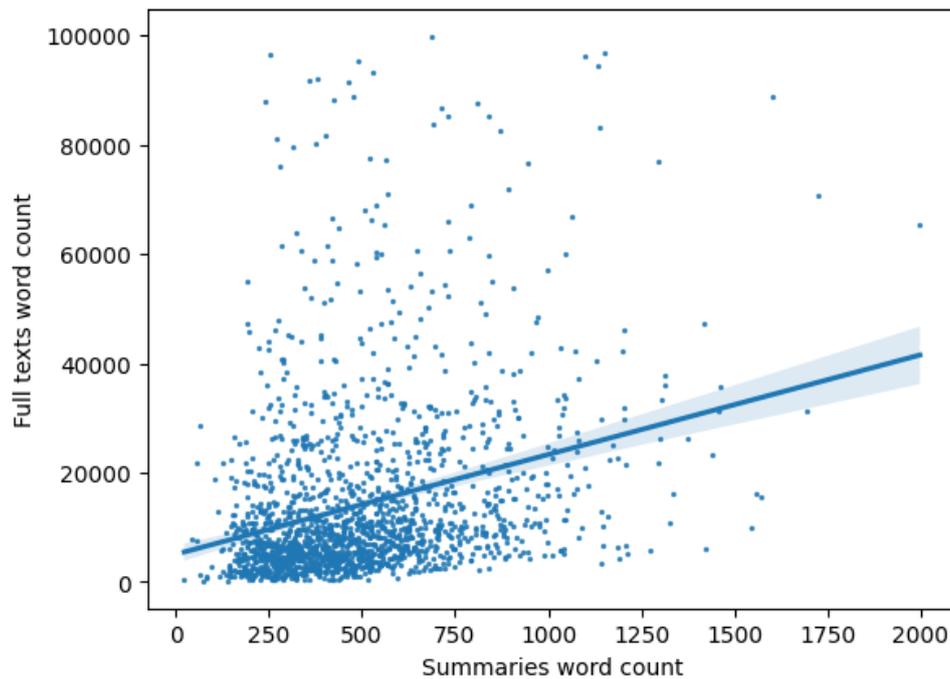


Figure 6: Each point represents the word count data of a summary and the corresponding word count of the full paper after removing outliers. The line and the space around it shows a linear regression model that best approximates the word count ratio between summaries and full papers.

number of tokens used. At the time of writing the paper the price for the model is \$0.0015/1K tokens for Input and \$0.002/1K tokens for output. Calculating all summaries for all languages would have cost \$600 with this model and \$14000 for the gpt-4 model. For price considerations we used the gpt-3.5-turbo model and generated summaries for 10% of the texts for all languages which cost us around \$60 with additional costs of around \$40 for various experiments.

The next challenge we faced was the ratio between the full text and the summary. As mentioned before we are aiming at generating summaries with the same length as the original summaries. On the other hand for GPT output we cannot precise the exact output length, only the maximum output length. When we call GPT API we have a maximum context of 4K tokens, which includes the input and the output. When we subtract the number of tokens from the prompt we are left with less than 4K for input text and output combined. What we tried first is to determine the ratio between the original full text and the summary. Let's say it is 9/1. We would then split the full text into chunks that are about 90% of the tokens allowed and leave 10% for the output. The problem we faced with this strategy is that in many cases the maximum

tokens allowed for the output is so small that the model cannot finish what it is trying to answer.

After experimenting with different maximum token counts we decided to fix the desired ratio between summary and full text to be 1/5 - roughly 600 tokens for summary, 3000 for full text. We multiply the length of the summary by 5. We additionally multiply this value by 1.5 because we cannot specify the exact number of tokens, but only max tokens, which means that we are aiming our input of full texts to be 7.5 times longer than the summaries we are trying to generate. If the original full text is longer than that we added a preprocessing step to reduce its content. We calculate the text rank scores for all sentences in the full text and remove the sentences with the lowest score until this ratio is achieved. This way our approach is combining extractive and abstractive methodologies for longer texts to achieve its goals. If the ratio is lower than 7.5 no preprocessing is done.

After the preprocessing step is executed, the text is split into chunks. Each chunk is a list of full sentences that do not reach the maximum allowed limit for tokens. Each chunk is summarized and results are concatenated

## 5 Experiments and results

In this section, our experiments are designed to comprehensively assess and compare the listed text summarization approaches across all 24 languages using the legislation dataset.

### 5.1 Experiments Design

The most widely used metric for the evaluation of text summarization is rouge (Recall-Oriented Understudy for Gisting Evaluation). Rouge is a set of metrics used for evaluating automatic summarization and machine translation software. The metrics compare an automatically produced summary to a human-produced summary. Rouge-N refers to the overlap of n-gram between the system and reference summaries. Rouge-L refers to Longest Common Subsequence (LCS) based statistics. The longest common subsequence problem considers sentence-level structure similarity naturally and identifies the longest co-occurring in sequence n-grams automatically. In particular rouge-1, rouge-2, and rouge-L F1 scores were used in the conducted experiments.

We started generating summaries that have the same length as the original summary. This way the precision and recall are the same. This way we avoid the problem of changing the F-score due to generating larger or smaller summaries and focus entirely on the relevance of the sentences and not on the length of the summary. For all experimented approaches we selected only sentences with at least 3 words.

### 5.2 Analysis of Results

Table 2 shows the results of the different approaches. The TF-IDF approach improves the baseline for all languages. TextRank and GPT outperform TF-IDF for all languages. For 8 of the languages, GPT outperforms TextRank, while for the others TextRank is the best.

| Lang | Baseline | TF-IDF | TextRank     | GPT          |
|------|----------|--------|--------------|--------------|
| BG   | 0.284    | 0.297  | 0.318        | <b>0.323</b> |
| CS   | 0.248    | 0.269  | 0.285        | <b>0.287</b> |
| DA   | 0.338    | 0.350  | 0.379        | <b>0.389</b> |
| DE   | 0.326    | 0.338  | <b>0.368</b> | 0.364        |
| EL   | 0.279    | 0.286  | 0.309        | <b>0.316</b> |
| EN   | 0.364    | 0.384  | <b>0.416</b> | 0.405        |
| ES   | 0.381    | 0.389  | <b>0.414</b> | 0.408        |
| ET   | 0.194    | 0.209  | <b>0.232</b> | 0.216        |
| FI   | 0.252    | 0.252  | <b>0.286</b> | 0.269        |
| FR   | 0.375    | 0.385  | <b>0.416</b> | 0.396        |
| GA   | 0.335    | 0.328  | <b>0.348</b> | 0.324        |
| HR   | 0.235    | 0.243  | <b>0.268</b> | 0.268        |
| HU   | 0.290    | 0.290  | <b>0.320</b> | 0.318        |
| IT   | 0.378    | 0.375  | <b>0.412</b> | 0.399        |
| LT   | 0.235    | 0.242  | <b>0.262</b> | 0.252        |
| LV   | 0.238    | 0.238  | <b>0.265</b> | 0.261        |
| MT   | 0.229    | 0.232  | <b>0.265</b> | 0.242        |
| NL   | 0.302    | 0.308  | 0.336        | <b>0.339</b> |
| PL   | 0.254    | 0.260  | <b>0.279</b> | 0.271        |
| PT   | 0.383    | 0.391  | <b>0.423</b> | 0.388        |
| RO   | 0.359    | 0.377  | <b>0.402</b> | 0.400        |
| SK   | 0.237    | 0.251  | 0.269        | <b>0.270</b> |
| SL   | 0.249    | 0.253  | 0.284        | <b>0.288</b> |
| SV   | 0.330    | 0.327  | 0.370        | <b>0.374</b> |

Table 2: Rouge 1 F1 scores for all 24 languages for all experiment types.

## 6 Conclusion

The paper introduces a new multilingual dataset of European legislative laws, encompassing 24 languages. This dataset, characterized by significant variations in document length and a limited number of documents per language, presents a valuable resource for the field of automatic summarization.

We conducted extensive experiments, employing three extractive summarization approaches, and introduced a novel two-step methodology that harnesses the capabilities of GPT models for summarization. Notably, our two-step approach outperforms existing methods in some languages, demonstrating its potential as an effective summarization technique. However, in certain cases, TextRank surpasses it in performance.

In addition to our summarization findings, we offer valuable insights into the practical considerations of using GPT, including detailed information on associated costs. This information is essential for researchers and practitioners looking to leverage state-of-the-art models in real-world applica-

tions.

By addressing the complex task of summarizing European legislative laws in diverse languages, our work contributes to the advancement of the field, offering a valuable resource and novel techniques for future research and applications in automatic summarization and multilingual natural language processing.

## 7 Acknowledgements

This research is partially supported by Project UNITE BG05M2OP001-1.001-0004 funded by the OP “Science and Education for Smart Growth” and also co-funded by the EU through the ESI Funds.

## References

- Paheli Bhattacharya, Kaustubh Hiware, Subham Rajgaria, Nilay Pochhi, Kripabandhu Ghosh, and Saptarshi Ghosh. 2019. A comparative study of summarization algorithms applied to legal case judgments. In *Advances in Information Retrieval*, pages 413–428, Cham. Springer International Publishing.
- Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. **LEGAL-BERT: The muppets straight out of law school**. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.
- Mohamed Elaraby, Yang Zhong, and Diane Litman. 2023. **Towards argument-aware abstractive summarization of long legal opinions with summary reranking**. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7601–7612, Toronto, Canada. Association for Computational Linguistics.
- Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. **Teaching machines to read and comprehend**.
- Yang Liu and Mirella Lapata. 2019. **Text summarization with pretrained encoders**.
- Usman Malik. 2019. **Text summarization with nltk in python**. <https://stackabuse.com/text-summarization-with-nltk-in-python/>.
- Tomas Mikolov, Kai Chen, G.s Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR*, 2013.
- National Institute of Standards and Technology. 2003. **DUC 2003 Dataset**. <https://duc.nist.gov/duc2003/tasks.html>.
- National Institute of Standards and Technology. 2008. **TAC 2008 Dataset**. <https://tac.nist.gov/2008/summarization/update.summ.08.guidelines.html>.
- Romain Paulus, Caiming Xiong, and Richard Socher. 2017. **A deep reinforced model for abstractive summarization**.
- Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. 2022. **Multilexsum: Real-world summaries of civil rights lawsuits at multiple granularities**.
- Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. **Pegasus: pre-training with extracted gap-sentences for abstractive summarization**.

# On a Hurltlex resource for Bulgarian

**Petya Osenova**

Department of Bulgarian Language  
Sofia University "St. Kl. Ohridski"

`petya@bultreebank.org`

## Abstract

The paper reports on the cleaning of the Hurltlex lexicon for Bulgarian as part of the multilingual Hurltlex resource.

All the challenges during the cleaning process are presented, such as: deleting strings or lexica that are clear errors from the automatic translation, establishing criteria for keeping or discarding a lexeme based on its meaning and potential usages, contextualizing the lexeme with the meaning through an example, etc. In addition, the paper discusses the mapping of the offensive lexica to the BTB-Wordnet as well as the system that has been used.

**Keywords:** Bulgarian, Hurltlex, mapping, wordnet, offensive language.

## 1 Introduction

The Hurltlex resource is a multilingual lexicon of hate/offensive words. This resource started with the Italian hate lexicon developed by Tullio De Mauro and organized in 17 categories. Later on, it was automatically extended to other languages – through links to available synset-based lexical thesauri, among them MultiWordNet (Pianta et al., 2002) and BabelNet (Navigli and Ponzetto, 2010). Also machine translation was used (Bassignana et al., 2018).

HurtLex is a lexicon of offensive, aggressive, and hate words in over 50 languages among which Bulgarian. The words were classified in 17 categories like the source area of usage (plants, animals) or the target qualities or groups (moral and behavioral defects, with potential negative connotations). Also, a category was added whether the word expresses a stereotype or not. Since stereotypes are culture specific and usually determined through specially designed questionnaires within sociolinguistic frameworks, for Bulgarian this information has not been gathered yet. Thus, for the

moment the inherited stereotypes from the Hurltlex resource are available without a focused localized justification..

The hate words belong also to two other categories: either *conservative* (where only words with offensive senses were translated from the original lexicon) or *inclusive* (where all the potentially relevant senses of the words in the original lexicon were translated). As a result from the automatic translation due to the non-exhaustive coverage of various linked multilingual wordnets and lexical resources, the translated counterparts might include ambiguous, incorrect or questionable words.

Thus, during the process of obtaining the Hurltlex lexicons in other languages, it became clear that there is noise in the resulting resources such as inappropriate, non-comprehensible or unclear in their offensive meanings words. As a result, some cleaning is needed over the obtained lexicon per language. Such a cleaning, for example, was already performed for Modern Greek by Stamou et al. (2022).

The authors detect Greek offensive language by cross-classifying words on three dimensions: context, reference, and thematic domain. They add to the categorization also the social and the cultural aspect which are language specific.

Here we present a work in progress on the cleaning and linking of the Bulgarian part of Hurltlex. Version 1.2 for Bulgarian was downloaded from the hurtlex repository<sup>1</sup> and then manually checked. Initially, it had altogether 2865 words but some were already initially deleted since there were repetitions of the wordforms of the same lexeme. Also, normalization was performed over the same words with varying spelling – mostly the same word with a capital and small letter.

At the moment the lexicon consists of 1370

<sup>1</sup><https://github.com/valeriobasile/hurtlex/tree/master/lexica/BG>

words which means that slightly more than 50 % of the original list of words were discarded before the human checks. At the same time, the lexicon is further enriched with the synonyms in the appropriate synsets through the sense mappings with BTB-Wordnet (Simov and Osenova, 2023). In addition, it has to be noted that Hurltex does not include only isolated words but also phrases. Despite this fact, the proper handling of multiword expressions remains for future work since some of them came as isolated words and other – as expressions.

The structure of the paper is as follows: the next section presents the system that has been used for cleaning and enriching the resource. Section 3 discusses in more detail the workflow and challenges when editing the data. Section 4 focuses on the mapping of the Hurltex lexicon to the BTB-Wordnet. Section 5 outlines the conclusions.

## 2 The resource editing system

The Hurltex list for Bulgarian was inserted into a customized version of the specially designed CLaDA-BG Dictionary Creation System (Angelov et al., 2022). The system presents a multifunctional editor that can be used for creating various types of lexical resources – thesauri like wordnets, or traditional types of dictionaries (specialized, explanatory, spelling, etc.). The system also provides possibilities of interlinking the available data depending on the goal.

In general, the following information is present in the customized version:

- the lexeme/phrase itself
- the part-of-speech of the lexeme or of the headword of the phrase
- the status of the entry completion (with labels *Ready*, *To check*, *Irrelevant*)
- indication of whether the word is present in the wordnet or not (with labels T(rue) or F(alse))
- the definition of the lexeme meaning (it comes from the wordnet, if it is present there, or from other available dictionaries – if the lexeme and/or its appropriate meaning is absent)
- comments on any related to the check-ups issues

A partial screenshot from the system is given in Fig. 1. In the first column from left to right, offensive words are given that relate to being lazy and being scruffy or dirty. The second column indicates the part of speech. Here the examples are mostly nouns as in the resource itself. The third column marks the status of the entry completion – in this case marked as *Ready*. In column 4 the Hurltex ID is given. Column 5 keeps track to the word as it came from the automatically generated Hurltex resource. This means that the word might have come with an unnecessary capital letter or in a certain wordform instead of a lemma, or within an ungrammatical expression, or as an error, etc. However, the edited final lexeme is in the first column form since there the words have been normalized to their lemmas and proper spelling. The last column that is shown here, outlines the mapping to the appropriate entry in BTB-Wordnet or, if missing there, the relation to a suggested definition.

An example of a word is the last one down left in the table мърморко, ‘marmorko-M.SG’ (a grumbler). An example for a phrase is социален аутсайдер, ‘sotsialen-M.SG autsayder-M.SG’ (a social outsider). Both get a noun as part of speech.

The status marking shows whether the checking is accomplished (label *Ready*) or it should be paid attention to later (label *To check*). It also indicates whether the word is considered non-appropriate for the lexicon (label *Irrelevant*). We are aware of the fact that such decisions are not easy to take since many words can become offensive in a given context. For that reason they are not removed but just marked as not appropriate. Thus, we start with the lexemes or phrases with offensive lexical meaning and gradually will consider also the context-bound cases.

If the word is already present in the wordnet with the offensive meaning, then its ID, category, definition, examples, etc. are copied into Hurltex. If however, the lemma/meaning is not present there, a definition is added by the editing linguist together with an example. The new information will be added to the BTB-Wordnet as well.

The linguist can leave comments of the following types: either there is no definition in the available dictionaries and other sources, or the definition is newly constructed, or the meaning is not offensive.

Despite the labels of readiness, additional systematic checks will be necessary after this first

|              |      |        |        |            |   |
|--------------|------|--------|--------|------------|---|
| мързел       | noun | Готово | BG8    | Мързел     | Id: 12508 noun.person LEMMA: мързел, мързеливец, мързеливк      |
| мързелан     | noun | Готово | BG82   | мързелан   | Id: 12508 noun.person LEMMA: мързел, мързеливец, мързеливк      |
| мързелив     | adj  | Готово | BG1519 | мързелив   | Id: 617 adj.all LEMMA: ленив, мързелив DEF: Който не обича и не |
| мързеливец   | noun | Готово | BG1622 | мързеливец | Id: 12508 noun.person LEMMA: мързел, мързеливец, мързеливк      |
| мърлява жена | noun | Готово | BG295  | мърляв жен | Id: 718 adj.all LEMMA: мръсен, изцапан, замърсен, нечист, мърг  |
| мърляч       | noun | Готово |        |            | Несръчен човек, неумел, лош работник или занаятчия              |
| мърляч       | noun | Готово | BG400  | мърляч     | Мръсен и противен тип.  |
| мърморко     | noun | Готово | BG2040 | мърморко   | Човек, който все мърмори и е недоволен.                         |

Figure 1: An example screenshot from the system.

phase of Hurltlex data cleaning. Also, a strategy is required on how to deal with words are not offensive through their lexical meaning but might become offensive in a certain expression or communicative situation. For example, the word *птица*, ‘ptitsa-F.SG’ is not offensive but in the expression *странна птица*, ‘stranna-F.SG ptitsa-F.SG’ (a strange person) it might become offensive.

Another observation is that many words are not offensive but they just refer to intolerable models in society or bear mostly neutral meanings. For example, such type of words are *фалш*, ‘falsh’ (falseness) or *мълва*, ‘malva-F.SG’ (a rumor). An additional obstacle is the fact that we had to work with words without senses and contextualizing examples.

Let us look more closely into the challenges of cleaning Hurltlex in the next section.

### 3 Cleaning Hurltlex: Workflow and Challenges

The workflow includes several steps and these are organized in the following way:

- The lexicon is available in the editing system where one keeps track to the initial variant of the resource but also manipulates the data accordingly.
- The content is checked - initially it is done by the alphabetically ordered words, and later it can be performed through various characteristics such as the level of completeness, the (non)-availability in the wordnet, the part of speech, etc.
- The words/phrases are categorized into three types: *Ready*, *To check* and *Irrelevant*. It should be noted that all of them are being checked later again, including the ones labeled *Ready*.
- The correctness of the part of speech is also checked and changed, if necessary. In case

of phrases, the part of speech is equal to the headword.

- If the lexeme is currently present in the wordnet with the appropriate meaning, it is marked with the boolean value *True (T)*, and then all the wordnet information is copied. Respectively, if the lexeme or the meaning is not present in the wordnet, the boolean value *False (F)* is selected. The inclusion of the missing words and/or meanings to the wordnet is envisaged as a future step. Also, the potential of the existing SentiWordNet (Baccianella et al., 2010) will be researched with the aim to see what part of the words/expressions with negative polarity intersect with the offensive lexica.
- Examples are added to each offensive meaning of the lexeme. Thus, the it can be seen in an appropriate context.

Thus, for the word *тиква*, ‘tikva’ (pumpkin) which is an offensive-oriented name of the head as part of the human body, there is an accompanying offensive synset in the wordnet, and the information is copied here. This entry can be seen in Figure 2. It provides the category *noun.body*, the set of synonyms as well as the definition. It should be noted that at this stage the synonyms in BTB-Wordnet are not marked as colloquial, dialectal, jargon and the like. These markings are envisaged for the future. Then comes the spelling of the word, its part of speech – noun, the status of the entry – in this case *Ready*, and the inclusion in wordnet – in this case *True*. The method of having obtained this word in Bulgarian is marked as *Inclusive*, i.e. all the potentially relevant senses were translated. The Hurltlex category is abbreviated as *or* – this means a relation to Plants. The lexeme is not considered stereotypical but as it was mentioned above, we should take this inherited information with a grain of salt since this dimension of information requires a special survey.

|                  |  |
|------------------|--|
| <b>ENTRY</b>     |  |
| Def:             | Id: 8217 noun.body LEMMA: главица, китара, глава, главичка, кратуна, тиква, чутура, куфалница DEF: Най-горната част на човешкото тяло или предната част на тялото на животно, където се намират мозъкът, устата и повечето сетивни органи. |
| Id:              | 2154   |
| Phrase:          | тиква  |
| Pos:             | noun   |
| Status:          | Готово   |
| Wordnet:         | .T.  |
| <b>INCLUSIVE</b> |  |
| Category:        | or   |
| Example:         | Арогантният Бойко надигна тиквата с помощта на продажната десница.   |
| Id:              | 3153   |
| Parent_id:       | 2154   |
| Stereotype:      | no   |

Figure 2: The information ticket for the word тиква, ‘tikva’ (pumpkin).

The identified challenges during the cleaning process are as follows: a) deleting strings or lexica that are clear errors resulting from the automatic translation, b) trying to establish criteria for keeping or discarding a word based on its lexical meaning and potential usages, c) contextualizing the word with an offensive meaning through an appropriate example, etc.

Let us look closer into each of these challenges:

**Errors** Here some errors are considered, such as collocations of arbitrary words or stop words. For the former some examples are: ябълка паркет, ‘yabalka-F.SG parket-M.SG’ (apple parquet) or томахавака-F.SG работа-F.SG ‘tomahavka rabota’ (tomahawk business); for the latter – с изключение на, ‘s-PREP izklyuchenie-N.SG na-PREP’ (with the exception of).

**Criteria** Here the main problem is the presence of isolated words or phrases that might have offensive meaning in some context or from a certain perspective but without this background information it is difficult to decide.

Among the data that is not considered as part of the Bulgarian Hurltex are the following ones: occupation like собственик на бизнес ‘sobstvenik-M.SG na-PREP biznes-M.SG’ (a business owner); специален пратеник, ‘spetsialen-M.SG pratenik-M.SG’ (a special correspondent); предприемач, ‘predpriemach-M.SG’ (an entrepreneur), портierer, ‘portier-M.SG’ (a porter); domains like растител-

ни биотехнологии, ‘rastitelni-PL biotehnologii-F.PL’ (plant biotechnologies); some common language words like преобличане, ‘preoblichane-N.SG’ (changing clothes); пропуск, ‘propusk-M.SG’ (a pass); нормален, ‘normalen-M.SG’ (normal).

The words/expressions that have to be considered carefully, are the ones that do not have offensive connotations per se. For example, the lexeme умен, ‘umen-M.SG’ (smart) can become offensive only in an ironic context where the expression Много си умен, ‘Mnogo-ADV si-AUX.2PERS.SG umen-M.SG!’ (You are very smart!) would mean the opposite statement Много си глупав!, ‘Mnogo-ADV si-AUX.2PERS.SG glupav-M.SG!’ (You are very stupid!). Needless to say, handling irony is an important part of all the tasks related to detecting offensive language and attitude. In such tasks very often lexicons with sentiment and offensive lexica become an integral part of the complex architectures – see for example in (Hernández Farías et al., 2015). At the same time, focused annotated data, embeddings at various levels as well as Large Language Models are used for modeling contexts and pragmatic conditions.

**Examples** The idea of adding appropriate examples to the offensive meanings is related to the problems, described above. Adding examples to only-offensive words is by no means very important for introducing the typical context of usage. For example, the qualitative word with an offensive nuance темерут, ‘temerut-M.SG’ (an antisocial person) that is classified as CDS - derogatory words, is illustrated with the following example from the media: Този с държавната работа е мързелив темерут, който е постоянно ритан и бутан от по-мързеливи и корумпирани темерути като него. (*The one with the government job is a lazy antisocial person who has been constantly kicked and pushed around by even lazier and more corrupted antisocial people like him.*). It should be noted that in this example also other offensive words were used, namely the adjectives мързелив, ‘marzeliv-M.SG’ (lazy), and корумпиран, ‘korumpiran’ (corrupted).

But examples are even more important in non-trivial cases like the ones where a positive word can be turned into an offensive one, or the word is offensive with its figurative meaning only, etc. For example, the word професор, ‘profesor’ (a professor) might be used ironically to offend someone

who claims to be talking with competence on many topics: Тоя миндъовец (демек, селянин от с. Миндя) се изказва по всяка тема - голям "професор" се извъди. (*This man mindyovets (i.e. from Mindya village) speaks out on every subject – a great "professor" he has become.*)

For the selection of appropriate examples in Hurltex we used internet and the Bulgarian part of the CLASSLA corpora – see more in (Ljubešić and Kuzman, 2024) – since it covers also non-standard communication in blogs, forums, etc.

#### 4 Mappings to BTB-Wordnet

Since no mapping to synsets in Bulgarian wordnets was available in the originally compiled Hurltex, we established our own mapping. In some cases, Hurltex is being enriched with the synonyms of the lexeme – either also offensive ones, or from other registers. In other cases, the lexeme is not present in the wordnet and thus has to be mapped through some strategy such as a link to an appropriate hyponym and/or a hypernym.

The most common cases of not having links to wordnet are the following:

- *The word is missing.* For example, the word дивак, ‘divak-M.SG’ (a savage) is missing. The same holds for the word дрипльо, ‘driplyo-M.SG’ (a ragamuffin) and many others.
- *The word is present but the appropriate sense is missing.* This factor is the most frequent one. For example, the word свиня, ‘svinya-F.SG’ (a swine) has only the meaning of a domestic animal, but not the meaning of a filthy or bad person. The same holds for the word червей, ‘chervey-M.SG’ (a worm). It has the meanings of the animal and the computer virus but not the one of an insignificant person.

Table 1 shows some statistics on the Bulgarian Hurltex data. The label *In Wordnet* indicates, as mentioned above, that more than half of the lemmas are not included there either as a lemma, or as an appropriate meaning.

The *Ready entry* means that there is no problem of identifying the word or phrase as an offensive one. More entries are like this but also a substantial part requires further elaboration. For example, the lexeme козел, ‘kozel-M.SG’ (a billy-goat) is

| Label         | True | False |
|---------------|------|-------|
| In Wordnet    | 641  | 729   |
| Nouns         | 1119 | 251   |
| Ready entry   | 817  | 553   |
| No definition | 72   | 1298  |

Table 1: Some statistics on the current status of the Bulgarian Hurltex resource.

marked as *Ready* but the link to the wordnet is to the meaning of man with a long sharp beard only. In fact, the examples show wider offensive usage towards an old man with an inappropriate behaviour.

*No definition* refers to cases where a definition in this meaning cannot be found on a regular basis in Bulgarian dictionaries. But such cases are very rare. Most cases here refer to words that do not have an offensive meaning (like празненство, ‘praznenstvo-N.SG’, (celebration)) and to words that show ironic meaning in the relevant context.

#### 5 Conclusions

This paper presents work in progress on the cleaning and enriching the Bulgarian Hurltex resource. The main steps in this ongoing work and the related challenges have been outlined. Another challenge that was not mentioned before is the difficulty to handle the contemporary cases with rapid change of some words from positive to negative or with the emergence of an offensive meaning in a specific context like the political arena.

The next steps include: a) the handling of context-dependent offensive lexica; b) the handling of multiword expressions and bigger contexts; c) the addition of the missing lemmas and meanings into BTB-Wordnet.

As future work we also consider the annotation of a corpus with offensive senses. Only in this way more precise criteria can be set with respect to the delimitation of the offensive meanings within their context-bound process of generation.

#### References

- Zhivko Angelov, Kiril Simov, Petya Osenova, and Zara Kancheva. 2022. The CLaDA-BG Dictionary Creation System: Specifics and Perspectives. In *CLARIN Annual Conference Proceedings*, pages 24–28, Prague, Czechia.
- Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. *SentiWordNet 3.0: An enhanced*

- lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).
- Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. **Hurtlex: A multilingual lexicon of words to hurt**. In *Italian Conference on Computational Linguistics*.
- Delia Irazú Hernández Farías, Emilio Sulis, Viviana Patti, Giancarlo Ruffo, and Cristina Bosco. 2015. **ValenTo: Sentiment analysis of figurative language tweets with irony and sarcasm**. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 694–698, Denver, Colorado. Association for Computational Linguistics.
- Nikola Ljubešić and Taja Kuzman. 2024. **CLASSLA-web: Comparable web corpora of South Slavic languages enriched with linguistic and genre annotation**. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3271–3282, Torino, Italia. ELRA and ICCL.
- Roberto Navigli and Simone Paolo Ponzetto. 2010. **BabelNet: Building a very large multilingual semantic network**. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden. Association for Computational Linguistics.
- Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. **Multiwordnet: developing an aligned multilingual database**. In *Proceedings of the First International Conference on Global WordNet*.
- Kiril Simov and Petya Osenova. 2023. **Recent developments in BTB-WordNet**. In *Proceedings of the 12th Global Wordnet Conference*, pages 220–227, University of the Basque Country, Donostia - San Sebastian, Basque Country. Global Wordnet Association.
- Vivian Stamou, Iakovi Alexiou, Antigone Klimi, Eleftheria Molou, Alexandra Saivanidou, and Stella Markantonatou. 2022. **Cleansing & expanding the HURTLEX(el) with a multidimensional categorization of offensive words**. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 102–108, Seattle, Washington (Hybrid). Association for Computational Linguistics.

# A Unified Annotation of the Stages of the Bulgarian Language. First Steps

**Fabio Maion**

Leopold-Franzens-Universität  
Innsbruck  
Fabio.Maion@uibk.ac.at

**Tsvetana Dimitrova**

Institute for Bulgarian  
Language, Bulgarian Academy  
of Sciences  
cvetana@dcl.bas.bg

**Andrej Bojadžiev**

St. Kliment Ohridski Sofia  
University  
aboy@slav.uni-sofia.bg

## Abstract

The paper reports on an ongoing work on a proposal of guidelines for unified annotation of the stages in the development of the Bulgarian language from the Middle Ages to the early modern period. It discusses the criteria for the selection of texts and their representation, along with some results of the trial tagging with an existing tagger which was already trained on other texts.

**Keywords:** tagging, historical data, Bulgarian language

## 1 Introduction

So far, research on Middle Bulgarian and early Modern Bulgarian texts has not followed systematically applied methods of corpus linguistics, and no attempts have been made to integrate the corpora with electronic descriptions (and editions) of texts and manuscripts. Separate corpora of texts reflecting different stages of the history of the Bulgarian language are part of other large corpora which also contain Old Russian and Old Serbian sources<sup>1</sup>, as well as historical material of other Indo-European languages (Greek, Latin, Armenian, among others<sup>2</sup>). Numerous texts have

been gathered in the Historical corpus of the Bulgarian language (from 10<sup>th</sup> to 17<sup>th</sup> century)<sup>3</sup>.

In addition, the annotation schemes followed by the research efforts (if available at all) differ quite a lot as they reflect specific research purposes and cannot be uniformly applied. This means that the results of indexing and data analysis cannot be based on uniform criteria and cannot be used for comparison purposes. In fact, some textual collections contain only searchable texts (concordances can be made using other external tools) without any further linguistic annotation, active links to dictionaries or other useful tools.

The present paper reports on the development of an ongoing project which aims at offering a proposal for a unified annotation for the Bulgarian texts of all stages in the history of the Bulgarian language. We have started from the annotation principles of the PROIEL project<sup>4</sup> to extend the linguistic annotation to be further used for Middle Bulgarian and early Modern Bulgarian texts while considering other recent efforts in this direction (Šimko, 2021; Šimko et al., 2021). A dictionary of words and their wordforms is generated, and an applicable model for description of texts and manuscripts<sup>5</sup> is to be integrated to allow for electronic publication of the texts along with applicable metadata. A similar project<sup>6</sup> for Middle Bulgarian dealing with the translation of Philip Monotrop's work *Dioptra* in the 14th century, is

---

<sup>1</sup> <http://www.manuscripts.ru/>

<sup>2</sup> <http://dev.syntacticus.org/#annotation-principles>

<sup>3</sup> <https://histdict.uni-sofia.bg/textcorpus/list>

<sup>4</sup> <https://github.com/proiel>

<sup>5</sup> <http://repertorium.obdurodon.org/>

<sup>6</sup> <https://m.pf.fwf.ac.at/en/research-in-practice/project-finder/57346>

being developed at the University of Innsbruck. In this project, the Slavonic translation and the Greek original are lemmatized, morphologically annotated and aligned on the word level, thus yielding a searchable bilingual corpus.

## 2 The approach

The goal of the project *A Unified Annotation of the Stages of the Bulgarian Language (AUSBUL)* is to create a model for infrastructure which will make the texts and the annotated data online accessible and user-friendly for researchers and other potential users. The infrastructure integrates several components:

1. A corpus of texts in Cyrillic that are formatted according to uniform criteria, suitable both for electronic publication and being enriched with linguistic annotation.
2. Linguistic annotation (morphological and syntactic annotation; plus lemmatization with reference to earlier (attested) and modern variants of the words) that follows standardized methods adopted in corpus linguistics and established by practice.
3. Linking the texts in the corpus with their electronic descriptions, along with a catalogue of their sources.
4. Metadata to the texts such as: information about the authors (and editions of the manuscripts and/or texts), references to dates and places found in the texts or other information that is necessary for understanding their context.

The project draws upon the idea that only such an integrated approach, devoted to sources from one era and resting on standardized practices and solutions, can provide new insights into the history of language and literature, to be further compared with similar phenomena from other cultures. The emphasis of our effort is to test for the possibility of applying the principles of corpus and computational linguistics to a selection of different types and genres of medieval texts. The result will be a unified model for describing the language and the texts of the Middle Bulgarian (13<sup>th</sup> – 14<sup>th</sup> centuries) and the Early Modern Bulgarian (17<sup>th</sup> – 18<sup>th</sup> centuries) periods with respect to both archaic and vernacular samples. By following uniform principles of electronic linguistic annotation, we may compare and analyze different phenomena in

the development of the grammatical system of the Bulgarian language.

At this first stage of the project, work is being done in several directions:

1. Selection of the texts to be included in the corpus.
- This part of the project requires a considerate analysis of the history of the texts. For example, some texts must be selected to make possible the comparison of the earlier (or archaic) variants with the Early Modern Bulgarian variants that reflect a more vernacular version of the language.
2. Brief description of the sources (manuscripts) in which they are found.
3. Compilation of a bibliography for the texts and their sources (manuscripts).
4. Trial annotation with texts of different genres and periods.

## 3 The texts

We will illustrate the approach with one of the texts currently being worked on, the Acts of the Apostle Thomas in India (BHG 1800 – 1801, CANT 245.II, Bonnet, 1903). The preliminary selection of the witnesses (sources) was made after first comparing the earlier witnesses from the 14<sup>th</sup> and 15<sup>th</sup> centuries with the copies from the 17<sup>th</sup> – 18<sup>th</sup> centuries. One representative of the text was selected – a manuscript from the Dragomirna Monastery No. 700 from the 15<sup>th</sup> century (Yufu, 1970; Nencheva, 2023). The manuscript contains a different, hitherto unknown redaction of the text (Velcheva, Bojadžiev, 2006), which would complement the observations made up to this point (Mitani, 2020). Transcripts from the 17<sup>th</sup> – 18<sup>th</sup> centuries are divided into two groups. The first group involving the archaic version of the text, dating back to the 14<sup>th</sup> – 15<sup>th</sup> century witnesses, which, is represented by the Damaskin of Kostenets (CHAI 503, second half of the 17<sup>th</sup> century, Petkanova-Toteva, 1965: 54; Hristova et al., 1982: 212). The second group contains the variants of the text according to the new Bulgarian damaskini and miscellanies from the 17<sup>th</sup> – 18<sup>th</sup> centuries. This second group is heterogeneous, with the following representatives:

1. *Koprivštitsa Damaskin*, second half of the 17<sup>th</sup> century (Miletič, 1908). The codex is now of unknown location; the text is used according to the Miletič's edition.

2. *Damaskin of Protopopintsi* from the 17<sup>th</sup> century (NBKM 708; Tsonev, 1923: 339–347).

3. *Miscellany of Joseph Bradati* from 1740 (NBKM 1058; Stoyanov, Kodov, 1963: 327–333; Dimitrova, Bojadziev, 2009);

4. *Damaskin of Pope Todor Vračanski* from 1789 (NBKM 1062; Stoyanov, Kodov, 1963: 349–355).

All these witnesses are representatives not only of different Early Modern Bulgarian versions of the text, but also reveal different directions in the development of the Bulgarian language in the 17<sup>th</sup> – 18<sup>th</sup> centuries.

#### 4 Resources and tests

We have experimented with applying linguistic annotation on two different versions of the texts – on the (archaic) *Damaskin of Kostenets* (Kosten) and the (vernacular) *Koprivštitsa Damaskin* (Kopriv). Before starting with the automatic linguistic annotation, we had to apply some preprocessing to our texts. The first step consisted of segmentation on the sentence level where we relied on the segmentation in the respective editions, i.e., we split the sentences wherever we found a punctuation mark indicating a sentence boundary (such as full-stop, colon, Georgian paragraph separator).

The problem of word segmentation (and tokenization) was more intricate as word segmentation in the manuscripts does not regularly correspond to modern practices. We adopt a method developed by (Šimko et al., 2021) for the edition of the Pop-Punčov Sbornik that allows us both to keep information on word boundaries in the manuscript and to provide the taggers with linguistic input coherent with modern practices. Special signs were added to the text indicating word boundaries:

- A vertical line means that a token is written together with the following token in the manuscript, but the tokens are analyzed as two units for the morphological annotation (e.g., цар | же lit. ‘king thus’, where же is a discourse particle, stands for царже in the manuscript but it is given as цар же in the annotated version).
- An underscore is added where a token which is analyzed as one unit in the morphological annotation, is divided into two tokens in the manuscript (e.g., the verb  $\bar{w}$  идѣть ‘to go’ with

the prefix  $\bar{w}$  stands for  $\bar{w}$  идѣть in the manuscript but  $\bar{w}$ идѣть in the annotated version).

As we are currently in the starting phase of the linguistic annotation and aim at finding out the best way to tag our data, we tested two different models for the tagging process. At first, we used the damaskini texts annotated by I. Šimko (2021) to train a model using them as training data. The tagging was performed using the Stanza tagger version, which was modified to use bidirectional character-level LSTM by default and specifically adjusted for parts-of-speech (for low-resource languages) by Y. Scherrer (2021). As this tagger does only perform part-of-speech tagging (POS-tagging) and morphological annotation but no lemmatisation, we had to use another tool for this purpose – Lemming (Müller et al., 2015). The annotated texts are stored in the CoNLL-U format and follow the conventions for Universal Dependencies (Petrov et al., 2012).

Before we could use the model based on the data of (Šimko, 2021) to tag our texts, we had to apply one further step of preprocessing. As all the training data was in the Latin alphabet, we created a script that transcribes the Cyrillic letters to their Latin counterparts and strips the texts of all the superscripts and diacritics. We, thus, performed the linguistic annotation on a graphically simplified version of the texts that matched the training data by (Šimko, 2021). The obtained results will henceforth be referred to as Tag1.

As the number of tokens in the training data was rather small (around 60.000 tokens), we performed a second round of tagging using other data from another source. In this second round, we used annotated Old Church Slavonic texts from the PROIEL (Eckhoff et al., 2018) and the TOROT (Eckhoff and Berdičevskis, 2015) corpora. The data is linguistically less similar to our texts than the data by (Šimko, 2021) but contains much more tokens (around 357.000). When we trained our model, we did not use the original data from the PROIEL and the TOROT corpora but adapted it to some linguistic peculiarities of the Bulgarian language. Prior research has shown that data in such an adapted format provides better results for Middle Bulgarian (Maion, 2022). The result of this second round of annotation will be referred to as Tag2.

## 5 Results

The results from the tagging using the two annotated datasets differ in elements that may have different linguistic interpretation depending on the purpose of the intended corpus (considering further annotation). The Pop-Punčov dataset (Tag1) follows the MULTEXT-East annotation guidelines with a stricter focus on morphology while the Dioptra dataset (Tag2) closely (although not entirely) follows the PROIEL/TOROT annotation principles (which were directed toward the next step in the syntactic annotation for the purpose of building the PROIEL treebank). Results with Tag1 and Tag2 differ in those elements (and parts-of-speech) which may have different syntactic functions – pronouns, adverbs, particles, auxiliaries. Table 1 and Table 2 below give the differences in marking with each dataset for Kosten and Kopriv.

| Element   | Tag1 (Pop-Punčov) | Tag2 (Dioptra) |
|---|-------------------|----------------|
| не ‘not’  | PART              | ADV            |
| же ‘thus’   | PART              | ADV            |
| бо ‘because’  | CCONJ             | ADV            |
| ли (interrogative particle)   | PART              | ADV            |
| Demonstrative pronouns (съ ‘this (over here)’, тъ ‘this’, онъ ‘that’) | PRON, ADJ, DET    | PRON, ADJ      |
| Possessive pronouns (мои ‘my’, твои ‘your’...)                        | ADJ               | PRON           |
| да ‘to’   | CCONJ             | ADV, SCONJ     |
| Auxiliaries   | AUX               | VERB           |
| Passive participles   | ADJ               | VERB; ADJ      |
| Proper names  | NOUN              | PROPN          |

Table 1: Kostenets

| Element   | Tag1 (Pop-Punčov) | Tag2 (Dioptra) |
|---|-------------------|----------------|
| не ‘not’  | PART              | ADV            |
| же ‘thus’   | PART              | ADV            |
| бо ‘because’  | CCONJ             | ADV            |
| ли (interrogative particle)   | PART              | ADV            |
| Demonstrative pronouns (съ ‘this (over here)’, тъ ‘this’, онъ ‘that’) | PRON, ADJ         | PRON, ADJ, DET |
| Possessive pronouns (мои ‘my’, твои ‘your’...)                        | ADJ               | PRON, ADJ      |
| да ‘to’   | CCONJ             | ADV, SCONJ     |
| Auxiliaries   | AUX               | VERB           |
| Passive participles   | ADJ               | VERB; ADJ      |
|   |                   |                |
| Proper names  | NOUN              | PROPN          |

Table 2: Kopriv

The accuracy of the tagger trained with the respective datasets on the two texts is given in Table 3.

| Text                               | POS    | Morphology |
|------------------------------------|--------|------------|
| Kosten – Tag1 (Pop-Punčov dataset) | 91.44% | 82.29%     |
| Kosten – Tag2 (Dioptra tagset)     | 92.36% | 89.56%     |
| Kopriv – Tag1 (Pop-Punčov dataset) | 95.03% | 93.62%     |
| Kopriv – Tag2 (Dioptra tagset)     | 73.97% | 65.18%     |

Table 3: Accuracy

Applied POS-tags were not considered erroneous when calculating the accuracy in the following cases:

a. When tags are not part of the dataset as in: PROPN that are marked as NOUN if there is no PROPN in the training dataset; AUX that are marked as VERB if there is no AUX in the training dataset.

b. Possessive pronouns that are marked as PRON or ADJ if there is such marking in the training dataset.

c. да ‘to’ when tagged as CCONJ, SCONJ or ADV.

d. The negative particle не when marked as PART or ADV.

e. Demonstratives marked as DET in noun phrases (mainly with the Dioptra dataset (Tag2)).

f. Various conjunctions that are tagged as CCONJ, SCONJ, ADV depending on the tagset and the dataset.

The tagger achieves the greatest accuracy with tagging the vernacular Kopriv when trained with the Pop-Punčov dataset (Tag1) and the lowest accuracy with Kopriv and trained with the Dioptra dataset (Tag2). The POS-tagging of the archaic Kosten was better when the tagger was trained on the Dioptra dataset (Tag2) than with the (vernacular) Pop-Punčov dataset (Tag1). When the tagger was trained with the Pop-Punčov dataset (Tag1) comprising texts from the same period, its results on both texts were much closer than when it was trained with the Dioptra dataset (Tag2).

Most errors on POS-level are found when the vernacular Kopriv text was tagged with the tagger trained with the Dioptra dataset (Tag2) – in the example below 4 of 9 wordforms are wrong.

|                           |           |
|---------------------------|-----------|
| И ‘and’                   | CCONJ     |
| непрѣстѣнно ‘ceaselessly’ | ADV       |
| бѣ ‘to-God’               | VERB NOUN |
| се ‘self’ (reflexive)     | DET PRON  |
| мѣаха ‘prayed’            | NOUN VERB |
| и ‘and’                   | CCONJ     |
| славѣха ‘praised’         | ADV VERB  |
| стаа ‘saint’              | ADJ       |
| трица ‘Trinity’           | NOUN      |

The results for morphological annotation are lower (and for lemmatisation are even lower) but they are also linked to the accuracy of the POS-tagging.

Our trial tagging has shown results that are similar to those from previous attempts at tagging early Slavic texts but are still lower due to the character of the texts (they are Bulgarian and from a later period). Except for the normalization method with statistical CRF-tagger MarMoT and a neural network tagger, (Scherrer et al., 2018) experimented with applying Modern Russian resources to pre-modern data to show that transfer experiments did not improve tagging performance significantly, but state-of-the-art taggers still

reached between 90% and more than 95% tagging accuracy even without normalization. J. Besters-Dilger (2021) applied neural network tagger CLStM to the Old Russian Žitie Evfimija Velikogo (GIM, Chud. 20), a copy of the second half of the 14<sup>th</sup> century. The tagger was successfully applied on non-normalised text with high accuracy – however, unknown words (which means those that had not been “seen” by the tagger before) still showed a higher error rate.

## 6 Ongoing work

The next step in our effort is the development of a tagset and annotation principles. At this point, we have decided to keep morphology oriented marking, with some additions that can be beneficial for the further mark-up levels. We have decided to keep PROP for proper names, and to mark all verbal forms as VERB (as in the Dioptra corpus) and all pronouns as PRON. Adverbials including pronominal adverbials will be marked as ADV while the pronominal adjectives formed with an adjectival suffix (as in вѣсѣкъ ‘each’, оногози ‘that’, etc.) are marked as ADJ. The forms with the article-like suffixes (such as жената ‘woman.DEF’ and црѣтомѣ ‘king.DEF’) will be marked as definite forms.

After correcting the results for POS, we expect to achieve better results with the morphological annotation but also with other texts that will be included in the database.

## 7 Acknowledgements

This research is carried out as part of the project “A Unified Annotation of the Stages of Bulgarian Language (AUSBUL)” funded by the Bulgarian National Science Fund under the Programme Bulgaria: Competitions for Financial Support for Bilateral Projects, Science & Technological Cooperation (WTZ) Austria / Bulgaria No. КП-06-Австрия / 2, 18.07.2023 / ОeAD-GsmBH (Österreichischer Austauschdienst) (BG 09/2023, WTZ Bulgarien S&T Bulgaria 2023-25).

## Abbreviations

BHG: Bibliotheca Hagiographica Graeca (Halkin, 1957; Halkin, 1984).

CANT: Clavis apocryphorum Novi Testamenti (Geerard, 1992: 144–150).

CHAI: Church-Historical and Archival Institute, Sofia

NBKM: National Library “St. Cyril and Methodius”, Sofia

## References

Juliane Besters-Dilger. 2021. Neural morphological tagging for Slavic: Strengths and weaknesses. *Scripta & e-Scripta*, 21: 79–92.

Maximilianus Bonnet. 1903. *Acta Apostolorum Apocrypha*. Post C. Tischendorf denuo ediderunt R. A. Lipsius et M. Bonnet. Partius alterus. Volumen alterum. Acta Philippi et Acta Thomae accendunt Acta Barnabae. Ed. M. Bonnet. Lipsiae: Apud Hermannum Mendelssohn [Phototypische Ausgabe. Darmstadt: Wissenschaftliche Buchgesellschaft, 1959].

Margaret Dimitrova and Andrej Bojadžiev. 2009. Apokrifat za apostol Toma v kasnata damaskinarska traditsia. *Godishnik na Asotsiatsia “Ongal”*, 8: 238–260.

Hanne Eckhoff, Kristin Bech, Gerlof Bouma, Kristine Eide, Dag Haug, Odd Einar Haugen, and Marius Jøhndal. 2018. The PROIEL treebank family: a standard for early attestations of Indo-European languages. *Language Resources and Evaluation* 52 (1): 29–65.

Hanne Eckhoff and Aleksandrs Berdičevskis. 2015. Linguistics vs. digital editions: The Tromsø Old Russian and OCS Treebank. *Scripta & e-Scripta*, 14–15: 9–25.

Maruritii Geerard. 1992. *Clavis apocryphorum Novi Testamenti*. Cura et studio Mauririi Geerard. Turnhout: Brepols

François Halkin. 1957. *Bibliotheca Hagiographica Graeca*. T. 2 (*Ioannes Calybita – Zoticus*). 3ème éd. (Subsidia Hagiographica 8a). Bruxelles: Société des Bollandistes.

François Halkin. 1984. *Novum Auctarium Bibliothecae Hagiographicae Graecae* (Subsidia hagiographica 65). Bruxelles: Société des Bollandistes.

Boryana Hristova, Darinka Karadzova, and Siyka Ikonomova. 1982. *Balgarski rakopisi ot XI do XVIII vek zapazeni v Bulgaria. Svoden katalog*. Sofia: Narodna biblioteka “Sv. sv. Kiril i Metodiy”.

Fabio Maion. 2022. Wege zur verbesserten automatischen Annotation des mittelbulgarischen Kirchenslawischen. *Scripta & e-Scripta*, 22: 365–390.

Keiko Mitani. 2020. Slavonic tradition of the Apocryphal Acts of Thomas in India and the MS 1789/700 of the Dragomirna Monastery (Moldavia, Romania). *Scripta & e-Scripta*, 20: 199–225.

Thomas Müller, Ryan Cotterell, Alexander Fraser, and Hinrich Schütze 2015. Joint lemmatization and morphological tagging with lemming. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, pages 2268–2274.

Elisaveta Nencheva. 2023. Panigirikat na dyak Andrey ot 1425 g. spryamo nay-blizkite po sastav yuzhnoslavjanski kodeksi. *Studia Literaria Serdicensia*, 3(5): 120–176.

[https://studialiteraria.eu/sites/default/files/ISSUE/ES/2023\\_issue\\_5/pdf/2023\\_sls\\_5\\_120\\_176.pdf](https://studialiteraria.eu/sites/default/files/ISSUE/ES/2023_issue_5/pdf/2023_sls_5_120_176.pdf) (Last access 2024-04-10)

Donka Petkanova-Toteva. 1965. *Damaskinite v balgarskata literatura*. Sofia: Izdatelstvo na BAN.

Slav Petrov, Dipanjanm, Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of LREC’2012*. Istanbul, pages 2089–2096.

Yves Scherrer, Achim Rabus, and Susanne Mocken. 2018. New developments in tagging pre-modern Orthodox Slavic texts. *Scripta & e-Scripta*, 18: 9–33.

Yves Scherrer. 2021. Adaptation of morphosyntactic taggers. In *Similar Languages, Varieties, and Dialects: A Computational Perspective, Studies in Natural Language Processing*, pages 138–166. Cambridge University Press.

Ivan Šimko. 2021. *Annotated Corpus of Pre-standardized Balkan Slavic Literature I.1* [Online]. Slovenian language resource repository CLARIN.SI.

<http://hdl.handle.net/11356/1441> (Last access 2024-04-08)

Ivan Šimko, Polina Mihova, Olivier Winistörfer, and Anastasia Escher. 2021. *Pop Punčov Sbornik – Digital Edition* [Online]. Zürich: UZH Institute of Slavic Studies.

<https://www.punco.uzh.ch/> (Last access 2024-04-10)

Manyo Stoyanov and Hristo Kodov. 1963. *Opis na slavyanskite rakopisi v Sofijskata narodna biblioteka*. Tom III. Sofia: Nauka i izkustvo.

Benyu Tsonev. 1923. *Opis na slavyanskite rakopisi v Sofijskata narodna biblioteka*. Tom II. Sofia: Izdanie na bibliotekata.

Boryana Velcheva and Andrej Bojadžiev. 2006. The Slavonic text of Acta Thomae in India. *Scripta & e-Scripta*, 3–4: 95–119.

Zlatka Yufu. 1970. Za desetomnata kolektsia Studion. (Iz arhiva na rumanskia izsledvach Yon Yufu). In *Prouchvania po sluchay II Kongres po balkanistika (Studia Balcanica 2)*. Sofia: BAN, pages 299–343.

# ChatGPT: Detection of Spanish Terms Based on False Friends

**Amal Haddad Haddad**  
University of Granada

amalhaddad@ugr.es

**Damith Premasiri**  
Lancaster University

d.dolamullage@lancaster.ac.uk

## Abstract

One of the common errors which translators commit when transferring terms from one language into another is erroneously coining terms which are based on a false friend mistake due to the similarity between lexical units forming part of terms. In this case-study, we use ChatGPT to automatically detect terms in Spanish which may be coined based on a false friend relation. To carry out this study, we implemented two experiments with GPT and compared the results. In the first, we prompted GPT to produce a list of twenty terms in Spanish extracted from the UN discourse, which are possibly based on false friend relation, and its English equivalents and analysed the veracity of the results. In the second experiment, we used an aligned corpus to further study the capabilities of the Language Model on detecting false friends in English and Spanish Text. Some results were significant for future terminological studies.

**Keywords:** ChatGPT; false friends; Translation of terms

## 1 Introduction

English is the lingua franca in science and technology (Faber, 2010), and most primary terms are coined in English and then transferred into other languages (Steurs and Kockaert, 2011; Depecker, 2015). The most common techniques for the transfer of primary terms are literal translation or calques (Quesada et al., 2002). One of the common errors which human translators may commit when transferring terms among languages is using false friends. False friends are defined as “lexemes in at least two different languages with identical or similar graphic/spoken forms whose semantics overlap either partially or not at all” (Szpila, 2006). When the overlap is partial, or the so-called semantic false friends (Dominguez and Nerlich, 2002),

only parts of the meaning are maintained, while other parts are different or lost, which may lead to semantic and conceptual loss. For example, the lexical units *eventually* in English and *eventualmente* in Spanish are considered partial false friends, as they share the meaning of *something to happen*, but in English, the adjective denotes something that will be realised for sure at the end of a process or activity, while the Spanish adjective denotes something that may or may not happen (Dominguez and Nerlich, 2002). In cases of no overlap, or the so-called chance false friends (Dominguez and Nerlich, 2002), or false cognates (Trishina, 2022) the meaning is totally different and leads to erroneous understanding. For example, the lexical units *sensible* in English and *sensible* in Spanish, meaning *sensitive*.

Previous studies showed that false friends are the source of miscommunication not only in common languages, but also in scientific languages (Lewis, 2020; Makayev et al., 2020). For example, Lewis gives a list of false friends between English and Croatian in the legal domain. Yetkin shows through corpus analysis the frequent use of partial false friends between English and Turkish in the diplomatic discourse and its terminology. In this context, we define a term based on false friend as a secondary term which was transferred from one language into another by committing a false friend mistake in the transfer of at least one lexical unit forming part of the term, and the false friend inequivalent meaning may be based on either a total false friend or a partial false friend.

Lewis (2020) highlights the need to carry out more research to show how frequent is the use of false friends in coining terms at interdisciplinary level, and to provide more tools to avoid this translation problem not only in common languages but also in scientific and technical discourse to prevent miscommunication.

Makayev et al. (2020) also highlight the need to investigate the influence of false friends and false cognates in the construction of domains. However, when a term is coined, it starts spreading all over different countries and when academics discover that it is erroneously coined, it is too late by then to control the spread of those terms. Even more, in cases of a false friend relations in a term, in some cases, the meaning becomes part of the language and becomes part of the official dictionaries, which may have influence at conceptual level and understanding of phenomena.

One of the difficulties, in relation to this problem, is that identifying the terms which are based on a false-friend relation is not an easy task. Firstly, it is because users read scientific discourse in one language, and no comparison is carried out by users unless it was the subject of comparative study. In second place, it is important to take into consideration that this phenomenon is not as common as other linguistic phenomena, and it usually requires high linguistic competence by the users to detect them. It also needs a detailed and cautious examination not only of the term itself, but also of the conceptual systems underlying the term, in order to detect those terms based on false friends in discourse and whether they impede or obstacle comprehension or not depending on the level of semantic equivalence between the original word in the first language and the acquired meaning in the second language. One of the approaches to detect those conceptual layers is Frame-Based Terminology (FBT) (Faber et al., 2012; Faber, 2015). There are AI models which try to detect false friends in discourse (Mitkov et al., 2007; Palmero Aprosio et al., 2020), but to the best of our knowledge, there are no automatic tools or programmes, or AI models which focus on the automatic detection of terms in specialised domains which are based on false friend relations.

To fill this gap, we propose the use of ChatGPT to identify possible terms based on false friend relations in specific specialised domains between English and Spanish, then we propose corpus analysis to verify how widely spread the use of such false friends in specialised discourse is.

This paper is organised as follows: in Section 2, we give a brief introduction about the use of ChatGPT and other Large Language Models (LLMs) to generate relevant results in the field of Translation and Terminology in relation to false friends

and its detection; in Section 3 we provide information about the methodology and the experiments carried out to prompt ChatGPT and to obtain relevant and significant results; in Section 4 we discuss the results, while in Section 5 we draw on several conclusions and highlight new trends for future research.

## 2 ChatGPT and LLMs

Using NLP to carry out research in Translation and Terminology is gaining increasing popularity. For example, studies like Mitkov et al. (2007) carried out an experiment to extract false friends and cognates from monolingual corpora in the linguistic combination of English-French, English-German and English-Spanish. Castro et al. (2018) use word vector representations to build a false friends classifier implementing their research in the combinations of Spanish and Portuguese, while studies like Al-Athwary (2021) used Large Language Models (LLMs) to detect false friends in less similar languages like Arabic and English. Several studies also focused on the detection of cognates (Gamallo and Garcia, 2012; Batsuren et al., 2022; Kanojia et al., 2021). Other studies used BERT for the detection of multiword expressions and term extraction (Rigouts Terryn et al., 2020). In this case study, we used ChatGPT (GPT) as a tool for the automatic detection of terms based on false friends in specific domains. To our knowledge, no previous studies focused on this aspect or proved the efficiency of GPT in carrying out those sorts of tasks.

## 3 Methods and experiments

In this small-scale case study, we use GPT to detect terms in Spanish which may be coined based on a false friend relation in some specialised domains. To carry out this study, first we asked GPT to produce a list of false friends between English and Spanish. We started with this query to limit the search to only a selected set of false friends to be used in the coming step. Then we asked GPT to produce a list of terms in Spanish that contain any of the words produced in the previous list. This step allowed for collecting a list of terms in the domain of study which are possibly based on a false friend relationship, but they would indeed still need further analysis to detect whether this assumption is true or false. Afterwards, we asked GPT to produce a list of terms in Spanish equivalent to the English

terms, which are possibly the result of a false friend relationship. This step gave more insightful results after prompting GPT with other queries based on previous research, which will be discussed in Section 3.1. Finally, we studied some of the terms and compared their use in Corpus, and we detected terms which are erroneously coined in the domains of study, which may lead to difficulties in communication or comprehension in specialised discourse. For the comparison of exact meaning of lexical units and to detect whether the false friend relation is correct or not, we looked out the meaning of those lexical units in the Cambridge Online Dictionary (Cambridge University Press, Accessed 2024) and the Real Academia Española (RAE) online dictionary (Real Academia Española, Accessed 2024). With respect to the resources used to verify the frequency of use of some terms or if they really exist in specialised discourse or not, we used the open parallel corpus of the United Nations (UN) (Kilgarriff et al., 2014), available on Sketch Engine, and a search in Google Scholar to detect the use of such terms in scientific articles. Regarding the second experiment (Section 3.2), we used the parallel corpus of the UN available in the UNPC webpage to further study the capabilities of the Language Model on detecting false friends in English and Spanish Text.

### 3.1 First Experiment

Our first prompt to GPT aimed at providing a definition of the concept of false friends to guarantee the coherence of the study. The next query was asking GPT to provide a list of false friends between English and Spanish. We obtained a list of 89 false friends with the explanation of the reason behind considering them as false friends by GPT (The whole list is provided in Appendix A). Next, GPT was asked to provide a list of terms in the domain of Climate Change in English containing one of the lexical units mentioned in the previous answer, that may be part of a term based on a false friend (Appendix B). We selected the domain of Climate Change in our prompt to GPT as the prompting of GPT is based on a term based on a false friend in this domain. Afterwards, GPT was asked to provide the translation of the provided terms (Appendix E). GPT explained after providing the list that “none of the translated terms provided are based on false friend translations. In this case, the translations are accurate and do not involve

words that are false friends between English and Spanish”.

In the next step, and in order to train GPT so that it provides a list with terms that it does consider as based indeed on false friends, we asked GPT a question based on the results of a previous case study in the domain of Climate Change, which shows, on the bases of FBT analysis (Faber et al., 2012; Faber, 2015) that the term *captura y secuestro del carbono* in Spanish is a calque from the English term *carbon capture and sequestration*, and that it is indeed a result of a false friend, as the lexical unit *secuestro* in Spanish means kidnapping and not storage as the meaning of the lexical unit *sequestration* implies in English (Haddad, 2022). Accordingly, we asked GPT the following question: “If we compared the terms *carbon capture and sequestration* and *captura y secuestro del carbono* is the term in Spanish based on a false friend?”. The answer of GPT was: “The terms *carbon capture and sequestration* and *captura y secuestro del carbono* are not based on a false friend translation. They are accurate translations from English to Spanish in the context of carbon capture and sequestration technology. *Capture* in English and *captura* in Spanish, as well as *sequestration* in English and *secuestro* in Spanish, maintain their intended meanings in this technical context without causing confusion or misunderstanding”. We provided the following information to train GPT: “But sequestration in English means *storage* while *secuestro* in Spanish means *kidnapping*”. GPT answered: “You are correct, and I apologize for the oversight. You have pointed out an important distinction between the two terms.”

After receiving this answer from GPT, we asked it to provide two other terms which it considers are based on a similar false friend relation, and it provided the terms *actual pollution* and *Library of Green Technology*. The first term, *contaminación actual*, appeared in some papers such as Herrera et al. (2012), detected in the abstract in English and its translation into Spanish, and in other documents found in the Repository of The National Autonomous University of Nicaragua. However, it is not widely used or coined as a term. With respect to the second term, *librería de tecnología verde*, it is not in use either, which means that the suggestions of GPT did not match our query.

Then we asked GPT to provide a larger list of terms, and consequently, we obtained a list of 20

terms. we explained that we need terms which exist in real discourse and that the identified lexical units are indeed clear and total false friends. GPT explained that it was a very challenging task to obtain such terms in the domain of environment, for this reason, we asked GPT to provide examples in other domains, such as UN discourse. This query led to more satisfactory results as GPT retrieved the term: *treaty ratification* and its Spanish equivalent *ratificación de tratados* among other terms that are discussed in Section 4.

### 3.2 Second experiment

On further investigation on the matter of detecting false friends in English and Spanish, we conducted a larger experiment with the UN parallel Corpus. We used nearly 10,000 aligned English and Spanish sentences to ask ChatGPT whether it can find any false friends in these texts. We created an informative prompt which had 4 main parts containing the instruction to follow. The prompt consisted of, 1. Definition of false friends; 2. Task explanation; 3. An example; 4. Instruction on response.

We decided to have the above parts in our prompt by performing trial-and-error mechanism. Final prompt can be found in Appendix C.

We used *gpt-3.5-turbo-0125* as our base model. We have listed the yielded parameter values in Table 1.

| Parameter       | Value                     |
|-----------------|---------------------------|
| Model           | <i>gpt-3.5-turbo-0125</i> |
| Temperature     | 0.1                       |
| response_format | <i>json_object</i>        |

Table 1: Parameters of ChatGPT experiments

Once we run the experiment, we saved all the data and conducted a manual evaluation if ChatGPT has detected false friends.

## 4 Results and Discussion

### 4.1 Results of first experiment

To analyse the obtained results of the first query, we checked all the list of terms in Spanish, whether they are considered as terms at first place, and we verified whether they exist or not. We categorised each term provided by GPT. Terms considered as based on false friend relation were tagged as (1); not a false friend relation but contains terms which are erroneously coined in Spanish due to translation errors or inaccuracy (tagged as 2), and terms or

words which are not considered relevant for this study (tagged as 3). For example, the term *bosque actual* and its equivalent *actual forest* were tagged as (1) as they show a false friend relation. The term *granja de paneles solares* as equivalent to Solar Panel Farm were tagged as (2), as GPT suggests the use of the term *planta de paneles solares* instead, explaining that it would be more accurate. The suggested term *señal de éxito* and its equivalent *exit sign*, were tagged as 3 as the term does not exist in Spanish. The list of all terms and their classification can be seen in Appendix D.

Out of 20 suggested terms by GPT, we obtained 2 terms based on a total false friend relation; 6 terms which are based on partial false friends and may be relevant for terminological studies related to the adequate coinage of terms; and 12 irrelevant terms.

### 4.2 Results of second experiment

In the second experiment, we provided GPT with the parallel corpus of the UN available in the UNPC webpage. GPT rendered 269 sentences which GPT considered as containing a term based on false friend relation. After analysing the results, and comparing the meaning of the lexical units and its lexical construction in comparison to its English equivalents, we classified the results as: (1) terms based on false friends; (2) terms and lexical units which are relevant to terminological and translation studies; (3) erroneous results. Under the first category, the results showed 17 sentences in the corpus of the UN with presence of false friends, based on the lexical units: *paradas*, *sensible*, *preservativo*, *comprensivo* and *secuestro*, as equivalents to *parades*, *sensible*, *preservative*, *comprehensive* and *sequestration*, and forming terms such as: *paradas mmilitares* as equivalent to *military parades* (*paradas* in Spanish means bus stops); *sensibles a los desastres naturales* as equivalent to *sensible to natural disasters* (*sensible* in Spanish means *sensitive* in English); *preservativos de vacunas* as equivalent to *vaccine preservatives* (*preservativos* in Spanish means condoms and the proper word is *conservantes*); and *marco comprensivo* as equivalent to *comprehensive approach* (*comprensivo* in Spanish means *understanding*). Under the second category, GPT detected terms which are not accurately translated or are relevant to other terminological and translation studies, but are not based on a false friend relation from English into

Spanish, such as the terms *paraíso fiscal* in Spanish and its English equivalent tax haven, which shows that the translation in Spanish is more similar to the word *heaven* than *haven*; *seguridad de la biotecnología* and its English equivalent *biosafety*, taking into consideration that the literal translation of the Spanish term would be *safety of biotechnologies*, which means that it is a possible candidate term for a future terminological research to verify the dimension of its adequacy. The results obtained were 14. The rest of the list rendered by GPT, 240 sentences, showed to be erroneous false friend and irrelevant, such as name of countries, similar cognates or proper lexical units translated from English into Spanish.

## 5 Conclusion

This case study uses ChatGPT to detect false friends in translation and terminological research, addressing a gap in computational linguistics and translation studies. It focuses on specialised domains such as the environmental discourse and UN terminology and provides insights into the intricacies of professional translation. The paper also sheds light on the challenges of translation in specialised fields, and demonstrates the potential use of LLMs for detecting false friends and terms based on false friends.

In spite of the scarcity of valid results obtained through the two experiments, and taking into consideration that it is difficult to detect terms based on false friends in discourse, we consider those results as useful for terminological research, and are considered evidence on the presence of transfer problems among languages in specialised discourse.

We also agree with Lewis (2020) who highlights the need for more research to show the ubiquitous presence of false friends in interdisciplinary Terminology and calls for a more systematic approach to identify them and avoid them. At European level, Lewis (2020) suggests creating a multilingual European database of false friends with corpus-based sentences to avoid falling in such errors across languages. We believe that the approach used in this study can be used as a starting point to obtain insightful preliminary results. In our view, GPT is one of the tools that can be used to detect a list terms and to create a database with possible false friends in any specialised domains; however all the results provided by GPT must be analysed manually to verify their veracity and suitability, as GPT

also provides erroneous results.

In a nutshell, the following conclusions can be also highlighted:

- GPT does not always provide a correct answer, and not all the provided terms fulfil the pre-established criteria. Prompting is very important, as well as the manual analysis of each term.
- GPT is useful to start a hypothesis and to obtain preliminary results to build upon them in more detailed research in the field of translation and terminology.
- GPT is a useful aid during the process of coining new terms, as once prompted, it provides insightful information about the adequacy of use, whether the suggested term is a false friend, as well as providing suggestions for a more accurate term.

Finally, we recommend a supervised and monitored use of GPT in terminology-related tasks, for the coinage of new terms and for the creation of new models that aim at creating new terms for low resources languages.

In future research, we aim at applying this methodology on other pairs of languages which are less similar, such as English and Arabic, as well as carrying out more experiments with other LLMs.

## Acknowledgments

This research was carried out as part of the projects PID2020-118369GB-I00 and A-HUM-600-UGR20, funded by the Spanish Ministry of Science and Innovation and the Regional Government of Andalusia.

## References

- Anwar AH Al-Athwary. 2021. False friends and lexical borrowing: A linguistic analysis of false friends between english and arabic. *Journal of Language and Linguistic Studies*, 17(S1):368–383.
- Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. 2022. A large and evolving cognate database. *Language Resources and Evaluation*, pages 1–25.
- Cambridge University Press. Accessed 2024. Cambridge dictionary. <https://dictionary.cambridge.org/>.

- Santiago Castro, Jairo Bonanata, and Aiala Rosá. 2018. A high coverage method for automatic false friends detection for spanish and portuguese. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 29–36.
- Loïc Depecker. 2015. How to build terminology science. 2015). *Handbook of Terminology*, 1:34–44.
- Pedro J Chamizo Dominguez and Brigitte Nerlich. 2002. False friends: their origin and semantics in some selected languages. *Journal of pragmatics*, 34(12):1833–1849.
- Pamela Faber. 2010. English as an academic lingua franca. *Revista Alicantina de Estudios Ingleses*, 23:19–32.
- Pamela Faber. 2015. Frames as a framework for terminology. *Handbook of terminology*, 1(14):14–33.
- Pamela B Faber et al. 2012. *A cognitive linguistics view of terminology and specialized language*. De Gruyter Mouton Berlin, Boston.
- Pablo Gamallo and Marcos Garcia. 2012. Extraction of bilingual cognates from wikipedia. In *International Conference on Computational Processing of the Portuguese Language*, pages 63–72. Springer.
- Amal Haddad Haddad. 2022. False friend in scientific and technical language: a corpus-based study in climate change discourse. In *Investigaciones de nuevo cuño en la academia*, pages 297–312. Thomson Reuters Aranzadi.
- Alan Humberto Herrera, Doctor Francisco Picado Pavón, Maestra Thelma Salvatierra, and Maestro Emilio Peña. 2012. Universidad nacional autónoma de nicaragua centro para la investigación en recursos acuáticos de nicaragua.
- Diptesh Kanojia, Kevin Patel, Pushpak Bhattacharyya, Malhar Kulkarni, and Gholamreza Haffari. 2021. Utilizing wordnets for cognate detection among indian languages. *arXiv preprint arXiv:2112.15124*.
- Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The sketch engine: ten years on. *Lexicography*, 1(1):7–36.
- Kristian Lewis. 2020. False friends in terminology: Croatian lost in translation. *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje*, 46(1):141–164.
- Khanif Makayev, Svetlana Takhtarova, Mery Gulkyan, Ilvira Kuznetsova, and Tatiana Kazymova. 2020. The urgency of considering “false friends of a translator” for construction specialists. In *IOP Conference Series: Materials Science and Engineering*, volume 890, page 012205. IOP Publishing.
- Ruslan Mitkov, Viktor Pekar, Dimitar Blagoev, and Andrea Mulloni. 2007. Methods for extracting and classifying pairs of cognates and false friends. *Machine translation*, 21:29–53.
- Alessio Palmero Aprosio, Stefano Menini, and Sara Tonelli. 2020. Adaptive complex word identification through false friend detection. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, pages 192–200.
- MGd Quesada, Fuertes Olivera, and Silvia Montero. 2002. Propuesta de estructura definicional terminográfica en ontoterm®. *Terminology*, 8(1):57–90.
- Real Academia Española. Accessed 2024. *Diccionario de la Real Academia Española*. Real Academia Española, Madrid.
- Ayla Rigouts Terry, Véronique Hoste, and Els Lefever. 2020. In no uncertain terms: a dataset for monolingual and multilingual automatic term extraction from comparable corpora. *Language Resources and Evaluation*, 54(2):385–418.
- Frieda Steurs and Hendrik Kockaert. 2011. Language planning and domain dynamics: challenges in term creation. In *Annual meeting of the Societas Linguistica Europaea (SLE): book of abstracts*, pages 275–276. Universidad de la Rioja.
- Grzegorz Szpila. 2006. False friends in dictionaries. bilingual false cognates lexicography in poland. *International Journal of Lexicography*, 19(1):73–97.
- Valentina Trishina. 2022. False cognates by spanish and russian speaking efl students from the perspective of cross-linguistic similarity in second language acquisition.
- Nihal Yetkin. 2011. Partial false friends in english-turkish translations: diplomatic texts. *Hacettepe Üniversitesi Edebiyat Fakültesi Dergisi*, 28(1).

**Appendixes**

**A List of false friends between English and Spanish provided by ChatGPT**

- Actual (Spanish) - Current, not "actual"
- Asistir (Spanish) - Attend, not "Assist"
- Carpeta (Spanish) - Folder, not "Carpet"
- Embarazada (Spanish) - Pregnant, not "Embarrassed"
- Ropa (Spanish) - Clothing, not "Rope"
- Sopa (Spanish) - Soup, not "Soap"
- Largo (Spanish) - Long, not "Large"
- Realizar (Spanish) - Accomplish, not "Realize"
- Éxito (Spanish) - Success, not "Exit"
- Librería (Spanish) - Bookstore, not "Library"
- Aparentemente (Spanish) - Apparently, not "Apparently"
- Constipado (Spanish) - Cold, not "Constipated"
- Sensible (Spanish) - Sensitive, not "Sensible"
- Pretender (Spanish) - Intend, not "Pretend"
- Fábrica (Spanish) - Factory, not "Fabric"
- Actualmente (Spanish) - Currently, not "Actually"
- Tuna (Spanish) - Cactus fruit, not "Tuna fish"
- Atender (Spanish) - Assist, not "Attend"
- Constipación (Spanish) - Cold, not "Constipation"
- Molestar (Spanish) - Bother, not "Molest"
- Colegio (Spanish) - School, not "College"
- Pretensión (Spanish) - Intention, not "Pretension"
- Mayor (Spanish) - Older, not "Mayor"
- Sano (Spanish) - Healthy, not "Sane"
- Ropa (Spanish) - Clothing, not "Roof"
- Abogado (Spanish) - Lawyer, not "Advocate"
- Molestar (Spanish) - Bother, not "Molest"
- Tren (Spanish) - Train, not "Trend"
- Asegurar (Spanish) - Ensure, not "Assure"
- Largo (Spanish) - Long, not "Largo"
- Mismo (Spanish) - Same, not "Mismo"
- Parentes (Spanish) - Relatives, not "Parents"
- Mayoría (Spanish) - Majority, not "Mayor"
- Suceder (Spanish) - Happen, not "Succeed"
- Actuar (Spanish) - Perform, not "Act"
- Apellido (Spanish) - Last name, not "Appellation"
- Ropa (Spanish) - Clothes, not "Rope"
- Ropa (Spanish) - Wardrobe, not "Rope"
- Pie (Spanish) - Foot, not "Pie"
- Falta (Spanish) - Lack, not "Fault"
- Gato (Spanish) - Cat, not "Gate"
- Mente (Spanish) - Mind, not "Mint"
- Cuento (Spanish) - Story, not "Count"
- Sensible (Spanish) - Sensitive, not "Sensible"
- Resumir (Spanish) - Summarize, not "Resume"
- Tarjeta (Spanish) - Card, not "Tarot"
- Exquisito (Spanish) - Delicious, not "Exquisite"
- Recordar (Spanish) - Remember, not "Record"
- Resto (Spanish) - Rest, not "Restaurant"
- Suceso (Spanish) - Event, not "Success"
- El embarazo (Spanish) - Pregnancy, not "Embarrassment"
- La carpeta (Spanish) - Folder, not "Carpet"
- La fábrica (Spanish) - Factory, not "Fabric"
- El éxito (Spanish) - Success, not "Exit"

- El colegio (Spanish) - School, not "College"
  - El abogado (Spanish) - Lawyer, not "Advocate"
  - El tren (Spanish) - Train, not "Trend"
  - La mayoría (Spanish) - Majority, not "Mayor"
  - El gato (Spanish) - Cat, not "Gate"
  - El acento (Spanish) - Accent, not "Asset"
  - La actualidad (Spanish) - Current events, not "Actuality"
  - El apodo (Spanish) - Nickname, not "Apology"
  - El edificio (Spanish) - Building, not "Editing"
  - El alquiler (Spanish) - Rent, not "Alkali"
  - La arena (Spanish) - Sand, not "Arena (sports venue)"
  - El barrio (Spanish) - Neighborhood, not "Bar (place to drink)"
  - El desayuno (Spanish) - Breakfast, not "Desert"
  - La lengua (Spanish) - Tongue, not "Language"
  - El libro (Spanish) - Book, not "Library"
  - El ruido (Spanish) - Noise, not "Rude"
  - El cuadro (Spanish) - Painting or picture, not "Quad"
  - El codo (Spanish) - Elbow, not "Code"
  - La carta (Spanish) - Letter (correspondence), not "Card"
  - La fecha (Spanish) - Date, not "Feta (cheese)"
  - La red (Spanish) - Network, not "Red (color)"
  - El rincón (Spanish) - Corner or nook, not "Rinse"
  - El suceso (Spanish) - Event, not "Success"
  - El traje (Spanish) - Suit, not "Trace"
  - El vaso (Spanish) - Glass, not "Vase"
  - El alquiler (Spanish) - Rent, not "Alkali"
  - La arena (Spanish) - Sand, not "Arena (sports venue)"
  - El barrio (Spanish) - Neighborhood, not "Bar (place to drink)"
  - El desayuno (Spanish) - Breakfast, not "Desert"
  - La lengua (Spanish) - Tongue, not "Language"
  - El libro (Spanish) - Book, not "Library"
  - El ruido (Spanish) - Noise, not "Rude"
  - El cuadro (Spanish) - Painting or picture, not "Quad"
  - El codo (Spanish) - Elbow, not "Code"
  - La carta (Spanish) - Letter (correspondence), not "Card"
- B List of terms in the domain of climate change containing a false friend from the previous list**
- "Carbon footprint reduction" - Contains "footprint," which in Spanish is "huella."
  - "Environmental impact assessment" - Contains "impact," which in Spanish is "impacto."
  - "Climate policy development" - Contains "policy," which in Spanish is "política."
  - "Renewable energy sources" - Contains "sources," which in Spanish is "fuentes."
  - "Climate data analysis" - Contains "data," which in Spanish is "datos."
  - "Emissions reduction goals" - Contains "emissions," which in Spanish is "emisiones."
  - "Sustainable development practices" - Contains "development," which in Spanish is "desarrollo."
  - "Biodiversity conservation efforts" - Contains "conservation," which in Spanish is "conservación."
  - "Climate adaptation strategies" - Contains "strategies," which in Spanish is "estrategias."

- "Natural disaster resilience" - Contains "disaster," which in Spanish is "desastre."
- "Climate change adaptation" - Incorporates "adaptation," which in Spanish is "adaptación."
- "Eco-friendly practices" - Includes "eco-friendly," with "eco" being a similar term in both languages.
- "Sustainable development goals" - Contains "sustainable," which in Spanish is "sostenible."
- "Carbon emissions reduction efforts" - Incorporates "reduction," similar to "reducción" in Spanish.
- "Environmental impact assessment" - Contains "impact," which in Spanish is "impacto."
- "Climate action plan implementation" - Incorporates "implementation," a similar term in both languages.
- "Renewable energy transition strategy" - Includes "transition," which in Spanish is "transición."
- "Climate resilience assessment" - Contains "resilience," similar to "resiliencia" in Spanish.
- "Biodiversity conservation initiatives" - Incorporates "conservation," which in Spanish is "conservación."
- "Sustainable energy solutions" - Contains "sustainable," which in Spanish is "sostenible."

## C ChatGPT prompts and parameters

### C.1 Prompt

A "false friend" is a linguistic term referring to words in different languages that look or sound similar but have different meanings. These similarities can often lead to confusion or misunderstandings for language learners or even native speakers who encounter them in unfamiliar contexts. For example, the English word "embarrassed" and the Spanish word "embarazada" sound similar, but "embarazada" means "pregnant" in Spanish, not "embarrassed." So, "embarazada" is a false friend for English speakers trying to communicate in Spanish.

You will be given sentences as pairs, the English sentence and the Spanish translation of it. Your task is to detect false-friends based terms in the given Spanish sentence and then return a JSON object saying if there is a false-friend term Yes/No and the false-friend Spanish term if you find any. Make sure you do not detect any country names, state names or any names.

For example : "captura y secuestro del carbono" is a term based on false friend translated from "carbon capture and sequestration", because the lexical unit "secuestro" is a false friend of "sequestration"

Example JSON object : {false\_friend\_term\_found : 'yes', false\_friend\_term = ['term1', 'term2']}

Do not provide any other explanation. Just return JSON object with the results. English : {sentence goes here}

## D List of all terms provided by GPT and their classification

- The term *bosque actual*, appears in the UN parallel corpus as the translation of existing forest. This use shows that it is the result of a false friend translation error. (1)
- The term *librería de estudios ecológicos* could not be found in online resources. (3)
- The term *señal de éxito de emergencia* could not be found in online resources. (3)
- The term *reserva natural* is correctly used in Spanish. (3)
- The suggested term *pesticidas de arma* does not exist in Spanish. (3)
- The term *granja de paneles solares* is a metaphor-based term, and it is a direct calque from English, not a result of a false friend; this term is used in academic research. Google Scholar search shows different articles using this term. However, GPT suggests the use of the term *planta de paneles solares* instead, explaining that it would be more accurate. (2)
- The term *gestión de recursos hídricos* is widely used in Spanish. The Real Academia Española dictionary (DRAE (Real Academia Española, Accessed 2024) defines *recurso* in one of its meanings as "Conjunto de elementos

disponibles para resolver una necesidad o llevar a cabo una empresa” [a set of available elements needed to cover a necessity or to launch a company], giving as example “Recursos naturales, hidráulicos, forestales, económicos, humanos”. Those examples coincide with the use of the term in English, hence, we do not consider this term as based on false friend relation. (3)

- The term *Departamento de estudios ambientales* is widely used in Spanish and refers to the same concept. We do not consider this pair as a false friend relation. (3)
- In the case of the term *transición de energía limpia*, which is widely used in scientific articles, it is not accurate according to GPT due to the preposition *de* which in its opinion, should be *hacia* in its place to achieve a more accurate meaning. In this case, we consider the false friend relation as valid and insightful due to the meaning attached to the preposition. (2)
- According to GPT, the use of the term *gestión de pescas sostenibles* is not accurate as *pesca* is the fishing activity, while *pesquería* is a more accurate term to be used instead of fishing, which refer to the “area of water where fish are caught so they can be sold” as defined in the Cambridge Online Dictionary (Cambridge University Press, Accessed 2024). In spite of understanding the reason of suggesting this more accurate term, the lexical units *pesca* and *pesquerías* are not false friends of fishing, however, this suggestion provided by GPT is considered insightful for terminological case-studies. (2)
- The term *control global de la contaminación* is suggested by GPT as the lexical unit *control* in Spanish may have more meanings in other contexts. For this reason, we do not consider this suggestion as valid as a false friend example, neither relevant for our study. (3)
- GPT considers the term *ratificación de tratados* as less false friend based term, as the lexical unit *ratificación* is a false friend of *ratification*. GPT explains that *treaty ratification* refers to the formal approval or acceptance of an international treaty by a country’s government. In Spanish, *ratification* can be a false

friend, as it is often translated as *ratificación*. However, *ratificación* primarily refers to the confirmation of an action or decision, which can lead to potential misunderstandings when discussing the formal acceptance of treaties. We checked the corresponding meaning in the Cambridge Dictionary and in the RAE Dictionary, and we have noticed that the answer provided by GPT was correct. (1)

- The term *conservación de especies en peligro de extinción* is not considered a false friend-based term as the meaning of *conservación* in Spanish according to the DRAE is equivalent to *conservation*, unlike what GPT suggests; however, according to GPT, *preservation* would be more accurate than *conservation* for the exact connotation of the meaning, and we consider this suggestion as relevant, as *perservar* is defined in DRAE as: “Proteger, resguardar anticipadamente a una persona, animal o cosa, de algún daño o peligro.” (Real Academia Española, Accessed 2024) (2)
- The term *infraestructura de energía limpia* as equivalent to Clean Energy Infrastructure is not steamed to be a false friend as *infraestructura* and *infraestructura* are not false friends. (3)
- The term *evaluación de impacto ambiental* is suggested by GPT as the lexical unit *impacto* has two meanings in Spanish. For this reason, we do not consider this term as based on a false friend, neither insightful for our case study. (3)
- The term *prácticas de agricultura sostenible* is suggested by GPT, as the lexical unit *sostenible* in Spanish may have more meanings in other contexts. For this reason, we do not consider this suggestion as valid as a false friend example, neither relevant for our study. (3)
- The term *metas de conservación ambiental* is not considered a false friend-based term as the meaning of *conservación* in Spanish according to the DRAE is equivalent to *conservation*, unlike what GPT suggests; however, according to GPT, *preservation*, would be more accurate than *conservation* for the exact connotation of the meaning, and we consider this

suggestion as relevant, as *perservar* is defined in RAE as: “Proteger, resguardar anticipadamente a una persona, animal o cosa, de algún daño o peligro.” (Real Academia Española, Accessed 2024). (2)

- The term *desarrollo de energía renovable* is not considered a false friend-based term, as *desarrollo* and *development* are not false friends and both involve the same meaning. (3)
- The term *conservación de la biodiversidad marina* is not considered a false friend-based term as the meaning of *conservación* in Spanish according to the RAE is equivalent to conservation, unlike what GPT suggests; however, according to GPT, *conservación* would be more accurate than *preservación* for the exact connotation of the meaning, and we consider this suggestion as relevant, as *perservar* is defined in DRAE as: “Proteger, resguardar anticipadamente a una persona, animal o cosa, de algún daño o peligro.” (Real Academia Española, Accessed 2024). In the case of the lexical unit *biodiversidad*, we do not agree with the suggestion of GPT as the meaning is equivalent in English and Spanish, and the Spanish term includes the connotation of variety of species according to the DRAE. (2)
- The term *evaluación de la sostenibilidad ambiental* is not considered a false friend-based term, as *sostenibilidad* and *sustainable* are not false friends and both involve the same meaning. (3)

**E Terms in English and their Equivalents in Spanish provided by GPT and the explanation of GPT why they are based on false friend relation**

| English Term                     | Spanish Term                         | GPT Explanation   |
|----------------------------------|--------------------------------------|---|
| Actual Forest                    | Bosque actual                        | "Actual" in Spanish means "current" or "present," not "actual" as in "real." So, "actual forest" in English might be mistakenly translated to "bosque actual" in Spanish, which means "current forest" rather than "real forest."   |
| Library of Ecological Studies    | Librería de estudios ecológicos      | As mentioned before, "library" in English might be mistranslated as "librería" in Spanish, which is commonly associated with a bookstore. So, "library of ecological studies" might be inaccurately translated as "librería de estudios ecológicos," while a more appropriate translation would be "biblioteca de estudios ecológicos."         |
| Emergency Exit Sign              | Señal de éxito de emergencia         | "Exit" in English might be mistakenly translated to "éxito" in Spanish, which means "success." So, "emergency exit sign" could be mistranslated as "señal de éxito de emergencia," while it should be "señal de salida de emergencia."  |
| Natural Reserve                  | Reserva natural                      | "Reserve" in English might be translated as "reserva" in Spanish, but it's essential to be aware that "reserva" in Spanish is often associated with booking or reserving something in advance. So, "natural reserve" might be inaccurately translated as "reserva natural," while "reserva natural" can also refer to a protected natural area. |
| Harmful Pesticides               | Pesticidas de arma                   | "Harmful" in English might be translated as "dañino" in Spanish. However, "harm" can be mistranslated as "arma" in Spanish, which means "weapon." So, "harmful pesticides" could be mistakenly translated as "pesticidas de arma," while it should be "pesticidas dañinos."   |
| Solar Panel Farm                 | Granja de paneles solares            | "Farm" in English might be translated as "granja" in Spanish, which usually refers to a place for agriculture or livestock. So, "solar panel farm" could be inaccurately translated as "granja de paneles solares," while "planta de paneles solares" would be a more appropriate translation.  |
| Water Resource Management        | Gestión de recursos hídricos         | "Resource" in English might be mistakenly translated as "recurso" in Spanish. However, "recurso" in Spanish can also mean a recourse or a means to solve a problem. So, "water resource management" might be mistranslated as "gestión de recursos hídricos," while it should be "gestión de la fuente hídrica."                                |
| Environmental Studies Department | Departamento de estudios ambientales | "Department" in English might be translated as "departamento" in Spanish. However, in this context, "departamento" may imply a residential unit. So, "environmental studies department" could be inaccurately translated as "departamento de estudios   |

|                                  |  |   |
|----------------------------------|--|---|
|                                  |  | ambientales," while "facultad de estudios ambientales" or "departamento de estudios del medio ambiente" might be more appropriate.  |
| Clean Energy Transition          | Transición de energía limpia                     | "Transition" in English might be translated as "transición" in Spanish. However, "transición" can also be used in the sense of transitioning from one gender to another. So, "clean energy transition" might be mistakenly translated as "transición de energía limpia," while "transición hacia la energía limpia" would be a more suitable translation.   |
| Sustainable Fisheries Management | Gestión de pescas sostenibles                    | "Fisheries" in English might be translated as "pescas" in Spanish. However, "pescas" can refer to fishing activities. So, "sustainable fisheries management" could be inaccurately translated as "gestión de pescas sostenibles," while "gestión sostenible de pesquerías" would be a more accurate translation.  |
| Global Pollution Control         | Control global de la contaminación               | "Control" in English might be translated as "control" in Spanish. However, "control" in Spanish can sometimes refer to a remote control, such as for a television. So, "global pollution control" could be mistakenly translated as "control global de la contaminación," while "control y regulación de la contaminación a nivel mundial" would provide a more accurate translation.   |
| Treaty ratification              | Ratificación de tratados                         | Treaty ratification & ratificación de tratados & "treaty ratification" refers to the formal approval or acceptance of an international treaty by a country's government. In Spanish, "ratification" can be a false friend, as it is often translated as "ratificación." However, "ratificación" primarily refers to the confirmation of an action or decision, which can lead to potential misunderstandings when discussing the formal acceptance of treaties. |
| Endangered Species Conservation  | Conservación de especies en peligro de extinción | "Conservation" in English might be translated as "conservación" in Spanish. However, "conservación" can sometimes be associated with preserving something in its current state, such as preserving food. So, "endangered species conservation" could be mistakenly translated as "conservación de especies en peligro de extinción," while "protección de especies en peligro de extinción" would be a more accurate translation.                               |
| Clean Energy Infrastructure      | Infraestructura de energía limpia                | "Infrastructure" in English might be translated as "infraestructura" in Spanish. However, "infraestructura" can sometimes be more broadly associated with physical structures and facilities. So, "clean energy infrastructure" could be inaccurately translated as "infraestructura de energía limpia," while "desarrollo de energías limpias" or "red de energía limpia" would provide a more accurate translation.   |

|   |   |  |
|---|---|--|
| Environmental Impact Assessment         | Evaluación de impacto ambiental           | "Impact" in English might be translated as "impacto" in Spanish. However, "impacto" can also refer to a collision or crash. So, "environmental impact assessment" could be mistakenly translated as "evaluación de impacto ambiental," while "evaluación de los efectos ambientales" might be a more suitable translation.   |
| Sustainable Agriculture Practices       | Prácticas de agricultura sostenible       | "Sustainable" in English might be translated as "sostenible" in Spanish. However, "sostenible" can sometimes be associated with maintaining or propping something up. So, "sustainable agriculture practices" could be inaccurately translated as "prácticas de agricultura sostenible," while "prácticas agrícolas sustentables" might be a more appropriate translation.   |
| Environmental Conservation Goals        | Metas de conservación ambiental           | "Conservation" in English might be translated as "conservación" in Spanish. However, as mentioned before, "conservación" can sometimes imply preservation rather than broader conservation efforts. So, "environmental conservation goals" could be mistakenly translated as "metas de conservación ambiental," while "objetivos de conservación del medio ambiente" would be a more accurate translation.                   |
| Renewable Energy Development            | Desarrollo de energía renovable           | "Development" in English might be translated as "desarrollo" in Spanish. However, "desarrollo" can also refer to the growth or expansion of something. So, "renewable energy development" could be inaccurately translated as "desarrollo de energía renovable," while "fomento de energías renovables" might provide a more appropriate translation.  |
| Marine Biodiversity Conservation        | Conservación de la biodiversidad marina   | "Biodiversity" in English might be translated as "biodiversidad" in Spanish. However, "biodiversidad" can sometimes be more narrowly associated with a wide variety of species without considering their variation. So, "marine biodiversity conservation" could be mistakenly translated as "conservación de la biodiversidad marina," while "preservación de la variabilidad marina" might be a more accurate translation. |
| Environmental Sustainability Assessment | Evaluación de la sostenibilidad ambiental | "Sustainability" in English might be translated as "sostenibilidad" in Spanish. However, "sostenibilidad" can sometimes be more closely linked to maintaining or propping something up. So, "environmental sustainability assessment" could be inaccurately translated as "evaluación de la sostenibilidad ambiental," while "evaluación de la viabilidad ambiental" might be a more suitable translation.                   |

# Deep Learning Framework for Identifying Future Market Opportunities from Textual User Reviews

Jordan Kralev

Department of Systems and Control / Technical University of Sofia  
Digital Computational Linguistics / Bulgarian Academy of Sciences  
jkralev@ieee.org

## Abstract

The paper develops an application of design gap theory for identification of future market segment growth and capitalization from a set of customer reviews for bought products from the market in a given past period. To build a consumer feature space, an encoded-decoder network with attention is trained over the textual reviews after they are pre-processed through tokenization and embedding layers. The encodings for product reviews are used to train a variational auto encoder network for representation of a product feature space. The sampling capabilities of this network are extended with a function to look for innovative designs with high consumer preferences, characterizing future opportunities in a given market segment. The framework is demonstrated for processing of Amazon reviews in consumer electronics segment.

**Keywords:** design gap theory, variational auto encoder, natural language processing

## 1 Introduction

In recent years, the large language models (LLM) such as GPT, PaLM, LLaMA demonstrate impressive performance on complex linguistic tasks in automatic text generation. There are a number of milestones marking this success in building such artificial neural network (ANN) models - encoder/decoder structures, probabilistic distribution encoding, embedding layers, long short-term memory cells, attention modeling - to name a few (Bordes et al., 2011; Lake et al., 2015; Paccanaro and Hinton, 2001). Since the growth of cloud computational capabilities and intensive research in the area, we see more claims that LLM are capable of outperforming or matching humans in number of language or image processing tasks. The so marked progress in ANN models allows development of more advanced expert systems to solve computational problems.

The focus of the present work is on developing a class of probabilistic predictive models for extracting information about consumer preferences in a particular online market, where the products in the market are divided in categories. As econometric knowledge explains, the consumer utility is the driving force for the current and future dynamics of the market. Hence, the model aim at identification of so-called "design gaps" in the selected market segments, which allows for economic decisions for developing new products in the given segment. Information about consumer preferences is probabilistically encoded based on an unknown utility function over consumer/design pairs together with prior distributions over customers and designs. The analytical form of the probability densities and embedding representation allow representing the consumer preference information as a multilayer smooth neural network, which allows employment of gradient based methods for learning of hyperparameters from a training data set.

The design gap models allow the prediction of consumer preference for an "unknown and not existing products". The prediction cannot exactly tell what will be these future products, even though an effort can be made to extract such information from the underlying feature space. The model output is a bounded subset of the design space or feature space, which will be favored by the customers. Such bounded subset can be contrasted with the unbounded set of all possible future designs. In the present research, we follow the approach from (Burnap and Hauser, 2018b) for identification of design gap into a diverse design and consumer spaces, where they identify possible future products in the automotive sector. Based on the existing market data, a probabilistic consumer utility model is estimated to account for the likelihood of a certain product design to be selected by a given consumer. They also introduce auxiliary feature dimensions

for product and the consumer to improve the efficiency in estimation of the probability distribution. Similarly, (Ren and Scott, 2017) present adaptive approach for identification of optimal product designs from a finite set of possibilities, proposing Group Identification Splitting Algorithm (GISA) that picks queries that halves the so defined part-worth space but preserves entire groups.

The contribution of the present paper is that it develops a way to recover consumer and design vector representations from the raw textual reviews connecting a particular consumer to a particular product. Therefore, our assumption is that a review text is that all product or consumer features can be extracted from enough number of reviews. The first component of the model is to build a simple natural language transformer trained for estimating the underlying feature vector from each textual review. The model relies on a long short-term memory (LSTM) cells and attention mechanism such to enable the decoder subsystem to predict the next word token in a sequence to match a given input sequence. A specific feature in this model is that we don't aim a high generative accuracy characteristic for LLM mentioned above. Instead, we need to ensure that the model is simple enough to capture in the low-dimensional state vector the essential review information and the state vector can be used to recover the essence of keywords in the input sequence.

The language transformer is used to process all the reviews from the selected market segments and allows for attaching a number of describing vectors to each product, for which we have a published review. On the other hand, we can characterize each customer by the state vectors of all reviews generated by him. Given this representation, we build a market transformer as a variational auto encoder (VAE) network. The VAE models a multivariate probability distribution over a small dimensional feature space, where all product designs from a given market segment are projected. The distribution can be used to generate high probability and low probability designs using a sampling layer. The VAE is tuned to recover the vectors describing a given market segment, and on the other hand to match the consumer preferences in this market segment.

The rest of the paper is organized as follows. Section 2 gives details on structure and tuning of the language transformer. Section 3 presents the

components of the VAE model, and Section 4 summarized the evaluation results in 5 product categories from the online shopping site Amazon.

## 2 Language Transformer

Since the information about consumer preferences is carried by the consumer reviews, which are in natural language, we design a review transformer. The transformer is a connection between an encoder and a decoder network. The purpose of the present research is not to achieve extremely accurate modelling of the natural language. Instead, we aim to ensure that basic meaning carried by review is recognized, especially with respect to tokens reflecting consumer preference, individual consumer features, and individual product features.

### 2.1 Embedding Layer

In the consumer review modelling, we take the smallest meaning carries to be individual words. Since consumer preference is commonly characterized in terms of keywords and grammatical content is from secondary importance, there is no need to fragment further the words into smaller units. The first step in digital language modeling is to select appropriate numerical representation of language lexemes. The common norm is to represent each token in a  $m$  dimensional vector space  $\mathbf{R}^m$  - called embedding space. The vector space can carry a rich algebraic structure - summation of vectors, scalar multiplication, scalar product, cross product, etc. When defining the embedding mapping operator  $E$ , usually we desire some important properties of the linguistic relations between tokens to be reflected into the algebraic structure of the vector space.

We assume a finite dictionary of words with size  $n$ , which is also justified for the purpose of consumer review processing. Therefore, the input tokens can be mapped to an integer between 0 and  $n$ , corresponding to the index of the token in the dictionary. With this assumption, the embedding operator  $E$  can be represented with a  $n \times m$  matrix, where each row contains the embedding vector for a particular token in the dictionary. The column values of a given embedding vector can be thought as a token features in selected vector space.

The training of embedding layer is not a goal of this paper. The task is usually performed over large datasets such as WordNet and Freebase, connecting words  $w_1$  and  $w_2$  with a structural relation  $R$ . In training, a cost is accounting for similarity mea-

tures between triplets and for negative examples for each relation, minimization is carried with a statistical gradient method. The relation  $R$  is numerically represented as a tuple of matrices acting respectively on the left and right word embeddings. Another part of the model is kernel density estimator (KDE) where a number of kernel triplets should be embedded close to triplets with at least one identical element. Interestingly entities that are close to each other in terms of 1-norm exhibit some complex similarities like ”\_lawn\_tennis” is close to other sports like ”baseball”, ”cricket”, etc. The method represent quite well relations for proper nouns, which is generally a difficult problem with other DLNN models. (Bordes et al., 2013; Burnap and Hauser, 2018a)

For the purpose of the current research, the Global Vectors for Word Representation (GloVe) embedding model is used (Pennington et al., 2014). Its training is based on modeling one-step correlation in the corpus or equivalently - the co-occurrence of word pairs. Therefore, words with higher probability of co-occurrence share closer embedding vectors. The model is able to assign vectors within same neighborhood to words with similar meaning. Also the model reflects the words with opposite meaning with more distanced in Euclidean norm vectors.

## 2.2 Encoder

The purpose of the encoder is to transform the input sequence of tokens into a sequence of  $n_e$  dimensional vectors reflecting the correlational structure of the processed consumer review. The encoder is implemented as an interconnection of an embedding layer and a long short-term memory (LSTM) node. This is a widely used RNN network used in natural language models due to its capability to conserve learned local correlations in the input sequence over long processing intervals, which is a feature for common natural grammatical structures such as gender and tense matching. The LSTM structure is beneficial, especially for backpropagation training, to minimize the vanishing of the gradient for long sequences.

The input sequence  $r$  of tokens in a given consumer review is represented with a vector

$$r = (r_1, r_2, r_3, \dots, r_L)^T \quad (1)$$

of  $L$  elements  $r_i \in [1, n]$  from a dictionary with  $n$  words. If the number of elements in the actual

review are less than  $L$ , then a terminating token is substituted for the missing elements. First, each input token is mapped to its embedding vector

$$r_i^{\text{emb}} = E^T \mathbf{1}(r_i), \quad (2)$$

where  $\mathbf{1}(\cdot)$  is a column vector of zeros with unit only at position  $r_i$ , hence it selects the embedding vector  $r_i^{\text{emb}} \in \mathbf{R}^m$  corresponding to the token  $r_i$ . The embedding vector is sent to a feedback interconnection of two LSTM layers to obtain the encoded sequence  $e_i$

$$\begin{aligned} x_{ff,i+1} &= F_{LSTM,ff} \left( x_{ff,i}, \begin{pmatrix} r_i^{\text{emb}} \\ e_{fb,i} \end{pmatrix} \right) \\ e_i &= G_{LSTM,ff} \left( x_{ff,i}, \begin{pmatrix} r_i^{\text{emb}} \\ e_{fb,i} \end{pmatrix} \right) \end{aligned}, \quad (3)$$

where  $x_{ff,i}$  is the current state of the feed forward (ff) LSTM node and  $e_{fb,i}$  is the current output of the LSTM node in the feedback (fb), while  $F_{LSTM,ff}$  and  $G_{LSTM,ff}$  are the state transition and output mappings of the feed forward LSTM encoder node. The respective output sequence of the encoder  $e$  is

$$e = (e_1, e_2, e_3, \dots, e_L)^T \quad (4)$$

- a sequence of  $n_e$  dimensional vectors with the same length as the input sequence in correspondence to the input tokens  $r_i$ . The feedback LSTM node action is expressed by

$$\begin{aligned} x_{fb,i+1} &= F_{LSTM,fb} (x_{fb,i}, e_i) \\ e_{fb,i} &= G_{LSTM,fb} (x_{fb,i}, e_i) \end{aligned}, \quad (5)$$

where  $x_{fb,i}$  is the current state of the feedback LSTM node,  $e_{fb,i}$  is the feedback signal produced from the layer, while  $F_{LSTM,fb}$  and  $G_{LSTM,fb}$  are the state transition and output mappings. The feedback signal is concatenated with the input token embedding, hence it acts as a model of an extracted input sequence context. As a result, the encoded sequence will model a cumulative extension of the context with the incoming input tokens. The purpose of training of the encoder is to ensure that information content of the encoded vector is increasing with each processed token.

$$I(e_{i+1}) > I(e_i) + I(r_{i+1}^{\text{emb}}|e_i), \quad (6)$$

where  $I(\cdot)$  is information operator taken over the current probability distribution of the encoded state.

### 2.3 Decoder

The decoder operates by continuously generating tokens till either a terminating token is generated or the maximal length of the sequence is reached. In every step the decoder takes the token generated in the previous step, the state vector of the decoder from the previous step and the whole encoded sequence. Also, in every step, the decoder state vector is combined with the encoded sequence using an attention mechanism. As a result, a context vector is produced which is combined with the embedding of the previous decoded token, and consequently fed to a LSTM node, which is initialized with the decoder states from the previous step. The LSTM node, in turn, produce a state vector consumed in the next step, and an output vector - a basis for token selection. The LSTM output is projected into a distribution over the input dictionary, and on this basis a maximum likelihood token is selected.

The purpose of the decoder is to generate an output sequence with no more than  $L$  tokens, hence it can be represented as a vector

$$y = (y_1, y_2, y_3, \dots, y_L)^T, \quad (7)$$

where  $y_i \in [0, n]$  represents a single token from the input dictionary mapped to an integer interval. Again, to represent the meaning of the token in relation to other tokens in the dictionary an embedding matrix  $E \in \mathbf{R}^{n \times m}$  is applied to produce an  $m$ -dimensional vector

$$y_i^{\text{emb}} = E^T \mathbf{1}(y_i), \quad (8)$$

where  $\mathbf{1}(\cdot)$  is a column vector of zeros with unit only at position  $y_i$ , hence it selects the embedding vector  $y_i^{\text{emb}} \in \mathbf{R}^m$  corresponding to the token  $y_i$ . The actual decoding is produced by an LSTM memory cell taking the embedding  $y_i^{\text{emb}}$  together with an  $m$ -dimensional context vector  $c_i$  to produce

$$\begin{aligned} x_{i+1} &= F_{LSTM,D} \left( x_i, \begin{pmatrix} y_i^{\text{emb}} \\ c_i \end{pmatrix} \right) \\ d_i &= G_{LSTM,D} \left( x_i, \begin{pmatrix} y_i^{\text{emb}} \\ c_i \end{pmatrix} \right) \end{aligned}, \quad (9)$$

where  $x_i \in \mathbf{R}^{2n_d}$  represents the current state of the LSTM node with  $n_d$  units in the forward and backward channels,  $F_{LSTM,D}$  and  $G_{LSTM,D}$  are state transition and output functions of the LSTM node, and  $d_i \in \mathbf{R}^{n_d}$  is the output signal. The output from the LSTM layer will be interpreted as distribution over dictionary  $\mathcal{L}$ . Hence, a dense

layer is trained to make a projection from LSTM  $n_d$  dimensional space into embedding  $m$  dimensional space,

$$\mu_i = \sigma(EW_3(d_i)), \quad (10)$$

where  $\sigma(\cdot)$  is a softmax operator translating the  $n$ -dimensional vector into a probability distribution

$$\mu_i : [0, n - 1] \rightarrow (0, 1) \quad (11)$$

over the input dictionary. The token at the position  $i + 1$  is then selected using maximal likelihood criteria as

$$y_{i+1} = \underset{z}{\operatorname{argmax}} \mu_i(z). \quad (12)$$

The attention mechanism is used to generate the context vector  $c_i$  into the embedding  $m$ -dimensional space. Being in that space, the context can be expressed in the words of input dictionary with the help of scalar product. First, the previous decoder LSTM state vector  $x_i$  is projected into  $n_d$  dimensional LSTM output space through a linear dense layer  $W_2 \in \mathbf{R}^{n_d \times 2n_d}$ . Similarly, every vector from the encoded sequence  $e$  with length  $L$  is projected into decoder LSTM output space with another linear mapping  $W_1 \in \mathbf{R}^{n_d \times n_e}$ . The attention distribution vector

$$a_i = (a_{i,1}, a_{i,2}, \dots, a_{i,L}), \quad (13)$$

where  $a_{i,j} \in (0, 1)$  is then produced with a softmax operation

$$a_i = \sigma(s_i) \quad (14)$$

from the scores vector  $s_i$  with elements  $s_{i,j} \in \mathbf{R}$  calculated as

$$s_{i,j} = V(W_1 e_j + W_2 x_i), \quad (15)$$

through a linear mapping  $V \in \mathbf{R}^{1 \times n_d}$  acting over the LSTM output space. As can be seen, the projection of current decoder state is used to bias the projections of the encoded tokens. This allows shifting of attention focus from one part of the input sequence to another, as a function of current decoder state. The attention values  $a_{i,j}$  function as weights over the elements of the encoded sequence producing the context vector

$$c_i = \sum_{j=1}^L a_{i,j} e_j. \quad (16)$$

## 2.4 Training

The training of the language transformer is performed over a random sample of sequences from the database of review texts between minimal and maximal word length. First, each input word sequence  $r$  is encoded into a sequence of vectors  $e$ . Then, each encoded vector is processed through the decoder till a sequence  $y$  with the same length  $L$  as input is produced. The loss function is calculated as

$$J_{LT} = - \sum_{i=1}^L E_{r_i}(\ln(\mu_i)), \quad (17)$$

where  $-E_{r_i}(\ln(\cdot))$  represents cross entropy operator. The training is performed for as many epochs as we see a satisfactory generated sequences from decoder over the train and validation sample.

## 3 Market Transformer

The market transformer is developed as a variational auto encoder, which acts as a model for the probability distribution

$$p(x_d|x_c), \quad (18)$$

where  $x_d \in \mathbf{R}^{n_e}$  is a vector of design features and  $x_c \in \mathbf{R}^{n_e}$  is a vector of consumer features for a particular market segment. The probability distribution contains the information of preferred designs by a particular consumer. Because the output dimension of the review encoder can become relatively large depending on the linguistic complexity of the textual reviews, we assume an underlying low dimensional feature space for the products  $h_d \ll n_e$  and for the consumers  $h_c \ll n_e$ .

### 3.1 Dataset Preparation

For the customer review dataset, we use publicly available data for Amazon Product Reviews available by product categories collected in 2023. The whole dataset contains about 48 million products, 571 million reviews and 54 million consumers (Hou et al., 2024). Additionally, a metadata is provided for each item describing user rating, item price, and other features. The dataset parsing is organized by focusing on a finite number of market subsegments identified by name. Data is processed on a monthly basis from a fixed starting year to a fixed ending year. The information is organized into table with columns - review text, sampling period, client identification, product identification, price, product category. The review rating is used

to filter only reviews with 4 and 5 stars, which mark the positive consumer preferences. Also, review texts are filtered by minimal and maximal length to keep consistency.

The review text is filtered by converting all words to lower case and discarding symbols which are not from the alphabet or digits. Each of the reviews in the selected market segments is processed through the encoder to produce the corresponding encoded sequence from  $n_e$  dimensional vectors. As noted in section (), due to the feedback structure of the encoder, the latest element in the encoded sequence will be with the highest information content. Hence, we can take only the last element from an encoded sequence, namely  $e_L$  to characterize the input sequence  $r$ .

The dataset information is in the form of triplets  $(C, R, P)$ , C standing for client, R standing for textual review and P standing for product. Since each textual review is encoded and represented with a vector  $e_L$ , we assume that the vector identifies partly both - the customer who generated the review and the product which receives such a review. In this respect, each product can be uniquely identified by all final encoder states produced by its reviews. Instead, on product level, the model is trained to work with a category of products. Such a generalization is justified to improve the statistical properties of the model. The encoding vectors  $e_L$  of a product category  $P$  form the design space for this category, denoted as

$$\hat{X}_P = (e_L(r^1), e_L(r^2), \dots, e_L(r^{N(P)})), \quad (19)$$

for each review  $r^l$  for a product in the category P. The number of reviews  $N(P)$  may vary for different product categories.

To each product category corresponds a customer category. A customer  $u$  is characterized with all the reviews he has generated collected in a matrix from final encoder states as

$$\hat{X}_u = (e_L(r^1), e_L(r^2), \dots, e_L(r^{N(u)})), \quad (20)$$

and the customer category contains all the characterizations from all the customers

$$\hat{X}_C = (X_{u,1}, X_{u,2}, \dots, X_{u,N(C)}), \quad (21)$$

with  $N(C)$  being the number of customers in a category C.

### 3.2 Consumer Utility

The consumer utility model characterizes the interaction between product category and consumer category with a multivariate probability distribution in a product feature space  $X_H$ . The product characterized by a matrix  $X_P$  is processed through a dense linear layer  $W_4$  into small dimensional representation as

$$Y_P = W_4 H_P. \quad (22)$$

Then the multivariate mean and multivariate standard deviations of the underlying feature space with dimension  $h_d$  are obtained by two layers with rectified linear activation as

$$\mu_p = W_\mu(y_P), y_P \in X_P \quad (23)$$

and

$$\sigma_p = W_\sigma(y_P). \quad (24)$$

This estimates a multivariate normal distribution with parameters  $N(\mu, \sigma)$ . After this a sampler layer generates a random sample from this distribution in the  $h_d$  dimensional space as

$$s_p \propto N(\mu, \sigma) \quad (25)$$

The random sample is sent to a dense decoder layer composed of linear hidden layer and output layer with  $n_e$  units to produce a estimate of the encoded product

$$\hat{y}_P = W_{dec}(s_p) \quad (26)$$

### 3.3 Training

The training of the market transformer is with respect to 3 objectives - minimize the difference between a given product vector  $e_p$  and the decoded vector  $\hat{e}_p$

$$J_{MT,vae} = E_{X_P}(e_p - \hat{e}_p), \quad (27)$$

minimize the cross entropy predicted by the model for a particular consumer  $p_M(e_p, e_c)$  with respect to the ground truth triplet  $\pi$  of whether this customer actually bought the item

$$J_{MT,ch} = E_\pi(-\ln(p_M)), \quad (28)$$

and a regularization term of not allowing underlying mean and variance vectors to grow unbounded

$$J_{MT,reg} = E_{X_C}(W_\mu(X_P, e_c)^2 + W_\sigma(X_P, e_c)^2). \quad (29)$$

The three objective are combined in the cost

$$J_{MT} = J_{MT,reg} + J_{MT,ch} + J_{MT,vae} \quad (30)$$

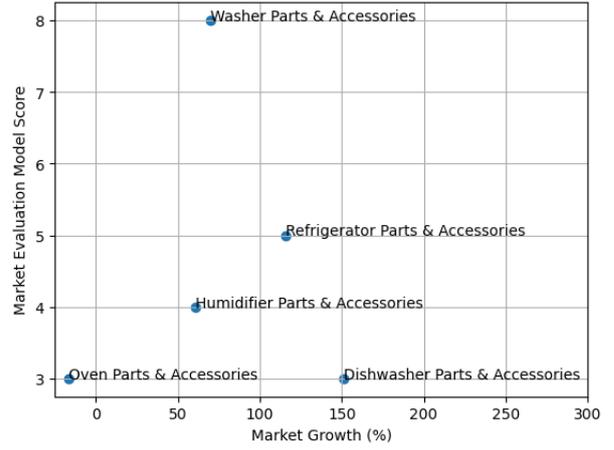


Figure 1: Model Prediction vs Market Capitalization Growth

## 4 Results

The following figures present the obtained results for 5 categories of Amazon reviews. The design gap in a given market segment is predicted by looking for low probability designs (i.e. not yet discovered) through a Monte-Carlo sampling of the underlying multivariate probability distribution. The low probability products are evaluated with respect to the consumer preference probability function  $p_M(x_d|x_c)$ . Hence in case of low probability design, which is highly preferable by customers we can conclude that in a given market segment we have unrealized potential for innovation.

To validate the correctness of such predictions, we compare the observed market growth in the 5 segments from 2013 to 2018 year. Figure 1 gives information about observed market capitalization growth and predicted design gap in the sectors. We observe positive correlation between models scores and capitalization growth, which means that the model correctly predicts future capital allocation in the observed sectors.

In Figure 2 we observe the product survival rate vs market prediction, where again we see a positive correlation with the model predictions. The product survive rate is characteristic to how much a given product is in demand during the observed period. In Figure 3 we examine correlation between market scores and actually appearing new products in the observed domains. Here also a positive correlation with the predicted design gap is observed.

For a reference of the size of the observed market segments, we calculate their market capitalization at the starting year in Figure 4.

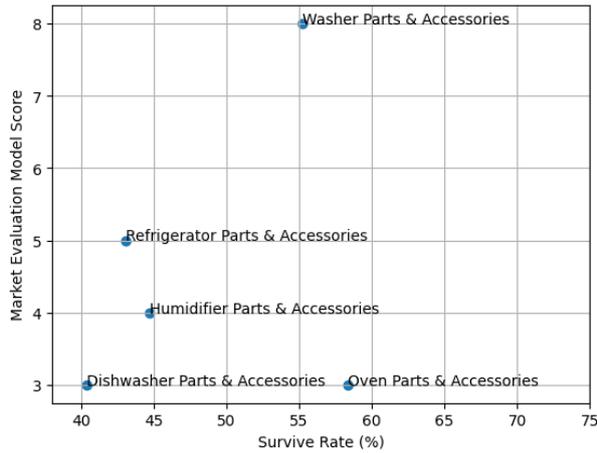


Figure 2: Model Prediction vs Survive Rate

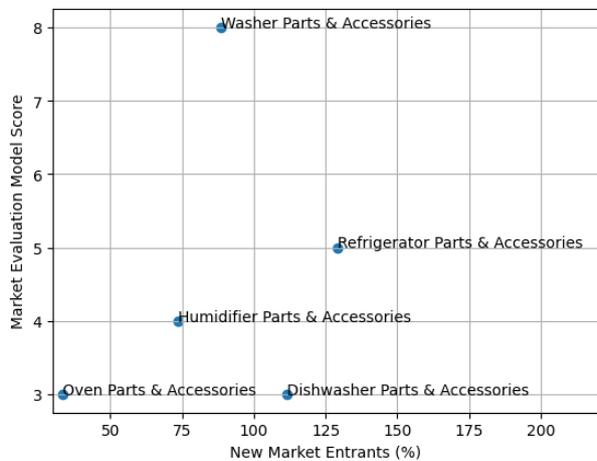


Figure 3: Model Prediction vs New Market Entrants

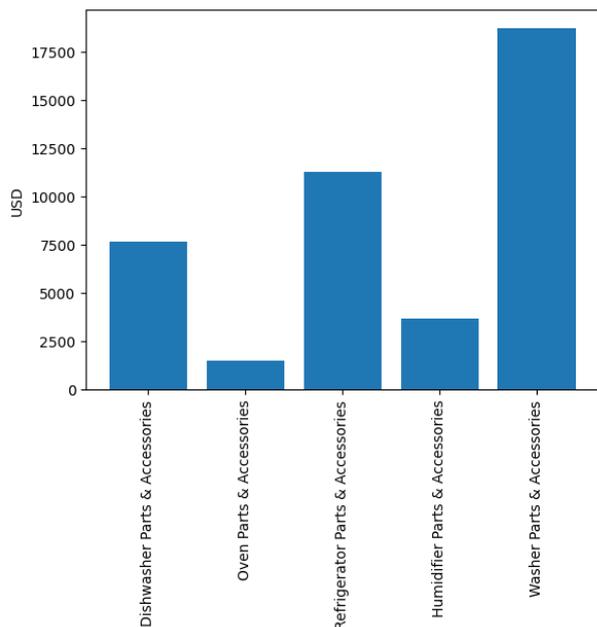


Figure 4: Market Capitalization

## 5 Conclusion

The present research combined key techniques from LLM research to build a simplified language transformer for encoding essential information in textual user reviews used to characterize the consumer preference in online marketing. The applicability of this encoded information is demonstrated for encoding actual consumer preferences using a variational auto encoder model. The developed market segment transformer allows prediction of possible low probability designs in a given market which might be highly favored by its eventual consumers.

## Acknowledgments

This study is financed by the European Union-NextGenerationEU through the National Recovery and Resilience Plan of the Republic of Bulgaria, project no. BG-RRP-2.004-0005.

## References

A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Neural Information Processing Systems (NIPS)*, pages 1–9.

A. Bordes, J. Weston, R. Collobert, and Y. Bengio. 2011. Learning structured embeddings of knowledge bases. In *in Proc. 25th AAAI Conference on Artificial Intelligence*, pages 301–306.

A. Burnap and J. Hauser. 2018a. Predicting “design gaps” in the market: Deep consumer choice models under probabilistic design constraints.

Alex Burnap and John Hauser. 2018b. Predicting “design gaps” in the market: Deep consumer choice models under probabilistic design constraints.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*.

B.M. Lake, R. Salakhutdinov, and J.B. Tenenbaum. 2015. Human-level concept learning through probabilistic program induction. *Science AAAS*, 350(6266):1332–1338.

A. Paccanaro and G.E. Hinton. 2001. Learning distributed representations of concepts using linear relational embedding. *IEEE Transactions on Knowledge and Data Engineering*, 13(2):232–244.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Max Yi Ren and Clayton Scott. 2017. Adaptive questionnaires for direct identification of optimal product design.

# Look Who’s Talking: The Most Frequently Used Words in the Bulgarian Parliament 1990-2024

Ruslana Margova

GATE Institute

Sofia University “St. Climent Ohridsky”

Bastiaan Bruinsma

Chalmers

University of Technology

ruslana.margova@gate-ai.eu

sebastianus.bruinsma@chalmers.se

## Abstract

In this study we identify the most frequently used words and some multi-word expressions in the Bulgarian Parliament. We do this by using the transcripts of all plenary sessions between 1990 and 2024 - 3,936 in total. This allows us both to study an interesting period known in the Bulgarian linguistic space as the years of “transition and democracy”, and to provide scholars of Bulgarian politics with a purposefully generated list of additional stop words that they can use for future analysis. Because our list of words was generated from the data, there is no preconceived theory, and because we include all interactions during all sessions, our analysis goes beyond traditional party lines. We provide details of how we selected, retrieved, and cleaned our data, and discuss our findings.

**Keywords:** corpus, parliament, most frequently used words, Bulgaria

## 1 Object and motivation

The political changes in Bulgaria in 1989 led to demands for greater transparency of political power, including the right of public access to information. As a result, transcripts of meetings of the National Assembly <sup>(1)</sup> were made public, but only after considerable public pressure. So far, these transcripts have mainly been used for qualitative analysis of individual debates on a case-by-case basis. They have rarely been considered as a corpus in their own right, most likely due to the considerable number of transcripts available (every transcript from 27 February 1879 - just 17 days after the National Assembly was established - and onwards is available) as well as the way they were

<sup>1</sup><https://www.parliament.bg/bg/plenaryst>

made accessible (each had to be downloaded individually).

Here, we will use natural language processing (NLP) methods to study this corpus as a whole, allowing us to identify the most frequently used words in the National Assembly between 1990 and 2024. We do this for both theoretical and methodological reasons. The theoretical reasons include gaining a better understanding of the topics discussed by parliamentarians. According to salience theory (Budge and Farlie, 1977), frequently used words are of greater importance to speakers and can provide insight into the interests of the National Assembly. This is particularly relevant as the period under study witnessed significant and structural changes in Bulgarian politics and society. Methodological reasons include our desire to generate a list of stop words that future researchers can use to further preprocess this corpus to better estimate any concepts of interest, as well as to provide an example of how this data can be used for other NLP-related problems.

## 2 Background

Transcripts of legislative debates are often used to study the opinions, positions and policy preferences of elected politicians (Abercrombie and Batista-Navarro, 2020). In the Bulgarian context, the focus is often on individual political speeches and the debate in which they were made. Thus, studies have been conducted on the use of foreign words (Rachev, 2023), the media behaviour of the political elite (Todorov, 2001; Yurukova, 2022), linguistic aggression (Uzanicheva, 2020; Milanov, 2021; Nenova, 2021), the appearance of European identity (Mavrodieva, 2014), the use of clichés, dialects

and factual errors (Milanov and Mihailova-Stalyanova, 2022), the quantitative ratio of words from one national assembly to another (Tarasheva, 2017), and the language of certain MPs (Tarasheva, 2015).

In addition, various attempts have been made to expand the current corpus. For example, (Osenova and Simov, 2012) provide an annotated version of part of the corpus; (Geneva et al., 2019) use the audio of the speeches to build a new corpus of Bulgarian speech suitable for training and evaluating modern speech recognition systems; and the Strazha Foundation will combine it with the duration of each session, the number of words by party, the average number of words per MP by party, the most verbose MPs and other related facts to discuss and comment on the current state of the National Assembly <sup>(2)</sup>.

Finally, transcripts from 2015 onwards have been made part of the ParlaMint dataset (Erjavec et al., 2023), in which each political speech is annotated with, among other things, the age, gender and political orientation of the speakers. As ParlaMint contains similar data from 17 European national parliaments, this allows for cross-country comparisons, as shown by Miok et al. (2023).

One thing these transcripts have not been used for is to examine the frequency of word choice. This is interesting, as this is often seen as one of the basic requirements for understanding the corpus (O’Keeffe and McCarthy, 2010). As a result, a domain-specific list of words that can be used as stop words is missing, as this requires recourse to the corpus one wishes to use (Sarica and Luo, 2021; Yang and Wilbur, 1996). Thus, the creation of such a list can help scholars to better deal with the data from these transcripts and make future analyses less complicated.

### 3 Data and Pre-processing

Each of the 3,936 minutes is structured in the same way. First, the chair and vice-chair and the secretary are identified, together with the date and time of the meeting. Then each speaker is identified individually and their remarks are listed. This includes both what they say and what else is happening in the meeting

<sup>2</sup><https://www.strazha.bg/>

at the same time. However, while noise or applause is included with general remarks, the specific insults and attacks from the floor are not (Tarasheva, 2017: cf.). The transcripts do not record the insults exchanged by the deputies in the chamber, but only those uttered from the gallery. The meetings themselves have no particular structure - sometimes votes are followed by further discussion; sometimes meetings begin with an agenda, but not always; sometimes they begin with proposals to change the agenda; and sometimes there are agenda items listed at the beginning.

After downloading the individual transcripts from the National Assembly website <sup>(3)</sup>, we convert them from HTML to TXT format. We fix any encoding problems and remove headers and footers. Next, we tokenise our words (this and all subsequent steps are performed using version 3.3.1 of the *quanteda* package in R (Benoit et al., 2018)), lowercase them, generate n-grams to capture common expressions, remove punctuation, symbols and numbers, and finally remove stop words as contained in the *BulTreeBank* corpus (Simov, 2014). This last step is crucial, as failure to do so would result in the identification of stop words that are common to Bulgarian in general, rather than those that are specific to the National Assembly. It also prevents our multi-word expressions (MWEs) from consisting solely of collections of frequently used words and expressions. This results in a corpus of 694,174 unique tokens. For our purposes here, we focus on the 250 most frequent words in this resulting data set (the last of which had a relative frequency of 0.033%), although this cut-off is necessarily arbitrary. Appendix A provides an overview of these words, together with an English translation.

### 4 Results

As a result, we get a list of words and some typical MWEs for parliamentary speeches. There is no specific study of MWEs in this analysis. However, MWEs and their derivatives play an important role in certain topics when NLP methods are used (Barbu Mititelu and Leseva, 2018). The list is rich with collocations typical for parliamentary life such as "уважаеми

<sup>3</sup><https://www.parliament.bg/bg/plenaryst>

дами господа народни представители” (Respectfully, ladies and gentlemen deputies), ”предложението прието” (The proposal is accepted).

Through a political-historical prism we can distinguish nine groups of meaning-functional types of words in the resulting list: a) legal terms; b) places and countries; c) financial; d) parliamentary behaviour; e) procedural; f) verbs; g) adverbs; h) party abbreviations; j) other. These types are not surprising. In an earlier study on the Bulgarian language in general, Koeva et al. (2012) found that the most commonly used nouns are those related to time, place and people.

The most common type (in terms of frequency) are words related to law, where the two abbreviations “ал” (paragraph) and “чл” (article) are the most common, followed by “закон” (law), “законопроект” (draft) and “предложение” (proposal). This is followed by geographical references. Unsurprisingly, the word “България” (Bulgaria) is the most frequently used, followed by related terms such as “страна” (country), “държава” (state), “република” (republic), “българските” (Bulgarian - adjectival) and “граждани” (citizen). Bulgarian as a nationality does not appear in this list of most frequently used words, but can be found instead in references to “общество” (society) or “хора” (people). More geographical references - such as “Европейският съюз” (European Union) and “София” (Sofia) - can also be found. It is noteworthy that Osenova and Simov (2012) found similar terms, suggesting that these terms have changed little in importance over time. Another common category is financial references - most often to the Bulgarian currency (“лв”). We also find words such as “пари” (money), “бюджет” (budget), “хиляди” (thousands) and “милиони” (million). Note that there are no references to other currencies. This suggests that the debate on the adoption of the euro as the official currency is not (yet) dominant during the period we are studying.

Next, we find words that demonstrate politeness and respect for colleagues (Osenova and Simov, 2012; Tarasheva, 2015: see also), where we find words such as “уважаеми” (dear), “моля” (please), and “благодаря” (thank you). This kind of politeness is often nothing more

than a set of linguistic conventions that operate independently of the current goal a speaker is trying to achieve (Christie, 2002). As such, this type of politeness is more operational, helping politicians to introduce themselves, rather than reflecting their opinions of each other. Related to this are words that refer to different parliamentary procedures, such as “решение” (decision), “гласуване” (voting), “комисия” (commission), “изказвания” (speeches), “предложения” (suggestions), “въпрос” (question), “процедура” (procedure), реплики (replies), and “текстове” (texts).

Two other categories are verbs and adverbs. Under the former, we find words like “мисля” (think), “казвам” (say), “смятам” (consider), “разбира” (understand), and “искам” (want), and under the latter words such as “всъщност” (in fact), “наистина” (really), “ясно” (clearly), “просто” (simply), “тоест” (i.e.), “действително” (actually), “изключително” (exceptionally), and “вярно” (truly). Interestingly, there are no verbs expressing insistence. Instead, the imperative particle “нека” (let us) is often used. Moreover, the tendency to use impersonal constructions also shows that parliamentarians seem to be trying to avoid personal responsibility, opting instead for general responsibility.

Finally, we find references to the parties. Interestingly, although the corpus consists of texts from more than 30 years, the word ГЕРБ - an abbreviation of one of the political parties - is also among the most frequently used words (“Граждани за европейско развитие на България” - Citizens for European Development of Bulgaria). And while the word “герб” can also refer to a coat of arms, in the parliamentary context here there is no doubt that the disambiguation of the word refers to the political party.

## 5 Conclusions and future work

The analysis of a corpus of Bulgarian parliamentary speeches reveals some interesting findings: Bulgarian politicians use Bulgaria prominently in their speeches; terms such as “European” are also important, but not as central as “Bulgarian”; the speeches also show linguistic politeness, presumably as a convention. Ab-

abbreviations related to law are common, as are terms describing procedures in legislative tasks. Verbs indicating cognitive effort are widespread, but the frequent use of the imperative particle “нека” (let us) suggests a tendency to defer decision-making or responsibility. The abbreviation for the Bulgarian currency is noteworthy, while the dominance of the abbreviation for the political party “ГЕПБ” reflects the dominance of this particular party, despite the presence of others in Parliament during the period analysed.

The generated list contains meaningful words such as “budget”, “decision”, “abstention”, “understand”, which are semantically relevant and essential and cannot be considered as stop words. However, the additional list provided can be used for specific purposes for further automated linguistic analysis with a different focus: for example, for more in-depth analysis of the main themes in the contemporary development of politics and public attitudes in Bulgaria after the beginning of the democratic changes. The large dataset allows for the study of how language has changed over the years, as well as for comparative analysis of the language of individual parties on particular issues. A more in-depth study can reveal the MWEs in parliamentary speech and their pragmatic role.

## 6 Acknowledgments

The results presented in this paper are part of the GATE Project. This project has received funding from the European Union’s Horizon 2020 WIDESPREAD- 2018-2020 TEAMING Phase 2 programme under Grant Agreement No. 857155.

## References

- Gavin Abercrombie and Riza Batista-Navarro. 2020. Sentiment and Position-Taking Analysis of Parliamentary Debates: A Systematic Literature Review. *Journal of Computational Social Science*, 3:245–270.
- Verginica Barbu Mititelu and Svetlozara Leseva. 2018. *Derivation in the Domain of Multiword Expressions*. In Manfred Sailer and Stella Markantonatou, editors, *Multiword Expressions Insights from a Multi-Lingual Perspective*, pages 215–246. Language Science Press, Berlin.
- Kenneth Benoit, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. 2018. *quanteda: An R Package for the Quantitative Analysis of Textual Data*. *Journal of Open Source Software*, 3(30):774.
- Ian Budge and Dennis Farlie. 1977. *Voting and Party Competition*. John Wiley & Sons, London.
- Chris Christie. 2002. *Politeness and the Linguistic Construction of Gender in Parliament: An Analysis of Transgressions and Apology Behaviour*. Sheffield Hallam Working Papers on the Web: Linguistic Politeness and Context.
- Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michal Rudolf, Matyáš Kopp, Starkadhur Barkarson, Steinthór Steingrímsson, Çağrı Çöltekin, Jesse de Does, Katrien Depuydt, Tommaso Agnoloni, Giulia Venturi, María Calzada Pérez, Luciana D. de Macedo, Costanza Navarretta, Giancarlo Luxardo, Matthew Coole, Paul Rayson, Vaidas Morkevičius, Tomas Krilavičius, Roberts Dargis, Orsolya Ring, Ruben van Heusden, Maarten Marx, and Darja Fišer. 2023. The ParlaMint Corpora of Parliamentary Proceedings. *Language Resources and Evaluation*, 57(1):415–448.
- Diana Geneva, Georgi Shopov, and Stoyan Mihov. 2019. *Building an ASR Corpus Based on Bulgarian Parliament Speeches*. In *Statistical Language and Speech Processing: 7th International Conference, SLSP 2019, Ljubljana, Slovenia, October 14–16, 2019, Proceedings*, pages 188–197, Berlin, Heidelberg. Springer-Verlag.
- 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.
- Ivanka Mavrodieva. 2014. The Concept of the Bulgarian “European Identity” Across the Bulgarian Language and Political Rhetoric in Bulgaria. *US-China Foreign Language*, 12(1):1–16.
- Vladislav Milanov. 2021. Българската политическа реч в балкански и славянски контекст. *Studia Philologica*, 40(1):265–278.
- Vladislav Milanov and Nadezhda Mihailova-Stalyanova. 2022. *Езикови портрети на български политици [Linguistic Portraits of Bulgarian Politicians]*. “St. Kliment Ohridski” University Publishing House, Sofia.
- Kristian Miok, Encarnacion Hidalgo-Tenorio, Petya Osenova, Miguel-Angel Benitez-Castro, and Marko Robnik-Sikonja. 2023. *Multi-aspect Multilingual and Cross-lingual Parliamentary Speech Analysis*.

Kristina Nenova. 2021. Езикът на омразата – изследване на политическата риторика и отражението ѝ върху качеството в медийната среда. In *Качествена журналистика и нова комуникационна среда [Quality journalism and a new communication environment]*, pages 295–304, Sofia. Faculty of Journalism and Mass Communication, Sofia University "St. Kl. Ohridski".

Anne O’Keeffe and Michael McCarthy. 2010. *The Routledge Handbook of Corpus Linguistics*. Routledge, London.

Petya Osenova and Kiril Simov. 2012. The Political Speech Corpus of Bulgarian. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, pages 1744–1747, Istanbul, Turkey. European Language Resources Association (ELRA).

Yordan Rachev. 2023. Чуждиците На Политиците. In *2023: Език, наука, комуникации и спорт – 60 години академично образование. Сборник с доклади [2023: Language, Science, Communications and Sport - 60 years of academic education. Collection of reports]*, pages 158–164, Varna. Varna Medical University Press.

Serhad Sarica and Jianxi Luo. 2021. [Stopwords in Technical Language Processing](#). *PLOS ONE*, 16(8):1–13.

Kiril Simov. 2014. [BulTreeBank Stopword List](#). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Elena Tarasheva. 2015. Изследване на дискурса на Волен Сидеров в 42. Народно събрание. *Rhetoric and Communications*, October(19):1–10.

Elena Tarasheva. 2017. [43 Народно събрание: Двойно повече приказки, Нинова говори пред изключен микрофон, Борисов спира руски проекти](#). Accessed: 12-04-2024.

Svetoslav Todorov. 2001. Медийната изява на лидерите на политически партии в България. *Социологически проблеми*, 33(3+4):137–153.

Yuliana Uzancheva. 2020. Стратегии на езикова агресия в парламентарната реч на България и Украйна. *Българска реч. Списание за езикознание и езикова култура*, 1:42–49.

Yiming Yang and John Wilbur. 1996. [Using Corpus Statistics to Remove Redundant Words in Text Categorization](#). *Journal of the American Society for Information Science*, 47(5):357–369.

Maria Yurukova. 2022. *Изборите За Еп През 2019 Г. Отразяване В Българските Онлайн Медии [EP Elections In 2019 Coverage In The*

*Bulgarian Online Media]*. "St. Kliment Ohridski" University Publishing House, Sofia.

## Appendix A Word List

| Term               | Translation              |
|--------------------|--------------------------|
| ал                 | paragraph                |
| чл                 | article                  |
| българия           | Bulgaria                 |
| заповядайте        | please                   |
| уважаеми колеги    | dear colleagues          |
| лв                 | BNG                      |
| против             | against                  |
| уважаеми господин  | dear_mr_president        |
| председател        |                          |
| отношение          | attitude                 |
| закон              | law                      |
| господин           | mr_president             |
| председател        |                          |
| става              | happen                   |
| мисля              | think                    |
| колеги             | colleagues               |
| наистина           | truly                    |
| начин              | a way                    |
| именно             | namely                   |
| разбира            | of course                |
| въпрос             | question                 |
| хора               | people                   |
| хората             | the people               |
| предложението      | the proposal is accepted |
| прието             |                          |
| тоест              | i.e.                     |
| комисията          | the commission           |
| държавата          | the country              |
| страна             | the country              |
| имаме              | we have                  |
| включително        | included                 |
| част               | part                     |
| връзка             | connection               |
| закона             | the law                  |
| изказвания         | statements               |
| народното събрание | parliament               |
| просто             | simply                   |
| предложение        | suggestion               |
| уважаема госпожо   | dear Mrs president       |
| председател        |                          |
| господин министър  | dear minister            |
| текст              | text                     |
| страната           | the country              |
| знаете             | you know                 |
| всъщност           | in fact                  |

continued on next page

| continued from previous page |                      |
|------------------------------|----------------------|
| Term                         | Translation          |
| смятам                       | I believe            |
| кажа                         | I say                |
| решение                      | decision             |
| реплики                      | replica              |
| правителството               | government           |
| комисията подкрепя           | the commission sup-  |
| предложението                | ports the proposal   |
| законопроект                 | draft bill           |
| ясно                         | clear, obvious       |
| относно                      | regarding            |
| виждам                       | i see                |
| свързани                     | linked, connected    |
| гласувайте                   | please, vote!        |
| средства                     | meanings             |
| госпожо                      | MRS president        |
| председател                  |                      |
| път                          | way                  |
| предложението                | suggestion           |
| следва                       | then                 |
| нека                         | let                  |
| процедура                    | procedure            |
| залата                       | the hall             |
| въпросът                     | the question         |
| стане                        | it will happen       |
| говорим                      | we are talking about |
| неща                         | things               |
| народни                      | MP                   |
| представители                |                      |
| времето                      | the time             |
| право                        | law                  |
| имате думата                 | you have the floor   |
| казвам                       | I say                |
| информация                   | information          |
| означава                     | it means             |
| пари                         | money                |
| съответно                    | thus                 |
| предложения                  | suggestion           |
| господин                     | Mr                   |
| лица                         | faces                |
| практика                     | practice             |
| гласуваме                    | we are voting        |
| работа                       | work                 |
| предлага                     | suggestion           |
| въпроси                      | questions            |
| уважаеми дами                | dear MP              |
| господа народни              |                      |
| представители                |                      |

continued on next page

| continued from previous page |                      |
|------------------------------|----------------------|
| Term                         | Translation          |
| законопроекта                | draft bill           |
| проблем                      | problems             |
| казва                        | i say                |
| става дума                   | it means             |
| възможност                   | possibility          |
| млн                          | millions             |
| място                        | place                |
| знам                         | i know               |
| думата                       | word                 |
| въздържали                   | abstention in voting |
| въпроса                      | question             |
| действително                 | really               |
| комисията подкрепя           | the commission sup-  |
| текста вносителя             | ports the proposal   |
| дейност                      | activity             |
| заменят                      | change               |
| човек                        | person               |
| друго                        | other                |
| народните                    | MP                   |
| представители                |                      |
| такава                       | such                 |
| рамките                      | frame                |
| член                         | article              |
| съжаление                    | regret               |
| уважаеми народни             | dear MP              |
| представители                |                      |
| комисия                      | commission           |
| случай                       | case                 |
| проект                       | project              |
| хил                          | thousand             |
| работи                       | work                 |
| имаше                        | had                  |
| необходимо                   | necessary            |
| надявам                      | hope                 |
| говори                       | speak                |
| бюджета                      | budget               |
| второ                        | second               |
| момент                       | moment               |
| става въпрос                 | it means             |
| предлагам                    | i suggest            |
| реплика                      | replica              |
| правим                       | we make              |
| европейския съюз             | European Union       |
| уважаеми господин            | dear MP              |
| министър                     |                      |
| текста                       | text                 |
| парламента                   | the parliament       |
| министерския съвет           | council of ministers |
| промени                      | change               |

continued on next page

| continued from previous page |                | continued from previous page |              |
|------------------------------|----------------|------------------------------|--------------|
| Term                         | Translation    | Term                         | Translation  |
| искате                       | want           | дейността                    | activity     |
| вносител                     | importer       | извършва                     | making       |
| цел                          | target         | промяна                      | change       |
| можем                        | we can         | вчера                        | yesterday    |
| правят                       | they do        | република България           | Bulgaria     |
| проблеми                     | problems       | абсолютно                    | total        |
| изключително                 | exceptional    | герб                         | GERB         |
| данни                        | data           | какви                        | which        |
| резултат                     | result         | казах                        | said         |
| министър                     | minister       | случая                       | case         |
| текстове                     | text           | каже                         | say          |
| смисъл                       | meaning        | значи                        | means        |
| достатъчно                   | enough         | решения                      | decision     |
| определени                   | particular     | оглед                        | meaning      |
| такова                       | such           | бюджет                       | budget       |
| трябвало                     | should         | българските                  | Bulgarians   |
| политика                     | politics       | граждани                     |              |
| срок                         | deadline       | нещата                       | things       |
| искам                        | want           | случи                        | happened     |
| общините                     | municipalities | другото                      | others       |
| случаи                       | cases          | създава                      | creates      |
| законът                      | law            | държава                      | country      |
| иначе                        | otherwise      | отсъства                     | are missing  |
| очевидно                     | obvious        | различни                     | different    |
| против въздържали            | against        | условия                      | cases        |
| приема                       | accept         | лицата                       | faces        |
| колегите                     | colleagues     | другите                      | others       |
| система                      | system         | решението                    | decision     |
| вниманието                   | attention      | имате                        | you have     |
| зала                         | hall           | документи                    | documents    |
| управление                   | government     | единствено                   | only         |
| думите                       | word           | страни                       | different    |
| мерки                        | measure        | едни                         | ones         |
| общо                         | general        | т.н                          | etc          |
| независимо                   | independent    | последните                   | last         |
| гласуване                    | voting         | програма                     | program      |
| работата                     | work           | струва                       | costs        |
| дейности                     | activities     | работят                      | work         |
| предложението                | suggestion     | правото                      | law          |
| приема                       |                | искаме                       | want         |
| същото                       | same           | членове                      | participants |
| контрол                      | control        | своите                       | their        |
| софия                        | capital        | разходи                      | costs        |
| направим                     | we make        | б                            | b            |
| процедурата                  | procedure      | искам кажа                   | want to say  |
| ред                          | order          | дава                         | gives        |
| възможността                 | possibilities  | цели                         | goals        |
| принцип                      | principal      | положение                    | position     |

continued on next page

continued on next page

continued from previous page

| <b>Term</b>       | <b>Translation</b> |
|-------------------|--------------------|
| лично             | personal           |
| системата         | system             |
| обществото        | society            |
| доклада           | report             |
| предвижда         | foresee            |
| средствата        | means              |
| действия          | works              |
| фонд              | fund               |
| казахте           | said               |
| началото          | at the beginning   |
| съгласно          | according to       |
| подкрепа          | supported          |
| тема              | topic              |
| нататък           | follow             |
| крайна сметка     | at the end         |
| приет             | accepted           |
| политически       | political          |
| някакви           | some               |
| води              | leads              |
| гражданите        | citizens           |
| възможно          | possible           |
| господин_димитров | Mr. Dimitrov       |
| вярно             | really             |
| трябваше          | it should be       |
| процес            | processes          |
| договор           | contract           |
| съответните       | respectively       |
| отговор           | answer             |

Table 1: Overview of the 250 most frequent words, their frequency and translation

# Estimating Commonsense Knowledge from a Linguistic Analysis on Information Distribution

Sabrina Mennella<sup>1,2</sup>, Maria Di Maro<sup>2,3</sup>, Martina Di Bratto<sup>2,3,4</sup>

<sup>1</sup> University of Catania, Italy

<sup>2</sup> URBAN/ECO Research Center, Italy

<sup>3</sup> University of Naples "Federico II", Italy

<sup>4</sup> Logogramma s.r.l., Italy

sabrina.mennella@phd.unict.it, maria.dimaro2@unina.it, martina.dibratto@unina.it

## Abstract

Commonsense Knowledge (CSK) is defined as a complex and multifaceted structure, encompassing a wide range of knowledge and reasoning generally acquired through everyday experiences. As CSK is often implicit in communication, it poses a challenge for AI systems to simulate human-like interaction. This work aims to deepen the CSK information structure from a linguistic perspective, starting from its organisation in conversations. To achieve this goal, we developed a three-level analysis model to extract more insights about this knowledge, focusing our attention on the second level. In particular, we aimed to extract the distribution of explicit actions and their execution order in the communicative flow. We built an annotation scheme based on FrameNet and applied it to a dialogical corpus on the culinary domain. Preliminary results indicate that certain frames occur earlier in the dialogues, while others occur towards the process's end. These findings contribute to the systematic nature of actions by establishing clear patterns and relationships between frames.

**Keywords:** Commonsense Knowledge, FrameNet, Semantic Annotation

## 1 Introduction

The development of high-quality Artificial Intelligence hinges on the critical challenge of equipping machines with Commonsense Knowledge (CSK) (McCarthy, 1959). This is essential for implementing systems that can elevate human-machine interaction to a more human-like level. The CSK was described as embodying the fundamental understanding of causal relationships, physical properties, social norms, and cultural references, crucial for effective communication and problem-solving in everyday situations (Cambria et al., 2009). Due to its multifaceted nature, CSK is generally taken for granted and is typically omitted in communication (written or oral) (Grice, 1975), except in cases

of ambiguity or when the listener requires clarification (Nguyen et al., 2022). In this regard, the field of Knowledge Representation (KR) has made significant contributions to the acquisition and application of CSK, leading to the design and construction of resources containing such information (Lenat, 1995; Sap et al., 2019). Nevertheless, since it is impossible to represent all human knowledge in one single resource (Brooks, 1991), we posit that the most intriguing aspect of CSK lies in its consideration as a *process*, rather than a static collection of information. Modelling the processes underlying CSK directly from linguistic data presents a more significant challenge compared to representing predefined knowledge. In this work, we propose a three-level analysis model to investigate the structure of CSK. Specifically, we focus on the second level, which entails a semantic analysis grounded in Frame Semantics (Fillmore et al., 1976) and FrameNet (Baker et al., 1998) applied to a dialogic corpus on the culinary domain. By analysing the frames distribution within the dialogues, we aim to extract insights that can, in future work, contribute to a more comprehensive understanding of CSK structure.

The paper is organised as follows: Section 2 provides a general overview of the state of the art and Section 3 outlines the scope of the study. Section 4 discusses the motivation behind selecting the culinary domain and the development of the knowledge base, which underpins the three-level analysis model detailed in Section 5. Section 6 describes the annotation scheme applied to the CookDial dialogical corpus and the methodology used for retrieving frames distribution. Finally, Section 7 presents the results obtained, followed by the Conclusions and Future Work in Section 8.

## 2 State of the Art

The CSK has been described as embodying the fundamental understanding of the world shared among

individuals, including (i) information about events that occur over time, (ii) the consequences of one's own and others' actions, (iii) the characteristics of physical objects, (iv) their perceptions, (v) their properties, and (vi) its interrelationships (Cankaya and Moldovan, 2009). A universally shared academic definition of CSK does not currently exist. Nevertheless, (Zang et al., 2013) attempted to limit the scope of the investigation by identifying the most representative characteristics that provide a complete description of this type of knowledge, such as (i) sharedness, (ii) fundamentality, (iii) implicitness. From a linguistic perspective, these features appear interesting as they recall some aspects of the Communal Common Ground (CCG) (Clark, 2015), one of the four typologies of Common Ground (CG) (Stalnaker, 2002). Despite the CCG and the CSK involving shared understandings and assumptions, these are essentially distinct concepts: CCG implies a specific connection between an individual and other members of a shared community, emphasising the interaction between the interlocutors; in contrast, CSK concerns an individual's interaction with the world at large, often shared implicitly and unconsciously (Zang et al., 2013). CCG involves active agreement between speakers, establishing shared beliefs and defining a common language for group identities and boundaries (MacWhinney and O'Grady, 2015). On the other hand, CSK does not require explicit agreement, assuming that it is already universally shared among speakers (Zang et al., 2013).

The ongoing need for advancements in equipping AI with robust and adaptable CSK capabilities has provided a significant stimulus for research in KR, which has contributed to the development of large-scale CSK databases (Lenat, 1995; Liu and Singh, 2004; Sap et al., 2019). Although significant progress has been made in this regard (Zhou et al., 2021a,b; Majumder et al., 2020), limitations persist in their ability to capture the open-ended semiotic process, where significance is continuously crafted, contested, and renegotiated within shifting horizons of understanding (Süerdem, 2024).

### 3 Objectives

Given the vast amount of information that CSK encapsulates and the limitations of aforementioned state-of-the-art approaches, we rather frame it as a *process*. In this case, knowledge is understood as a process that generates structured relationships

between actions and entities resulting from recurrent interactions stored in a database, and not as a mere repository of pre-existing facts. Indeed, it is more intriguing to analyse the processes by which this knowledge is formed rather than dwelling on its representation. Our goal is, therefore, to uncover the processes that comprise this knowledge, introducing a three-level analysis designed to extract more detailed information about CSK structure. For the scope of this work, the focus is on the second level, where we aim to identify the distribution of explicit semantic information within the communicative flow in the culinary domain. In future work, this analysis will facilitate the identification and schematisation of implicit information in a given domain. In particular, this will be possible by considering CSK as the result of the analysis of graph patterns and their probability.

### 4 Data sources

For our investigation, we took into account the culinary domain, guided by two main factors: (i) culinary practices are presumed to be highly familiar due to their everyday nature – most people routinely prepare meals; (ii) the domain exhibits strong action co-occurrences, as individual actions are linked (e.g., the action of *beating eggs* implies the action of *cracking eggs*). To facilitate the identification of the entities and actions involved in recipe instructions along with their relationships and co-occurrences, the initial step involves constructing our knowledge graph. We employed three main resources: the Recipe1M+ dataset (Marin et al., 2021), FlavorDB (Garg et al., 2018), and the Epic-Kitchens dataset (Damen et al., 2018), collectively representing the knowledge base of ingredients, recipe titles with instructions, food flavours, and information about daily activities performed in the kitchen that are not explicitly mentioned in recipe instructions (e.g., *take eggs - crack eggs - throw eggshell into bin*). The domain construction follows the methodology described in (Origlia et al., 2022), where multiple sources were integrated into Neo4J (Webber, 2012). This data organisation facilitates the cross-referencing of information, enabling the establishment of intricate relationships within the domain.

For carrying out the linguistic analysis, the Cook-Dial dialogue corpus (Jiang et al., 2023) is employed. The corpus comprises 260 human-to-human English dialogues based on the culinary

domain, in which an agent, given a recipe document extracted from the RISEC corpus (Jiang et al., 2020), guides the user to cook a dish. Data were collected by applying the experimental *Wizard-of-Oz* method (Fraser and Gilbert, 1991), involving two participants interacting via a live chat platform. The application setup simulated the interaction between a voice assistant (agent) and a user. The agent had full access to the text of the recipe, while the user only knew its title. From this corpus was possible to identify actions relevant to the preparation of dishes, analysing their distribution within the dialogue flow.

## 5 Analysis Model

The foundation of a good communication is a set of regulative principles that facilitate its success, managing dialogue in accordance with logical and relevant criteria, as well as respecting the principle of cooperation between speakers. The maxim of *quality* (Grice, 1975) states that the contribution to the conversation should be as informative as is required. Therefore, a speaker is not expected to provide an excess or deficiency of information; rather, they will offer only the necessary information. Consequently, people typically *assume* a division of the knowledge they share (Whiting and Watts, 2024). Although some information is explicitly introduced into the discourse, some other is assumed and not explicitly discussed, agreed upon, or questioned (Amaral et al., 2011). Knowing what it can be presupposed and what must be made explicit, in other words, showing communicative competence (Hymes et al., 1972), still represents a challenge for conversational agents.

For this reason, we propose to classify this knowledge into three typologies: *Foreground knowledge*, *Background knowledge* and *Presupposed knowledge*. This classification allows us for a more structured approach to managing information, thereby facilitate a clearer understanding and more effective analysis of the data. We define foreground knowledge as information explicitly expressed in both oral communication and written texts. In contrast, background knowledge refers to basic fundamental information about entities often left omitted. Lastly, following the semantic-pragmatic approach to presuppositions (Stalnaker et al., 1977), we categorise as presupposed knowledge the implicit information automatically inferred by speakers. These typologies are interrelated, as the former facilitates

the accurate interpretation of the latter. Though instructions for *whisking the eggs* may not explicitly mention it, we inherently infer essential presupposed knowledge, including prior actions like *egg-breaking* and the use of a tool (e.g., a fork) for the beating process, as long as the background knowledge about the nature of eggs themselves (e.g., eggs are liquid and can be beaten).

To uncover the processes underlying the foreground information, a three-level analysis is presented and summarised in Figure (1).

1. I Level. This level relates to the comprehensive ontological knowledge about entities and actions. This knowledge is represented by sources integrated into the graph database described in Section 4. The action *whisk the eggs* assumes that the knowledge of the entity *egg* is already available for the hearer, regardless of whether it has been explicitly described in the dialogue or not. This assumption is based on the fact that the knowledge of the object is part of the shared understanding of the world. This ontological information can be retrieved by querying the database when necessary (e.g., I need to know the state of an ingredient to perform actions).
2. II Level. The focus is on the action and the entity involved in a foreground event. This level refers to a semantic analysis applied to each sentence of the dialogue, employing an annotation scheme based on FrameNet (Section 6). An example is represented by the sentence *whisk the eggs*, where the action of *whisking* evokes the frame *cause\_to\_amalgamate* described in FrameNet.
3. III Level. The focus is on the presupposed action and entities that enable the frame identified in the II level to take place. This level pertains to a probabilistic analysis on the Epic Kitchens dataset, containing co-occurrences of nouns and verb classes, ultimately aiming to predict the core action that defines the frame itself. For instance, the action of *cracking the eggs*, which does not appear explicitly in the dialogue, is implied in the action *whisking the eggs*, semantically marked as *cause\_to\_amalgamate*. By applying the probabilistic calculus on entity relationships within the database, it will be possible to extract the most likely co-occurrences of actions within

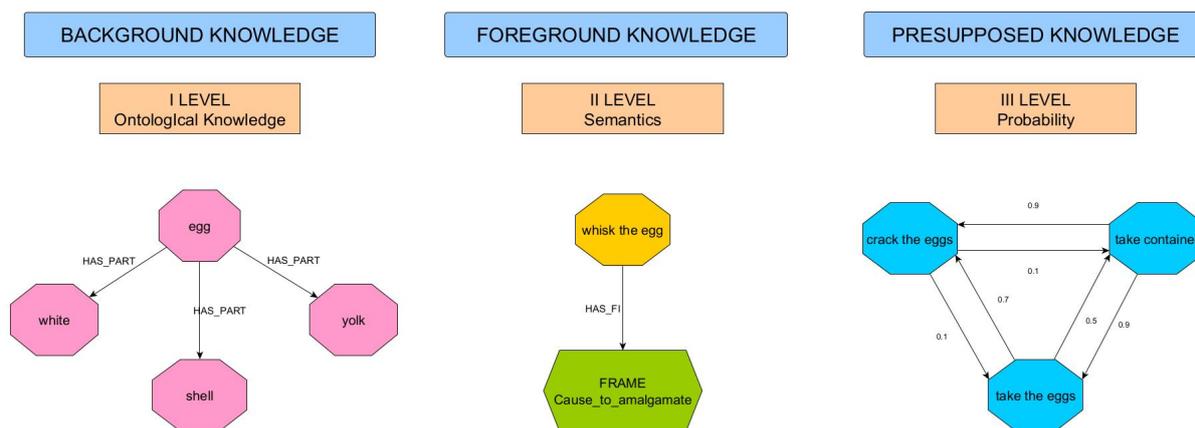


Figure 1: Analysis model with the example of the instruction *whisk the eggs*. At the first level, the model includes ontological knowledge of entities (e.g., *egg*) and their subparts (e.g., *shell*, *yolk*). At the second level, the action *whisk the eggs* invokes the *cause\_to\_amalgamate* frame. At the third level, the action of *whisking* implies a series of action chains (e.g., *take container*, *crack eggs*), determined by the probability of their occurrences represented as relationship properties.

a given semantic context, while avoiding the verbalisation of presupposed actions.

Due to the complexity of the analysis, the present work will focus only on the second level, exploring the distribution of foreground semantic information within the communication flow. In future work, these results will allow us to deepen the analysis of background and presupposed knowledge.

## 6 Methodology

To identify the semantic characteristics of foreground information, an annotation scheme was developed using the FrameNet lexical database (Baker et al., 1998) based on *Frame Semantics* (Fillmore et al., 1976) for describing word senses. A semantic frame is defined as a coherent structure of concepts which evokes a situation, an event or a state along with its participants. In FrameNet, each concept (*frame*) is schematised with its definition, its examples, and its *frame elements*, which represents the semantic roles required by the *lexical unit* (LU) evoking the frame. The sentence *Bake the cookies at 350 degrees* corresponds to *Apply\_heat* frame described as follows:

A Cook applies heat to *Food*, where the *Temperature\_setting* of the heat and *Duration* of application may be specified. A *Heating\_instrument*, generally indicated by a locative phrase, may also be expressed.

*Cook*, *Food*, *Temperature\_setting*, *Heating\_instrument*, *Duration* are the FEs of the frame. Words such as *fry*, *bake* or *boil* represent the LUs evoking the frame. In this work, we identified 29

| Frame Intents (FI)       | Transcript  | Frame Elements (FE) |
|--------------------------|---|---------------------|
| Taking                   | take a <i>knife</i>                                 | Theme               |
| Cause_change_of_phase    | melt 1/4 cup butter in a <i>medium-size pan</i>     | Container           |
| Cause_to_continue        | keep the <i>chicken warm</i>                        | State               |
| Cause_temperature_change | could you preheat your oven to <i>400 degrees</i> ? | Temperature_goal    |
| Soaking                  | marinate it <i>during the night</i>                 | Duration            |

Table 1: FI example for corpus annotation along with their FE. Due to limited space, only 5 out of 29 FIs and one FE for each are reported.

domain-based frames (defined as Frame Intent, FI) along with their FE, as shown in Table 1. We chose to label frames as FIs as they determine the explicit actions expressed by users. Once the dialogues are annotated, they will be integrated into the database and connected to existing resources.

To gain frame recurrences and their positions within dialogues, we annotated 46 dialogues using Label Studio (Tkachenko et al., 2020-2022) (2), an open-source data labelling platform which facilitates the creation of annotated datasets. Two annotators were engaged to annotate the first ten dialogues, ensuring the annotation agreement. The MASI (Measuring Agreement on Set-valued Items) distance (Passonneau, 2006) was employed as it is particularly useful for handling multiple labels for a single item, ranging from 1 to indicate identical sets, to 0 to indicate completely disjointed sets. Additionally, The Krippendorff’s Alpha (Passonneau, 2004) was applied to assess the annotation quality, calculating the metric of weighted agreement. Results show an agreement value of 0.75, confirming the validity of the annotation scheme.



Figure 2: Label Studio interface. Highlighted text segments within the dialogue correspond to the assigned labels.

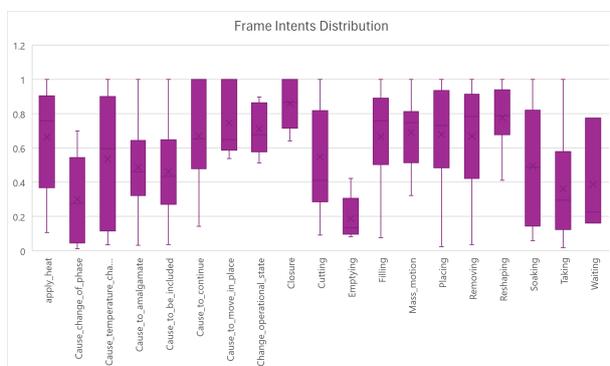


Figure 3: FI distribution within the dialogues. Only 19 out of 29 FI are taken into account for our analysis.

## 7 Results

Following the completion of the annotation phase, we extracted the dialogues from the platform, executing a Python script to ascertain the FI's distribution within the dialogue stream. For enhanced visualisation purposes, the data was then converted into a graphical representation, as illustrated in (3). From 29 Frames, we identified 19 relevant for our analysis.

Results show that certain FI as *Taking* (e.g. take a bowl), *Soaking* (e.g. soak the chicken), *Emptying* (e.g. drain the turkey), *Cause\_temperature\_change* (e.g. preheat the oven to 400 degrees), and *Cause\_change\_of\_phase* (e.g. melt 1/4 cup butter) occur earlier, while *Cause\_to\_continue* (e.g. keep the chicken warm), *Cause\_to\_move\_in\_place* (e.g. turn the pancake), *Reshaping* (e.g. roll up each crepes), *Placing* (e.g. put the chicken on plate) and *Closure* (e.g. seal the bag) oc-

cur towards the process's end. This distribution reflects the natural flow of a culinary task, where initial steps involve preparing ingredients (*Taking*, *Soaking*, *Emptying*) and manipulating temperature (*Cause\_temperature\_change*, *Cause\_change\_of\_phase*), while later stages focus on cooking food (*Cause\_to\_move\_in\_place*), monitoring progress (*Cause\_to\_continue*), modelling the shape (*Reshaping*) and finalising the process (*Placing*, *Closure*). The specific action sequences that frequently occur at particular points in the dialogue enable a deeper investigation into presupposed knowledge and facilitate the extraction of action co-occurrences semantically implied by the foreground knowledge.

## 8 Conclusions and Future Works

In this paper, we proposed a three-level analysis for deepen the investigation of CSK structure. In particular, we focused on the second level, annotating 46 dialogues extracted from the CookDial corpus to calculate FI recurrences and their positions within dialogues. The analysis revealed that there are FI predominantly appeared in the initial stages of the dialogue and others towards the end of it, reflecting the natural flow of a cooking process. Those results hold significant importance as they contribute to the systematic nature of this information by establishing clear patterns and relationships between frames. A further study is underway on Epic Kitchens, allowing us to identify *presupposed* actions that can be omitted from recipe instructions without impacting completion.

## References

- Patrícia Amaral, Chris Cummins, and Napoleon Katsos. 2011. Experimental evidence on the distinction between foregrounded and backgrounded meaning. In *Proceedings of the 2011 ESSLI Workshop on Projective Content*, pages 1–7. Citeseer.
- Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.
- Rodney A Brooks. 1991. Intelligence without representation. *Artificial intelligence*, 47(1-3):139–159.
- Erik Cambria, Amir Hussain, Catherine Havasi, and Chris Eckl. 2009. Common sense computing: From the society of mind to digital intuition and beyond. In *Biometric ID Management and Multimodal Communication: Joint COST 2101 and 2102 International Conference, BioID\_MultiComm 2009, Madrid, Spain, September 16-18, 2009. Proceedings 2*, pages 252–259. Springer.
- Hakki C Cankaya and Dan Moldovan. 2009. Method for extracting commonsense knowledge. In *Proceedings of the fifth international conference on Knowledge capture*, pages 57–64.
- Eve V Clark. 2015. Common ground. *The handbook of language emergence*, pages 328–353.
- Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. 2018. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736.
- Charles J Fillmore et al. 1976. Frame semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conference on the origin and development of language and speech*, volume 280, pages 20–32. New York.
- Norman M Fraser and G Nigel Gilbert. 1991. Simulating speech systems. *Computer Speech & Language*, 5(1):81–99.
- Neelansh Garg, Apuroop Sethupathy, Rudraksh Tuwani, Rakhi Nk, Shubham Dokania, Arvind Iyer, Ayushi Gupta, Shubhra Agrawal, Navjot Singh, Shubham Shukla, et al. 2018. Flavordb: a database of flavor molecules. *Nucleic acids research*, 46(D1):D1210–D1216.
- HP Grice. 1975. Logic and conversation. *Syntax and Semantics*, 3:43–58.
- Dell Hymes et al. 1972. On communicative competence. *sociolinguistics*, 269293:269–293.
- Yiwei Jiang, Klim Zaporjets, Johannes Deleu, Thomas Demeester, and Chris Develder. 2020. Recipe instruction semantics corpus (risec): Resolving semantic structure and zero anaphora in recipes. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 821–826.
- Yiwei Jiang, Klim Zaporjets, Johannes Deleu, Thomas Demeester, and Chris Develder. 2023. Cookdial: a dataset for task-oriented dialogs grounded in procedural documents. *Applied Intelligence*, 53(4):4748–4766.
- Douglas B Lenat. 1995. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.
- Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.
- Brian MacWhinney and William O’Grady. 2015. *The handbook of language emergence*. John Wiley & Sons.
- Bodhisattwa Prasad Majumder, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Julian McAuley. 2020. Like hiking? you probably enjoy nature: Persona-grounded dialog with commonsense expansions. *arXiv preprint arXiv:2010.03205*.
- Javier Marin, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. 2021. Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):187–203.
- John McCarthy. 1959. Programs with common sense.
- Tuan-Phong Nguyen, Simon Razniewski, Julien Romero, and Gerhard Weikum. 2022. Refined commonsense knowledge from large-scale web contents. *IEEE Transactions on Knowledge and Data Engineering*.
- Antonio Origlia, Martina Di Bratto, Maria Di Maro, and Sabrina Mennella. 2022. A multi-source graph representation of the movie domain for recommendation dialogues analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1297–1306.
- Rebecca Passonneau. 2004. Computing reliability for coreference annotation.
- Rebecca Passonneau. 2006. Measuring agreement on set-valued items (masi) for semantic and pragmatic annotation.
- Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019.

- Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3027–3035.
- Robert Stalnaker. 2002. Common ground. *Linguistics and philosophy*, 25(5/6):701–721.
- Robert Stalnaker, Milton K Munitz, and Peter Unger. 1977. Pragmatic presuppositions. In *Proceedings of the Texas conference on performatives, presuppositions, and implicatures*. Arlington, VA: Center for Applied Linguistics, pages 135–148. ERIC.
- Ahmet Süerdem. 2024. The challenges and opportunities in large language models: Navigating the perils of stochastic and scholastic parrots in artificial understanding and common sense. *AI and Common Sense*, pages 195–212.
- Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020–2022. **Label Studio: Data labeling software**. Open source software available from <https://github.com/heartexlabs/label-studio>.
- Jim Webber. 2012. A programmatic introduction to neo4j. In *Proceedings of the 3rd annual conference on Systems, programming, and applications: software for humanity*, pages 217–218.
- Mark E Whiting and Duncan J Watts. 2024. A framework for quantifying individual and collective common sense. *Proceedings of the National Academy of Sciences*, 121(4):e2309535121.
- Liang-Jun Zang, Cong Cao, Ya-Nan Cao, Yu-Ming Wu, and Cun-Gen Cao. 2013. A survey of commonsense knowledge acquisition. *Journal of Computer Science and Technology*, 28(4):689–719.
- Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayathnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. 2021a. Commonsense-focused dialogues for response generation: An empirical study. *arXiv preprint arXiv:2109.06427*.
- Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayathnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. 2021b. Think before you speak: Explicitly generating implicit commonsense knowledge for response generation. *arXiv preprint arXiv:2110.08501*.

# Pondera: A Personalized AI-Driven Weight Loss Mobile Companion with Multidimensional Goal Fulfillment Analytics

**George Pashev**  
University of Plovdiv “Paisii  
Hilendarski”  
georgepashev@uni-  
plovdiv.bg

**Silvia Gaftandzhieva**  
University of Plovdiv “Paisii  
Hilendarski”  
sissiy88@uni-  
plovdiv.bg

## Abstract

The global obesity epidemic is a significant challenge to public health, necessitating innovative and personalized solutions. This paper presents Pondera, an innovative mobile app revolutionizing weight management by integrating Artificial Intelligence (AI) and multidimensional goal fulfillment analytics. Pondera distinguishes itself by supplying a tailored approach to weight loss, combining individual user data, including dietary preferences, fitness levels, and specific weight loss objectives, with advanced AI algorithms to generate personalized weight loss plans. Future development directions include refining AI algorithms, enhancing user experience, and validating effectiveness through comprehensive studies, ensuring Pondera becomes a pivotal tool in achieving sustainable weight loss and health improvement.

**Keywords:** AI; Weigh-Loss; Mobile application.

## 1 Introduction

Healthcare chatbots have significantly advanced medical technology by providing personalized, accessible, and engaging solutions in various domains such as mental health, chronic disease management, and weight loss. These chatbots deliver tailored dietary and exercise recommendations, essential for effective weight management. According to systematic reviews and meta-analyses by Franz et al. (2007) and Young et al. (2012, 2014), tailored interventions are crucial for sustained weight loss, underscoring the potential of chatbots in this area. By utilizing user data, chatbots enhance engagement and motivation through adaptive interactions and constant availability, which is crucial for users

seeking weight loss support. The integration with wearable technology further personalizes feedback and recommendations, enhancing intervention effectiveness.

Despite their potential, deploying healthcare chatbots involves overcoming challenges related to information accuracy, user trust, and behavioral change. Ensuring the reliability of chatbot-provided information is critical, given the potential for negative health outcomes from inaccuracies. This necessitates rigorous sourcing and verification processes, ensuring information is derived from reputable, evidence-based medical sources (Franz et al., 2007). Furthermore, maintaining algorithmic transparency and mitigating bias are essential to ensure that chatbots provide unbiased, medically sound advice (Young et al., 2014).

Compliance with regulatory and ethical guidelines is vital for user trust and data security. In the U.S., the FDA regulates healthcare chatbots that offer diagnostic or therapeutic advice, detailing criteria for software oversight based on intended use and potential patient risks. The GDPR in the EU imposes strict data handling requirements, impacting chatbots that process personal health information. Similarly, the U.S.'s HIPAA mandates the protection of sensitive patient data, with additional international standards from ISO ensuring the reliability and safety of healthcare chatbots globally.

Healthcare chatbots are poised to revolutionize weight management and broader health interventions through their ability to provide personalized, dynamic support. However, realizing this potential requires continuous improvement, adherence to regulatory standards, and integration into comprehensive digital health ecosystems. Future advancements in AI will further enhance the personalization capabilities of

healthcare chatbots, making them indispensable tools in promoting healthier lifestyles and managing weight effectively.

Pondera's multidimensional analytics engine surpasses WeightMentor's (Holmes et al, 2019) basic goal-setting by simultaneously analyzing weight, diet, exercise, sleep, and stress. This holistic approach enables nuanced insights and personalized interventions. Our machine learning algorithms identify correlations between lifestyle factors, allowing for targeted goal adjustments. As an adaptive system, Pondera incorporates feedback loops and dynamic goal adjustment mechanisms. It continuously refines user goals and interventions based on real-time data, dynamically adjusting recommendations and support strategies. This adaptive architecture ensures personalized, effective support throughout the user's weight loss journey, optimizing outcomes and engagement.

## 2. Pondera: design and development

### 1.1 Pondera functionalities and components

#### Overview of Pondera Development Goals

Pondera aims to lead in weight management by effectively using AI and personalized analytics, as outlined in these specific goals:

G1 - Comprehensive Personalization: Utilize AI to analyze user data points to craft customized weight loss plans that evolve based on feedback.

G2 - Interactive User Assessment: Improve quizzes to understand users' weight loss goals and challenges, including psychological factors.

G3 - Nutrition and Fitness Integration: Offer diverse dietary and fitness options tailored to individual preferences and needs.

G4 - Behavioral Change Support: Implement habit formation, motivation, and progress tracking tools to encourage lasting changes.

G5 - User Engagement and Community Building: Develop features allowing users to share experiences and motivate each other.

G6 - Data Privacy and Security: Ensure robust data protection measures adhering to GDPR and HIPAA regulations.

G7 - Adaptive Learning and Feedback Loops: Continuously refine plans based on user feedback and changing circumstances.

G8 - Comprehensive Health Integration: Track and improve overall health metrics, promoting holistic well-being.

G9 - Partnerships with Health Professionals: Collaborate with experts to enhance credibility and effectiveness.

G10 - Continuous Research and Innovation: Stay at the forefront of AI, machine learning, and nutrition/fitness developments.

These goals guide Pondera's development to not only assist users in weight management but also support broader health and well-being objectives. By focusing on these goals, Pondera can truly revolutionize weight management, offering users a unique and effective tool to achieve their weight loss and health objectives.

For the *development of Pondera*, a mobile application designed for personalized weight management and training plans, a comprehensive software architecture involving multiple technologies is required.

It contains 4 software components:

- User Interface (UI): This layer includes the presentation and interaction layer of the application, built with HTML, JavaScript, and Bootstrap. It allows users to input their goals, preferences, and other required details (see Fig. 1).
- Front-End: The front-end is responsible for sending requests to the back-end via AJAX calls and updating the UI based on the data received. It is built using JavaScript and interacts with the Flask CORS back-end for data processing.
- Back-End (API): The Flask application serves as the back-end, handling API requests from the front-end, processing data, interacting with the SQLite3 database, and communicating with external APIs like GPT-3.5. It employs Cross-Origin Resource Sharing (CORS) to enable secure cross-origin requests and responses.
- Database (SQLite3): This database stores all the static data required by Pondera, including user information, goals, user groups, available resources (meals, training sets), and their associated parameters (Zone diet blocks, calories, vegetarian index, ketogenic index, HIIT index, etc.).

Figure 1. The data input form for Plan Generation

### 1.2 Software prototype

This system provides personalized training and eating plans tailored to user goals, dietary preferences, and exercise intensity, developed using Python Flask CORS, JavaScript, HTML, Bootstrap, and SQLite3.

The Pondera database includes eight key entities: Users, Goals, UserGroups, UserGoals, Resources, Meals, Trainings, and UserPlans. Users have attributes like UserID, Username, and Weight; Goals include GoalID and GoalDescription; Resources and Meals detail items such as ResourceID and Calories; Trainings and UserPlans track elements like VideoURL and daily assignments.

Relationships within the database include One-to-Many between Users and UserGoals, Many-to-Many between Users and UserGroups, and One-to-One between Resources and either Meals or Training. UserPlans detail the many-to-many relationships between Users, Meals, and Training, organizing daily meal and training assignments.

The app integrates with the GPT-3.5 API to update meal and training data dynamically, ensuring complete information for generating personalized plans.

Upon receiving user inputs (weight, desired training intensity, dietary preferences, and goal), the system utilizes a multidimensional vector space model to match and recommend a diverse yet *personalized set of meals and training plans* that align with the user's inputs and the Zone diet principles. Fig. 3 presents the process flow diagram of Pondera.

This architecture supports the dynamic generation of personalised weight management plans, leveraging the power of AI for data

completion and offering users a tailored approach to achieving their weight loss goals:

- **User Interaction:** Users interact with the UI to enter their personal information, goals, and preferences.
- **Data Processing:** The front-end sends this data to the back-end via AJAX.
- **API Logic:** The Flask back-end processes the request, queries the SQLite3 database for matching resources, and communicates with the GPT-3.5 API as needed to complete missing data (see Fig. 4).

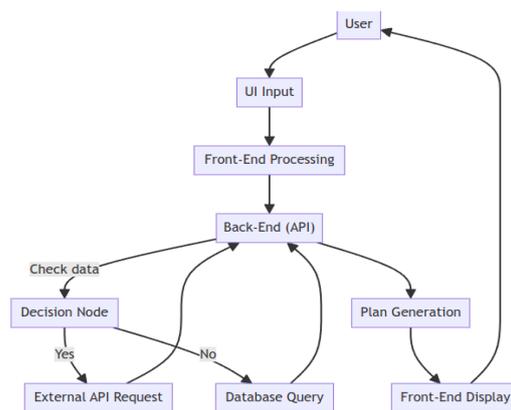


Figure 3. Process flow diagram of Pondera

```

from flask import Flask, request, jsonify
from your_plan_generator import generate_personalized_plan # Placeholder for your actual function
app = Flask(__name__)
@app.route('/api/generate_plan', methods=['POST'])
def generate_plan():
    try:
        # Extracting user input from the request
        user_data = request.json
        weight = user_data.get('weight')
        training_intensity = user_data.get('training_intensity')
        daily_blocks = user_data.get('daily_blocks')
        desired_weight = user_data.get('desired_weight')
        # generate_personalized_plan returns a dict
        plan = generate_personalized_plan(weight, training_intensity,
        daily_blocks, desired_weight)
        return jsonify(plan), 200
    except Exception as e:
        return jsonify({'error': str(e)}), 400
if __name__ == '__main__':
    app.run(debug=True)
    
```

Figure 4. Python Flask API method endpoint for plan generation

Using algorithms, the back-end calculates the best match of meals and training sets based on the user's inputs and the multidimensional vector space model.

The vectorization process for Pondera's data involves converting the structured data from the

database into numerical vectors, which can be processed by machine learning models for generating personalized plans. This process includes:

- **Encoding Categorical Data:** Attributes like DietaryRestrictions, FitnessLevel, and GoalType are categorical and can be converted into numerical vectors using techniques like one-hot encoding or label encoding.
- **Normalizing Numerical Data:** Attributes such as Age, Height, Weight, Calories, Proteins, Carbs, and Fats should be normalized to ensure they're on a similar scale, typically between 0 and 1, to prevent any one feature from dominating the model's behaviour.
- **Text Vectorization:** For textual data, such as Ingredients in meals, techniques like TF-IDF (Term Frequency-Inverse Document Frequency) can be used to convert text into a meaningful vector of numbers.
- **Aggregating Data:** User profiles might need to aggregate data from Goals, Training Sets, and Meals based on user activity. This aggregated data can then be vectorized as a part of the user's profile vector.
- **Dimensionality Reduction:** After vectorization, dimensionality reduction techniques such as PCA (Principal Component Analysis) can be applied to reduce the number of features, if necessary, to simplify the model without losing significant information.

```
import pandas as pd
from sklearn.preprocessing import OneHotEncoder, MinMaxScaler
def preprocess_data(user_profiles, goals, training_sets, meals):
    # Fill missing values
    user_profiles.fillna(user_profiles.mean(), inplace=True) # Numerical
    columns
    user_profiles.fillna('unknown', inplace=True) # Categorical columns
    # Convert categorical data to numerical format
    encoder = OneHotEncoder(sparse=False)
    categorical_columns = cols # Example columns
    encoded_features =
encoder.fit_transform(user_profiles[categorical_columns])
    encoded_df = pd.DataFrame(encoded_features,
columns=encoder.get_feature_names(categorical_columns))
    user_profiles = pd.concat([user_profiles.drop(categorical_columns,
axis=1), encoded_df], axis=1)
    # Normalize numerical values
    scaler = MinMaxScaler()
    numerical_columns = ['Age', 'Weight'] # Example columns
    user_profiles[numerical_columns] =
scaler.fit_transform(user_profiles[numerical_columns])
```

```
# Similar preprocessing would be done for goals, training_sets, and
meal data frames
return user_profiles # This function would return all preprocessed
data frames in a real scenario
```

Figure 5. Data Preprocessing Process

This structured approach enables the creation of a comprehensive vector space that represents the multifaceted data involved in personalizing weight loss plans. With the vectors ready, machine learning algorithms can then be applied.

Moving forward to the *Feature Extraction step*, we'll build upon the preprocessed data. The goal of feature extraction is to convert the raw data into a set of features that can be used for creating machine learning models. This involves identifying which attributes of the data are most relevant to the problem you're trying to solve and possibly creating new features from the existing ones to better capture the underlying patterns in the data. Feature extraction in Pondera involves:

- Selecting relevant nutritional information from meals (e.g., calories, proteins, carbs, fats) that aligns with dietary goals.
- Extracting key attributes from training sets (e.g., difficulty level, duration, calories burned).
- Incorporating user-specific goals and progress metrics into the features.

Fig. 6 shows a part of the code for implementing feature extraction in Python.

```
def extract_features(user_profiles, goals, training_sets, meals):
    # Example of extracting nutritional features from meals
    nutritional_features = meals[['Calories', 'Proteins', 'Carbs', 'Fats']]
    # Example of extracting training set features
    training_features = training_sets[['DifficultyLevel', 'Duration',
'CaloriesBurned']]
    user_goals_features = pd.merge(user_profiles, goals, on='UserID',
how='left')
    return nutritional_features, training_features, user_goals_features
```

Figure 6. Feature extraction

In the Feature Extraction step following data preprocessing, we select key attributes from the data for vectorization to help craft personalized weight loss plans:

- **User Profiles:** Important features include age, current weight, dietary restrictions, and fitness level, which influence meal and workout recommendations.

- **Goals:** Factors like goal type (weight loss, muscle gain), target weight, and target date are crucial for plan personalization and need precise quantification.

- Training Sets: Attributes such as difficulty level, duration, and calories burned are vital for aligning workout plans with user goals and fitness levels.

- Meals: Essential nutritional details include calories, proteins, carbs, fats, and compatibility with dietary restrictions, critical for meal plan formulation.

**The Vectorization Process** requires three steps:

- **Numerical Features:** Numerical features like age, weight, calories, proteins, carbs, and fats are already in a suitable format for most machine learning algorithms. However, they might require normalization to ensure all features are on the same scale.
- **Categorical Features:** Categorical features, especially those that have been one-hot encoded, are already in a vectorized form. However, it's essential to ensure that the vectorization is consistent across the dataset to match the one-hot encoding schema used during training.
- **Combining Features:** Once all features are in a numerical format, we combine them into a single vector for each user profile and meal. This vector represents the input to our machine-learning models.

Fig. 7 presents a simple vectorization process for features implemented in Pondera where **user\_features** and **meal\_features** are pandas DataFrames containing our preprocessed and extracted features.

```
import numpy as np
def vectorize_features(user_features, meal_features):
    user_vectors = user_features.to_numpy()
    meal_vectors = meal_features.to_numpy()
    return user_vectors, meal_vectors
```

Figure 7. Vectorization process

Normalization and dimensionality reduction are essential for optimizing machine learning algorithms. Normalization adjusts each feature to scale uniformly, typically with a mean of 0 and standard deviation of 1, or within a range like 0 to 1. This uniform scaling reduces bias from features with larger scales, enhancing algorithm performance and speeding up algorithms like gradient descent.

Dimensionality reduction, often through methods like Principal Component Analysis (PCA), reduces the number of variables, retaining the most critical information with minimal data loss. This process simplifies models, decreases

overfitting, and reduces computational demands, ultimately transforming the data into a lower-dimensional space that captures significant variance.

The use of tools like scikit-learn's StandardScaler and PCA in Python exemplifies these processes. StandardScaler normalizes features, while PCA reduces dimensions, preserving 95% of the data's variance to maintain essential information for model effectiveness.

To optimize daily meal and exercise plans, we employ a greedy algorithm that iteratively selects the optimal combination to balance calorie intake and expenditure while meeting nutritional goals. This method focuses on minimizing the difference between daily calorie consumption and burn, ensuring all nutritional needs (proteins, carbs, fats) are met within specified limits.

```
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
def normalize_and_reduce_dimensions(data):
    scaler = StandardScaler()
    normalized_data = scaler.fit_transform(data)
    pca = PCA(n_components=0.95) # retain 95% of the variance
    reduced_data = pca.fit_transform(normalized_data)
    return reduced_data
```

Figure 8. Normalization and dimensionality reduction

- The implementation of this algorithm involves:
- Sorting meals and exercises by their caloric and nutritional values.
  - Selecting meals that meet daily nutritional requirements without exceeding caloric limits.
  - Choosing exercises that address any caloric surplus or help achieve calorie deficit goals.

This approach allows for practical daily planning, making decisions that are sufficient for day-to-day progress without needing to be globally optimal. The greedy algorithm simplifies decision-making by focusing on immediate objectives, efficiently balancing the diet and exercise routine to meet the user's health and weight loss targets.

Optimal meals are chosen from a sorted list based on their ability to meet daily nutritional needs without surpassing caloric limits. This selection is iterative, adding meals that enhance the nutritional profile, and updating remaining nutritional needs after each selection.

Exercises are selected to either match or exceed remaining caloric needs after meal choices. They aim to address any caloric surplus from meals or

create a deficit, selected for their high caloric burn to efficiently meet targets.

Placeholder functions help sort and select meals and exercises by calculating nutritional scores and adjusting needs based on meals' nutritional content. These functions, crucial for the algorithm, allow for informed decisions in the daily plan.

This algorithm iteratively selects meals and exercises that balance caloric intake and expenditure, aligning with nutritional requirements, to effectively meet fitness goals and cater to user preferences.

```
def optimize_daily_plan(selected_meals, selected_exercises,
daily_calorie_needs, daily_nutritional_needs):
    # Sort and select meals and exercises based on nutritional and
    calorie needs
    sorted_meals = sorted(selected_meals, key=lambda x:
calculate_nutritional_score(x, daily_nutritional_needs))
    optimal_meals, remaining_needs = select_optimal_meals
(sorted_meals, daily_calorie_needs, daily_nutritional_needs)
    optimal_exercises = select_optimal_exercises(sorted(selected_exercises,
key=lambda x: x['calories_burned'], reverse=True),
remaining_needs['calories'])
    return optimal_meals, optimal_exercises
def select_optimal_meals(meals, calorie_needs, nutritional_needs):
    # Select meals that match nutritional and calorie requirements
    optimal_meals, remaining_needs = [], nutritional_needs.copy()
    for meal in meals:
        if meets_needs(meal, remaining_needs):
            optimal_meals.append(meal)
            update_remaining_needs(meal, remaining_needs)
    return optimal_meals, remaining_needs
def select_optimal_exercises(exercises, calorie_target):
    # Select exercises to fulfill or exceed remaining calorie needs
    optimal_exercises, calories_burned = [], 0
    for exercise in exercises:
        if calories_burned < calorie_target:
            optimal_exercises.append(exercise)
            calories_burned += exercise['calories_burned']
    return optimal_exercises
# Assume implementation details for placeholder functions
def calculate_nutritional_score(meal, needs): pass # Calculate closeness
to nutritional needs
def meets_needs(meal, needs): pass # Check meal meets remaining
nutritional needs
def update_remaining_needs(meal, needs): pass # Update needs based
on meal
```

Figure 9. Part of the code for optimizing the daily plan

As a response, the personalized plan is sent back to the front-end, where it is displayed to the user.

## Your Personalized Plan

### Day 1

#### Meals

- Breakfast: Scrambled eggs with spinach (2 Zone blocks)
- Lunch: Grilled chicken salad (3 Zone blocks)
- Dinner: Salmon with asparagus (4 Zone blocks)

#### Training

Description: 30-minute HIIT session

Calories Burned per kg: 0.5

[Workout Video](#)

Figure 10. A view of an Item of a Generated Plan

The User Interface (UI) for displaying a generated plan in Pondera is designed to provide a clear and concise overview of a user's personalized training and eating plan. Utilizing Bootstrap's grid system, the UI is responsive, ensuring it looks good on various devices. Each day of the plan is clearly labeled (e.g., "Day 1"), with meals and workouts detailed under specific subheadings like "Meals" and "Training." These sections list meal names, their nutritional content like Zone diet block count, and a brief description of the day's exercise routine, including estimated calories burned and links to workout videos.

This structured presentation aids users in navigating their plan, enhancing readability and interaction. The UI also includes visual cues like icons or progress bars to provide feedback on the user's progress. Overall, the UI is not just informative but engaging, encouraging users to actively participate in their health regimen by following through with meals and workouts and interacting dynamically with the plan through features like marking activities as completed or navigating between plan days. This interface acts as a bridge between the complex backend algorithms and the user, making the personalized plans accessible, understandable, and actionable.

## 2 Experiment: Sample Request for Menu Generation

**Objective:** To evaluate the effectiveness of Pondera's AI-driven personalization algorithm in generating a week-long tailored meal plan for a user with specific dietary preferences and nutritional goals.

**Methodology:** A request was submitted to the Pondera system with the following user profile inputs:

**Dietary Preferences:** Vegetarian, Allergies: None, Nutritional Goals: Weight loss, with a focus on high protein and low carbs, Physical Activity Level: Moderate

Fig. 11 presents a generated menu.

|   |
|---|
| <p><b>Day 1</b><br/> <b>Breakfast:</b> Scrambled tofu (150g) with spinach (50g) and avocado (100g) on whole-grain toast (2 slices).<br/> <b>Lunch:</b> Quinoa salad with chickpeas (100g), cucumbers (50g), tomatoes (50g), and feta cheese (30g).<br/> <b>Dinner:</b> Lentil soup (250g) with a side of roasted broccoli (75g) and carrots (75g).<br/> <b>Snacks:</b> Almonds (30g) and Greek yogurt (150g).</p> <p><b>Day 2</b><br/> <b>Breakfast:</b> Greek yoghurt (150g) with mixed berries (100g) and a sprinkle of chia seeds (15g).<br/> <b>Lunch:</b> Whole grain pasta (100g cooked) with pesto sauce (30g) and roasted vegetables (100g).<br/> <b>Dinner:</b> Grilled portobello mushroom (100g) with quinoa (100g) and steamed green beans (75g).<br/> <b>Snacks:</b> Sliced apple (150g) with peanut butter (15g).</p> |
|---|

Figure 11. Generated Menu Sample

### Comments on the Generated Menu

**Pros: Personalization:** The menu adheres to the user's dietary preferences and nutritional goals, showcasing Pondera's ability to tailor recommendations. **Nutritional Balance:** Meals are well-balanced, providing a good mix of protein, healthy fats, and complex carbohydrates, aligning with the weight loss goal.

**Cons: Repetition:** The generated menu may lack variety over a week, potentially leading to diet fatigue. Including more diverse ingredients and cuisines could improve user satisfaction.

**Practicality:** Some meals might require significant preparation time, which could be a barrier for users with busy schedules. Suggesting quicker options or meal prep tips could enhance usability.

The experiment indicates that while Pondera's menu generation feature is effective in creating personalized and nutritionally balanced meal plans, further refinement is needed in diversifying meal options and considering practicality for users with varying lifestyles.

More experiments indicated that sometimes the menu may contain incompatible food.

### 3 Conclusion

This paper highlights the development of Pondera, a mobile app designed for personalized

weight management using AI and goal fulfillment analytics. It outlines how AI algorithms, user assessment, and the integration of nutrition and fitness methodologies are utilized to create tailored weight loss plans. The paper emphasizes the importance of extensive AI testing, user experience design, and validation studies to ensure the app's effectiveness in real-world scenarios. It also points to the need for scaling the app to serve a diverse user base and integrating continuous feedback mechanisms. Looking ahead, further advancements in AI and digital health ecosystems could significantly boost the effectiveness of healthcare chatbots in managing weight and promoting healthier lifestyles, making them vital tools in combating obesity.

### Acknowledgments

The paper is financed by the Scientific Research Fund at the University of Plovdiv "Paisii Hilendarski", project № MUPD23-FTF-019.

### References

M. Franz, J. VanWormer, A. Crain, J. Boucher, T. Histon, W. Caplan, ... & N. Pronk. 2007. Weight-loss outcomes: A systematic review and meta-analysis of weight-loss clinical trials with a minimum 1-year follow-up. *Journal of the American Dietetic Association*, 107(10): 1755-1767.

M. Young, D. Lubans, C. Collins, R. Callister, R. Plotnikoff, P. Morgan, ... & T. Burrows. 2014. Behavioral interventions to reduce sedentary behavior in children and adolescents: Systematic review and meta-analyses. *British Journal of Sports Medicine*, 48(3): 147-155.

M. Young, R. Plotnikoff, C. Collins, R. Callister, P. Morgan. 2012. Social cognitive theory and physical activity: A systematic review and meta-analysis. *Obesity Reviews*, 13(12):, 1100-1111.

FDA. Digital Health Guidance Documents, <https://www.fda.gov/medical-devices/digital-health-center-excellence/guidances-digital-health-content>.

European Commission. 2016. General Data Protection Regulation, <https://eur-lex.europa.eu/eli/reg/2016/679/oj>.

U.S. Department of Health & Human Services. 2021. HIPAA for Professionals, <https://www.hhs.gov/hipaa/for-professionals/index.html>.

ISO. 2021. ISO/TS 22272:2021 Health Informatics - Methodology for analysis of business and information needs of health enterprises to support standards based architectures,

<https://www.iso.org/standard/78905.html>.

Y. Oh, J. Zhang, M. Fang, Y. Fokushida. 2021. A systematic review of artificial intelligence chatbots for promoting physical activity, healthy diet, and weight loss. *Int J Behav Nutr Phys Act*, 18(160).

B. Sears. 2024. <https://drsears.com/the-zone-diet/>.

A. Paoli. 2014. Ketogenic diet for obesity: friend or foe?, *Int. Journal of Environmental Research and Public Health*, 11(2): 2092-2107.

K. Varady. 2011. Intermittent versus daily calorie restriction: Which diet regimen is more effective for weight loss?, *Obesity Reviews*, 12(7): e593-e601.

R. Estruch, E. Ros, J. Salas-Salvadó, M. Covas, D. Corella, F. Arós, ..., M. Martínez-González. 2018. Primary Prevention of Cardiovascular Disease with a Mediterranean Diet Supplemented with Extra-Virgin Olive Oil or Nuts. *New England Journal of Medicine*, 378(25):e34.

M. McMacken, S. Shah. 2017. A plant-based diet for the prevention and treatment of type 2 diabetes, *Journal of Geriatric Cardiology*, 14(5):342.

A. Rahmanti, H. Yang, B. Bintoro, A. Nursetyo, M. Muhtar, S. Syed-Abdul, Y. Li. 2022. SlimMe a Chatbot With Artificial Empathy for Personal Weight Management: System Design and Finding." *Frontiers in Nutrition*, 9:870775

S. Holmes, A. Moorhead, R. Bond, H. Zheng, V. Coates, M. McTear. 2019. WeightMentor, bespoke chatbot for weight loss maintenance: Needs assessment & Development, 2019 IEEE Int. Conference on BIBM, 2845-2851.

# Mitigating Hallucinations in Large Language Models via Semantic Enrichment of Prompts: Insights from BioBERT and Ontological Integration

Stanislav Penkov

Sofia University “St. Kliment Ohridski”

spenkov101@gmail.com

## Abstract

The advent of Large Language Models (LLMs) has been transformative for natural language processing, yet their tendency to produce "hallucinations"—outputs that are factually incorrect or entirely fabricated—remains a significant hurdle. This paper introduces a proactive methodology for reducing hallucinations by strategically enriching LLM prompts. This involves identifying key entities and contextual cues from varied domains and integrating this information into the LLM prompts to guide the model towards more accurate and relevant responses. Leveraging examples from BioBERT for biomedical entity recognition and ChEBI for chemical ontology, we illustrate a broader approach that encompasses semantic prompt enrichment as a versatile tool for enhancing LLM output accuracy. By examining the potential of semantic and ontological enrichment in diverse contexts, we aim to present a scalable strategy for improving the reliability of AI-generated content, thereby contributing to the ongoing efforts to refine LLMs for a wide range of applications.

**Keywords:** Large Language Models (LLMs), Semantic Prompt Enrichment, Hallucination Mitigation, Domain-Specific Ontologies, BioBERT Entity Recognition

## 1. Introduction

Large Language Models (LLMs) have revolutionized numerous sectors by enabling machines to parse, understand, and generate human-like text, thus becoming cornerstones of

modern AI applications. However, despite their sophistication, LLMs face a critical challenge that threatens their reliability and ethical application: the generation of "hallucinations"—“the creation of factually erroneous information spanning a multitude of subjects” (Rawte et al., 2023). This issue, while technical, carries profound ethical implications, particularly when these models are deployed in information-sensitive areas such as healthcare, law, or education (Martino et al., 2023; Rawte et al., 2023; Feldman et al., 2023; Wan et al., 2024; Peng et al., 2023). It raises questions about the trustworthiness and dependability of AI-generated information, emphasizing the need for a solution that ensures LLM outputs are not just coherent but also factually accurate.

To address this, our paper introduces a novel, comprehensive methodology that pre-emptively mitigates these hallucinations by integrating domain-specific knowledge and semantic information directly into LLM prompts. This approach indirectly complements existing techniques like Retrieval-Augmented Generation (RAG) (Lewis et al., 2021) and prompt engineering. Drawing on tools like BioBERT<sup>1</sup> for biomedical entity recognition and the ChEBI<sup>2</sup> (Chemical Entities of Biological Interest) database for chemical ontologies, we propose a versatile approach to enhancing the accuracy of LLM outputs. While the current submission focuses on the theoretical framework and initial findings, we outline a detailed plan for future empirical evaluation to rigorously validate our methodology and provide concrete evidence of its effectiveness.

---

<sup>1</sup> [GitHub - dmis-lab/biobert: Bioinformatics2020: BioBERT: a pre-trained biomedical language representation model for biomedical text mining](https://github.com/dmis-lab/biobert)

<sup>2</sup> [Chemical Entities of Biological Interest \(ChEBI\)](https://www.ebi.ac.uk/ChEBI/)

## 2. Background

The versatility of general-purpose LLMs in handling tasks across various languages and domains introduces significant complexities in effectively evaluating and mitigating hallucinations (Zhang et al., 2023). These models, while robust, often display limitations in dynamic or culturally diverse contexts where nuances of language and factual accuracy are paramount. Current strategies such as dataset curation (Wan et al., 2024), model fine-tuning (Martino et al., 2023), and the integration of external knowledge bases (Peng et al., 2023) typically address these issues only after inaccuracies have been identified, which is not sufficient in preventing the initial occurrence of hallucinations.

Our methodology addresses these critical gaps by leveraging cutting-edge tools like BioBERT (Lee et al., 2019) and ChEBI, which anchor LLM outputs in verifiable facts, thereby not only reacting to but pre-emptively correcting potential inaccuracies. By embedding deeper semantic understanding directly into LLM prompts, our approach extends the ethos behind RAG, enhancing its capability to improve LLM reliability. This proactive integration, aligned with the principles demonstrated by Brown et al. (2020) in “Language Models are Few-Shot Learners”, sets a new benchmark for developing robust and ethical AI systems. The innovative approach ensures that LLMs operate within ethical bounds, enhancing their reliability across diverse applications and reducing the risk of misinforming users.

## 3. Proposed Methodology

To pre-emptively address the challenge of hallucinations in LLM outputs, our methodology employs a multi-layered approach, uniquely combining BioBERT for entity recognition, ChEBI for structured chemical ontologies, and direct API interactions with LLMs to guide the generation of accurate and relevant textual responses.

### 3.1 Technical Setup and Frameworks

**3.1.1 BioBERT<sup>3</sup>:** In our methodology, BioBERT is not merely used for its strong biomedical entity recognition capabilities. Instead, it serves a crucial role in the initial phase of our semantic enrichment process, where it precisely identifies and categorizes biomedical entities within LLM prompts. This specificity is vital as it ensures that the subsequent enrichment steps are accurately informed, targeting the most relevant semantic contexts required for each prompt. This targeted recognition is critical for minimizing errors in the generated text, especially in complex scenarios involving medical terminologies and contexts.

**3.1.2 libChEBI API<sup>4</sup>:** Similarly, the ChEBI database is utilized through the libChEBI API not just as a repository of chemical ontologies but as an integral component of our semantic enrichment framework. By programmatically accessing detailed ontological data, we enrich LLM prompts with deep semantic information that comprehensively describes chemical entities and their interactions. This enrichment goes beyond basic ontology integration by dynamically adjusting the context of the prompts to reflect current and precise chemical knowledge, significantly reducing the likelihood of generating inaccurate chemical data in LLM outputs.

**3.1.3 LLM API:** Interaction with LLMs, such as OpenAI's<sup>5</sup> GPT models, is achieved through their respective APIs.

### 3.2 Step-by-Step Implementation

**3.2.1 Setting Up the Environment:** The implementation begins with configuring the environment for BioBERT and setting up the API interactions with ChEBI and LLMs. This preparation ensures that all components are ready for the subsequent steps of entity recognition and prompt enrichment.

**3.2.2 Pre-processing and Entity Recognition with BioBERT:** Using BioBERT, we pre-process and analyze the input text to identify key biomedical entities. This identification is crucial for determining the specific entities around which the prompt enrichment revolves.

<sup>3</sup> <https://huggingface.co/dmis-lab/biobert-v1.1>

<sup>4</sup> <https://github.com/libChEBI/libChEBIpy>

<sup>5</sup> [OpenAI API | OpenAI](https://openai.com)

**3.2.3 Retrieving Ontological Data from ChEBI:** Once entities are identified, we query the ChEBI database to fetch ontological information related to these entities. Such information provides a layer of semantic depth to the prompt, grounding the LLM's output in factual accuracy.

**3.2.4 Enriching LLM Prompts:** The core of our methodology lies in the enrichment of LLM prompts. By integrating the entity recognition from BioBERT with the ontological data from ChEBI, we craft prompts that are rich in domain-specific knowledge and context. This integration is pivotal for guiding the LLM towards generating responses that are not only relevant but also anchored in factual accuracy.

**3.2.5 Generating Responses with LLM API:** With enriched prompts in hand, we interact with the LLM via its API, feeding it the prompts and analyzing the generated responses. This phase tests the effectiveness of our semantic prompt enrichment, with the expectation that the model's outputs will exhibit a marked reduction in hallucinations.

### 3.3 Planned Empirical Evaluation

Due to time and resource constraints, we were unable to conduct a full empirical evaluation for this submission. However, we outline our planned evaluation framework to be conducted in future work to rigorously validate our methodology.

**3.3.1 Data Collection:** We plan to collect a diverse dataset of biomedical queries and responses to ensure comprehensive evaluation.

**3.3.2 Experimental Design:** Our future work will implement a controlled study comparing the outputs of LLMs using original versus enriched prompts. This will help us understand the impact of semantic enrichment on reducing hallucinations.

**3.3.3 Expert Review:** Domain experts will be involved to review and score the responses based on accuracy and relevance. This qualitative assessment will provide valuable insights into the effectiveness of our approach.

**3.3.4 Statistical Analysis:** We will analyze the collected data to determine the statistical significance of improvements in accuracy and

reduction in hallucinations. Metrics such as precision, recall, and F1-score will be used to quantify the benefits of our methodology.

By conducting this empirical study in future work, we aim to provide concrete evidence of the effectiveness of our methodology. This planned evaluation will not only validate our approach but also contribute valuable insights to the ongoing efforts to enhance the reliability of AI-generated content.

### 3.4 Integration Flow Example

Example: Describing the interaction between Aspirin and blood pressure.

Input Text: "Describe the interaction between Aspirin and blood pressure."

1. BioBERT identifies entities
  - Output: "Aspirin", "blood pressure"
2. Retrieve ontological data from ChEBI
  - Output: "Aspirin (acetylsalicylic acid, C9H8O4)"
3. Enrich LLM prompt
  - Output: "Describe the interaction between Aspirin (acetylsalicylic acid, C9H8O4) and blood pressure, considering its properties as an anti-inflammatory agent and its effects on blood coagulation."
4. Generate LLM response
  - Output: Detailed and accurate response based on enriched prompt

The integration flow example illustrates the practical application of combining BioBERT for biomedical entity recognition with ChEBI for chemical ontology, facilitated by direct interactions with LLM APIs. This strategic combination not only enhances the input to the models but also guides them towards generating outputs that are both accurate and contextually relevant. By embedding such enriched prompts, we pave the way for the development of LLMs capable of producing more accurate and trustworthy AI-generated content across a broad spectrum of domains. Our methodology

demonstrates a scalable and versatile approach to improving the fidelity of LLM outputs, addressing the critical challenge of hallucinations and setting a new standard for the reliability of AI in sensitive and information-intensive fields.

#### 4. Pre-emptive Implementation and Comparative Analysis

A pivotal innovation of our approach is the introduction of a pre-emptive layer of semantic depth, a concept designed to mitigate potential hallucinations even before they occur. Unlike traditional methods that react to inaccuracies post-generation, our methodology proactively incorporates domain-specific ontologies and entity recognition into the enrichment of LLM prompts. This pre-emptive strategy is fundamental in setting our work apart from existing techniques such as the Retrieval-Augmented Generation (RAG) system highlighted by Kang et al. (2023). While RAG enhances LLM responses by integrating external knowledge before generation, our approach extends these capabilities by embedding a deeper layer of semantic understanding, offering a novel solution to the challenge of hallucinations in LLM outputs.

Our technical framework aligns with broader hallucination mitigation tools, drawing parallels with initiatives like LLM-Augmenter and FreshPrompt (Peng et al., 2023; Vu et al., 2023). By incorporating BioBERT and ChEBI for structured prompt enrichment, we not only align with but also advance the principles of augmenting LLMs with external knowledge and feedback mechanisms. This theoretical exploration sets a new benchmark for mitigating hallucinations and opens avenues for future research to explore pre-emptive measures over reactive ones.

#### 5. Discussion and Broader Implications

Our proactive methodology not only shifts the paradigm of semantic enhancements (Zhang et al., 2023) but also sets a new standard for reliability in AI-generated content. By incorporating empirical validations across various domains, we aim to substantiate our methodology's robustness. Exploring additional domain-specific ontologies will refine and expand our approach, enhancing

the adaptability of LLMs. This is crucial as it allows LLMs to operate effectively within diverse fields, pushing the boundaries of current AI capabilities. Future research will customize and enhance the semantic accuracy of LLM outputs, ensuring AI-generated content is reliable and contextually appropriate across disciplines. This holistic approach mitigates the risks associated with hallucinations and advances the development of ethically sound and universally dependable AI systems.

#### 6. Conclusion

This paper details an innovative, proactive methodology designed to tackle the challenge of hallucinations in LLM outputs from the ground up, marking a significant stride towards the ethical deployment and operational reliability of these advanced models. While our current findings are based on theoretical frameworks, we have outlined a comprehensive plan for empirical evaluation in future work. This planned study will rigorously validate our approach, providing the concrete evidence needed to establish the effectiveness of semantic prompt enrichment in improving LLM output accuracy and reliability. By establishing this new standard for the development of AI systems, we underscore our methodology's potential to significantly reduce hallucinations and enhance the factual integrity of AI-generated content.

#### References

- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language Models are Few-Shot Learners. <https://doi.org/10.48550/arXiv.2005.14165>
- Feldman, P., Foulds, J. R., & Pan, S. (2023). Trapping LLM Hallucinations Using Tagged Context Prompts. <https://doi.org/10.48550/arXiv.2306.06085>
- Kang, H., Ni, J., & Yao, H. (2024). Ever: Mitigating Hallucination in Large Language Models through

Real-Time Verification and Rectification.  
<https://arxiv.org/abs/2311.09114>

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. <https://arxiv.org/abs/2005.11401>

Martino, A., Iannelli, M., & Truong, C. (2023). Knowledge Injection to Counter Large Language Model (LLM) Hallucination. In Proceedings of the European Semantic Web Conference (ESWC). Yext New York NY.

Peng, B., Galley, M., He, P., Cheng, H., Xie, Y., Hu, Y., Huang, Q., Liden, L., Yu, Z., Chen, W., & Gao, J. (2023). Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback. <https://doi.org/10.48550/arXiv.2302.12813>

Rawte, V., Priya, P., Islam Tonmoy, S. M., Zaman, S. M. M., Sheth, A., & Das, A. (2023). Exploring the

Relationship between LLM Hallucinations and Prompt Linguistic Nuances: Readability Formality and Concreteness. AI Institute University of South Carolina USA. <https://doi.org/10.48550/arXiv.2309.11064>

Vu, T., Iyyer, M., Wang, X., Constant, N., Wei, J., Wei, J., Tar, C., Sung, Y.-H., Zhou, D., Le, Q., & Luong, T. (2023). FreshLLMs: Refreshing Large Language Models with Search Engine Augmentation. <https://doi.org/10.48550/arXiv.2310.03214>

Wan, F., Huang, X., Cui, L., Quan, X., Bi, W., & Shi, S. (2024). Mitigating Hallucinations of Large Language Models via Knowledge Consistent Alignment. <https://doi.org/10.48550/arXiv.2401.10768>

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, A. T., Bi, W., Shi, F., & Shi, S. (2023). Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. <https://doi.org/10.48550/arXiv.2309.01219>

# Commercially Minor Languages and Localization

Maria Ivanova Todorova  
Sofia University  
mivanovat@uni-sofia.bg

## Abstract

This paper offers a perspective of languages with a less significant volume of digital usership as minor in the context of globalization and localization. With this premise, the risks this status poses to the quality of localized texts, the substantiality of genre conventions, the public image of professional translators, and the users' linguistic competence in these languages is explored. Furthermore, the common lack of established or clear conventions in the localization of digital products into commercially minor languages (and in the digital product genres) is highlighted as one of the factors amplifying these risks. These perspectives are contextualized with the Bulgarian language with examples of errors encountered in Bulgarian digital content localized from English and more specifically – errors and problems related to gender neutrality and register.

**Keywords:** minor languages, localization quality, linguistic competence, English-Bulgarian translation

## 1 Introduction

In localization, a major language is considered a language with high global distribution and commercial use, and a minor one – the opposite. In this context, not only all minority languages become minor, but also the understanding of “minority” in the globalized context should be re-examined. As early as 20 years ago, Cronin makes one crucial claim that is denotative for linguistic inequality in an era of technology and globalization, by declaring that “the hegemony of English in the fastest-growing areas of technological development means that all other

languages become in this context, minority languages.” (2003: 146). Speaking about this extended idea of “minority”, Cronin lists several negative consequences of “the absence of a critical self-reflective activity in the translation enterprise”: the vulnerability of minority languages to interference in majority-to-minority translation, the alteration of minority languages in translation, and the predominant focus on majority languages in Translation Theory (2003: 149). This paper applies Cronin's logic by extension to Bulgarian, a globally and commercially minor language, and, considering the all-pervasiveness of digital content in modern everyday life, explores the potential issues such a lesser status can lead to in the context of digital product localization.

## 2 Localizing digital content in Bulgarian

In the globalized present, every digital-market player strives to bring their product to as many users as possible – a crucial step in this process is localization, the adaptation of said product to the target market's locale<sup>1</sup>. When international enterprises enter the small Bulgarian market, they are first faced with one dilemma – to localize or not to localize. The inevitability of this question lies in “the quantitative prism of return on investment”, through which business executives perceive localization (Dunne, 2006: 1).

One can easily imagine that with fewer than 5 million users who access the Internet on a monthly basis (Miniwatts Marketing Group, 2023) Bulgaria is not such an enticing venture. Nevertheless, out of 22 identified English-source websites in the 100 most visited by Bulgarian users in August 2021, 12 have been localized into Bulgarian. What is more, out of the 7 originally English ones with over 1 million unique visits by Bulgarian users in the same

---

<sup>1</sup> ‘not just a language, but usually a particular variety of a language, plus local conventions regarding currency, date and hour settings, presentation of numbers, right through to such things as symbolic color coding’ (Pym, 2004: 2).

period, only 1 has not been localized (SimilarWeb, 2021).

What does this mean for the Bulgarian-speaking usership? Firstly, a great share of Bulgarian users find it easy to navigate online content and services in English, with localization remaining a lower priority than the significance of the service offered through this content. Secondly, despite the existence of such a big fraction being able to navigate in English, more than half of the companies which have their services based on English still choose to localize their content in Bulgarian. Among those are the giants Microsoft & Google each of which own dozens of digital service software products, as well as most of the “big players” with significant market reach, such as Booking, or eSky. They do so regardless of the uncertain or potentially small benefit from this investment. The reason behind this could be an effort towards achieving a universal and inclusive market reach. The lesser commercial status, however, combined with the complexity and multimodality of localization as a process, puts Bulgarian and other commercially minor languages in a volatile position when it comes to the investment in and, respectively, quality control on localized content.

What is more, when discussing the pragmatic quality of translation solutions in localization, Jimenez-Crespo refers to the conventionality in genres and translation adequacy in finding corresponding conventions in source and target languages (Jimenez-Crespo, 2009: 68). Even if some conventions in the digital product genres exist in Bulgarian and other minor languages, however, they have been heavily influenced by localization practices with “the majority of software and web applications [were] being developed [in English] in the United States” during the dawn of localization (Esselink, 2000: 4). In other words, most underlying genre conventions have originally appeared in English and then have been established in other languages through finding functionally equivalent solutions to already existing paradigms in English originals. When there is no full correspondence between these paradigms and those of the target language, translators in localization projects end up in the critical position of the ones to determine and establish the good practices.

As Postolea puts it, “especially when their target language is lesser known and terminologically

standardised, specialised translators, just like journalists or authors, are creators of language too [...] From this perspective, specialised translators have responsibilities towards the target language too and they should be wary of the risk of introducing loose translations and poor style into the target context.” (2016: 63) This responsibility, however, rarely corresponds to proportionate resources (both economic and informational) in localization to a minor language, which inevitably leads to diffuse accountability and, respectively, questionable quality. It is also a prerequisite for companies to rely more heavily on automation and AI-powered machine translation (MT) – an approach which boosts inclusivity, but is still far from ideal for smaller and low-resource languages (Pym, Ayvazyan, Prioleau, 2022: 13).

### **3 Translation errors in digital products localized into Bulgarian**

Translation errors in localized websites, mobile applications, or even operation systems are certainly not uncommon and we often laugh about them in professional circles and even blame them on machine translation. MT, however, is rarely exclusively used in the localization context. In order to analyze and systematize the types of errors found in products localized from English into Bulgarian, I collected 150 mistakes found in 42 such products by browsing and using my mobile devices in Bulgarian for a little over a year, excluding mistakes from products that were highly likely to have used non-post-edited MT exclusively or errors that cannot be linked to factors specific to the context of localization. Some of the errors I encountered could still be attributed to MT, especially when it comes to short ambiguous or polysemous user-interface text units. The ones included in the corpus, however, were found in products whose companies have shared information about their localization processes publicly, often showing how they approach and value quality assurance – among those are various Microsoft and Google products (including Android), Revolut, Glovo, AirBnB, and others. Regardless of whether and to what extent MT was used in the processes that led to these mistakes, their existence and nature bring up concerns about the average level of quality in localization from English into Bulgarian and the potential consequences thereof.

The errors encountered range from problems related to text types or communicative situations typical of the digital product genres (37 %) to problems linked to the technological macrostructure behind the texts showing in digital products (41 %) and problems resulting from various levels of ambiguity of the succinct source texts (22 %) (Appendix A). Due to volume constraints, for the purposes of this paper, I will focus on the former type – genre-related errors, and more specifically, those related to gender neutrality and register (Appendix B), while some examples of the other categories can be found in Appendix C.

The mistakes associated with those two aspects occur mainly due to the asymmetry between English and Bulgarian in the categories of gender and number, and more particularly in lexemes referring to the user in the second person. Judging by the mistakes encountered, there does not seem to be a universal approach in translating digital product text types with such lexemes, which leads to inconsistency (Image 1) and grammatically incorrect solutions which hundreds of thousands of users witness on a daily basis. Considering the minor status of Bulgarian in localization, motivation to tackle these issues would unlikely come from the top (product owners) and Postolea’s statement that translators become “creators of language” rings all the more true.

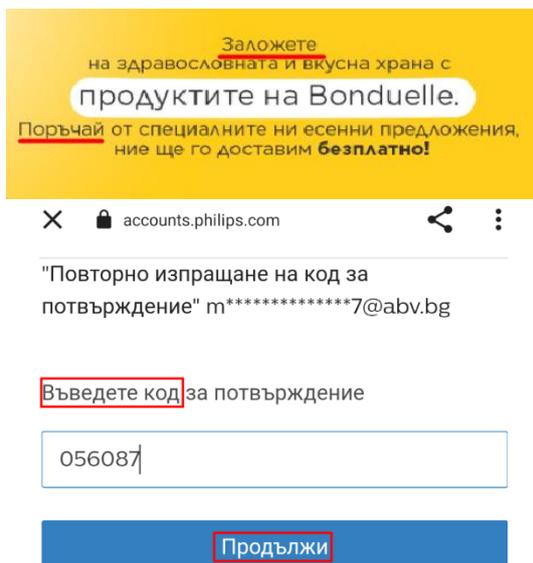


Image 1: Register inconsistency in Glovo and Philips Home

What happens though when there does not seem to be a consensus among stakeholders? Regardless of how common such translation inconsistencies

and mistakes can be in localized products, it is unlikely they would lead to any drastic linguistic changes on their own. In some of the linguistics aspects they are associated with, however, faltering of the norms (or uncertainty where one lacks) is observed on a greater scale.

The difficulties of adopting gender-neutral language in Bulgarian, for instance, often spark debates in professional translators’ groups. It might seem that we are far from codifying a solution, but translators have to find one on a regular basis – be it when localizing a product addressing an unknown individual user or a book or movie with non-binary characters. In localization, the lack of established strategies to apply in such cases mainly leads to ungrammatical solutions and inconsistency. If one solution were to prevail, however, it can slowly become an “unspoken rule” in localization practice and in digital product genres, if not in more universal use. In such context, as owners of the target language, translators involved in localization projects become decision-makers with the two most obvious choices of either disregarding the original neutrality or adapting the respective paradigm in Bulgarian to match it. As a result of this reality, the target texts in the minor language often end up as either too conservative and sometimes outright incorrect (Appendix B: Lines 21–24) or the target language simply gives way to the paradigm of the English source, using plural forms when referring to an individual user (Appendix B: Lines 10–20). Undoubtedly, these are not the only solutions when a variety of translation shifts can be adopted to bypass the asymmetry – e.g. instead of “Ще бъдете пренасочени” [You will be redirected(pl.)], a structural shift can be adopted to avoid the use of participle which will have to be marked for gender in the singular: “Ще Ви пренасочим” [We will redirect you(sg.)]. Such strategies, however, might not be universally applicable and often introduce additional limitations in the localization context that already implies numerous constraints on its own as seen in the second error category in Appendix A. Nevertheless, they can be applied in most of the context of the observed errors but have not, which lends the industry and academia at least one actionable solution – to include such strategies in translator training for the purposes of localization.

While the matter of linguistic gender-neutrality is yet to be solved in Bulgarian, the polite form of

address has already been codified. Despite the various adjustments over the years, we now have clear rules in regard to the use of capitalized pronouns and plural forms. And even though the past faltering of this form has weakened its use in present day and might have led to some of the related errors found in localized products, there are consistent patterns found that are likely due to something more than a disregard of the rules. In fact, Bulgarian localizers often oscillate between the informal singular and formal addresses, with the majority of products resorting to a “hybrid” non-codified form, characterized by the use of lowercase pronouns and plural adjectives and past participles when referring to an individual user. (Todorova, 2023: 122–123) What is more, judging by Google and Microsoft’s localized texts, this hybrid version is consistent with the conglomerates’ style guides and requirements. Unfortunately, Google does not provide open access to its documentation, while Microsoft only mentions the matter in passing, not singling it out, perhaps due to its lack of official codification (Todorova, 2023: 124).

#### 4 User reception

The reception of such translation errors would depend on the linguistic competence of the user. If they are a proficient speaker, such errors could only erode the image of professional translators. On the one hand, if the user is aware that a human is indeed behind such an error, it raises the question of how well professional translators are trained and how much these services are valued by investors in the face of digital product owners. On the other hand, if the user assumes the mistake results from the use of MT, the notion that professional translation services are not valued and sought after is reinforced. And in the cases where the user’s linguistic competence is already compromised, encountering such mistakes in the everyday use of their phone, for instance, could only deepen their lack of understanding or solidify previously adapted incorrect paradigms.

Additionally, if users commonly encounter such unfortunate localization instances, the question stands of how many people would actually choose to navigate digital content in their native tongue instead of in English when this is possible. In a country such as Bulgaria with a digitally minor language and an above-average rate of English proficiency (Education First, 2023: 4), proficient

English speakers would likely prefer to use digital products in English. As a result, even though they might have a better than average command of Bulgarian as well, their digital experience will be shaped by English modalities.

What is more, in a study on digital users’ reception, Taanonen concludes that “the [user] group with competent English skills is relatively more critical towards translated content and its quality than the group with weaker English-language skills.” (2014: 94) Losing even a part of this proficient and critical towards translation quality group, decreases the chance of motivation for improvement coming from the users. Additionally, by turning to English, users of higher proficiency decrease the market value of Bulgarian further and, respectively, the incentive for product owners to invest in improving the quality of localization.

#### 5 Final remarks

In the economic enterprise which localization essentially is, the quality of target texts in a language of a commercially minor status such as Bulgarian is often of lesser investment concern. This leaves translators at the vulnerable position of decision-makers within a low-resourced specialized context. Changes to this reality in the localization process, however, can mainly be expected to stem from the stakeholders considered owners of the minor language – the translators and translator trainers. A step in this direction could be the forming of a better understanding of translation problems in localization from English into Bulgarian, which would lead to the establishing of practical solutions and their integration in translation training.

On a grander scale, with the rise of AI-powered MT, which outperforms any human linguist in terms of speed and cost-efficiency, all the more focus should be put on finding ways to educate users and ensure their linguistic skills are strong enough to sift digital content through and identify quality localized content. Additionally, efficient processes should be found to guarantee that stakeholders in Bulgarian localization, such as translators and translation agencies, receive proper training, are equipped with sufficient resources, and enjoy professional prestige that would motivate higher responsibility and performance.

## References

- Michael Cronin. 2003. *Translation and Globalization*. London and New York: Routledge.
- Anthony Pym. 2004. *The Moving Text.: localization, translation, and distribution*. Amsterdam/Philadelphia: John Benjamins Publishing.
- Keiran J. Dunne. 2006. Introduction: A Copernican revolution. In *Perspectives on Localization*, pages 1–11. Amsterdam/Philadelphia: John Benjamins Publishing.
- Miniwatts Marketing Group. 2023. *Data for the 27 Member States of the European Union*. Accessed on 14 April 2024. <https://www.internetworldstats.com/europa.htm#bg>
- Sorina Postolea. 2016. Translating in a Specialised Context: Challenges and Risks. In *Buletinul Institutului Politehnic din Iași* (Bulletin of the Polytechnic Institute of Iasi), LXII(LXVI), f.1, pages 51–66.
- SimilarWeb. 2021. *Industry Leaders Data*. Accessed on 22 September 2021. <https://pro.similarweb.com/>
- Miguel A. Jimenez-Crespo. 2009. The evaluation of pragmatic and functionalist aspects in localization: towards a holistic approach to Quality Assurance. In *The Journal of Internationalization and Localization*, 1(1): 60–93. Amsterdam/Philadelphia: John Benjamins Publishing.
- Bert Esselink. 2000. *A Practical Guide to Localization*. Amsterdam/Philadelphia: John Benjamins Publishing.
- Anthony Pym, Nune Ayvazyan, Jonathan Prioleau. 2022. Can raw machine translation ensure inclusion? The case of public-health information in Catalonia. *1<sup>st</sup> International Conference Right to Languages: Linguistic Policies and Translation and Interpreting in Public Services and Institutions*.
- Maria Todorova. 2023. Forms of Address in Website and Software Localization From English into Bulgarian. In: *Studia Philologica Universitatis Tarnovenssis (series annua)*, 42. Veliko Tarnovo: Veliko Tarnovo University Publishing House, 119–126.
- Education First. 2023. *EF English Proficiency Index. A Ranking of 113 Countries and Regions by English Skills*. Accessed on 14 April 2024. <https://www.ef.com/assetscdn/WIBIwq6RdJvcD9bc8RMd/cefcom-epi-site/reports/2023/ef-epi-2023-english.pdf>
- Mari Taanonen. 2014. *Targeting Language Markets through Web Localization: A Case Study of a Digital Cloud Service*. Master's Thesis. Espoo: Aalto University

**Appendix A Problem categories and respective error types found in products localized into Bulgarian**

| CATEGORY                    | SUBCATEGORY                     | TYPE  |
|-----------------------------|---------------------------------|---|
| GENRE-RELATED               | FORM OF ADDRESS                 | INCONSISTENCY   |
|                             |                                 | LACK OF PRONOUN CAPITALISATION IN THE POLITE ADDRESS                        |
|                             |                                 | ADJECTIVE OR PARTICIPLE PLURALISATION                                       |
|                             | USER GENDER                     | ADJ. OR PARTICIPLE MASCULINISATION  |
|                             |                                 | ADJ. OR PARTICIPLE NEUTERISATION  |
|                             |                                 | INCONSISTENCY   |
|                             | CALL-TO-ACTION NAVIGATION TEXTS | AMBIGUITY   |
|                             |                                 | INCONSISTENT TERMINOLOGY  |
|                             | TERMS & NEOLOGISMS              | INCONGRUITY AND/OR AMBIGUITY OF TERMS                                       |
|                             |                                 | WORD ORDER INTERFERENCE   |
| MACROSTRUCTURAL LIMITATIONS | VARIABLES IN THE SOURCE CODE    | VARIABLE-SPELLING OF BULGARIAN PREPOSITIONS                                 |
|                             |                                 | ASYMMETRY IN GENDER   |
|                             |                                 | ASYMMETRY IN NUMBER   |
|                             |                                 | NON-VISIBLE USER-INTERFACE (UI) TEXT DISRUPTING THE PRODUCT'S FUNCTIONALITY |
|                             | CHARACTER LIMITATIONS           | VISIBLE UI TEXT DISRUPTING THE PRODUCT'S DESIGN                             |
|                             |                                 | UNCLEAR ABBREVIATIONS   |
|                             |                                 | DURATION  |
|                             | LOCALE CONVENTIONS              | CURRENCY  |
|                             |                                 | CAPITALIZATION  |
|                             |                                 | POLYSEMY  |
| INSUFFICIENT CONTEXT        | & SOURCE AMBIGUITY              | FUNCTIONAL POLYSEMY   |
|                             |                                 | HETEROSEMY  |
|                             |                                 | SYNCRETISM  |
|                             |                                 | SYNONYMY  |
|                             |                                 | ERRORS DUE TO ASYMMETRY IN GENDER   |
|                             | & TARGET PARADIGMATIC VARIATION | ERRORS DUE TO ASYMMETRY IN NUMBER   |
|                             |                                 | ERRORS DUE TO ASYMMETRY IN DEFINITENESS                                     |
|                             |                                 | FUNCTIONAL POLYSEMY IN TARGET TEXT  |
|                             |                                 | PUNCTUATION INTERFERENCE  |

Appendix B Some of the errors related to register and gender neutrality

| No. | ENGLISH SOURCE   | BULGARIAN TARGET  | PRODUCT       | TYPE   |
|-----|--|---|---------------|--|
| 1.  | Thanks for reviewing your stay, Maria!   | Благодарим <b>ви</b> , че оценихте престоя си, Maria!   | Booking       | LACK OF CAPITALISATION IN SINGULAR POLITE ADDRESS                |
| 2.  | Log in as Maria... Not <b>you</b> ?  | Влизане като Maria... Не сте <b>вие</b> ?   | Facebook      | LACK OF CAPITALISATION IN SINGULAR POLITE ADDRESS                |
| 3.  | Manage <b>your</b> Google Account  | Управление на профила <b>ви</b> в Google  | Google Chrome | LACK OF CAPITALISATION IN SINGULAR POLITE ADDRESS                |
| 4.  | Maria your photo book draft is about to expire   | Maria, черновата на албума <b>ви</b> е на път да изтече   | Google Photos | LACK OF CAPITALISATION IN SINGULAR POLITE ADDRESS                |
| 5.  | This app is an internal test version that has been shared with <b>you</b> . If <b>you</b> install it, <b>your</b> email address will be shared with the developer. | Това приложение е вътрешна версия, която е споделена с <b>вас</b> . Ако го <b>инсталирате</b> , имейл адресът <b>ви</b> ще бъде споделен с програмиста. | Google Play   | LACK OF CAPITALISATION IN SINGULAR POLITE ADDRESS                |
| 6.  | {username} liked <b>your</b> post  | {username} хареса публикацията <b>ви</b>  | Instagram     | LACK OF CAPITALISATION IN SINGULAR POLITE ADDRESS                |
| 7.  | Autofill <b>your</b> card details at checkout  | Автоматично попълване на данните на картата <b>ви</b> за плащане  | Revolut       | LACK OF CAPITALISATION IN SINGULAR POLITE ADDRESS                |
| 8.  | An activation email was sent to <b>your</b> email address.   | Имейл за активация беше изпратен до <b>вашия</b> имейл адрес.   | Xiaomi Home   | LACK OF CAPITALISATION IN SINGULAR POLITE ADDRESS                |
| 9.  | Every night when you are on Wi-Fi, we'll automatically download <b>your</b> favourite music...   | Всяка вечер при наличие на връзка с Wi-Fi ще изтегляме любимата <b>ви</b> музика...   | YouTube Music | LACK OF CAPITALISATION IN SINGULAR POLITE ADDRESS                |
| 10. | You will be <b>forwarded</b> back to the home screen   | Ще бъдете <b>пренасочени</b> към началния екран.  | BigBlueButton | ADJECTIVE OR PARTICIPLE PLURALISATION IN SINGULAR POLITE ADDRESS |
| 11. | Are you <b>sure</b> you want to follow this link?  | <b>Сигурни</b> ли сте, че искате да последвате тази връзка?   | Facebook      | ADJECTIVE OR PARTICIPLE PLURALISATION                            |
| 12. | Are you <b>sure</b> you want to log out?   | <b>Сигурни</b> ли сте, че искате да излезете?   | Facebook      | ADJECTIVE OR PARTICIPLE PLURALISATION                            |
| 13. | {user} commented on a photo that you're <b>tagged</b> in   | {user} коментира снимка, в която <b>сте отбелязани</b>  | Facebook      | ADJECTIVE OR PARTICIPLE PLURALISATION                            |
| 14. | Have an account already?   | Вече <b>сте регистрирани</b> ?  | Glovo         | ADJECTIVE OR PARTICIPLE PLURALISATION                            |
| 15. | Are you <b>hungry</b> ?  | <b>Гладни</b> ли сте?   | Glovo         | ADJECTIVE OR PARTICIPLE PLURALISATION                            |
| 16. | <b>Unstoppable!</b>  | <b>Неудържими</b> сте!  | Google Fit    | ADJECTIVE OR PARTICIPLE PLURALISATION                            |

|     |   |   |               |   |
|-----|---|---|---------------|---|
| 17. | {user}, are you ready to print your photo book?                                   | Мария, <b>готови</b> ли сте да отпечатате фотоалбума си?  | Google Photos | ADJECTIVE OR PARTICIPLE PLURALISATION   |
| 18. | {user} commented on a post you are tagged in                                      | {user} коментира публикация, в която <b>сте отбелязани</b>  | Instagram     | ADJECTIVE OR PARTICIPLE PLURALISATION   |
| 19. | Do you know and trust {user}?<br>If you are <b>unsure</b> , don't pay them...     | Ако не сте <b>сигурни</b> , не му плащайте...   | Revolut       | ADJECTIVE OR PARTICIPLE PLURALISATION   |
| 20. | –   | Здравейте, Мария, Имаме страхотни новини: официално сте <b>проверени</b> . Това означава, че вече можете да използвате своята карта на Revolut... | Revolut       | ADJECTIVE OR PARTICIPLE PLURALISATION   |
| 21. | We've contacted you because you opted-in to receive newsletters from Takeaway.com | Свързахме се с теб, защото си се <b>съгласил</b> да получаваш новини от Takeaway.com  | Takeaway      | ADJECTIVE OR PARTICIPLE MASCULINISATION |
| 22. | Maria, are you ready to...  | Мария, готов ли си да се впуснеш...   | Takeaway      | ADJECTIVE OR PARTICIPLE MASCULINISATION |
| 23. | Followed by {user}  | Последвано от {user}  | Instagram     | ADJECTIVE OR PARTICIPLE NEUTERISATION   |
| 24. | Do you know and trust {user}?<br>If you are unsure, don't pay <b>them</b> ...     | Познавате ли и имате ли доверие на <b>{female user's name}</b> ?<br>Ако не сте сигурни, не му плащайте...   | Revolut       | ADJECTIVE OR PARTICIPLE MASCULINISATION |

Appendix C Some examples of errors form other categories

| ENGLISH SOURCE                     | BULGARIAN TARGET                   | PRODUCT              | CONTEXT   |
|------------------------------------|------------------------------------|----------------------|---|
| Weekly on {day of the week}        | Ежеседмично в вторник              | Google Calendar      | Incorrect spelling of the preposition <i>o</i> in the settings of an event due to the variability of the succeeding noun.                                 |
| 1B+ Downloads                      | 1 млрд.+ <b>Изтегляния</b>         | Google Play          | Use of English capital case in Bulgarian.   |
| {number} key moments in this video | 7 ключови моменти в този видеоклип | Google + YouTube     | Use of incorrect plural form of a masculine noun due to asymmetry between English and Bulgarian in the category of number in combination with a variable. |
| {number} timer set                 | 1 таймер е зададен                 | Android              | English word order in info text potentially due to hardcoded macrostructure.  |
| {number} days left                 | 17 дни остават                     | Fundraising platform | English word order in info text potentially due to hardcoded macrostructure.  |
| Order no.                          | Номер на поръчка                   | Obuvki.bg            | Omission of the definite article in Bulgarian   |
| Hi {user},                         | Здравей Мария,                     | Takeaway             | Adopting the English source's punctuation in the context of a variable element.   |

---

**SPECIAL SESSION ON  
WORDNETS, FRAMENETS AND ONTOLOGIES**

---

# Semantic features in the automatic analysis of verbs of creation in Bulgarian and English

Ivelina Stoyanova

Department of Computational Linguistics  
Institute for Bulgarian Language, Bulgarian Academy of Sciences  
iva@dc1.bas.bg

## Abstract

The paper focuses on the semantic class of verbs of creation as a subclass of dynamic verbs. The objective is to present the description of creation verbs in terms of their corresponding semantic frames and to outline the semantic features of the frame elements with a view to their automatic identification and analysis in text. The observations are performed on Bulgarian and English data with the aim to establish the language-independent and language-specific features in the semantic description of the analysed class of verbs.

**Keywords:** verbs of creation, frame semantics, Bulgarian, English

## 1 Introduction

The paper shows ongoing research on the semantic and conceptual properties of verbs with a view to their realisation in various languages. In particular, here we focus on the class of creation verbs and their possible automatic analysis and disambiguation in text.

For the purpose of the automatic classification of creation verbs according to their aspectual class (state, activity, accomplishment and achievement, as discussed below), we consider a set of features at lexical, semantic, syntactic and valence level which determine the realisation of the verb in the sentence.

While some verbs can be clearly classified to one aspectual class at the lexical level, others cannot be classified and can have different realisations depending on the context. Some verbs can express an activity (e.g., *He has been writing all morning*), accomplishment (e.g. *He wrote a letter*), or even a state (e.g., *He writes illegibly*, i.e. the quality / state of his writing is such that it cannot be easily read).

As a consequence of these observations, the aspectual classes are considered as realised in

text rather than at the lexical level. The aspectual properties are therefore described as a combination of lexical, morphosyntactic and valence properties.

An additional layer for consideration is the lexical aspect in Bulgarian and other Slavic languages. Examples 1 – 6 show use cases of the Bulgarian verbs *готвя* ‘cook’ (imperf.) and *сготвя* ‘cook’ (perf.), the latter derived from the former using prefixation, and both translated into English as ‘cook’. Relevant to the study of the semantic properties of verbs of creation is the fact that perfective verbs are always telic (Examples 5 and 6) and the limitations in their interpretation stemming from that (e.g., Example 6). Ambiguity at the lexical level between the telic and atelic interpretation of verbs can only occur with imperfective verbs (Examples 1, 2 vs. 4).

- (1) *Той готви часове наред.*  
He cooks hours in sequence.  
‘He cooks for hours on end.’ (*готвя* ‘cook’, imperf.; atelic)
- (2) *Той готви ястието часове наред.*  
He cooks dish-DEF hours in sequence.  
‘He cooks the dish for hours.’ (*готвя* ‘cook’, imperf.; atelic)
- (3) *Той готви това ястие за един час.*  
He cooks this dish for one hour.  
‘It takes him one hour to cook this dish.’ (*готвя* ‘cook’, imperf.; telic)
- (4) *Той сготви ястието.*  
He cooked dish-DEF.  
‘He cooked the dish.’ (*сготвя* ‘cook a complete dish’, perf.; always telic)
- (5) *Той сготви ястието за един час.*  
He cooked dish-DEF for one hour.  
‘He cooked the dish in an hour.’ (*сготвя* ‘cook a complete dish’, perf.; always telic)
- (6) \**Той сготви ястието часове наред.*  
He cooked dish-DEF hours in sequence.

\*‘He (completely) cooked the dish for hours.’ (*сготовля* ‘cook a complete dish’, perf.; always telic)

With a view to establishing testing criteria and approaches for the automatic classification of verbs in terms of the aspectual classes, the semantic class of verbs of creation is a good illustrative class since creation verbs, as a subclass of verbs of change, presuppose reaching an end point – a new state, an object coming into existence as a result of the performed action, thus have a clear telic interpretation as achievements and accomplishments. However, it is interesting to observe their use in specific contexts as activities, and based on illustrative material to perform contrastive analysis and observe the semantic, morphosyntactic and valence features determining their interpretation and allowing their disambiguation.

## 2 Relevant works

The review of relevant works goes in two directions: (a) an overview of general aspectual verb classes in order to derive the set of lexical-semantic, morphosyntactic and valence properties determining the classification of predicates; and (b) an outline of the particular class under observation, the class of creation verbs, and its place among the general classification of verbs.

The observations presented here are based on the aspectual classes of activities, accomplishments, achievements and states (Vendler, 1957, 1967; Dowty, 1979), distinguished with respect to the following four features: [ $\pm$  static], [ $\pm$  dynamic], [ $\pm$  telic] and [ $\pm$  punctual]. In particular, we focus on dynamic verbs characterised by [ $-$  static] and [ $+$  dynamic], and analyse the class of verbs of creation within dynamic verbs.

Rappaport Hovav (2008: 14–16) discusses the event-denoting predicates and argument realisation with respect to the lexicalized meaning and the lexical properties that determine the aspectual meaning, as well as the aspectual class of the larger unit, the VP, that the verb appears in.

Dowty (1979: 132–135) and Rappaport Hovav (2008: 16–17) state that all dynamic predicates ([ $-$  static]) are characterised by a change, and thus they refer to an interval as the change occurring involved two moments in time (initial and final moment). With respect to aspectual properties of

verbs, Dowty (1979: 132–135) and Rappaport Hovav (2008: 16–17) distinguish between scalar (e.g., *fall*, *warm*) and non-scalar change (e.g., *scribble*, *exercise*). Scalar verbs require an ordered set of values for a particular attribute, with the order from the initial towards the final value. Non-scalar verbs involve a complex change which cannot be expressed in terms of a scale on a single attribute.

Moreover, Beavers (2008: 245) argues that the telicity of dynamic predicates arises from a homomorphism between the event and some bounded participant in the event. In particular, the boundedness of the event is related to the boundedness of the incremental theme, a property, or a path. Beavers (2008: 257) gives the following definition:  $P$  is a dynamic predicate iff  $P$  predicates over an event  $e$ , a force-recipient  $x$ , a scale of change  $s$ , and possibly other entities.

Further, Rappaport Hovav (2008: 17–18) argues that the change specified by activity verbs is usually more complex than that of achievements and accomplishments. In line with the observations of Beavers (2008: 250–257), Rappaport Hovav (2008: 17–18) discusses two-point scales (where there are only two states, e.g. *die*, *reach*, *shut*) and multi-point scales (for gradual change, e.g. *warm*, *lengthen*, *flatten*). Two-point scales are inherently bounded, thus the predicates associated with them are telic and punctual. On the other hand, for multi-scale predicates an additional distinction is made between closed scales (where there is an end point), e.g. *flatten* (until  $x$  becomes flat) and open scales (where there is no end point), e.g. *lengthen* (potentially indefinitely).

Beavers (2008) proposed a unified approach to the determination of telicity by considering three verb classes: incremental theme verbs; true change-of-state verbs with an argument exhibiting a gradable property of the argument; and inherently directed motion verbs for which the path of motion is a measurable feature. Most relevant to the study of creation verbs is the incremental theme, whose scalability is with respect to volume, area, height, etc. Further, Levin (2010: 1–2) introduces the feature **Scale** and considers it as an integral part of the verb semantics and conceptual structure.

With a view to verbs of creation, there are several particularly relevant studies that outline the place of the verb class under investigation within

general classifications of verb predicates.

The classification of predicates and the semantic relations between predicates and their arguments have been studied within various theoretical approaches, based in general on their syntactic properties and behavior (Levin (1993); Pinker (1989), among others), the thematic structure (Van Valin and LaPolla, 1997) or the semantics of frames (Fillmore, 1982). Levin (1993) analyses creation and transformation verbs as one whole class as they exhibit similar alternations. In particular, verbs of creation enter into the Material/Product alternation (e.g., *He carved the wood into a flute / He carved the flute out of wood*), the Raw Material subject alternation (e.g., *She baked bread from the wholemeal flour / This wholemeal flour bakes a good bread*), the Instrument subject alternation (e.g., *I draw nice sketches with this pencil / This pencil draws nice sketches*), etc.

In his analysis of creation and depiction verbs, Forbes (2006) talks about event semantics and the “unfinished object problem” in sentences such as *Jack was building a house, Jack was drawing a circle*. Although there is vagueness with respect to what is considered a house (even unfinished), a circle is not a circle unless it is complete. However, the author assumes that the event in the progressive implies existential neutrality, neither requiring it nor forbidding the existence of the object.

In Bulgarian and other Slavic languages the verb aspect is a lexical category and the perfective and imperfective verbs are considered different words with different lexical meanings derived as a result of word formation. The properties of the verb aspect have been studied extensively (Andreychin, 1944; Ivanchev, 1971; Nitsolova, 2008; Kutsarov, 2007; Koeva, 2011; Charalozova, 2021).

In summary, we adopt Vendler’s classification of activities, achievements, accomplishments and states with the relevant features and test to distinguish between them. The aspectual classification in Bulgarian is more complex as it also takes into account the verb’s lexical aspect and the additional restrictions stemming from it. Verbs of creation within the class of verbs of change fall into the categories of activities, achievements and accomplishments as they express a transition from one state (non-existence) to another (existence), so there is an inherent end point but it is not necessarily implied when the

activities are discussed.

### 3 Verbs of creation in WordNet and BulNet

The focus is on the representation of verbs of creation in the hierarchical structure of Princeton WordNet and the Bulgarian WordNet, and how it is reflected by the system of the semantic frames from FrameNet that the verbs evoke.

The lexical-semantic network WordNet (Miller, 1995; Fellbaum, 1998) represents the lexicon in the form of a network of synonym sets (synsets) interconnected by semantic, lexical and other relations. One of the main relations building the hierarchical structure of WordNet is hypernymy (and its opposite relation – hyponymy), which organises the vocabulary of a given semantic field into a tree.

WordNet, as well as its Bulgarian counterpart BulNet (Koeva, 2006, 2021), is the main resource used in the study. The semantic description of verb predicates in WordNet also includes their classification into general semantic classes based on assigned semantic primitives (Miller and Fellbaum, 2007), e.g. verbs of motion, verbs of emotion, verbs of communication, verbs of creation, etc.

In Princeton WordNet there are 834 synsets labelled with the semantic primitive verb.creation, covering a total of 1898 verb literals in English. Out of them, 453 synsets are linked to verb synsets in the Bulgarian WordNet covering a total of 2073 verb literals in Bulgarian.

In general, the class of verbs of creation includes, among others:

- Verbs of intentional creation of a physical object – rooted at the synsets eng-30-01685313-v {create} ‘pursue a creative activity; be engaged in a creative activity’; eng-30-01617192-v {make, create} ‘make or cause to be or to become’ and eng-30-01753788-v {create} ‘bring into existence’.
- Verbs of creating a depiction – stemming at eng-30-01686956-v {picture, depict, render, show} ‘show in, or as in, a picture’.
- Verbs of building – starting from eng-30-01654628-v {construct, build, make} ‘make by combining materials and parts’ and from eng-30-01656788-v {assemble, piece, put

together, set up, tack, tack together} ‘create by putting components or members together’.

- Verbs of decorating – stemming at eng-30-01675963-v {decorate, adorn, grace, ornament, embellish, beautify} ‘make more attractive by adding ornament, colour, etc.’.
- Verbs of authoring, or textual creation – stemming at eng-30-01698271-v {write, compose, pen, indite} ‘produce a literary work’.
- Verbs of manufacturing – starting from the synset eng-30-01621555-v {produce, make, create} ‘create or manufacture a man-made product’.
- Verbs of cooking – rooted at eng-30-01664172-v {cook, fix, ready, make, prepare} ‘prepare for eating by applying heat’.
- Verbs of processing – stemming at synset eng-30-01668603-v {work, work on, process} ‘shape, form, or improve a material’.
- Verbs of abstract, cognitive and/or emotional creation – rooted at eng-30-01631534-v {create by mental act, create mentally} ‘create mentally and abstractly rather than with one’s hands’ or at eng-30-01646866-v {provoke, evoke, call forth, kick up} ‘evoke or provoke to appear or occur’.
- Verbs of performing – stemming from the synset eng-30-01714208-v {perform} ‘give a performance (of something)’.
- Verbs of musical performance – stemming at synset eng-30-01726172-v {play} verb.creation ‘perform music on (a musical instrument)’.

#### 4 Semantic frames in FrameNet describing verbs of creation

The study of the systematic semantic relations in each of the resources, as well as the characteristics determining the relationship between their basic units (synsets and semantic frames) and the relations between them, supports the enrichment of synsets in WordNet with conceptual information. By assigning frames to the synsets in WordNet, we aim at defining semantic classes of verbs based

on similar lexical semantics, but more importantly, that evoke the same or similar (related) frames which exhibit similar configurations of frame elements.

FrameNet (Baker et al., 1998) is a system of semantic frames with their frame elements. Frames are schematic descriptions of the conceptual structure of situations through actors, circumstances, and other conceptual roles called frame elements. Koeva (2010) discusses the properties of the resource BulFrameNet – a corpus-based lexicon giving an exhaustive account of the semantic and syntactic combinatory properties of Bulgarian verbs. Koeva and Doychev (2022) present BulFrame – a web-based system for the extensive description of verbs using semantic frames offering a unified theoretical model for the formal presentation of frames and frame elements.

Lexical units in FrameNet, in particular verbs, are grouped in semantic frames based on common semantics, formalised through a common set of participants and circumstances (frame elements) and the relations between them (Fillmore, 1982, 1985, 2003; Fillmore and Baker, 2009; Ruppenhofer et al., 2016) with valence patterns inductively derived from corpus evidence.

There has been considerable work on mapping WordNet and FrameNet automatically (Shi and Mihalcea, 2005; Tonelli and Pighin, 2009; Leseva and Stoyanova, 2020), with synsets in WordNet being assigned semantic frames from FrameNet.

The system of FrameNet semantic frames that described verbs of creation is presented in Table 1.

While usage examples in FrameNet are illustrating the use of English words, the valence patterns are largely applicable to other languages and moreover, to some degree, the most frequent and typical syntactic configurations are also transferable (or can be adapted) cross-linguistically. For example, the frame **Building** has as most frequent configurations the following: NP.Ext<sub>Agent</sub> VERB NP.Obj<sub>Created\_entity</sub> and NP.Ext<sub>Agent</sub> VERB NP.Obj<sub>Created\_entity</sub> PP[from]<sub>Components</sub>. Example 7 illustrates that these are valid for typical usage examples both in Bulgarian and in English, up to language specific lexical selections (e.g., prepositions).

(7) NP.Ext<sub>Agent</sub> VERB NP.Obj<sub>Created\_entity</sub>  
He BUILT a house.

NP.Ext<sub>Agent</sub> VERB NP.Obj<sub>Created\_entity</sub>  
Той ПОСТРОИ къща.

|  |                           |                           |                      |
|--|---------------------------|---------------------------|----------------------|
| Frame                                      | Frame elements            |                           |                      |
|  | Agent                     | Created entity            | Components           |
| General frames                             |                           |                           |                      |
| <b>Creating</b>                            | <b>Creator / Cause</b>    | <b>Created entity</b>     |                      |
| → <b>Intentionally create</b>              | <b>Creator</b>            | <b>Created entity</b>     | <b>(Components)</b>  |
| Frames with physical <b>Created entity</b> |                           |                           |                      |
| → <b>Building</b>                          | <b>Agent</b>              | <b>Created entity</b>     | <b>Components</b>    |
| → <b>Create physical artwork</b>           | <b>Creator</b>            | <b>Representation</b>     |                      |
| → <b>Manufacturing</b>                     | <b>Producer / Factory</b> | <b>Product</b>            | <b>(Resource)</b>    |
| → <b>Cooking creation</b>                  | <b>Cook</b>               | <b>Produced food</b>      | <b>(Ingredients)</b> |
| → <b>Text creation</b>                     | <b>Author</b>             | <b>Text</b>               |                      |
| Frames with abstract <b>Created entity</b> |                           |                           |                      |
| → <b>Achieving first</b>                   | <b>Cognizer</b>           | <b>New idea</b>           |                      |
| → <b>Coming up with</b>                    | <b>Cognizer</b>           | <b>Idea</b>               |                      |
| Frames of creating a <b>Performance</b>    |                           |                           |                      |
| <b>Performers and roles</b>                | <b>Performer</b>          | <b>Role / Performance</b> |                      |

Table 1: FrameNet frames representing the class of creation verbs. None-core frame elements are in brackets.

NP.Ext<sub>Agent</sub> VERB NP.Obj<sub>Created\_entity</sub>  
 He BUILT a garden bridge  
 PP[from]<sub>Components</sub>  
 from recycled pallets.

NP.Ext<sub>Agent</sub> VERB NP.Obj<sub>Created\_entity</sub>  
 Той НАПРАВИ мост в градината  
 PP[from, of]<sub>Components</sub>  
 от рециклирани палети.

## 5 Usage examples

Usage examples illustrating the use of verbs of creation and their syntactic realisation are mostly drawn from the SemCor and BulSemCor – both annotated with WordNet senses.

SemCor (Miller et al., 1993, 1994; Landes et al., 1998) is manually annotated corpus developed by the Princeton WordNet team. Open-class single words and multiword expressions are assigned unique WordNet senses. SemCor contains a total of 226,040 sense annotations.

BulSemCor (Koeva et al., 2006, 2011; Koeva, 2012) is modelled after SemCor aiming to ensure good coverage of general lexis. In addition to open-class words, in BulSemCor closed-class words (preposition, conjunctions, particles) are also

annotated. BulSemCor contains about 100,000 annotated units.

The use of SemCor and BulSemCor is motivated by the fact that verbs are annotated with unique word senses from WordNet, so they are good source of comparable examples in English and Bulgarian.

A total of 220 semantically annotated sentences have been collected in both Bulgarian and English, illustrating the use of creation verbs.

## 6 Towards automatic analysis and identification of aspectual classes

In this section we discuss the first steps towards the automatic analysis of the usage examples of creation verbs with a view to their automatic classification into Velnder’s aspectual classes. There are two aspects of the approach: (a) defining the set of lexical, semantic, morphosyntactic and valence features determining the classification; and (b) defining a set of procedures for the automatic classification based on the comprehensive analysis of the VP phrase the verb is realised in. While the general

features in (a), which we discuss below, are to a large extent language-independent, the concrete procedures in (b) may rely on language-specific analysis and while we aim at flexibility, we also recognise the limitation of the approach with regards to its application for languages other than Bulgarian and English.

The research presented here is just the first step towards designing methods for automatic identification of aspectual classes, and thus, disambiguation of verbs of creation in text. We aim at providing insights into the possible features that determine the behaviour of verbs, that can further be used, if properly formalised, for training in machine learning.

From the analysis of the related works and the examples extracted from corpora, we can summarise the following observations:

- Creation verbs are agentive. Their semantic frames are characterised by concretisations of the **Agent** frame element, e.g. **Creator**, **Author**, **Cook**. Somewhat different are the frame elements **Cognizer** in the frames representing creation of abstract (mental) entity. In the frame **Manufacturing** the frame elements **Producer** or **Factory** can represent a person, organisation, plant, etc. which produce the product.
- The **Created entity** can be of several semantic types – either an artefact, a man-made physical object (in frames such as **Building**, **Manufacturing** or **Cooking creation**), a text or some language representation (in **Text creation**), a representation or depiction of either a physical object, event or idea (in **Performers and roles** and **Create physical artwork**), or an abstract entity (as in the frames **Achieving first** or **Coming up with**).
- In general, creation verbs are always associated with a **Created entity** and are essentially telic. Their analysis requires two moments in time: a moment before the **Created entity** comes into existence, and the moment it becomes real. However, the **Created entity** can allow for a scalable, gradual interpretation (a whole building, but also half-built, or almost built building), which then in turn allows for considering the activity within the event at any one

interval after the beginning and before the completion, without the telic interpretation (no clear end point and while the result / completion is not implied in the sentence's meaning, thus the event is not necessarily bounded) and using the verb as an activity rather than an accomplishment (or, more rarely, an achievement).

- The incremental property (scalability) of the **Theme**<sup>1</sup> can be at the lexical level (a lexical characteristic of the **Created entity**) or be specified at the sentence level by an interval realised as a prepositional or an adverbial phrase.
- A separate temporal characteristic of the situation (expressed as a prepositional phrase, an adverbial or a clause) can also introduce into the semantics atelicity giving the verb the interpretation as an activity rather than an accomplishment.

Examples 8 – 11 illustrate cases of atelic interpretations of verbs of creation evoking different frames. Example 8 shows a case where the verb's meaning allows for the atelic interpretation as an activity. Example 9 illustrates a combination of the lexical properties of the verb *строя* 'build' and the incremental theme 'highway' (incremental as it is built in stages / sections). In Example 10 forming a caretaker government implies a two-point scale (either not formed or formed), the use of a time / duration phrase *вече цяла седмица* 'for a whole week already' suggests that the act of formation is not punctual but durational, thus making it possible to have an atelic interpretation as an activity. The last Example 11 shows a case of iterative aspect where while the invention of a new technology is a bounded and punctual event (an achievement), the constant iteration of the event brings a possible atelic interpretation when considering not the individual occurrence but the complex iterative activity. However, some authors consider the iterative interpretation as a separate aspectual class which requires further analysis.

In FrameNet DNI (definite null instantiations), INI (indefinite null instantiations), CNI

<sup>1</sup>Here, we use the most general meaning of 'Theme' as the most generalised interpretation of the participant in the situation which determines the stages in the event – the frame element that is being created.

(constructional null instantiations) mark cases where a frame element is not overtly expressed but assumed within the structure of the situation described by the frame. However, here in Example 8 we mark the empty core element with ‘X’ as it is expected from the evoked frame but not when the verb is used as an activity (non-telic, non-bounded) rather than as accomplishment (telic, bounded).

(8) Frame: **Text creation**

[Ванцаров]<sub>Author</sub> пише, за да задоволява  
Vaptsarov writes so that fulfill  
творческия си плам. [X]<sub>Created entity</sub>  
his creative flame.

‘Vaptsarov writes in order to fulfil his creative drive.’

(9) Frame: **Building**

[Държавата]<sub>Creator</sub> ви строи  
State-DEF you builds  
[модерен път]<sub>Created entity</sub>,  
modern road  
а вие недоволствате.  
and you complain.

‘The government builds you a modern highway and you are complaining.’

(10) Frame: **Intentionally create**

[Президентът]<sub>Creator</sub> вече цяла седмица  
President-DEF already whole week  
съставя [служебно  
forms caretaker  
правителство]<sub>Created entity</sub>.  
government.

‘For a whole week already the president has been trying to form a caretaker government.’

(11) Frame: **Achieving first**

[Учените]<sub>Creator</sub> постоянно изобретяват  
Researchers-DEF constantly invent  
[нови технологии]<sub>Created entity</sub>.  
new technologies.

‘Researcher constantly keep inventing new technologies.’

In the observed examples, we can summarise several different realisations of the frame element **Created entity** which denote atelic semantic interpretation of the verb (as an activity) rather than a telic one (as an accomplishment). Let us

illustrate these with variations to Example 8 above, where the **Created entity** is not specified. In order to achieve an atelic interpretation, the **Created entity** can be expressed as: (a) an uncountable, collective, abstract or generalised entity (e.g., poetry, Example 12); and (b) plural and indefinite (e.g., poems, Example 13).

(12) Frame: **Text creation**

[Ванцаров]<sub>Author</sub> пише [поезия]<sub>Created entity</sub>,  
Vaptsarov writes poetry  
за да задоволява творческия си плам.  
so that fulfill his creative flame.

‘Vaptsarov writes poetry in order to fulfil his creative drive.’

(13) Frame: **Text creation**

[Ванцаров]<sub>Author</sub> пише  
Vaptsarov writes  
[стихотворения]<sub>Created entity</sub>, за да  
poems so that  
задоволява творческия си плам.  
fulfill his creative flame.

‘Vaptsarov writes poems in order to fulfil his creative drive.’

However, Examples 14, 15 and 16 demonstrate different cases of the expression of telicity. The **Created entity** is definite, however may not be entirely fixed or fully determined, e.g. we may not know the beginning of the set *the last poems* or the end of the set of poems collectively named *the first poems*. While in Examples 14 and 15 additional circumstances are revealed, e.g. the *Place* (prison; which also refers to the time of the situation and can be interpreted as *while he was in prison before his death*) or the *Time* (in his youth), then in Example 16 such time-related details are not present. Moreover, Example 16 poses the question whether the presence of a span defined by a (sub)set of the **Created entity** (e.g., first poems) or a time interval (e.g., in his youth), is sufficient for the telic interpretation, as compared to Example 17 (similar also to Example 8) which does not imply an end point although in fact also refers to a limited time span (Vaptsarov’s life).

(14) Frame: **Text creation**

[Ванцаров]<sub>Author</sub> пише  
Vaptsarov writes  
[последните си стихове]<sub>Created entity</sub>  
his last poems-DEF  
[в затвора]<sub>Place</sub>.  
in prison.

‘Vaptsarov writes his last poems in prison.’

(15) Frame: **Text creation**

[Вапцаров]<sub>Author</sub> пише  
Vaptsarov writes  
[първите си стихове]<sub>Created entity</sub>  
his first poems-DEF  
[в младежките си години]<sub>Time</sub>  
in his youth years.

‘Vaptsarov writes his first poems in his youth.’

(16) Frame: **Text creation**

[Вапцаров]<sub>Author</sub> пише  
Vaptsarov writes  
[първите си стихове]<sub>Created entity</sub>  
his first poems-DEF  
[под влияние на поезията на Яворов]<sub>Explanation</sub>  
influenced by Yavorov’s poetry.

‘Vaptsarov writes his first poems under the influence of Yavorov’s poetry.’

(17) Frame: **Text creation**

[Вапцаров]<sub>Author</sub> пише [X]<sub>Created entity</sub>  
Vaptsarov writes  
[под влияние на поезията на Яворов]<sub>Explanation</sub>  
influenced by Yavorov’s poetry.

‘Vaptsarov writes under the influence of Yavorov’s poetry.’

## 7 Future work

A set of principles can be derived for the consistent semantic description of verbs of creation through FrameNet semantic frames. In particular, the identification of inconsistencies and gaps in the hierarchical structure in each of the two resources, WordNet and FrameNet, can be beneficial – such as frames which are not defined, e.g. we can see that WordNet represents a more diverse and fine-grained subclasses within the class of verbs of creation, while FrameNet does not offer a complete system of frames to cover all nuances; this will also apply to the system of frame elements to adequately reflect the level of specialisation and concretisation of meanings.

Moreover, the scalability, or gradability of the frame element **Created entity**, e.g. the so-called incremental theme, is an essential semantic feature which to a large degree determines the syntactic realisation of the verb, so it is an important semantic feature to implement in frame element

description. It can be introduced as a separate semantic feature on the frame element, and the relevant syntactic realisations should be explored in more details.

Further, other non-core frame elements within the semantic frames evoked by verbs of creation are also key in determining the telic and atelic realisation of the verb in context. As shown by the examples, these can be **Time** (expressed as prepositional phrases, adverbial phrases, clauses, etc.), **Place** (in some cases with temporal references as well). Questions regarding the possible interpretations arise when the sentence is focused on describing additional aspects of the situation, e.g. **Explanation**, **Purpose**, **Manner**, etc., and whether and when these cases can be considered as atelic.

The current study outlines some of the main specific features of the verbs denoting creation and does not aim at completeness and extensive coverage of all semantic classes, their description, or comprehensive representation of the features governing their realisation in a sentence. As is evident from the data, the class of verbs of creation covers a wide range of semantically diverse verbs. A more in-depth analysis is required to uncover the specific features of certain subclasses within the class and be able to fully describe their syntactic realisation and alternations. Moreover, this will be a substantial step towards their automatic identification and processing.

## Acknowledgments

This paper is carried out as part of the scientific programme under the project *An Ontology of Activity Predicates – Linguistic Modelling with a Focus on the Bulgarian Language* funded by the Bulgarian National Science Fund (Grant Agreement No. KP-06-N80/9 of 8.12.2023).

## References

- Lyubomir Andreychin. 1944. *Osnovna balgarska gramatika*. Hemus Publishing House, Sofia.
- Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *COLING-ACL '98: Proceedings of the Conference. Montreal, Canada*, pages 86–90.
- John Beavers. 2008. [Scalar complexity and the structure of events](#). In Johannes Dölling, Tatjana Heyde-Zybatow, and Martin Schäfer, editors, *Event Structures in Linguistic Form and Interpretation*,

- volume 5 of *Language, Context and Cognition*, pages 245–266. De Gruyter, Berlin, Boston.
- Katya Charalozova. 2021. Kategoriyata vid na glagola i metodicheski aspekti na prepodavaneto y v balgarskite uchilishta zad granitsa. *Bulgarian Language*, 68:13–21.
- David R. Dowty. 1979. *The Semantics of Aspectual Classes of Verbs in English*. In *Word Meaning and Montague Grammar: The Semantics of Verbs and Times in Generative Semantics and in Montague's PTQ*, pages 37–132. Springer Netherlands, Dordrecht.
- Christiane Fellbaum, editor. 1998. *WordNet: an electronic lexical database*. MIT Press, Cambridge, MA.
- Charles J. Fillmore. 1982. Frame Semantics. In *Linguistics in the Morning Calm (Ed. by The Linguistic Society of Korea)*, pages 111 – 137. Seoul: Hanshin.
- Charles J. Fillmore. 1985. Frames and the semantics of understanding. *Quaderni Di Semantica*, 6:222–254.
- Charles J. Fillmore. 2003. Valency and semantic roles: the concept of deep structure case. In Vilmos Ágel, Ludwig M. Eichinger, Hans Werner Eroms, Peter Hellwig, Hans Jürgen Heringer, and Henning Lobin, editors, *Dependenz und Valenz: Ein internationales Handbuch der zeitgenössischen Forschung*, pages 457 – 475. Walter de Gruyter.
- Charles J. Fillmore and Collin F. Baker. 2009. A frames approach to semantic analysis. In B. Heine and H. Narrog, editors, *The Oxford handbook of linguistic analysis*, pages 313–340. Oxford: Oxford University Press.
- Graeme Forbes. 2006. Verbs of creation and depiction. In *Attitude Problems: An Essay On Linguistic Intensionality*. Oxford University Press.
- Svetomir Ivanchev. 1971. *Problemi na aspektualnostta v savyanskite ezitsi*. Sofia.
- Svetla Koeva. 2006. BulNet (leksikalno-semantichna mreza na balgarskiya ezik) — chast ot svetovната leksikalno-semantichna mreza. *Balgarski ezik*, pages 19 – 32.
- Svetla Koeva. 2010. *Balgarskiyat Freymnet*. Institute for Bulgarian Language, Sofia.
- Svetla Koeva. 2011. Verb aspect, alternations and quantification. *Cognitive Studies | Études cognitives*, 11:125–134.
- Svetla Koeva. 2012. Bulgarian sense-annotated corpus – between the tradition and novelty. *Cognitive Studies | Études cognitives*, 12:181–198.
- Svetla Koeva. 2021. *The Bulgarian WordNet: Structure and specific features*. *Papers of Bulgarian Academy of Sciences*, 8(1):47–70.
- Svetla Koeva and Emil Doychev. 2022. *Ontology Supported Frame Classification*. In *Proceedings of the 5th International Conference on Computational Linguistics in Bulgaria (CLIB 2022)*, pages 203–213, Sofia, Bulgaria. Department of Computational Linguistics, IBL – BAS.
- Svetla Koeva, Svetlozara Leseva, Borislav Rizov, Ekaterina Tarpomanova, Tsvetana Dimitrova, Hristina Kukova, and Maria Todorova. 2011. Design and development of the Bulgarian sense-annotated corpus. In *Information and communications technologies: present and future in corpus analysis: Proceedings of the III International Congress of Corpus Linguistics*, pages 143 – 150.
- Svetla Koeva, Svetlozara Leseva, and Maria Todorova. 2006. Bulgarian sense tagged corpus. In *Proceedings of LREC 2006*, pages 79 – 86.
- Ivan Kutsarov. 2007. *Teoretichna gramatika na balgarskiya ezik. Morfologiya*. Publishing house of Plovdiv University, Plovdiv.
- Shari Landes, Claudia Leacock, and R. Teng. 1998. Building Semantic Concordances. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*. MIT Press.
- Svetlozara Leseva and Ivelina Stoyanova. 2020. Beyond Lexical and Semantic Resources: Linking WordNet with FrameNet and Enhancing Synsets with Conceptual Frames. In S. Koeva, editor, *Towards a Semantic Network Enriched with a Variety of Semantic Relations*, pages 21 – 48. Sofia: Prof. Marin Drinov Academic Publishing House of the Bulgarian Academy of Sciences.
- Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. Chicago and London: The University of Chicago Press.
- Beth Levin. 2010. Lexicalized scales and verbs of scalar change. <https://web.stanford.edu/~bclewin/cls10change.pdf>.
- George A. Miller. 1995. WordNet: A lexical database for English. *Commun. ACM*, 38(11):39–41.
- George A. Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G. Thomas. 1994. *Using a Semantic Concordance for Sense Identification*. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.
- George A. Miller and Christiane Fellbaum. 2007. WordNet Then and Now. *Language Resources and Evaluation*, 41:209 – 214.

- George A. Miller, Claudia Leacock, Randee Teng, and Ross T. Bunker. 1993. *A Semantic Concordance*. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.
- Ruselina Nitsolova. 2008. *Balgarska gramatika. Morfologiya*. Publishing House of Sofia University, Sofia.
- Steven Pinker. 1989. *Learnability and Cognition: The Acquisition of Argument Structure*. MA: MIT Press.
- Malka Rappaport Hovav. 2008. Lexicalized meaning and the internal temporal structure of events. In Susan Rothstein, editor, *Theoretical and Crosslinguistic Approaches to the Semantics of Aspect*, pages 13 — 42. John Benjamins Publishing Company.
- Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher. R. Johnson, Collin. F. Baker, and Jan Scheffczyk. 2016. *FrameNet II: extended theory and practice*. International Computer Science Institute, Berkeley, California.
- Lei Shi and Rada Mihalcea. 2005. Putting pieces together: combining FrameNet, VerbNet and WordNet for robust semantic parsing. In A. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing. CICLing 2005. Lecture Notes in Computer Science*, volume 3406. Springer, Berlin, Heidelberg.
- Sara Tonelli and Daniele Pighin. 2009. New Features for FrameNet – WordNet Mapping. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL'09)*, Boulder, USA.
- Robert D. Van Valin and Randy J. LaPolla. 1997. *Syntax: Structure, meaning and function*. Cambridge: Cambridge University Press.
- Zeno Vendler. 1957. Verbs and Times. *Philosophical Review*, pages 143 – 160.
- Zeno Vendler. 1967. *Linguistics in Philosophy*. Ithaca, NY: Cornell University Press.

# A ‘Dip-dive’ into Motion: Exploring Lexical Resources towards a Comprehensive Semantic and Syntactic Description

Svetlozara Leseva

Institute for Bulgarian Language, Bulgarian Academy of Sciences

zarka@dcl.bas.bg

## Abstract

In this paper I illustrate the semantic description of verbs provided in three semantic resources (FrameNet, VerbNet and VerbAtlas) in comparative terms with a view to identifying common and distinct components in their representation and obtaining a preliminary idea of the resources’ interoperability. To this end, I provide a comparison of a small sample of motion verbs aligned with semantic frames and classes in the three resources. I also describe the semantic annotation of Bulgarian motion verbs using the framework defined in the Berkeley FrameNet project and its enrichment with information from the other two resources, which has been enabled by the mapping between: (i) their major semantic units – FrameNet frames, VerbNet classes and VerbAtlas frames, and (ii) their ‘building blocks’ – frame elements (FrameNet) and semantic roles (VerbNet, VerbAtlas).

**Keywords:** semantic annotation, semantic frames, verb classes, FrameNet, VerbNet, VerbAtlas, WordNet

## 1 Introduction

In this paper I embark on a comparison of the semantic description of verbs (using WordNet as a lexical inventory) provided in three semantic resources, FrameNet, VerbNet and VerbAtlas, to the end of identifying common and distinct components in their representation and obtaining a preliminary idea of the resources’ interoperability. The study is based on a small sample of motion verbs aligned with semantic frames and classes in the three resources. The analysis is applicable both to theoretical studies of the semantic components of meaning and to the practical task of validating the automatic assignment of semantic information from the resources employed to WordNet synsets. In the second part of the paper I describe the semantic annotation of Bulgarian motion verbs using

the framework defined in the Berkeley FrameNet project and its enrichment with information from the other two resources, which has been enabled by the mapping between: (i) their major semantic units – FrameNet frames, VerbNet classes and VerbAtlas frames, and (ii) their ‘building blocks’ – frame elements (FrameNet) and semantic roles (VerbNet, VerbAtlas). The two tasks have been implemented simultaneously, focusing on motion verbs. Part of the work has also been employed in the development of a FrameNet-like semantic resource for Bulgarian.

## 2 Language resources

Several large-coverage lexical-semantic and syntactic resources as well as a couple of annotated corpora have been employed in this study. In this Section I briefly present them and comment on how they are integrated with each other.

### 2.1 Lexical semantic resources

#### 2.1.1 WordNet

The work described below makes use of the Princeton WordNet and its counterpart for Bulgarian, the Bulgarian WordNet.

The Princeton WordNet<sup>1</sup>, PWN, (Miller, 1995; Fellbaum, 1998) is a large lexical database that represents comprehensively conceptual and lexical knowledge as a richly populated network whose nodes denote cognitive synonyms (synsets) linked to each other by means of a number of conceptual-semantic and lexical relations such as hypernymy, meronymy, antonymy, etc.

The Bulgarian Wordnet (Koeva, 2021)<sup>2</sup> has been developed by translation and adaptation of the PWN synsets and further enriched with language-specific synsets and features. The counterparts in

<sup>1</sup><https://wordnet.princeton.edu/>

<sup>2</sup>The Bulgarian WordNet may be browsed at: <https://dcl.bas.bg/bulnet/>

the two wordnets are related to each other through unique interlingual identifiers, which also provide links to the respective synsets in all other wordnets that support them.

The two wordnets provide the verb inventories (for each language) that have been enriched with semantic and syntactic information through the mapping with the other lexical resources employed.

### 2.1.2 FrameNet

FrameNet (Baker et al., 1998; Baker, 2008) is a lexical resource which couches lexical and conceptual knowledge in a framework originating in the theory of Frame Semantics (Fillmore, 2003; Ruppenhofer et al., 2016). Frames are conceptual structures describing particular types of objects, situations, or events along with their components (frame elements, FEs) (Baker et al., 1998; Ruppenhofer et al., 2016). Each frame in FrameNet is represented by means of a definition that describes schematically the situation and the way in which at least the most essential FEs are involved in it. Each FE is also supplied with a definition that further clarifies its semantics and its interaction with other FEs.

Frame elements have different status reflecting their role in the description of a given frame, i.e. core, peripheral or extra-thematic (Ruppenhofer et al., 2016: 19–20). Core FEs instantiate conceptually obligatory components, while making the frame unique and different from other frames. Peripheral FEs make reference to notions such as Time, Place, Manner, Means, Degree, etc. Extra-thematic FEs characterise an event against a backdrop of another state of affairs. Frames are expressed, or 'evoked', by a set of lexical units (LUs), which are pairings of a word and a meaning whose conceptual semantics is represented by the frame.

FrameNet frames form a network by means of a number of frame-to-frame relations (cf. Ruppenhofer et al. (2016: 81–84) for a detailed account). Some of them, the strictest one being Inheritance, define relations involving the handing down of properties among hierarchically organised frames. In particular, Inheritance is defined as a relation between a more general (parent) frame and a more specific (child) frame where there should be a rather strict correspondence between the semantic characteristics of the parent and the child frame (Petrucci, 2015). An illustration of this relation is represented by the sister frames `Self_motion`, `Fluidic_motion`, among others, which share the main configuration of frame elements defined

for their parent frame `Motion`, but in addition provide a further specification of the `THEME` as an entity moving under its own power and will, i.e. a `SELF_MOVER` (in `Self_motion`), or as a `FLUID` (in `Fluidic_motion`).

Apart from their linguistic significance, Inheritance and other, primarily hierarchical relations play a role in some of the proposed FrameNet-to-WordNet mapping procedures.

### 2.1.3 VerbNet

VerbNet (Kipper-Schuler, 2005; Kipper et al., 2008) is a hierarchical network of English verbs which represents their syntactic and semantic patterns<sup>3</sup>. It is organised into 274 classes extending Levin's classification (Levin, 1993) through refining and adding subclasses so as to provide better syntactic and semantic coherence among members of a class. VerbNet explicitly projects semantic relations onto syntactic structures and encodes information about thematic roles, arguments' selectional restrictions and syntactic frames. While the syntactic dimension of the resource is more specific to English, the semantic roles and the selectional restrictions employed provide well-motivated semantic generalisations cross-linguistically.

### 2.1.4 VerbAtlas

VerbAtlas (Di Fabio et al., 2019) is a lexical-semantic resource representing the semantic description of the verb synsets in BabelNet. BabelNet is a very large, richly populated multilingual semantic network (covering more than 500 languages) which integrates lexicographic and encyclopaedic knowledge from WordNet and Wikipedia (Navigli and Ponzetto, 2010)) as well as from other sources. Each verb synset in VerbAtlas is assigned a frame corresponding to its prototypical predicate-argument structure. Obligatory components are described using semantic roles and the semantic properties governing their compatibility.

The alignment of these resources harnesses WordNet's rich lexical knowledge (glosses, examples, semantic primitives and semantic relations) and the detailed semantic and syntactic description implemented in FrameNet, VerbNet and VerbAtlas. Section 3 below describes the logic underlying the alignment efforts.

<sup>3</sup><https://verbs.colorado.edu/verbnet/>

### 3 Inter-resource mapping

This work employs a WordNet-to-FrameNet mapping described in [Leseva and Stoyanova \(2020\)](#) which makes use of and builds upon several previously implemented ones, in particular: the direct mappings provided within FrameNet ([Baker and Fellbaum, 2009](#)), eXtendedWordFrameNet ([Laparra and Rigau, 2010](#)) and MapNet ([Tonelli and Pighin, 2009](#)), supplemented with additional indirect mapping through VerbNet ([Palmer, 2009](#); [Palmer et al., 2014](#)). The combination of these previous mappings resulted in 4,306 unique synset-to-frame alignments, amounting to 30.5% of all the verb synsets ([Leseva and Stoyanova, 2020](#)).

Different methods for expanding the coverage between the two resources have been proposed. Two of the well-known approaches involve discovering a possibly suitable frame for a verb in WordNet not featured in FrameNet based on its semantic relation(s) to verbs having a correspondence in FrameNet ([Burchardt et al., 2005](#)) or applying graph-based algorithms to identify relevant senses of verbs evoking certain semantic frames ([de La calle et al., 2016](#)).

The procedure used in this paper is based on the exploration of the structural properties of WordNet and FrameNet ([Leseva and Stoyanova, 2020](#)). Verbs in a WordNet synset generally exhibit the same or very similar meaning, which implies that they are associated with the same semantic frame. Moreover, both resources are hierarchically organised based on the notion of inheritance from a more general to a more specific synset or frame. The alignment between the resources reflects the notion of inheritance – a more specific concept should be associated with the frame of its hypernym or with a more specific frame elaborating on (and possibly inheriting from) or otherwise related to the mother frame, although this is not borne out consistently in practice. The method involves: (1) manual assignment of frames to root verb synsets to ensure accurate mapping at the top hierarchical level (and thus alleviate error propagation); (2) automatic assignment of the hypernym’s frame onto hyponyms which were not previously mapped; and (3) implementation of a set of enhancing procedures involving post-editing ([Leseva and Stoyanova, 2020](#)).

As a result, the coverage has been increased to 13,104 synset-to-frame alignments (94% of the verb synsets), of which more than 6,000 have been validated manually.

Finally, VerbAtlas is aligned with WordNet by design, as its verb frames have been defined specifically to provide appropriate semantic description in terms of predicate-argument structures and ensure complete coverage of the WordNet synsets.

#### 3.1 Dataset selection

Two Bulgarian corpora have been employed as a source of examples for semantic role annotation: a semantically annotated corpus, BulSemCor ([Koeva et al., 2006, 2011](#)) designed according to the general methodology of the original SemCor ([Miller et al., 1993](#); [Landes et al., 1998](#)) coupled with criteria for ensuring appropriate coverage of contemporary general lexis. The size of the corpus is close to 100,000 annotated units. As BulSemCor is manually annotated with WordNet senses, it provides disambiguated examples for the studied verbs.

As discussed above, most verb synsets have been assigned a FrameNet frame via one or another of the mapping efforts described in Section 3. While the assignment itself needs manual validation (which has also been underway), once it is verified, a SemCor-like corpus turns into a corpus available for semantic role annotation. This is exactly the approach adopted in this paper: the BulSemCor sentences containing verb synsets from the studied domain whose assigned FrameNet frames have been validated are extracted for annotation.

Where the number of examples is found to be insufficient, the dataset has been supplemented with sentences from the Bulgarian National Corpus, a corpus of 1.2 billion words of running Bulgarian text distributed in 240,000 text samples spanning the second half of the 20th century and the beginning of the 21st century ([Koeva et al., 2012](#)). As the corpus is not sense-disambiguated, the examples excerpted from it have been selected manually so as to correspond to the studied senses.

The resulting collection of example sentences has been annotated according to the FrameNet guidelines. The phrases realising syntactically the core frame elements have been explicitly marked in a similar fashion to the annotation in the Berkeley FrameNet corpus. The selection covers 893 annotated clauses or sentences distributed as follows across five frames: *Motion* – 149; *Self\_motion* – 262; *Arriving* – 182; *Departing* – 178; *Traversing* – 122.

#### 4 Preliminaries: Scratching the surface of the Motion domain

The examples in the following part of the paper feature motion verbs, in particular verbs evoking the FrameNet frame *Self\_motion*, which is a direct descendant of the ‘prototypical’ frame for this semantic domain: *Motion*. This latter frame, as well as the ones that inherit its properties involve the translational motion of different entities. Its semantic description is presented below in more detail, as it serves as a point of departure in defining the more specific frames (e.g. *Self\_motion*), while also showing the resemblances with the way motion is described in VerbNet and VerbAtlas.

##### FrameNet definition: *Motion*

Some entity, the THEME starts out in one place SOURCE and ends up in some other place, the GOAL, having covered some space between the two (PATH). Alternatively, the AREA or DIRECTION in which the THEME moves or the DISTANCE covered may be mentioned.

**Core frame element set:** THEME, SOURCE, GOAL, PATH, AREA, DIRECTION, DISTANCE.

The THEME<sup>4</sup> is the entity that changes location either by moving on its own and/or under its own power or by being moved, carried, etc. by another entity or force. Its semantic specification includes **animate beings** and **physical objects**.

The SOURCE is a location or an entity occupying space that serves as the starting point or landmark where the moving entity is at before it starts changing location. Its semantic specification spans various **locations**, including but not limited to geological and other natural formations, geographical points, celestial bodies or **physical objects**, including man-made structures, such as buildings, constructions, facilities, etc.

The GOAL is the location or another entity serving as the point at which the moving entity ends up as a result of the motion. It has the same semantic specification as the SOURCE.

The PATH is any trajectory of motion confined between the SOURCE and the GOAL, (a part of) the ground over which the moving entity travels or a landmark by which it travels. Its specification encompasses **locations** or **physical objects** that

<sup>4</sup>For better generalisation, the definition of each frame element takes into account its description provided in the prototypical *Motion* frame and in other frames related to it.

may be construed as having extent along which the motion takes place, various **media** (water, air), etc.

The AREA identifies the setting in which the motion does not occur along a single linear trajectory. In consequence, it cannot co-occur with SOURCE, GOAL and PATH. The semantic specification of the AREA refers to **locations** or **physical objects** that may be construed as comprising some spatial expanse within which the motion takes place, various **media**, etc.

The DIRECTION indicates the general spatial orientation of the motion along a line (the PATH) from the deictic centre towards a (possibly implicit) reference point that is neither the GOAL of the posture change nor a landmark along the way of the moving part of the body. Its specification includes compass points (north, east, south, west), body relative directions (left, right, back, front, backward, forward, up, down), coordinates, etc.

The DISTANCE characterises the spatial extent of the motion, the distance travelled by the moving entity. Its specification includes **distance denotations** expressed either in a give system of measurement or in relative terms (farther, closer), etc.

*Self\_motion* is an elaboration of the *Motion* frame and related to it by means of an Inheritance relation. It involves a similar configuration of core FEs with some further restrictions.

##### FrameNet definition: *Self\_motion*

The SELF\_MOVER, a living being, moves under its own direction along a PATH. Alternatively or in addition to PATH, an AREA, DIRECTION, SOURCE, or GOAL for the movement may be mentioned.

The most important distinction between this frame and *Motion* is the capability of the SELF\_MOVER to change location by exercising their own will and power by the coordinated movement of their bodies, which is not a necessity with the *Motion* THEME. By metaphorical extension, SELF\_MOVERS may be self-directed entities such as vehicles. Its semantic specification thus includes **animate beings** and **vehicles**.

The remaining core FEs in this frame are the ones defining the elements and aspects of the route of movement as defined for the parent frame *Motion* and have the same semantic specification as their counterparts there<sup>5</sup>.

<sup>5</sup>DISTANCE is not defined as a core FE, but I will treat it on a par with its equivalent in the *Motion* frame.

## 5 Motivation and prerequisites: What do we gain from resource alignment?

One of the goals of this paper is to illustrate the logic and motivation underlying the comparison undertaken towards the identification of common and distinct components of the semantic description of motion verbs in FrameNet, VerbNet and VerbAtlas. The purpose is to obtain a preliminary idea of the interoperability of the resources and the optimal ways to harness their strengths towards their further enrichment. To this end I compare a small sample of motion verbs and their semantic descriptions across resources. The inventory of verbs employed in the study are WordNet synsets to which FrameNet frames, VerbAtlas frames and possibly VerbNet classes have been assigned as described in the Sections 2.1 and 3.

The purpose of the study is twofold: (i) to illustrate the procedure of validating the automatic assignments of FrameNet frames to WordNet synsets and to suggest how the comparison of the semantic descriptions provided in different resources may give insights into the data; (ii) to describe the semantic annotation undertaken for Bulgarian using the validated FrameNet frame assignments and information from the other two semantic resources (VerbNet and VerbAtlas). The two tasks are implemented simultaneously, focusing on a particular class of verbs, in this case – motion predicates.

Let us illustrate the actual result of the alignment among the resources and what is gained by it. In most simple terms, the mapping between a WordNet synset and a FrameNet frame, on the one hand, or a WordNet synset and a VerbNet class, on the other, is based primarily on an existing correspondence between (i) one or more WordNet synset members (literals) and one or more FrameNet lexical units (LUs); or (ii) between one or more WordNet synset members and one or more verbs belonging to a VerbNet class. This lexical correspondence is the backbone of the inter-resource alignment, which is further expanded by more refined procedures aimed at obtaining greater and more perfect coverage. In the case of the mapping between WordNet and FrameNet, these include calculation of similarity between the definitions in different resources or between the definitions and other fragments of knowledge such as a frame’s name (consider the correspondence between the gloss of the synset {change:2} ‘undergo a change...’ and the frame’s name *Undergo\_change*), the utilisation of

the resources’ structure, especially semantic inheritance, causative/inchoative relations, etc.

Figure 1 shows a straightforward mapping of the synset {walk:1} ‘use one’s feet to advance; advance by steps’ to the FrameNet LU *walk.v*, which evokes the frame *Self\_motion*. Each LU is supplied with a dictionary definition, in this case one borrowed from the Concise Oxford Dictionary, and the comparison to the synset gloss confirms the correspondence between the units in the two resources. The accuracy of the alignment with the VerbNet class run-51.3.2 is borne out by the correspondences between pairs of VerbNet semantic roles and FrameNet frame elements<sup>6</sup>, and/or by VerbNet lemmas sharing class membership with *walk*, which have as their WordNet counterparts either hyponyms – e.g. {limp:1}, {shuffle:1}, {amble:1}, or sisters – e.g. {swim:1}, {run:34} – of {walk:1} or are otherwise close to this synset in the WordNet structure (and hence, have strong similarity).

Unlike FrameNet and VerbNet where the alignment with WordNet is a secondary feature implemented on the basis of the already developed resources – semantic frames and lexical units (FrameNet) and semantic classes and lists of verb lemmas (VerbNet) – VerbAtlas has been developed specifically to describe the verb synsets in BabelNet (of which WordNet is a part), both in terms of their lexical semantics and predicate argument structure. The 13,767 verb synsets in WordNet are covered by 466 frames. WordNet’s {walk:1} is assigned the VerbAtlas frame 0137f GO-FORWARD. The definition provided for this frame describes the relationship among a set of semantic roles, part of which may be straightforwardly aligned to frame elements in the frame *Self\_motion* and to semantic roles in the VerbNet class run-51.3.2<sup>7</sup>.

Given the very different premises of how the frames or classes are assigned, there is not and cannot be a one-to-one correspondence among the resources. In a test set of 206 manually validated synsets mapped to the FrameNet frame *Self\_motion*, 88 are aligned with the VerbNet class run-51.3.2 and 135 are assigned one of 4 VerbAtlas frames as shown in Table 1, while the

<sup>6</sup>The colour coding on Figure 1 shows the matching frame elements and semantic roles that have been associated to each other empirically.

<sup>7</sup>Part of the correspondences, e.g. the one between FrameNet’s PATH / VerbNet’s TRAJECTORY and the VerbAtlas LOCATION are not straightforward and have been established by annotating example sentences on the VerbAtlas web platform: <https://verbatlas.org/>.

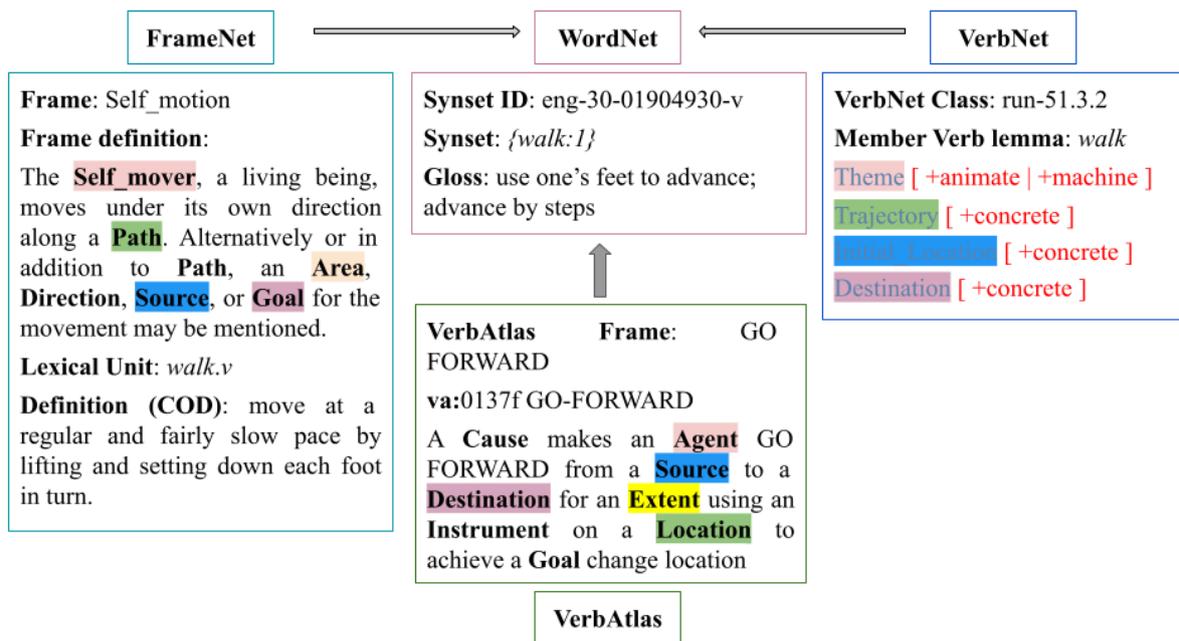


Figure 1: An illustration of the alignment between a WordNet synset and its counterparts in FrameNet, VerbNet and VerbAtlas, with the empirically confirmed frame element-to-semantic role correspondences.

remaining synsets are assigned one of 20 other Verbatlas frames in fewer numbers<sup>8</sup>.

The examples below illustrate the semantic and syntactic annotation across the different resources. For each verb in the examples I have selected a sentence from the manually annotated FrameNet corpus and have run it (possibly simplified) through the Verbatlas SRL platform<sup>9</sup> and the VerbNet SLR parser<sup>10</sup>. The resulting automatic annotation, with corrections where deemed necessary, is given below along with the original FrameNet sentence.

**Example 1: AMBLE**

**FrameNet frame: Self\_motion**

[He]<sub>SELF\_MOVER</sub> AMBLED [down the long winding corridor]<sub>PATH</sub>.

**Verbatlas frame: MOVE-ONESELF**

[He]<sub>THEME</sub> AMBLED [down the long winding corridor]<sub>DESTINATION</sub>.

**VerbNet class: run-51.3.2**

[He]<sub>THEME</sub> AMBLED [down the long winding corridor]<sub>TRAJECTORY</sub>.

**Example 2: LUMBER**

**FrameNet frame: Self\_motion**

[He]<sub>SELF\_MOVER</sub> LUMBERED [down the steep narrow staircase]<sub>PATH</sub>.

**Verbatlas frame: TRAVEL**

[He]<sub>THEME</sub> LUMBERED [down the steep narrow staircase]<sub>DESTINATION</sub>.

**VerbNet class: run-51.3.2**

[He]<sub>THEME</sub> LUMBERED [down the steep narrow staircase]<sub>TRAJECTORY</sub>.

Testing such sets of sentences gives a glimpse into the similarities and differences in the annotation and how these differences may be remedied.

Below I also examine some differences in the definition of FrameNet frames and Verbatlas frames. As shown in Table 1, {lumber:1} is assigned the frame TRAVEL in the latter resource, while FrameNet and VerbNet assign it to the same frame/class as {walk:1}. Looking at the definitions of the two Verbatlas frames, one notes several differences. One major distinction is that TRAVEL’s roleset does not define a SOURCE role, but it does define a GOAL, and possibly a PATH (the role LOCATION may be construed in this way based on the automatic annotation)<sup>11</sup>. The FrameNet Travel frame specifies a similar set of elements of the route of motion as its parent frame Self\_motion: SOURCE, PATH, GOAL, AREA, DIRECTION. Example 3 below presents an annotated FrameNet

<sup>8</sup>Not all synsets are assigned a VerbNet class.

<sup>9</sup><https://verbatlas.org/>

<sup>10</sup><https://verbnetslr.com/>

<sup>11</sup>For the sake of clarity I will refrain from analysing the remaining roles defined in the GO-FORWARD frame but not in the TRAVEL frame.

| Synset     | FrameNet    | VerbNet    | VerbAtlas    | Synsets | VerbAtlas frames  |
|------------|-------------|------------|--------------|---------|---|
| {lumber:1} | Self_motion | run-51.3.2 | TRAVEL       | 45      | A <b>cause</b> makes an <b>agent</b> TRAVEL on a <b>location</b> to a <b>destination</b>  |
| {walk:1}   | Self_motion | run-51.3.2 | GO-FORWARD   | 39      | A cause makes an <b>agent</b> GO FORWARD from a <b>source</b> to a <b>destination</b> for an <b>extent</b> using an <b>instrument</b> on a location to achieve a goal Change location                     |
| {amble:1}  | Self_motion | run-51.3.2 | MOVE-ONESELF | 29      | An <b>agent</b> makes a <b>theme</b> MOVE ONESELF on a <b>location</b> for an <b>extent</b> from a <b>source</b> to a <b>destination</b> to a <b>patient</b> having a <b>result</b> (+ <b>attribute</b> ) |
| {run:34}   | Self_motion | run-51.3.2 | RUN          | 22      | An <b>agent</b> make a <b>theme</b> RUN with a <b>co-theme</b> on a <b>location</b> or from a <b>source</b> to a <b>destination</b> for an <b>extent</b> for a <b>purpose</b> using an <b>instrument</b>  |

Table 1: The most frequent correspondences between the FrameNet frame *Self\_motion*, the VerbNet frame *run-51.3.2* and the corresponding VerbAtlas frames.

sentence for the verb *journey.v*, which is assigned the FrameNet frame *Travel*, the VerbAtlas frame *TRAVEL* and the VerbNet class *run-51.3.2*. In this case, a *SOURCE* (*from Heathrow*) and a *GOAL* (*to Edinburgh*) are expressed, but both are annotated as *DESTINATION* by the VerbAtlas SRL system, possibly because the role *SOURCE* is not part of the description of the VerbAtlas frame. Likewise, *from Heathrow* is parsed by the VerbNet parser not as *INITIAL\_LOCATION*, the role which has empirically been confirmed to correspond to the FrameNet *SOURCE* (as illustrated in Figure 1), but as *DIRECTION*. After inspecting the list of syntactic frames of *run-51.3.2*, one can speculate that this can be attributed to the fact that there is not an explicit syntactic frame associated with this VerbNet class or its subclasses that allows the *INITIAL\_LOCATION* and the *DESTINATION* to be expressed simultaneously (there is, however a syntactic frame describing the simultaneous expression of the *INITIAL\_LOCATION*, the *TRAJECTORY* and the *DESTINATION*). In addition, the role *DIRECTION* is assigned not from VerbNet but from the PropBank roleset (marked by an underline in the examples). The same goes for the *MANNER* role,

roughly corresponding to the FrameNet frame element *MODE\_OF\_TRANSPORTATION*, which does not have a (clear-cut) counterpart in the corresponding VerbAtlas frame and VerbNet class.

**Example 3: JOURNEY**

**FrameNet frame: Travel**

[He]<sub>TRAVELER</sub> JOURNEYED [from Heathrow]<sub>SOURCE</sub> [to Edinburgh]<sub>GOAL</sub> [by overnight coach]<sub>MODE\_OF\_TRANSPORTATION</sub>.

**VerbAtlas frame: TRAVEL**

[He]<sub>AGENT</sub> JOURNEYED [from Heathrow]<sub>DESTINATION</sub> [to Edinburgh]<sub>DESTINATION</sub> [by overnight coach]<sub>ATTRIBUTE</sub>.

**VerbNet class: run-51.3.2-1**

[He]<sub>THEME</sub> JOURNEYED [from Heathrow]<sub>DIRECTION</sub> [to Edinburgh]<sub>DESTINATION</sub> [by overnight coach]<sub>MANNER</sub>.

Similarly, in Example 4, the peripheral frame element *COTHEME* specified in the *Self\_motion* frame does not have a counterpart in the respective rolesets in the other two resources. As a result, it is not labelled or, in the case of the VerbNet parser, is assigned the PropBank role *COMITATIVE*.

**Example 4: WALK****FrameNet frame: Self motion**

*I hope you won't mind if [I]<sub>SELF\_MOVER</sub> WALK [home]<sub>GOAL</sub> [with you]<sub>COTHEME</sub>.*

**VerbAtlas frame: GO FORWARD**

*I hope you won't mind if [I]<sub>AGENT</sub> WALK [home]<sub>DESTINATION</sub> [with you].*

**VerbNet class: run-51.3.2.2-1**

*I hope you won't mind if [I]<sub>THEME</sub> [WALK home]<sub>DESTINATION</sub> [with you]<sub>COMITATIVE</sub>.*

The comparison of the semantic descriptions provided by the three lexical resources points to the conclusion that at least some of the differences in the construals of verb senses may need to be reconsidered in a cross-resource perspective. The correspondences shown in Table 1 and the analysis of the examples demonstrate that some fine-grained distinctions made in a resource may lead to inaccurate predictions. For instance, the role COTHEME described in the VerbAtlas RUN frame (represented by verbs such as *run*) is just as applicable in the GO FORWARD frame (e.g. *walk*, Example 4), although it is not defined there, and the same is true for the RUN role PURPOSE. The *Self motion* frame, which is evoked by both *run* and *walk* in the relevant senses, accounts for the respective fragments of the semantic description via the peripheral elements by the same names (Examples 4, 5).

Thanks to the alignment between VerbNet and PropBank, the VerbNet parser suggests an accurate label for the PURPOSE in Example 5, just as it does for the COTHEME in Example 4. In this case, the VerbNet-to-PropBank mapping provides a notable example of a richer and more robust semantic description obtained through combining the strengths of different resources.

**Example 5: WALK****FrameNet frame: Self motion**

*[He]<sub>SELF\_MOVER</sub> WALKED [round]<sub>PATH</sub> [to stare out of the window]<sub>PURPOSE</sub>.*

**VerbNet class: run-51.3.2.2-1**

*[He]<sub>THEME</sub> WALKED [round]<sub>DESTINATION</sub> [to stare out of the window]<sub>PURPOSE</sub>.*

Cross-resource comparison also matters, as it may point to possible distinctions that have gone unnoticed in one or more resources. Consider the synsets {dance:1} ('move in a graceful or rhythmic way') and {dance:2} ('move in a pattern; usually to musical accompaniment; do or perform a dance'). The former, like {walk:1} is assigned the

VerbAtlas frame GO-FORWARD as it describes a manner of motion along a route, e.g. *She danced into the room*. The latter is described by the frame DANCE defined as 'An agent DANCES with a co-agent...', e.g. *He was dancing tango with his girlfriend on the shining dance floor*. The only lexical unit *dance.v* in FrameNet evokes the *Self motion* frame. As attested by the Concise Oxford Dictionary definition provided for it, however, it encompasses two distinct dictionary senses: 'move rhythmically to music, typically following a set sequence of steps' and '2a: move in a quick and light or lively way', which may be considered as corresponding to the distinction made in VerbAtlas. Through this comparison, one is alerted to the fact that a reconsideration of the FrameNet description of *dance* may be necessary so as to account for the different senses.

In such a way, the detailed exploration of resemblances and differences in the assignments provides insights into the points of divergence in the construal of verb semantics and presents a challenge and an intriguing topic for research in both theoretical and practical terms.

**6 A corpus of annotated examples**

The corpus of annotated Bulgarian examples (Section 3.1) consists of sentences illustrating the use of a number of representative verbs from 5 highly frequent motion frames: *Motion*, *Self motion*, *Traversing*, *Arriving* and *Departing*.

FrameNet provides a number of valence patterns for each lexical unit evoking a frame. Valence patterns represent co-occurring combinations of frame elements attested in the FrameNet corpus, i.e. the actual realisations of a lexical unit in context. Following the Berkeley FrameNet approach, each example in the Bulgarian corpus is supplied with annotation that marks both the frame elements realised in a sentence and their syntactic category and function. In such a way, information is obtained about: (i) the combinatorial properties of the annotated lexical unit, i.e. the particular combinations of both core and non-core frame elements that actually occur in the examples; (ii) the syntax-semantic interface, i.e. the regularities in the syntactic projection of the frame elements into syntactic positions; (iii) the syntactic groups realising the frame elements (along with their morpho-syntactic properties) and the prepositions, complementisers and other lexical items that introduce them.

Example 6 illustrates the manual annotation of a Bulgarian sentence, with explicit marking of the boundaries of the constituents and the type of phrase realising each core frame element and selected non-core ones. The English original of the sentence is provided for reference.

### Example 6

#### FrameNet frame: *Self motion*

*A* [*po vodnite alei*]<sub>PATH:PP</sub> *PLU-VAHA* [*kucheshkata*]<sub>MANNER:ADV</sub> [*oshte igove*]<sub>SELF\_MOVER:NP.EXT</sub>.

*And* [*other Igs*]<sub>SELF\_MOVER:NP.EXT</sub> *were DOG-PADDLING* [*along the water lanes*]<sub>PATH:PP</sub>.

An important feature of the FrameNet methodology and by extension of the annotation adopted in the Bulgarian corpus, is that syntactically non-overt but semantically obligatory frame elements are also marked (as null instantiations, NIs) cf. Ruppenhofer et al. (2016: 28–30), as illustrated below.

### Example 7

#### FrameNet frame: *Self motion*

[ ]<sub>SELF\_MOVER:DNI:NP.EXT</sub> *VARVYAHME* [*veshsto kato ledena zala*]<sub>PATH:PP</sub>.

[*We*]<sub>SELF\_MOVER:DNI:NP.EXT</sub> *were WALKING* [*through a kind of ice hall*]<sub>PATH:PP</sub>.

Table 2 (see the Appendix) illustrates the distribution of the most frequent valence patterns and their syntactic expression across the five studied frames as attested in the corpus of examples. For generalising purposes, only the configurations involving core frame elements have been presented.

Table 3 illustrates the typical valence patterns discovered for the verb *varvya* (one of the Bulgarian counterparts of the synset {walk:1}). Each of them is represented in a separate row along with the syntactic realisations attested in the dataset<sup>12</sup>.

The corpus is being expanded both in terms of the verb coverage and the number of examples. As certain valence patterns are much more frequent than others, in order to obtain an accurate picture of the combinatorial properties of verbs and the syntactic realisation of frame elements other selection procedures are being tested.

A further extension inspired by the alignment with VerbNet and VerbAtlas that has been initi-

<sup>12</sup>To avoid redundancy, nominal (NP.EXT) and pro-drop subjects are represented as aggregates in both tables. Pro-drop subjects are treated as an instance of definite null instantiation (DNI), i.e. as syntactically non-overt elements whose referent is retrievable from the broader context.

ated involves the automatic assignment of the corresponding VerbNet and VerbAtlas semantic roles based on the identification of the counterparts of the FrameNet frame elements in the respective role-sets, as illustrated by the example in Table 4. Based on previous efforts (Leseva and Stoyanova, 2022), this mapping is also assisted by the automatic annotation of test sentences using the VerbAtlas SRL system and the VerbNet parser in order to confirm or reject possible frame element-to-semantic role alignments. This procedure will translate the interoperability of the resources into interoperability of the annotation. The Bulgarian corpus is currently made available for online search<sup>13</sup>.

## 7 Conclusion

The work described in this paper represents a cursory dip-dive into verb semantics, aiming to delineate the specifics of: (i) a cross-lingual 'deep dive' study implemented through the adaptation of semantic resources developed for English to Bulgarian; (ii) a cross-resource analysis implemented through the exploration of the conceptualisation of verb semantics in resources constructed according to different theoretical frameworks. This undertaking has been enabled through two types of alignment: (i) a mapping between verbs (in WordNet) and their semantic descriptions represented as FrameNet semantic frames, VerbAtlas predicate-argument structure frames and VerbNet verb classes and predicate-argument structures; (ii) an alignment between equivalent facets of the semantic description, e.g. between corresponding frame elements and semantic roles or between a semantic frame and a predicate-argument structure. The end goal is the devising of a synergistic approach to the semantic description of verbs and the harnessing of the wealth of linguistic information to semantic annotation, semantic role labelling and similar tasks.

## Acknowledgments

This paper is carried out as part of the scientific programme under the project *An Ontology of Activity Predicates – Linguistic Modelling with a Focus on the Bulgarian Language* funded by the Bulgarian National Science Fund (Grant Agreement No. KP-06-N80/9 of 8.12.2023).

<sup>13</sup>The data are deployed at: <https://dcl.bas.bg/corpus-data-semantic-frames-2024/>.

## References

- Colin F. Baker and Christiane Fellbaum. 2009. WordNet and FrameNet as Complementary Resources for Annotation. In *Proceedings of the Third Linguistic Annotation Workshop (ACL-IJCNLP '09)*, Association for Computational Linguistics, Stroudsburg, PA, USA, pages 125–129.
- Collin F. Baker. 2008. FrameNet, present and future. In *The First International Conference on Global Interoperability for Language Resources*, Hong Kong. City University, City University.
- Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *COLING-ACL '98: Proceedings of the Conference. Montreal, Canada*, pages 86–90.
- Aljoscha Burchardt, Katrin Erk, and Anette Frank. 2005. A WordNet detour to FrameNet. In *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen*, volume 8 of *Computer Studies in Language and Speech*. Lang, Frankfurt, Germany.
- Andrea Di Fabio, Simone Conia, and Roberto Navigli. 2019. Verbatlas: a novel large-scale verbal semantic resource and its application to semantic role labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China, November 3 – 7, 2019*, page 627 – 637. Association for Computational Linguistics.
- Christiane Fellbaum, editor. 1998. *WordNet: an electronic lexical database*. MIT Press, Cambridge, MA.
- Charles Fillmore. 2003. Valency and semantic roles: the concept of deep structure case. In Vilmos Ágel, Ludwig M. Eichinger, Hans Werner Eroms, Peter Hellwig, Hans Jürgen Heringer, and Henning Lobin, editors, *Dependenz und Valenz: Ein internationales Handbuch der zeitgenössischen Forschung*, pages 457 – 475. Walter de Gruyter.
- Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2008. Language Resources and Evaluation. *Commun. ACM*, 42(1):21–40.
- Karin Kipper-Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. PhD Thesis. Computer and Information Science Dept., University of Pennsylvania. Philadelphia, PA.
- Svetla Koeva. 2021. *The Bulgarian WordNet: Structure and specific features*. *Papers of Bulgarian Academy of Sciences*, 8(1):47–70.
- Svetla Koeva, Svetlozara Leseva, Borislav Rizov, Ekaterina Tarpomanova, Tsvetana Dimitrova, Hristina Kukova, and Maria Todorova. 2011. Design and development of the Bulgarian sense-annotated corpus. In *Information and communications technologies: present and future in corpus analysis: Proceedings of the III International Congress of Corpus Linguistics*, pages 143 – 150.
- Svetla Koeva, Svetlozara Leseva, and Maria Todorova. 2006. Bulgarian sense tagged corpus. In *Proceedings of LREC 2006*, pages 79 – 86.
- 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.
- Maddalen Lopez de Lacalle, Egoitz Laparra, Itziar Aldabe, and German Rigau. 2016. *A Multilingual Predicate Matrix*. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2662–2668, Portorož, Slovenia. European Language Resources Association (ELRA).
- Shari Landes, Claudia Leacock, and R. Teng. 1998. Building Semantic Concordances. In *WordNet: An Electronic Lexical Database*.
- Egoitz Laparra and German Rigau. 2010. eXtended WordFrameNet. In *Proceedings of LREC 2010*, pages 1214–1219.
- Svetlozara Leseva and Ivelina Stoyanova. 2020. Beyond lexical and semantic resources: linking WordNet with FrameNet and enhancing synsets with conceptual frames. In *Towards a Semantic Network Enriched with a Variety of Semantic Relations*. Prof. Marin Drinov Academic Publishing House of the Bulgarian Academy of Sciences.
- Svetlozara Leseva and Ivelina Stoyanova. 2022. Linked Resources towards Enhancing the Conceptual Description of General Lexis Verbs Using Syntactic Information. In *Proceedings of the Fifth International Conference Computational Linguistics in Bulgaria (CLIB 2022)*, pages 214 – 223. Institute for Bulgarian Language.
- Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. Chicago and London: The University of Chicago Press.
- George A. Miller. 1995. WordNet: A lexical database for English. *Commun. ACM*, 38(11):39–41.
- George A. Miller, Claudia Leacock, Randee Teng, and Ross T. Bunker. 1993. *A Semantic Concordance*. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.
- Roberto Navigli and Simone Paolo Ponzetto. 2010. *BabelNet: Building a Very Large Multilingual Semantic Network*. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden. Association for Computational Linguistics.
- Martha Palmer. 2009. Semlink: linking PropBank, VerbNet and FrameNet. In *Proceedings of the Generative Lexicon Conference*. 9–15.

Martha Palmer, Claire Bonial, and Diana McCarthy. 2014. SemLink+: FrameNet, VerbNet and event ontologies. In *Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore (1929–2014)*, Baltimore, Maryland USA, June 27, 2014, pages 13–17. Association for Computational Linguistics.

Miriam R. Petruck. 2015. The Components of FrameNet. Tutorial at NAACL 2015. [Http://naacl.org/naacl-hlt-2015/tutorial-framenet-data/FNComponentsMRLP.pdf](http://naacl.org/naacl-hlt-2015/tutorial-framenet-data/FNComponentsMRLP.pdf).

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher. R. Johnson, Collin. F. Baker, and Jan Scheffczyk. 2016. *FrameNet II: extended theory and practice*. International Computer Science Institute, Berkeley, California.

Sara Tonelli and Daniele Pighin. 2009. New features for framenet – wordnet mapping. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL'09)*, Boulder, USA.

Appendix A

| Valence pattern  | Frame       | Occurrences |
|--|-------------|-------------|
| [NP.EXT] <sub>THEME</sub> [PP] <sub>PATH</sub>                           | Motion      | 55          |
|  | Self_motion | 86          |
|  | Traversing  | 23          |
| [NP.EXT] <sub>SELF_MOVER</sub> [ ] <sub>PATH.DNI</sub>                   | Self_motion | 45          |
| [NP.EXT] <sub>THEME</sub> [PP] <sub>GOAL</sub>                           | Motion      | 25          |
|  | Self_motion | 35          |
|  | Arriving    | 75          |
|  | Departing   | 35          |
| [NP.EXT] <sub>THEME</sub> [PP] <sub>AREA</sub>                           | Motion      | 15          |
|  | Self_motion | 37          |
| [NP.EXT] <sub>THEME</sub> [ ] <sub>SOURCE.DNI</sub>                      | Departing   | 53          |
| [NP.EXT] <sub>THEME</sub> [NP.OBJ] <sub>SOURCE</sub>                     | Departing   | 38          |
| [NP.EXT] <sub>THEME</sub> [PP] <sub>GOAL</sub> [ ] <sub>SOURCE.DNI</sub> | Departing   | 35          |
| [NP.EXT] <sub>THEME</sub> [PP] <sub>SOURCE</sub>                         | Departing   | 15          |
| [NP.EXT] <sub>THEME</sub> [NP.OBJ] <sub>PATH</sub>                       | Traversing  | 79          |
| [NP.EXT] <sub>THEME</sub> [ ] <sub>GOAL.DNI</sub>                        | Arriving    | 36          |

Table 2: Distribution of the most frequent valence patterns across the studied motion frames.

| SELF_MOVER | PATH    | GOAL  | SOURCE | DIRECTION | MANNER | TIME | COTHEME | No. |
|------------|---------|-------|--------|-----------|--------|------|---------|-----|
| NP.EXT     |         | PP    |        |           |        |      |         | 3   |
| NP.EXT     |         | ADV P |        |           |        |      |         | 3   |
| NP.EXT     | PP      |       |        |           |        |      |         | 11  |
| NP.EXT     | PP      |       |        |           | ADV P  |      |         | 3   |
| NP.EXT     | PP + PP |       |        |           |        |      |         | 1   |
| NP.EXT     | PP      |       |        |           |        |      | PP      | 1   |
| NP.EXT     | PP      |       |        | PP        |        |      |         | 1   |
| NP.EXT     |         |       |        | ADV P     |        |      |         | 1   |
| NP.EXT     | PP      |       |        | ADV P     |        |      |         | 1   |
| NP.EXT     | PP      |       |        |           | PP     |      |         | 1   |
| NP.EXT     | PP      |       |        |           | ADV P  |      |         | 1   |
| NP.EXT     |         |       |        |           | S      | NP   |         | 1   |
| NP.EXT     |         |       |        |           |        | NP   |         | 1   |

Table 3: Distribution of the valence patterns and their syntactic realisations for the verb *varvya* in the synset {*varvya*:9; *hodya*:6} ({*walk*:1}). The data include 40 annotated sentences.

|                                     |  |
|-------------------------------------|--|
| <b>Annotation</b>                   | <b>An annotated Bulgarian sentence and its English translation annotated in the styles of FrameNet, VerbNet and VerbAtlas</b>  |
| Corpus of BG examples               | [ <i>Togava</i> ] <sub>TIME:ADVP</sub> [ <i>grazhdanite</i> ] <sub>SELF_MOVER:NP.EXT</sub> <i>BYAGAHA</i> [ <i>s chanti i kufarcheta</i> ] <sub>DEPICTIVE:PP</sub> [ <i>po stramnite i tamni ulichki</i> ] <sub>PATH:PP</sub> [ <i>kam dvata tunela pod grada</i> ] <sub>GOAL:PP</sub> .                   |
| <b>FN frame:</b><br>Self_motion     | [ <i>Then</i> ] <sub>TIME:ADVP</sub> [ <i>the inhabitants</i> ] <sub>SELF_MOVER:NP.EXT</sub> <i>RAN</i> [ <i>purses and briefcases in hand</i> ] <sub>DEPICTIVE:SABS</sub> [ <i>down the steep dark alleys</i> ] <sub>PATH:PP</sub> [ <i>to the two tunnels under the city</i> ] <sub>GOAL:PP</sub> .      |
| <b>VerbNet class:</b><br>run-51.3.2 | [ <i>Then</i> ] <sub>TIME:ADVP</sub> [ <i>the inhabitants</i> ] <sub>THEME:NP.SUBJ</sub> <i>RAN</i> [ <i>purses and briefcases in hand</i> ] <sub>ADVERBIAL</sub> [ <i>down the steep dark alleys</i> ] <sub>TRAJECTORY:PP</sub> [ <i>to the two tunnels under the city</i> ] <sub>DIRECTION:PP</sub> .    |
| <b>VerbAtlas frame:</b> RUN         | [ <i>Then</i> ] <sub>TIME:ADVP</sub> [ <i>the inhabitants</i> ] <sub>THEME:NP.SUBJ</sub> <i>RAN</i> [ <i>purses and briefcases in hand</i> ] <sub>ATTRIBUTE</sub> [ <i>down the steep dark alleys</i> ] <sub>DESTINATION:PP</sub> [ <i>to the two tunnels under the city</i> ] <sub>DESTINATION:PP</sub> . |

Table 4: An example of a manually annotated sentence according to the FrameNet methodology, its English translation (for reference) and the automatic assignment of the roles from the relevant VerbNet class (if available) and VerbAtlas frame, using the VerbNet parser and the VerbAtlas SRL platform. The underlined semantic roles are assigned from PropBank.

# Multilingual Corpus of Illustrative Examples on Activity Predicates

Ivelina Stoyanova, Hristina Kukova, Maria Todorova, Tsvetana Dimitrova

Department of Computational Linguistics  
Institute for Bulgarian Language  
Bulgarian Academy of Sciences  
{iva, hristina, maria, cvetana}@dcl.bas.bg

## Abstract

The paper presents the ongoing process of compilation of a multilingual corpus of illustrative examples to supplement our work on the syntactic and semantic analysis of predicates representing activities in Bulgarian and other languages. The corpus aims to include over 1,000 illustrative examples on verbs from six semantic classes of predicates (verbs of motion, contact, consumption, creation, competition and bodily functions) which provide a basis for observations on the specificity of their realisation. The corpus of illustrative examples will be used for contrastive studies and further elaboration on the scope and behaviour of activity verbs in general, as well as its semantic subclasses.

**Keywords:** activity predicates, semantic frames, multilingual corpus

## 1 Introduction

The paper discusses the ongoing work on the compilation, analysis and annotation of a corpus of examples in several languages along with the challenges it poses. The task-specific dataset comprises examples of activity verbs and is tailored to serve as language data for contrastive conceptual analysis of verbs. Initially, verbs belonging to six semantic classes were extracted (verbs of motion, contact, consumption, creation, competition, and bodily functions) from the Bulgarian WordNet (BulNet), which were further manually filtered on account of their appurtenance to the Vendlerian aspectual class of activity verbs. Examples from monolingual (semantically annotated) and parallel corpora were then excerpted illustrating the use of the verbs in context. The verbs under observation are assigned FrameNet frames mapped to the relevant WordNet synsets.

The main objective of our work is to construct a demonstration corpus of annotated examples illustrating the usage of activity verbs. Moreover, we aim at: (a) linking various lexical, semantic and conceptual resources in order to provide comprehensive description of verbs; (b) partial (or full, in case of parallel examples) cross-language alignment in terms of verb translational equivalents based on WordNet, and in terms of participants in their semantic frames; (c) flexibility of corpus design to allow data from other languages to be added; and (d) flexibility of annotation to facilitate the expansion of the scope and variability of the examples in the corpus.

At present, the corpus includes illustrative examples for the use of activity verbs in two languages – Bulgarian and English. Further work on the corpus will include at least two more languages – Russian (a Slavic language) and Italian (a Romance language). This will provide linguistic material for observations on both closely related and more distant languages.

The methodology for constructing the corpus, the description of verbs and the annotation of examples is largely language-independent and can be applied to extract and compile datasets of various languages.

The remainder of the paper is structured as follows. Section 2 contains a theoretical overview of previous work. Section 3 gives a description of the resources used in the study. Then we discuss the linguistic data and the process of its selection (in Section 4) along with some illustrative examples (in Section 5). Section 6 involves a description of how multilingual data is represented as well as some disputable points, and Section 7 proposes an annotation schema with its main principles and steps. Section 8 summarises the main findings of the study with a view to the future work.

## 2 Relevant works

Activity verbs are members of Vendler's aspectual classification (Vendler, 1957, 1967), where verbs are divided into activities, states, achievements, and accomplishments. This classification has been subsequently elaborated by Dowty (1979) and Van Valin and LaPolla (1997). They propose four key semantic features which define the scope of the aspectual classes: [ $\pm$  static], [ $\pm$  dynamic], [ $\pm$  telic], and [ $\pm$  punctual]. Punctuality distinguishes achievements (which are punctual) from accomplishments (which are non-punctual).

The classes of non-stative verbs are distinguished by: dynamism; presence or absence of an internal limit – “proceed towards a terminus” (Vendler, 1957) regarding accomplishments, and time limitation – “achievements occur at a single moment” (Vendler, 1957); terminativeness (Maslov, 1982); boundedness (Paducheva, 2009). The temporal limitation is often equated with the presence of an intrinsic end-point or an instantaneous climax. Some verbs can be classified in more than one aspectual class depending on their use, for example some verbs can be both activities and accomplishments (e.g. *walk* / *walk to the store*), and states and achievements (e.g. the ambiguity of many mental state verbs such as *recognize*, *understand*, *know*). Thus, the aspectual classes are considered at the VP level rather than at the lexical level, which means that the aspectual properties are expressed in a complex lexical, morpho-syntactic and valence-related way by the verb and its arguments (Rappaport Hovav, 2008: 16–20).

The class of active predicates are broadly described in the grammars of the Bulgarian language (Maslov, 1982; Gramatika, 1983; Barkalova, 1997). Different types of verbs of states, processes, actions, activities and events have been the subject of various studies on Bulgarian. Koeva et al. (2022) offer a detailed ontological description of predicates and predicatives of state in Bulgarian, Russian and English comparatively, based on large lexical resources and corpus data. Kostova (2010) describes lexical-semantic groups of basic verbs of motion; Vateva (2005) examines lexical-semantic relations between verbs of movement in Bulgarian and their metaphorical use in different functional styles; Dekova (2006) explores particular groups of motion verbs in comparison with English verbs.

A semantic description of verbs of change is proposed in connection with their representation in the semantic frames of FrameNet (Leseva and Stoyanova, 2021). A classification and semantic description of verbs of contact has been proposed by Todorova (2023).

Different features of Bulgarian verbs and classifications have been proposed. Vlahova (2004) describes systemic dependencies between predicate types, semantic roles, grammatical categories and grammatical transformations of verbs. Koeva (2006b) offers a typology of Bulgarian verbs based on their (in)ability to form diatheses and alternations.

As far as we know, there is neither a complex research on the typology, semantic and syntactic properties of activities in Bulgarian in comparison with other languages, nor a specific database with particular selection of activity predicates.

In our work we step on the analysis of Kolokovska (2005) which is close to the interpretation of Lyons (1977) and Bulygina (1982) who consider activities as a conceptual-semantic category, characterised by the following semantic features: localisation in time; change of at least one participant in the situation and activity. We also rely on the aspectual understanding of the imperfective verbs in Bulgarian, representing the activity in its progress (Stankov, 1980: 6, 43).

## 3 Resources

### 3.1 Lexical-semantic resources

The selection of lexical entries and their lexical and semantic description are extracted from two main resources – WordNet and FrameNet.

WordNet (Miller, 1995; Fellbaum, 1998) represents the lexicon in the form of a network of synonym sets (synsets) interconnected by semantic, lexical and other relations. The main structural relation is hypernymy (and its opposite relation – hyponymy), by which the vocabulary of a given semantic field is organised into a tree, the beginning (root) being the most general or abstract concept of the corresponding field.

The semantic description of verb predicates in WordNet also includes their classification into general semantic classes based on assigned semantic primitives (Miller and Fellbaum, 2007), e.g. verbs of motion, verbs of emotion, verbs of communication, verbs of change, etc. We focus on several semantic classes of verbs in WordNet that

are representative of the class of activities: verbs of contact, verbs of motion, verbs of consumption, verbs of competition, verbs of body, and verbs of creation.

WordNet ensures vast lexical coverage of the English lexicon structured and enriched with lexical and semantic information in the form of synset glosses, usage examples, notes on the usage or grammatical specificities, and a rich network of semantic relations. The Bulgarian counterpart – the Bulgarian WordNet (BulNet) (Koeva, 2006a, 2021), is linked to the Princeton WordNet through interlingual index, and serves as the main resource for the extraction of Bulgarian verbs representing activities.

However, WordNet encodes no explicit semantic information about the participants in the situations described by the predicates and only limited information about their syntactic behaviour. Moreover, WordNet does not consistently reflect the different lexical meanings of verbs that can be referred to more than one aspectual class.

FrameNet (Baker et al., 1998) is a system of semantic frames which are schematic descriptions of the conceptual structure of situations through actors, circumstances, and other conceptual roles presented as frame elements. The frames are organised using a number of relations – hierarchical (Inheritance, Use, Subframe, etc.) and other types (for example, Causation).

Lexical units in FrameNet, in particular verbs, are grouped in semantic frames based on common semantics, formalised through a common set of participants and circumstances (frame elements) and the relations between them (Fillmore, 1982, 1985, 2003; Fillmore and Baker, 2009; Ruppenhofer et al., 2016). A set of valence patterns derived from corpus evidence characterises each lexical unit. Valence patterns show the configurations of frame elements in the realisation of the verb with their respective syntactic function.

There have been efforts to construct a FrameNet-based resource for Bulgarian – BulFrameNet – a corpus-based lexicon giving an exhaustive account of the semantic and syntactic combinatorial properties of Bulgarian verbs (Koeva, 2010), while Koeva and Doychev (2022) present a web-based system for the extensive description of verbs using semantic frames

offering a unified theoretical model for the formal presentation of frames and frame elements.

In our work, FrameNet description of verbs is used to design an annotation schema for the corpus which will ensure unified representation in terms of semantic frames, frame elements and syntactic function of the elements in the realisation of the frame. The description of verb semantics using FrameNet can contribute to the unified description of lexical-aspectual classes of verbs and the analysis of their specific syntactic realisation, in particular to allow comparisons between the realisation of activities with respect to other aspectual classes.

### 3.2 Corpora

Examples are extracted from a number of different corpora using a range of techniques: (i) monolingual semantically annotated corpora where the verbs are disambiguated and assigned a WordNet sense, thus making it possible to extract sentences illustrating particular verb meaning; (ii) parallel corpora which allow the extraction of parallel examples illustrating the usage of verbs in two or more languages; however, they require additional annotation, filtering and disambiguation; (iii) additional monolingual resources to collect examples for language specific usage or less frequent linguistic phenomena.

Monolingual semantically annotated corpora such as SemCor (Miller et al., 1993) and BulSemCor (Koeva et al., 2011) are used in order to extract illustrative examples for English and Bulgarian, respectively. Words are annotated with WordNet senses which enables the extraction of examples.

Parallel corpora are a useful source of examples which illustrate the use of verbs with equivalent or similar meaning in more than one language in aligned sentences.

The Bulgarian-English Sentence- and Clause-Aligned Corpus (BulEnAC)<sup>1</sup> (Koeva et al., 2012a) is a parallel corpus of aligned Bulgarian and English sentences and clauses with annotation of the syntactic relation between clauses. The corpus contains 366,865 tokens (176,397 tokens in Bulgarian and 190,468 tokens in English). The texts in BulEnAC cover five categories: administrative texts, fiction, journalism, science, informal texts. Texts for each language have been

<sup>1</sup>[https://dcl.bas.bg/en/resources\\_list/bulenac/](https://dcl.bas.bg/en/resources_list/bulenac/)

annotated with sentence and clause boundaries and then semi-automatically aligned at the clause level (automatic alignment followed by manual validation). Having aligned clauses is useful for extracting parallel examples as, on one hand, the clause provides the minimal scope for realising the verb's arguments, and on the other hand, aligned clauses ensure easier matching of translational equivalents, both for the verb and for its arguments.

The Bulgarian National Corpus is the largest corpus for Bulgarian: it consists of a monolingual (Bulgarian) part and 47 parallel corpora and amounts to 5.4 billion tokens. The Bulgarian part includes about 1.2 billion tokens of running text distributed in 240,000 text samples<sup>2</sup> (Koeva et al., 2012b). The Bulgarian-English parallel corpus within the Bulgarian National Corpus covers over 100,000 parallel texts and 260 mln. tokens. There is also a large parallel corpus with Italian, and a small Bulgarian-Russian parallel corpus of fiction.

Additional sources of illustrative examples can also be used. For English, the corpus of examples in FrameNet<sup>3</sup> (Burchardt and Pennacchiotti, 2008) annotated with explicit and implicit frame elements supplies extensive empirical evidence about the syntactic realisations of semantic frames that is valuable not only for linguistic generalisations about the target language (English) but as a point of departure for cross-linguistic observations. Semantically annotated corpora exist for many languages, including for our target set of languages (Russian and Italian)<sup>4</sup>.

In addition, examples from other parallel corpora can be added. All target languages are covered in the parallel subcorpora within the Russian National Corpus (Savchuk et al., 2024) from which additional examples can be extracted and manually validated.

#### 4 Selection of verbs

As a first step, we extracted from the Bulgarian WordNet (which has been developing as a parallel resource to the Princeton WordNet (Koeva, 2021)) all (single) verbs from the synsets belonging to the semantic classes verb.motion, verb.creation, verb.contact, verb.competition, verb.consumption, and verb.body (Miller, 1995; Fellbaum et al.,

2009). We analysed the verbs and selected only those that represent activities. The rationale behind using WordNet, and the Bulgarian WordNet in particular, as a source for verbs selection is that: (a) we are able to extract sets of verbs from the observed semantic classes; (b) we have access to the lexical and semantic description of the verbs in the Bulgarian WordNet; (c) we can also extract translational equivalents in other languages (linked to Princeton WordNet) and further collect illustrative examples for their use in corpora.

Along with the semantic information, the Bulgarian WordNet comprises also some lexicogrammatical information encoding the Bulgarian verb aspect. Using this, we filtered the extracted verbs and selected only the imperfective ones. Then we selected manually the verbal set so far leaving aside the prefixed verbs, as well as those that refer to states, accomplishments, and achievements (Vendler, 1967). The verbs that were selected refer to continuing activities, may have human or human-like volitional subjects, do not have a terminal point and a (tangible) result. More or less we step in the process of selection on the preliminary criteria offered in Koeva and Ivanova (2024).

For example, the verb *write* 'produce a literary work', which is classified as verb.creation, would refer to an activity (ongoing in the past) in Example 1a, but also to an accomplishment (with a tangible result) in Example 1b. In Bulgarian, two different verbs would be used – a non-prefixed imperfective one *пише* (Example 1c) and the prefixed perfective verbs *напиша* (Example 1d).

- (1) **a.** En: He WROTE novels, short stories, lyrics, essays, plays for almost 70 years.
- b.** En: He did the research, and he WROTE the book.
- c.** Bg: Той ПИШЕ романи, кратки разкази, поезия, есета и пиеси в продължение на почти 70 години.
- d.** Bg: Той направи проучването и НАПИСА книгата. 'He did the research and WROTE the book.'

Idiomatic and phrasal verbs (as in *ходя на лов* 'hunt; run; hunt down; track down' with a definition 'pursue for food or sport (as of wild animals)'), light verbs (as in *търси отговор* 'looking for an answer') and English verbs with no lexicalisation in Bulgarian (e.g., *крада база* 'steal'

<sup>2</sup><https://search.del.bas.bg/>

<sup>3</sup><http://framenet.icsi.berkeley.edu/>

<sup>4</sup><http://globalwordnet.org/resources/wordnet-annotated-corpora/>

with a definition ‘steal a base’ in the domain of basketball) were also discarded (*ловувам* ‘hunt’ which is the synonym of *ходя на лов*, however, would be left in the verb dataset).

| semantic class   | Synsets | Literals |
|------------------|---------|----------|
| verb.contact     | 265     | 334      |
| verb.motion      | 199     | 296      |
| verb.creation    | 89      | 109      |
| verb.consumption | 49      | 69       |
| verb.competition | 70      | 100      |
| verb.body        | 24      | 33       |

Table 1: Selected verbs according to their semantic class

The resulting set consists of 941 verbs, most of which are members of synsets classified as verb.contact (334 (single) verbs, e.g. *чистя* ‘clean, make clean’ with a definition ‘make clean by removing dirt, filth, or unwanted substances from’) and verb.motion (296 verbs, e.g. *плувам* ‘swim’ with a definition ‘travel through water’); among the selected verbs there are 100 verbs classified as verb.competition (e.g. *воювам* ‘war’ defined as ‘make or wage war’); 109 verbs are classified as verb.creation (e.g. *свирия* ‘play’ with a definition ‘play on an instrument’); relatively smaller are the class verb.consumption – 69 verbs (e.g. *ям* ‘eat’ defined as ‘eat a meal; take a meal’), and verb.body – 33 verbs (e.g. *душа* ‘choke; strangle’ defined as ‘constrict (someone’s) throat and keep from breathing’).

## 5 Selection of illustrative examples

The verbs are used to automatically extract examples from: (i) semantically annotated corpora and (ii) parallel corpora with no semantic annotation.

From semantically annotated corpora we have extracted automatically examples using the predetermined verb set. The verbs are disambiguated, and assigned a particular WordNet sense (see section 4). Example 2 shows two sentences from semantically annotated corpora showing two words belonging to the same WordNet sense.

From parallel corpora with no semantic annotation, we extract Bulgarian examples containing verbs from the predetermined set with all their possible WordNet senses, which are then aligned to a sentence in the other language

(English) in which the verb is also identified. Example 3 shows two aligned parallel sentences; all possible senses of the identified Bulgarian verb are given; for each of the possible senses it is checked whether the English verb can be found among the literals of that synset, and in this way the most likely candidate sense(s) are identified.

- (2) Examples from SemCor and BulSemCor for verbs of the same synset (sentences are not parallel, only verbs are linked to WordNet synset: eng-30-01698271-v / пиша ‘write’, verb.creation, ‘produce a literary work’).

a. En: Mr. Sansom WRITES of foreign parts with a dedication to decoration worthy of a pastry chef creating a wedding cake. (SemCor)

b. Bg: Баща ѝ ПИШЕШЕ криминални романи, които се славеха с особен успех и тя ги четеше с удоволствие. (BulSemCor)

- (3) Examples from BulEnAC for verbs in aligned parallel sentences.

a. Bg: Колко пъти съм СЛАГАЛА камъни да варя, за да не разберат съседките, че дни наред нямаме какво да сложим в тенджерата!

En: How many times have I PUT stones to boil, so that the neighbours won’t know that days after days we have nothing to put in the pot!

Possible WordNet senses for *put*:

# eng-30-01493380-v / verb.contact / ‘place temporarily’

# eng-30-01494310-v / verb.contact / ‘put into a certain place or abstract location’

# eng-30-00050652-v / verb.body / ‘put clothing on one’s body’

# eng-30-01500372-v / verb.contact / ‘cause to sit or seat or be in a settled position or place’

# eng-30-01465921-v / verb.contact / ‘arrange or fix in the desired order’

b. Several times I’ve had to PUT stones on to boil so the neighbors wouldn’t know that we often go for many days without putting on the pot.

Then the possible verb senses / WordNet synsets were aligned to their BulNet correspondences:

- # eng-30-01493380-v / Not aligned
- # eng-30-01494310-v / **Aligned**
- # eng-30-00050652-v / Not aligned
- # eng-30-01500372-v / Not aligned
- # eng-30-01465921-v / Not aligned

Next, we selected manually the appropriate illustrative examples for Bulgarian and for English. We disregard examples with verbs used in figurative (metaphorical) contexts (as in *The world’s best golfer, shooting below par, came to the last hole of the opening round...*), in multiword expressions (as part of light verb constructions or idioms, as in *eat humble pie*), verbs in passive constructions, or in other uses which do not refer to ongoing activities in the present or in the past (thus, certain verb forms can be excluded, such as past simple, past perfect, etc.).

A set of corpus examples was selected with the distribution shown in Table 2. So far there are 245 examples selected for Bulgarian and 257 for English. The aim is to achieve a dataset of over 1000 examples for each language.

| Semantic class   | Bulgarian |          | English |          |
|------------------|-----------|----------|---------|----------|
|                  | All       | Selected | All     | Selected |
| verb.contact     | 116       | 50       | 3485    | 54       |
| verb.motion      | 153       | 96       | 2867    | 81       |
| verb.creation    | 52        | 35       | 1060    | 40       |
| verb.consumption | 50        | 31       | 1005    | 37       |
| verb.competition | 24        | 14       | 576     | 20       |
| verb.body        | 38        | 19       | 604     | 25       |

Table 2: Selected examples from BulSemCor (Bulgarian) and SemCor (English)

## 6 Challenges in the process of selection

The main challenge was, first, to differentiate between the different senses of the verb, which in some cases are very close (Example 4), and second, to distinguish activities and accomplishments and achievements with verbs that select direct configuration of arguments or arguments with particular semantic characteristics (Example 5). Some verbs may be categorised differently with respect to the intentional elements – such as *плача* ‘to cry’, which can refer to a non-intentional, as well as to an intentional act.

- (4) Examples from BulEnAC for verbs in aligned parallel sentences with several possible closely related WordNet senses.

- a. Bg: През целия си живот СИ ТЪР-СИЛ това съкровище, само за да получиш уважението на историците.
- b. En: You’ve spent your entire life SEARCHING for this treasure, only to have the respect of the historical community.

Possible WordNet senses:

- # eng-30-01317533-v / търся / verb.contact / ‘go in search of or hunt for’
- # eng-30-01315613-v / търся / verb.contact / ‘try to locate or discover, or try to establish the existence of’

- (5) Examples from BulEnAC for verbs in aligned parallel sentences with several possible closely related WordNet senses.
  - a. Since he couldn’t sleep anyway, he might as well stand their watches for them or WRITE their reports. (telic, thus classified as an accomplishment)
  - b. All his life her father WAS WRITING poems and novels. (generalised atelic; or telic + iterative)
  - c. In his poems he WROTE about the beauty of the countryside.

WordNet synset:

- eng-30-01698271-v / write / verb.creation / ‘produce a literary work’

In addition, mismatches are observed in the extracted parallel examples on various levels as illustrated in the examples below.

- (6) Example from BulEnAC for verbs in aligned parallel sentences with mismatch in translation.
  - a. We’re not even out.
  - b. Дори не СЕ РАЗХОЖДАМЕ. ‘We do not even walk around.’
- (7) Example from the Russian-Bulgarian parallel corpus within RusNC<sup>5</sup> for verbs in aligned parallel sentences with different verbs. The example was extracted manually.
  - a. Ru: Внизу проплывали игрушки-парусники... (Alexander Belyaev. The Ruler of the World. Russian edition: 1940.
  - b. Bg: Отдолу ПЛАВАХА играчки платноходки. (Bulgarian translation:

<sup>5</sup>The examples are taken from The Russian National Corpus (<https://ruscorpora.ru>).

1988, trans. by Assen Trayanov)  
‘Toy sailboats floated from below...’

This initial work on data selection served as a starting point for laying out the theoretical basis in determining the scope of **activity predicates**, their specific features and the possible approaches for the distinction between different senses of a verb with a view to its realisation in text.

## 7 Proposed annotation schema

The corpora used for extracting examples are supplied with basic annotation such as sentence splitting, POS tagging, lemmatisation, performed for both English and Bulgarian (Koeva et al., 2020).

The annotation aims at identification and description of the following syntactic components: (a) the verb, its WordNet sense, and the semantic frame it evokes; (b) noun phrases matched to frame elements and serving as external argument (NP.ext) or direct object (NP.Obj); (c) prepositional phrases (PP) matched to frame elements; (d) subordinate clauses marked with different conjunctions, direct quotes usually marked using punctuation, and other lexical elements that realise frame elements. In particular, as a minimum we aim to identify and annotate core frame elements, but in some cases non-core frame elements which are essential for the interpretation of the verb, are also annotated (e.g., when an element is essential to distinguish a verb as an activity rather than other aspectual class).

Here we present the main principles and steps for the annotation of the illustrative corpus of examples which aim to ensure the consistency of the annotation as well as the flexibility allowing for its expansion in terms of including more languages, more examples and more levels of annotation. The annotation is ongoing.

### 7.1 Matching a verb to a WordNet sense

Each verb is matched to a WordNet sense in order to: (1) provide cross-language linking to translational equivalents of the verb in different languages by linking them through the interlingual index of WordNet; and (2) provide linking to FrameNet and assign a FrameNet semantic frame to the verb, so that we can investigate the syntactic realisation of the verb and the frame elements in its evoked semantic frame.

As seen above, in semantically annotated corpora verbs have already been disambiguated and assigned a WordNet sense. However, these corpora are limited in size and coverage, and other more general corpora are also used for deriving examples. Additional semi-automatic procedures are applied to identify the WordNet verb sense. For example, in parallel corpora the two verbs within the aligned sentences can be used for additional automatic validation whenever possible. For monolingual corpora and other cases of verb ambiguity, manual validation has been performed.

### 7.2 Identification of the FrameNet frame evoked by a verb

After the verb has been matched to a particular WordNet sense, it can be assigned the FrameNet frame that characterises the verbs of the respective synset. For this purpose we rely on the mapping between WordNet synsets and FrameNet frames ((Shi and Mihalcea, 2005), (Tonelli and Pighin, 2009), (Palmer et al., 2014), (Leseva et al., 2018), among others).

By identifying the frame evoked by the verb, we are able to analyse the configurations of frame elements in the example sentences and to make observations on the verb class based on its realisation.

### 7.3 Identification of the syntactic components corresponding to core frame elements

We take as a point of departure the valence patterns as a collective set for all lexical units evoking a given frame. The generalised valence patterns show the possible configurations of frame elements for the evoked semantic frame and their corresponding syntactic realisations.

- (8) Valence patterns for the FrameNet frame **Text creation** evoked by *write* ‘produce a literary work’.

[NP.Ext]<sub>Author</sub> [NP.Obj]<sub>Text</sub>  
 [NP.Ext]<sub>Author</sub> [NP.Obj]<sub>Text</sub> [PP]<sub>Time</sub>  
 [NP.Ext]<sub>Author</sub> [NP.Obj]<sub>Text</sub> [ADVP]<sub>Time</sub>  
 [NP.Ext]<sub>Author</sub> [NP.Obj]<sub>Text</sub> [PP]<sub>Manner</sub>  
 [NP.Ext]<sub>Author</sub> [NP.Obj]<sub>Text</sub> [ADVP]<sub>Manner</sub>

#### 7.4 Identification of the valence pattern associated with the example

The FrameNet valence patterns describe all the co-occurrence combinations of frame elements (both core and non-core) attested for each annotated lexical unit in the FrameNet annotated corpus. The set of the identified and annotated frame elements is matched against the set of possible valence patterns associated with the semantic frame of the verb. Priority is given to valence patterns containing only core frame elements than to more elaborate patterns, as well as to more frequent patterns (frequency is extracted from the dataset of annotated examples in English from FrameNet).

There may be a mismatch in the syntactic category across languages, e.g., a certain frame element may be a direct object in one language and a prepositional object in another. Languages may also differ in terms of the overtiness of syntactic information, i.e. the possibility to leave an obligatory element non-explicit (null instantiations retrievable from the context or the grammatical construction); language-specific diatheses, constructions, word order, morphosyntactic features, etc. The inventory of means that introduce certain frame elements such as prepositions, conjunctions, wh-words, etc. are also language-specific.

In annotating the data we pay attention to the cases of null instantiation, where the frame element is not overt – definite (e.g. pro-drop in Bulgarian), indefinite when the frame element represents a generalised non-specific entity (e.g. with communication verbs that are not directed to an addressee but such is implied), constructional when the lexical omission is licensed by certain constructions (e.g. imperative), and incorporated frame element – where the meaning of the frame element is incorporated in the meaning of the verb, and thus not expressed in the sentence.

Original patterns from FrameNet are generalised in order to allow cross-language match with the Bulgarian data. Particular attention is paid to examples which are not matched to a pattern in order to identify patterns characteristic for Bulgarian that do not appear in FrameNet or for English in general.

### 8 Applications and future work

The corpus aims at providing illustrative examples for the usage of activity verbs – a large and

diverse class of verb predicates which shows various specific characteristics in contrast to other aspectual classes. Further, the aspectual properties of many verbs are not realised on the lexical level (in the lexical meaning) but within the larger unit (as in VP). Thus, the corpus will be a useful source of examples for studying the syntactic realisation of activity verbs.

The work on the corpus demonstrates the principles of information transfer: (a) by linking different resources in terms of scope, coverage, description layers, granularity of semantic categories, to provide a basis for comprehensive description of verb semantics; and (b) across languages to facilitate the development of resources and language processing tools for low-resourced languages such as Bulgarian.

Moreover, the corpus can be used to study the syntax and valence patterns across languages, thus facilitating comparative studies on conceptual structure. The corpus will provide empirical material for the comparative study between languages with lexical aspect (such as Bulgarian, Russian and other Slavic languages) and those without lexical aspect (such as English). The flexible structure and the annotation scheme allow the corpus to be expanded with more examples, languages and annotation.

The collection of parallel data of activity predicates is aimed at abstract ontological description and will allow the comparison of the features of conceptualisation, lexicalisation and grammaticalisation of activities in Bulgarian and other languages. The parallel collection will be used for theoretical comparison in the conceptualisation of different types of activities that correlate with the grammatical structure in individual languages. This will make possible the typological description of activity predicates and the highlighting of language-specific and universal features at the semantic and syntactic level.

### Acknowledgments

This research is carried out as part of the project *Ontology of Activity Predicates – Linguistic Modelling with a Focus on Bulgarian* funded by the Bulgarian National Science Fund, Grant Agreement No KII-06-H80/9 from 8.12.2023.

## References

- Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *COLING-ACL '98: Proceedings of the Conference. Montreal, Canada*, pages 86–90.
- Petia Barkalova. 1997. *Balgarskiat sintaksis – poznat i nepoznat*. Plovdivsko Universitetsko Izdatelstvo.
- Tatiana V. Bulygina. 1982. K postrojeniju tipologiji predikatov v russkom jazuke [to build typologies predicates in russian]/Tv buligina. *Semanticheskiji tipu predikatov/[otv. red. ON Seliverstova].–M.: Nauka*, pages 7–85.
- Aljoscha Burchardt and Marco Pennacchiotti. 2008. **FATE: a FrameNet-Annotated Corpus for Textual Entailment**. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. European Language Resources Association (ELRA), Marrakech, Morocco.
- Rositsa Dekova. 2006. *Lexical encoding of verbs in English and Bulgarian*. Det historisk-filosofiske fakultet.
- David R. Dowty. 1979. *The Semantics of Aspectual Classes of Verbs in English*, pages 37–132. Springer Netherlands, Dordrecht.
- Christiane Fellbaum, editor. 1998. *WordNet: an electronic lexical database*. MIT Press, Cambridge, MA.
- Christiane Fellbaum, Anne Osherson, and Peter E. Clark. 2009. Putting Semantics into WordNet's "Morphosemantic" Links. In *Proceedings of the Third Language and Technology Conference, Poznan, Poland*. [Reprinted in: *Responding to Information Society Challenges: New Advances in Human Language Technologies. Springer Lecture Notes in Informatics*], volume 5603, pages 350–358.
- Charles J. Fillmore. 1982. Frame Semantics. In *Linguistics in the Morning Calm (Ed. by The Linguistic Society of Korea)*, pages 111–137. Seoul: Hanshin.
- Charles J. Fillmore. 1985. Frames and the semantics of understanding. *Quaderni Di Semantica*, 6:222–254.
- Charles J. Fillmore. 2003. Valency and semantic roles: the concept of deep structure case. In Vilmos Ágel, Ludwig M. Eichinger, Hans Werner Eroms, Peter Hellwig, Hans Jürgen Heringer, and Henning Lobin, editors, *Dependenz und Valenz: Ein internationales Handbuch der zeitgenössischen Forschung*, pages 457–475. Walter de Gruyter.
- Charles J. Fillmore and Collin F. Baker. 2009. A frames approach to semantic analysis. In B. Heine and H. Narrog, editors, *The Oxford handbook of linguistic analysis*, pages 313–340. Oxford: Oxford University Press.
- Gramatika. 1983. *Gramatika na savremenen balgarski knijoven ezik T3. Sintaksis*. Sofia: Akademichno izdatelstvo „Prof. Marin Drinov“.
- Svetla Koeva. 2006a. BulNet (leksikalno-semantichna mrežha na balgarskiya ezik) — chast ot svetovnata leksikalno-semantichna mrežha. *Balgarski ezik*, pages 19–32.
- Svetla Koeva. 2006b. *Sintaktichni transformatsii*, pages 106–138. SemaRSh, Sofia.
- Svetla Koeva. 2010. *Balgarskiyat Freymnet*. Institute for Bulgarian Language, Sofia.
- Svetla Koeva. 2021. **The Bulgarian WordNet: Structure and specific features**. *Papers of Bulgarian Academy of Sciences*, 8(1):47–70.
- Svetla Koeva and Emil Doychev. 2022. **Ontology supported frame classification**. *Proceedings of the Fifth International Conference Computational Linguistics in Bulgaria*, pages 203–214.
- Svetla Koeva and Elena Ivanova. 2024. Izsledvane na lingvistichni testove za razgranichavane na aspektualnite klasove (s fokus varhu balagrski i ruski). In *Derzhavinski cheteniya, Moskva*.
- Svetla Koeva, Elena Ivanova, Yovka Tisheva, and Anton Zimmerling, editors. 2022. *Ontologiya na situatsiite za sastoyanie – lingvistichno modelirane. Sapostavitelno izsledvane za balgarski i ruski*. Prof. Marin Drinov Publishing House of Bulgarian Academy of Sciences.
- Svetla Koeva, Svetlozara Leseva, Borislav Rizov, Ekaterina Tarpomanova, Tsvetana Dimitrova, Hristina Kukova, and Maria Todorova. 2011. Design and development of the Bulgarian sense-annotated corpus. In *Information and communications technologies: present and future in corpus analysis: Proceedings of the III International Congress of Corpus Linguistics*, pages 143–150.
- Svetla Koeva, Nikola Obreshkov, and Martin Yalamov. 2020. **Natural language processing pipeline to annotate Bulgarian legislative documents**. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6988–6994, Marseille, France. European Language Resources Association.
- Svetla Koeva, Borislav Rizov, Ekaterina Tarpomanova, Tsvetana Dimitrova, Rositsa Dekova, Ivelina Stoyanova, Svetlozara Leseva, Hristina Kukova, and Angel Genov. 2012a. Bulgarian-English Sentence- and Clause-Aligned Corpus. In *Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2)*, page 51–62. Lisboa: Colibri.
- Svetla Koeva, Ivelina Stoyanova, Svetlozara Leseva, Rositsa Dekova, Tsvetana Dimitrova, and Ekaterina Tarpomanova. 2012b. **The Bulgarian National Corpus: theory and practice in corpus design**. *Journal of Language Modelling*, 0(1):65–110.

- Siya Kolokovska. 2005. Semantika i motiviranost na terminite za protsesi v savremennia balgarski knizhoven ezik. *Elektronno spisanie LiterNet 11*.
- Nadezhda Kostova. 2010. *Osnovni glagoli za dvizhenie v balgarskiya ezik*. Avangard Prima, Sofia.
- Svetlozara Leseva and Ivelina Stoyanova. 2021. Semantichno opisanie na glagoli za promyana i yerarhichna organizatsiya na kontseptualnite freymove. In *Proceedings from the Annual Conference of the Institute for Bulgarian Language*, volume 2.
- Svetlozara Leseva, Ivelina Stoyanova, and Maria Todorova. 2018. Classifying verbs in WordNet by harnessing semantic resources. In *Proceedings of CLIB 2018, Sofia, Bulgaria*.
- John Lyons. 1977. *Semantics*. Cambridge University Press.
- Yuriy Maslov. 1982. *Gramatika na balgarskiya ezik*. Nauka i izkustvo, Sofia.
- George A. Miller. 1995. WordNet: A lexical database for English. *Commun. ACM*, 38(11):39–41.
- George A. Miller and Christiane Fellbaum. 2007. WordNet Then and Now. *Language Resources and Evaluation*, 41:209 – 214.
- George A. Miller, Claudia Leacock, Randee Teng, and Ross T. Bunker. 1993. *A Semantic Concordance*. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.
- Elena V Paducheva. 2009. Leksicheskaya aspektualynosty i klassifikatsia predikatov po maslovu-vendleru. *Voprosi yazikoznania*, (6):3–20.
- Martha Palmer, Claire Bonial, and Diana McCarthy. 2014. SemLink+: FrameNet, VerbNet and event ontologies. In *Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore (1929–2014), Baltimore, Maryland USA, June 27, 2014*, pages 13–17. Association for Computational Linguistics.
- Malka Rappaport Hovav. 2008. Lexicalized meaning and the internal temporal structure of events. In Susan Rothstein, editor, *Theoretical and Crosslinguistic Approaches to the Semantics of Aspect*, pages 13 — 42. John Benjamins Publishing Company.
- Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher. R. Johnson, Collin. F. Baker, and Jan Scheffczyk. 2016. *FrameNet II: extended theory and practice*. International Computer Science Institute, Berkeley, California.
- Svetlana O. Savchuk, Timofey Arkhangel'skiy, Anastasiya A. Bonch-Osmolovskaya, Ol'ga V. Donina, Yuliya N. Kuznetsova, Ol'ga N. Lyashevskaya, Boris V. Orekhov, and Mariya V. Podryadchikova. 2024. Nacional'nyj korpus ruskogo jazyka 2.0: novye vozmozhnosti i perspektivy razvitija [russian national corpus 2.0: new opportunities and development prospects. *Voprosy Jazykoznanija*, 2:7–34.
- Lei Shi and Rada Mihalcea. 2005. Putting pieces together: Combining framenet, verbnet and wordnet for robust semantic parsing. In *Computational Linguistics and Intelligent Text Processing: 6th International Conference, CICLing 2005, Mexico City, Mexico, February 13-19, 2005. Proceedings 6*, pages 100–111. Springer.
- Valentin Stankov. 1980. *Glagolnizat vid v balgarskiya ezik*. Nauka i izkustvo, Sofia.
- Maria Todorova. 2023. *Semantic annotation of common lexis verbs of contact in Bulgarian*. In *Proceedings of the 19th Joint ACL-ISO Workshop on Interoperable Semantics (ISA-19)*, pages 11–17, Nancy, France. Association for Computational Linguistics.
- Sara Tonelli and Daniele Pighin. 2009. New Features for FrameNet – WordNet Mapping. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL'09), Boulder, USA*.
- Robert D. Van Valin and Randy J. LaPolla. 1997. *Syntax: Structure, meaning and function*. Cambridge: Cambridge University Press.
- Veselina Vateva. 2005. *Balgarskite glagoli za dvizhenie v semantichen i stilistichen aspekt*. Burgas: Diamant.
- Zeno Vendler. 1957. Verbs and Times. *Philosophical Review*, pages 143 – 160.
- Zeno Vendler. 1967. *Linguistics in Philosophy*. Ithaca, NY: Cornell University Press.
- Radka Vlahova. 2004. Kam strukturno-semantichnata harakteristika na nyakoi prefigirani glagoli v savremennia balgarski ezik. *Godishnik na SU. T. Ezikoznanie*, 89:5 – 29.

# Large Language Models in Linguistic Research: the Pilot and the Copilot

Svetla Koeva

Institute for Bulgarian Language, Bulgarian Academy of Sciences

svetla@dcl.bas.bg

## Abstract

In this paper, we present two experiments focussing on linguistic classification and annotation of examples, using zero-shot prompting. The aim is to show how large language models can confirm or reject the linguistic judgements of experts in order to increase the productivity of their work. In the first experiment, new lexical units evoking a particular FrameNet semantic frame are selected simultaneously with the annotation of examples with the core frame elements. The second experiment attempts to categorise verbs into the aspectual classes, assuming that only certain combinations of verbs belonging to different aspectual classes evoke a semantic frame. The linguistic theories underlying the two experiments, the development of the prompts and the results of the experiments are presented.

**Keywords:** LLMs, linguistic research, prompt engineering

## 1 Introduction

Until recently, natural language processing (NLP) relied on specialised language resources such as monolingual, bilingual and multilingual corpora as well as lexical and conceptual resources to develop functional applications. However, breakthroughs in artificial intelligence and the emergence of large language models (LLM) have changed the field, enabling the successful completion of a variety of NLP tasks in a completely different way.

Widely studied applications of LLMs include document intelligence tasks, such as sentiment analysis (Krugmann and Hartmann, 2024), text classification (Sun et al., 2023), risk prediction (Cao et al., 2024), information extraction (Peng et al., 2024), and many others. In addition, LLMs are used for machine translation (Zhu et al., 2024), content creation tasks such as creative writing, automatic sentence completion, paraphrasing, personalised decision making and code generation. LLMs

play a crucial role in virtual assistants that facilitate various applications such as language understanding, speech generation and speech recognition (Wang et al., 2022; Arora et al., 2024).

The increasing use of large language models can be expected not only for traditional NLP tasks, but also to tackle typical linguistic challenges. The recent application of LLMs in linguistic research can be outlined in several directions. Most attempts aim at using LLMs for linguistic annotation to facilitate corpus-based linguistic studies by automatically annotating texts with targeted linguistic information (Kuzman et al., 2023; Gilardi et al., 2023), among others, but there are also attempts to use LLMs for theoretical investigations within specific linguistic frameworks (Beguš et al., 2023; Torrent et al., 2024).

In this paper we offer two experiments focusing on the linguistic classification and annotation of examples. The aim is to show how some newly released LLMs can confirm or reject the linguistic hypotheses of experts in order to increase the productivity of their work. In the first experiment, new proposals for lexical units evoking a particular FrameNet semantic frame are classified simultaneously with the annotation of examples with the core frame elements. In the second experiment, verb lexical units are categorised into different semantic classes, whereby it is assumed that a certain semantic frame can only be evoked by verbs that belong to certain combinations of semantic classes. The linguistic theories on which the two experiments are based, the development of the prompts and the results of the experiments are briefly presented.

The paper deals with the following topics: Section 2 gives an overview of the use of large language models in linguistic research. Section 3 briefly introduces the large language models used in the experiments. Section 4 describes experiments with two types of prompts: a) for the si-

multaneous augmentation of FrameNet semantic frames with new lexical units and annotated examples and b) for the classification of lexical units into semantic classes. The presentation of both experiments includes a description of the linguistic theories, the prompt structures and an overview of the results. Section 5 contains a conclusion on the benefits and challenges of using LLMs in linguistic research based on the experiments conducted.

## 2 LLMs and linguistic research

Prompt engineering has already been used to solve linguistically relevant tasks. In few-shot learning, only a few examples are given to the model during inference and task definition (Brown et al., 2020: 5-7). One-shot learning and zero-shot learning are similar to few-shot learning (Wei et al., 2023), but as the names suggest, only one or zero demonstrations are allowed to formulate a task description in natural language.

Recently, the ability of LLMs to recognise and classify specific language constructions, to analyse data within a theoretical framework, to annotate texts with relevant linguistic information to support corpus-based linguistic studies, etc. has been investigated. The following is a brief overview of the applications of LLMs in the field of linguistics, formally divided into: recognition of language constructions, analysis of data within a theoretical framework and linguistic annotation, as some of the studies combine more than one approach.

### 2.1 Recognition of language constructions

GPT-3 (Brown et al., 2020) knowledge of rare constructions with semantic and syntactic constraints, such as the construction *indefinite article + adjective + numeral + noun* in English (e.g. “a lovely five days”), was assessed (Mahowald, 2023). The acceptability judgements of the GPT-3 for this construction were compared with human judgements on a range of sentences to conclude that the GPT-3 judgements are broadly similar to human judgements.

Another study investigated the extent to which the word frequency data of LLMs match the data of a large general corpus; the collocation data of LLMs match a large general corpus; and LLMs can recognise lexico-grammatical patterns and perform genre categorisation (Uchida, 2024). ChatGPT 3.5 showed a high agreement with the COCA (Corpus of Contemporary American English) ranking in

terms of word frequency, but varied when analysing certain word types due to more repetitions. The overall agreement for the collocation patterns tested was 42.8%. When examining open slots in grammatical patterns, more than half of the words in ChatGPT 3.5 matched the top 20 of COCA, and more than 65% were within the top 40, showing effective verification of lexico-grammatical patterns. However, the hit rate for genre identification was low at both word and text level.

The book *Copilot for Linguists* introduces the concept of using LLM chatbots as a tool for trained linguists and shifts the focus from what these chatbots can achieve with language to how they can support linguists in their work (Torrent et al., 2024). Experiments are presented in which LLM chatbots were prompted to analyse grammatical constructions and to enrich FrameNet semantic frames in both English and Brazilian Portuguese. Prompt engineering techniques derived from these experiments were shared, and the potential of LLMs to act as copilots for construction grammarians in linguistic research was explored, highlighting their ability to recognise instances of fully developed constructions, analyse their syntax and understand their meaning. The study aims to demonstrate the valuable role that LLM chatbots can play in supporting the analytical endeavours of linguists. However, the limitations are also pointed out, e.g. when analysing constructions in languages other than English and when understanding the semantics of language constructions.

### 2.2 Formal analysis of linguistic data

Efforts have been made to show that LLMs can produce coherent and valid formal analysis of linguistic data (Beguš et al., 2023). The metalinguistic capabilities of GPT-3.5 and GPT-4 (OpenAI, 2023) were tested, focussing on three subdomains of formal linguistics: syntax, phonology and semantics. It was shown that GPT-4 is able to analyse both relatively simple and more complex syntactic structures largely correctly, while GPT-3.5 performs worse on the same tasks. GPT-4 was tested on two phonological problems by prompting the model with small datasets: a palatalisation process in Korean and a spirantisation process in an artificial language, and GPT-4 copes well with both. The ability of GPT-4 to produce lambda calculus analyses of English sentences, including cases of scopal ambiguity (e.g. *Every student likes a classmate*), was

tested. It was found that the model works well with simpler sentences and understands scopal ambiguity, but makes some significant errors when using the lambda calculus formalism. Overall, GPT-4 is largely (but not perfectly) able to produce coherent analyses of simple problems in each of the three domains tested, detect ambiguity, correct its own analytical errors, and comment on the feasibility of multiple solutions.

### 2.3 Linguistic annotation

The performance of ChatGPT and the multilingual XLM-RoBERTa language model (Conneau et al., 2019) was evaluated on the genre identification task to determine which model is best suited to enrich large web corpora for English and Slovenian with genre information (Kuzman et al., 2023). The two models are compared in three scenarios, switching between the languages of the prompts and the test datasets. The results show that Chat-GPT outperforms the fine-tuned model when applied to a completely new test dataset. However, when the model is fully prompted in Slovenian, performance drops significantly, demonstrating the limitations of using ChatGPT in smaller languages at the time of the experiments.

It was evaluated how ChatGPT performs subjective tasks related to social norms and cultural context, such as identifying implicit hate speech online and providing explanations for it (Huang et al., 2023). The results show that ChatGPT correctly identifies 80% of implicit hate tweets in the experimental setup, demonstrating its potential as a data annotation tool with a simple prompt design. However, it was noted that there is a risk of misleading non-experts if the model's decisions are incorrect.

Another study showed that ChatGPT classifications with zero-shots are better than MTurk (Amazon Mechanical Turk, a crowdsourcing marketplace) annotations (Gilardi et al., 2023). The analysis was performed on a sample of 6,183 documents, including tweets and news articles. Several annotation tasks were implemented, e.g. *relevance* to determine whether a tweet relates to content moderation or politics; *topic detection* to determine whether a tweet falls into one of six predefined topics; *stance detection* to determine whether a tweet supports, opposes or remains neutral on a US law, etc. The performance of ChatGPT was evaluated based on accuracy and intercoder agreement. The

results showed that ChatGPT's zero-shot accuracy outperformed the crowd workers by an average of 25 percentage points, while ChatGPT's intercoder agreement outperformed both the crowd workers and the trained annotators on all tasks.

The performance of LLMs was evaluated on the task of annotating local grammars, focussing on the speech act of apology in English (Yu et al., 2024). The analysed corpus contained 5,539 instances of the word *sorry*, extracted from the Spoken British National Corpus 2014. The experimental setup involved few-shot prompting techniques and included the following steps: comparing the performance of GPT-4 based Bing Chatbot and ChatGPT 3.5 when annotating 50 instances and comparing the performance of Bing Chatbot and a human annotator when annotating 1000 instances. The results show that the Bing Chatbot performs better than ChatGPT 3.5. Although the human annotator achieved slightly higher accuracy than the Bing Chatbot, the latter showed robust abilities to understand both the semantic and pragmatic aspects of the language.

The conclusions that can be drawn from the brief overview of the use of LLMs for linguistic (or related) tasks are that they have the greatest application in the recognition of language units and their classification, i.e. in linguistic annotation. Automatic annotation of texts was not absolutely accurate even before LLMs (with varying degrees of success for different annotation tasks) (Liao and Zhao, 2019), but has been successfully used for applications requiring large annotated training data. Currently, the refinement of the prompting process to obtain correct information from the LLMs and the progress of the LLMs themselves can lead to positive results in several directions: creation of large semantic language resources and development of benchmarking datasets to evaluate LLMs for various linguistic tasks.

In this study, we will analyse the potential of current LLMs for the creation of **semantic language resources**, in particular for the enrichment of FrameNet's semantic frames with new lexical units and annotated examples, as well as for the semantic classification of lexical units.

## 3 Large language models used in the experiments

There are already several surveys on large language models describing the history of pre-training and

breakthroughs (Han et al., 2021; Zhou et al., 2023), the scaling and impact of pre-trained models (Wang et al., 2022), the prompting methods in natural language processing (Liu et al., 2023), the multimodal pre-trained models (Wang et al., 2023), the recent advances in LLMs with introduction to the background, key results and mainstream techniques (Zhao et al., 2023), the evaluation methods for LLMs (Chang et al., 2023), the comparison between some of the most popular LLMs, including the three families: GPT, LLaMA and PaLM, and the discussion of their features, contributions and limitations (Minaee et al., 2024), among others.

Here we will briefly introduce the LLMs that were used in the experiments to illustrate the capabilities of LLMs to assist the creation of large semantic language resources: (in alphabetical order) **Claude 3.5 Sonnet**, **Gemini 1.5 Pro**, **GPT-4o**, **GPT-4o mini**. The models were last accessed on 21 and 22 July 2024.

It is not our aim to compare these LLMs for several reasons: a) models are evolving very fast and new models and updates of existing models are constantly appearing; b) some of the models have been improved in a clearly defined direction; c) for some of the selected LLMs, detailed information about architecture, training data, model size, etc. is not available. Therefore, we do not focus on comparing models, but on exploring a possible way to work with LLMs on a specific linguistic task.

The **Claude 3** is a family of multimodal models released by Anthropic<sup>1</sup>: Claude 3 Opus, Claude 3 Sonnet and Claude 3 Haiku. Claude 3 has been trained with a mixture of public and proprietary data and is subject to rigorous cleaning and filtering methods. All models have image processing and show good performance in logical reasoning, maths and coding (Anthropic, 2023). Claude 3, especially the Opus model, reportedly outperforms other state-of-the-art models in various evaluation benchmarks such as the Google-Proof Question-Answering Benchmark (GPQA), Measuring Massive Multitask Language Understanding (MMLU) and others.

**Sonnet** has been updated to **Claude 3.5**, the latest version of the Claude LLMs<sup>2</sup>. In Anthropic's tests, the Claude 3.5 Sonnet outperforms some of the latest LLMs and other Claude models. The Claude 3.5 Sonnet has powered the Claude chatbot.

<sup>1</sup><https://www.anthropic.com/news/claude-3-family>

<sup>2</sup><https://www.anthropic.com/news/claude-3-5-sonnet>

It is also available via the Anthropic API.

The **Gemini 1.5** family of Google DeepMind aims to retrieve and analyse millions of context tokens, including multiple long documents, video and audio materials (Gemini Team, Google, 2024). To achieve this, the models are trained on multiple 4096-chip pods of TPUv4 accelerators, using a wide range of multimodal and multilingual data that includes image, audio and video content.

**Gemini 1.5 Pro**<sup>3</sup> is a sparse Mixture-of-Expert (MoE) Transformer-based model that builds on the advances of Gemini 1.0 (Gemini Team, Google, 2023) in a variety of multimodal tasks such as visual understanding, classification, summarisation and content creation from image, audio and video. The extensive evaluations with diagnostic multimodal long-context benchmarks show that Gemini 1.5 Pro is able to retrieve and understand large amounts of data. The model is available via the Gemini Chatbot and the Gemini API.

**GPT-4** is a large multimodal model that accepts both image and text input and generates text output (OpenAI, 2023). It is reported that GPT-4 understands and generates text in more languages compared to its predecessor and outperforms GPT-3.5 and other large language models in a number of traditional NLP benchmarks and the MMLU benchmark.

The next generation, **GPT-4o (omni)**<sup>4</sup>, has been integrated into the text, vision and audio modalities through techniques such as filtering training data and refining model behaviour through post-training. According to OpenAI, GPT-4o matches the performance of GPT-4 Turbo in the areas of text, reasoning and coding intelligence evaluated with traditional benchmarks, and offers multilingual, audio and vision processing. GPT-4o is available via the ChatGPT Plus and the OpenAI API.

The **GPT-4o mini**<sup>5</sup> is a smaller version of the GPT-4o model. GPT-4o mini has reportedly outperformed GPT-3.5 Turbo in several LLM benchmarks and is trained with an instruction hierarchy method (Wallace et al., 2024) that improves the model's resistance to jailbreaks and system prompt extraction. In ChatGPT, GPT-3.5 was replaced by GPT-4o mini. It is also available via the OpenAI API.

<sup>3</sup><https://deepmind.google/technologies/gemini/pro/>

<sup>4</sup><https://www.infoq.com/news/2024/05/openai-gpt4o/>

<sup>5</sup><https://platform.openai.com/docs/models/>

## 4 Prompting-based experiments to facilitate linguistic research

We present two experiments with zero-shot prompting aimed at linguistic classification and annotation of examples. The aim is to demonstrate the effectiveness of LLMs in making linguistic decisions, thus increasing the expert’s confidence in creating semantic resources with reliable information without the need for a second or third expert to perform the same activities.

The first experiment (Augmenting FrameNet semantic frames with lexical units and annotations) simultaneously aims to select new lexical units that evoke a particular FrameNet semantic frame, suggest relevant examples and annotate them with the core elements of the frame. The second experiment (Classification of FrameNet lexical units into semantic classes) aims to categorise verbs that evoke a particular FrameNet semantic frame into relevant semantic classes.

The general framework can be summarised as follows: a) formulation of a linguistic task; b) selection of suitable LLMs (mainly according to two criteria: novelty and accessibility); c) formulation of a prompt template aimed at fulfilling the linguistic task; d) execution of the experiment and collection of data from the selected LLMs; e) human evaluation of the results obtained.

### 4.1 Augmenting FrameNet semantic frames with lexical units and annotations

The annotation of new examples for existing frames with the syntactic realisation of frame elements and lexical units and the discovery of new lexical units for existing frames is the most common extension to improve lexical coverage and representation (Torrent et al., 2024).

An approach to expanding lexical units that evoke a semantic frame utilises the links between lexical units that evoke a frame and lexical units in other resources to discover potential new lexical units. For this task, vector representations of lexical units and clustering techniques are used (Yong and Torrent, 2020).

Our experiment aims at a simultaneous extension of the semantic frames of FrameNet with new lexical units and annotated examples. For the expansion of lexical units, we use a list of potential candidates that are accepted or rejected by LLMs, in contrast to other approaches that use direct instructions to LLMs to propose new lexical units

(Torrent et al., 2024). Furthermore, we aim at a full-text annotation with the core frame elements of the synthetic examples provided by the LLMs.

#### 4.1.1 Frame semantics in brief

FrameNet is based on the theory of frame semantics (Fillmore, 1976, 1982).

The central idea of frame semantics is that word meanings are described in terms of semantic frames, which are schematic representations of the conceptual structures and patterns of beliefs, practices, institutions, images, etc. that provide a foundation for meaningful interaction in a given speech community (Fillmore et al., 2003: 235).

FrameNet<sup>6</sup> is a collection of semantic frames that contain a common abstract semantic representation for a set of lexical units and valency patterns that represent semantic and syntactic descriptions based on the annotation of examples. The semantic frame in FrameNet includes the following components: frame name; informal definition of the situation that the frame represents; semantic type of the frame (optional); set of frame elements associated with the frame (core and non-core: peripheral, extrathematic and core-unexpressed); relations between frame elements, if any; frame-to-frame relations, if any; and the lexical units that evoke the frame.

The frame element information includes the name of the frame element, its informal definition, the semantic type (optional), and examples illustrating the use of the frame element (optional).

The information on the lexical units includes a definition, a semantic type (optional), annotated examples and derived valency patterns that show the correspondence between the frame elements and their syntactic realisation.

An excerpt from the semantic frame **Arriving** is shown to illustrate this. **Arriving** frame in FrameNet has the definition: “An object **Theme** moves in the direction of a **Goal**. The **Goal** may be expressed or it may be understood from context, but it is always implied by the verb itself.” The core frame elements are: **Theme** defined as “**Theme** is the object that moves. It may be an entity that moves under its own power, but it need not be” and **Goal** defined as “**Goal** is any expression that tells where the **Theme** ends up, or would end up, as a result of the motion”.

The verbs that evoke this frame are: appear.v, ap-

<sup>6</sup><http://framenet.icsi.berkeley.edu>

proach.v, arrive.v, come.v, crest.v, descend (on).v, enter.v, find.v, get.v, hit.v, make it.v, make.v, reach.v, return.v, visit.v.

#### 4.1.2 Prompt design

For our experiment, we use semantic and lexical information from FrameNet, in particular from the semantic frame **Arriving**, focussing on lexical units of verbs. There are several preparatory steps: the selection of potential new lexical units for the **Arriving** frame and the experimental formulation of an appropriate prompt, aiming at a simultaneous evaluation of the selected lexical units and suggestions for annotated examples to illustrate the core elements of the frame.

We consider WordNet (Miller et al., 1990) as a natural source of new lexical units for existing frames in FrameNet, since WordNet contains about 117 000 synsets<sup>7</sup>.

The first step is to match the lexical units of a particular semantic frame with the corresponding synsets from WordNet. Although previous mappings between FrameNet and WordNet are known (Shi and Mihalcea, 2005), the method used is relatively simple and combines the exact matching of lexical units from FrameNet with literals from WordNet and the evaluation of the similarity of the definitions. This approach takes into account the fact that WordNet contains many literals with the same form but different meanings. After mapping, a human judgement is made to decide whether a particular mapping between a lexical unit in FrameNet and a WordNet literal is correct.

Among the lexical units that evoke a particular semantic frame, there are synonyms, hypernyms and hyponyms (troponyms for verbs), although the existing semantic relations between lexical units are not explicitly labelled. All troponym synsets up to a hypernym synset whose literal(s) are mapped with a lexical unit from FrameNet are considered potential candidates. If the mapping is rejected by an expert, the corresponding troponyms are also ignored.

The results of the mapping show that one lexical unit is not present in WordNet, three mapped synsets have no troponyms and in two mappings the synset appears as a troponym of another mapped synset. The number of potential new lexical units evoking the semantic frame **Arriving** is 137.

Each prompt contains the name of the targeted

semantic frame, its definition, the core elements of the frame with their definitions and a list of the new lexical units provided with definitions from WordNet. The desired output format is also included. The prompt template has the following structure:

*The semantic frame “frame name” in FrameNet has the definition: “frame definition”. The core frame elements are: “name of core frame element defined as “frame element definition”.*

*Indicate which of the following verbs: “verb”: “definition”, evoke the semantic frame “frame name” and those that do not; give three examples of each verb and annotate the core frame elements in the examples. Use this pattern:*

“verb”: “definition”

*Example 1: “example” “core frame element name”: “annotation”*

The Appendix A illustrates a prompt with four verbs.

Preliminary tests were carried out to determine the optimal format for the prompts according to human judgement, with the aim of achieving a satisfactory level of completeness of responses.

#### 4.1.3 Results and discussion

The prompt requires an assessment of the relevance of the lexical units to a semantic frame, the provision of relevant examples and the annotation of the examples with the core elements.

The classification of lexical units according to whether or not they evoke the semantic frame **Arriving** is linked to the correct interpretation of the verb meaning by the LLMs. In some cases, the use of the lexical units in the synthetic examples that do not illustrate the intended meaning is classified, e.g. *The company **scaled** back production to match declining demand.* However, some unintended verb meanings that are illustrated with examples are correctly categorised as not evoking the semantic frame **Arriving**, e.g. *The new policy aims to **get at** the root causes of inequality.* In some cases, models provide clues that can be helpful for an expert’s final decision, e.g. ***To get at** implies reaching an abstract or physical destination making it relevant to the **Arriving** frame as it involves a **Theme** moving towards a defined **Goal**.*

The examples given are most likely synthetic, but it is also possible that they are contained in the training data. Regardless, all examples are grammatically correct and sound natural. They are structured to consist of parts interpreted as **Theme** and **Goal**, but usually do not contain more than one

<sup>7</sup><https://wordnet.princeton.edu>

| LLM                      | Lexical units |      |      | Annotation |      |      |
|--------------------------|---------------|------|------|------------|------|------|
|                          | P             | R    | F1   | P          | R    | F1   |
| <b>Claude 3.5 Sonnet</b> | 0.87          | 0.58 | 0.73 | 0.77       | 0.53 | 0.63 |
| <b>Gemini 1.5 Pro</b>    | 0.80          | 0.66 | 0.72 | 0.80       | 0.40 | 0.53 |
| <b>GPT-4o</b>            | 0.87          | 0.53 | 0.66 | 0.86       | 0.48 | 0.61 |
| <b>GPT-4o mini</b>       | 0.58          | 0.87 | 0.69 | 0.50       | 0.75 | 0.60 |

Table 1: Results from the experiment “Augmenting FrameNet semantic frames”).

clause and do not illustrate complex syntactic structures.

The annotation of core frame elements is correct in most cases, as long as the examples represent the given lexical unit and not others with a different meaning, e.g. *The hikers surmounted the peak and gazed out at the breathtaking view* (**Theme:** hikers, **Goal:** peak). Again, the use of simple syntactic structures, which are always explicit, is a shortcoming when it comes to illustrating the varied use of language.

Table 1 shows the calculated results compared to the manual annotation of 20% of the outputs. As mentioned above, these calculations cannot be considered relevant for the evaluation of the LLMs and some of the results are not readily comparable. The manual annotation for lexical units simultaneously takes into account the correct classification of the lexical unit as part of the semantic frame **Arriving** and the correct suggestion of examples; and for the annotation – the correct suggestion of examples and the correct annotation of the core frame elements. With this approach of simultaneous evaluation of two components, the precision and recall values are reduced. For example, the combination of a correct classification of a verb and an example that illustrates a different meaning is scored as a true negative, as is the correct annotation of a core frame element in an inappropriate example.

The partial manual annotation of the LLM output shows that the proposed approach cannot completely replace the manual development of semantic resources or manual annotation: both the resulting verb list, the examples and the annotations have to be manually evaluated and in some cases rejected or re-annotated. However, the experiment shows that the LLMs can be used as a second annotator when enriching the FrameNet with new lexical units (different models can be selected as more suitable for different linguistic tasks). Furthermore, since the annotation with the core frame elements was correct in a large number of examples reflecting the meaning of the verbs, it is possible to suc-

cessfully use LLMs for the automatic enrichment of FrameNet annotations for examples selected by experts. There is also great scope for improving the prompt(s) with instructions on exactly how to annotate the frame elements (e.g. inclusion of modifiers and articles).

## 4.2 Classification of FrameNet lexical units into semantic classes

The second experiment aims to categorise the verbs that evoke a certain semantic frame into semantic classes, in this case aspectual classes (related to situation types, also called eventuality types): *states*, *activities*, *accomplishments* and *achievements*. This classification is relevant for the differentiation of activities, accomplishments and achievements in frames that represent events, as the states in FrameNet are grouped in separate semantic frames.

### 4.2.1 Verb aspectual classes in brief

The categorisation of verbs into aspectual classes is based on the following characteristics (Vendler, 1957): **change** with the values dynamicity and stativity; **temporal extent** with the values durativity and punctuality; **defined endpoint** (homogeneity) with the values telicity and atelicity, and comprises four situation types:

- **states** – continuous stative situations that are atelic;
- **activities** – continuous dynamic situations that are atelic;
- **accomplishments** – continuous dynamic situations that are telic;
- **achievements** – punctual dynamic situations that are telic.

This classification was followed by numerous (more detailed) classifications of situation types, including (Kenny, 2003; Dowty, 1997; Piñón, 1997), which aim to describe in a conventional way the situations that are important for the semantic representation of verbs and their argument structure. The grouping of verbs within a semantic frame ac-

ording to aspectual classes is also reflected in the grouping of the frame elements and the valency patterns associated with the verbs.

The proposed experiment aims to classify verbs that evoke a particular semantic frame into three semantic classes: activities, accomplishments and achievements.

#### 4.2.2 Workflow and prompt design

The identification of semantic classes of verbs is intended to show whether activities, accomplishments and achievements can coexist in semantic frames.

The workflow includes the selection of verbs to be tested from FrameNet, the selection of suitable linguistic tests and the creation of the prompt(s).

The selected verbs are verbs that evoke the semantic frame **Arriving** from FrameNet, as well as verbs that were added by mapping with WordNet and evaluated by an expert (20 in total). Due to their similarity in meaning, they can be conditionally divided into two groups: the group of **arrive** verbs and the group of **approach** verbs.

There are several linguistic tests that have been formulated to distinguish between different aspectual classes of verbs (Dowty, 1997). We have selected tests that are only relevant for distinguishing between activities, accomplishments and achievements:

- The verb occurs with expressions such as *for an hour*, which means that it occurs at any point in the hour. When the result is positive, activities are clearly distinguished from achievements and accomplishments.

- The verb can occur with the verb *finish*. If the result is positive, accomplishments are clearly distinguished from activities and achievements.

- The verb can occur with the verb *stop*. If the result is negative, achievements are clearly distinguished from accomplishments and activities.

After some experiments, the prompt is structured as follows:

*There is a list of verbs:*

“verb”: “definition”;

“verb”: “definition”.

*Give examples with each of the verbs from the list in the following constructions if the examples are grammatically correct.*

“Theme” is verb-ing “Goal” *for an hour*.

“Theme” *finish* verb-ing “Goal”.

“Theme” *stop* verb-ing “Goal”.

The Appendix B illustrates a prompt. The experiment was conducted with the same LLMs.

#### 4.2.3 Results and discussion

The list of verbs for the experiment contains only verbs that can be categorised as activities, accomplishments and achievements. Some models, such as Claude 3.5 Sonnet and Gemini 1.5 Pro, correctly distinguish between two classes: activities/accomplishments (the group of **approach** verbs) and achievements (the group of **arrive** verbs), assuming that activity verbs can co-occur with the *for*-expression and the verb *stop*, accomplishment verbs can co-occur with the verbs *finish* and *stop*, while achievement verbs cannot co-occur with the *for*-expression and the verbs *finish* and *stop*. The output also contains some explanations:

Claude 3.5 Sonnet: **Note:** *Verbs that describe punctual actions (arrive, drive in, enter, get, come) or don’t imply a gradual process (come on, go up) don’t fit well with these constructions in their given meanings.*

Gemini 1.5 Pro: **Approach:** *Indicates movement towards a goal, but doesn’t necessarily mean reaching it. Arrive: Indicates the completion of a journey, reaching the intended destination.*

The other models give examples for all constructions, even if they are not always grammatically correct. All models provide some examples that are correct but illustrate either a different meaning or a different construction.

A clear distinction between activities and accomplishments cannot be made in this experiment.

The group of **approach** verbs is provided with some examples with *for*-expressions that are typically combined with activities, such as:

*The ship approached the island for an hour.* (This implies a slow, gradual approach.)

*The car neared the city for an hour.* (Similar to approach, implies a gradual movement.)

The *for*-expressions, like the *in*-expressions, are treated semantically as a way of measuring the scope of eventualities. An *in*-adverbial measures the time span in which eventualities expressed by telic predicates culminate, while a *for*-adverbial measures the temporal duration of eventualities denoted by atelic predicates (Filip, 2011). In our experiment, the combination with a *for*-expression should signal that the verbs are activities and not accomplishment or achievements. However, there can be shifts between telic and atelic interpretations of a verb and in some cases the *for*-expression

can be combined with accomplishments. In the sentence *The guests **arrived** at the wedding **for an hour***, the *for*-expression means that the guests intended to stay at the wedding for an hour, and the interpretation is telic.

Many linguistic works emphasise that aspectual distinctions are distinctions between linguistic expressions and not properties of events (Rothstein, 2004). For example, accomplishment verbs can differ in their telicity according to the properties of their direct objects (Verkuyl, 1972, 1993). An accomplishment verb is usually the head of a telic verb phrase, but the verb phrase is atelic if the direct object is a bare plural or a mass noun. When a verb is an activity, the properties of the direct object have no influence on the telicity of the verb phrase.

One of the main differences between activities and accomplishments are the **goals** incorporated in the meaning of accomplishments, which is also a feature of the **approach** verbs.

The linguistic tests with the verbs *finish* and *stop* aim to distinguish achievement verbs that do not normally come after the verbs *finish* and *stop* because they do not describe any kind of process. The achievement verbs can occur with *stop* if they only express a habitual, repetitive event, e.g. *The visitors stopped coming to the museum* (every day). In the experiment, examples are regularly given with both *finish* and *stop* with the **approach** verbs:

*The hikers finished drawing close to the summit.*

*The boat stopped coming near the shore.*

To summarise, the results of the experiment show that the verbs in the **approach** group can be defined either as activities or accomplishments, depending on the context in which they are used, while the verbs in the **arrive** group can be defined as achievements.

Different examples, even if they express the same meaning of a verb, may refer to different situation types, so the linguistic tests may or may not work, mainly because they were not constructed for LLMs. While an expert may be able to make a relatively quick decision about a particular example, the task of classifying verbs based on multiple constructions in which they may appear with different meanings should be further refined.

For the reasons mentioned above and because of the nature of the prompt, this experiment cannot be used to evaluate the models. However, the results may lead the expert to make the correct de-

cision based on the examples and interpretations provided by the LLMs. The fact that one and the same semantic frame can be evoked by both activity/accomplishment verbs such as **approach** and achievement verbs such as **arrive** could raise the question of the reorganisation of some frames (the general meaning of the frame **Arriving** is that of achievement). Such an indication could arise from the fact that other verbs that are among the selected potential candidates for the frame **Arriving** are actually part of other frames, e.g. the verb *land* evokes the frame **Vehicle landing** with the definition: “A flying **Vehicle** comes to the ground at a **Goal** in a controlled fashion, typically (but not necessarily) operated by an operator”.

## 5 Conclusions

In general, LLMs cope quite successfully with the linguistic annotation of words or phrases expressing a given FrameNet core frame element, so it can be expected that the automatic annotation of the (core) frame elements for previously selected examples can be successfully performed by LLMs. Furthermore, the use of LLMs as a bank of examples illustrating different linguistic phenomena can be perfected with more specific instructions in the prompts.

Some difficulties in using LLMs for linguistic classification into aspectual classes may arise from the nature of linguistic tests that only work for a particular linguistic context, while there may even be a shift between aspectual classes within a verb meaning for different constructions and contexts. This means not only that linguistic tests need to be further elaborated in order to work with LLMs, but also that a theoretical justification needs to be provided for the contextual conditions under which aspectual class shifts occur, as has already been done for a number of cases in English.

The use of more than one LLM with the same prompts allows the expert to confirm or reject an initial hypothesis, i.e. a second and third annotator may be superfluous. On the other hand, carefully analysing the errors of the LLMs can also help the expert to make one or the other decision.

To summarise, it can be said that the use and importance of LLMs in linguistic work, as in many other areas, will increase. In any case, LLMs are useful to linguists as interlocutors who can surprise them with unexpected linguistic usages.

## Acknowledgments

The present study is carried out within the project *Ontology of Activity Predicates – Linguistic Modelling with a Focus on Bulgarian* funded by the Bulgarian National Science Fund, Grant Agreement No KP–06–H80/9 from 8.12.2023.

## References

- Anthropic. 2023. [The Claude 3 Model Family: Opus, Sonnet, Haikus](#).
- Siddhant Arora, Ankita Pasad, Chung-Ming Chien, Jionghao Han, Roshan Sharma, Jee weon Jung, Hira Dharmyal, William Chen, Suwon Shon, Hung yi Lee, Karen Livescu, and Shinji Watanabe. 2024. [On the Evaluation of Speech Foundation Models for Spoken Language Understanding](#). *ArXiv*.
- Gašper Beguš, Maksymilian Dabkowski, and Ryan Rhodes. 2023. [Large Linguistic Models: Analyzing theoretical linguistic abilities of LLMs](#). *ArXiv*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language Models are Few-Shot Learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Yupeng Cao, Zhi Chen, Qingyun Pei, Fabrizio Dimino, Lorenzo Ausiello, Prashant Kumar, K. P. Subbalakshmi, and Papa Momar Ndiaye. 2024. [RiskLabs: Predicting Financial Risk Using Large Language Model Based on Multi-Sources Data](#). *ArXiv*.
- Yu-Chu Chang, Xu Wang, Jindong Wang, Yuanyi Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Weirong Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qian Yang, and Xingxu Xie. 2023. [A Survey on Evaluation of Large Language Models](#). *ACM Transactions on Intelligent Systems and Technology*, 15:1 – 45.
- 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.
- David R. Dowty. 1997. [Word Meaning and Montague Grammar: The Semantics of Verbs and Times in Generative Semantics and in Montague’s PTQ](#). *Studies in Linguistics and Philosophy*. Springer Dordrecht.
- Hana Filip. 2011. [Lexical aspect](#). In Robert Binnick, editor, *The Oxford Handbook of Tense and Aspect*, pages 721–752. Oxford University Press, Oxford.
- Charles J. Fillmore. 1976. [Frame semantics and the nature of language](#). In *Annals of the New York Academy of Sciences: Conference on the origin and development of language and speech*, volume 280:1, pages 20–32. New York.
- Charles J. Fillmore. 1982. [Frame semantics](#). In *Linguistics in the morning calm*, page 111–137. Hanshin Publishing, Seoul.
- Charles J. Fillmore, Christopher R. Johnson, and Miriam R.L. Petruck. 2003. [Background to FrameNet](#). *International Journal of Lexicography*, 16(3):235–250.
- Gemini Team, Google. 2023. [A Family of Highly Capable Multimodal Models](#).
- Gemini Team, Google. 2024. [Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context](#).
- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. [ChatGPT outperforms crowd workers for text-annotation tasks](#). *Proceedings of the National Academy of Sciences*, 120(30).
- Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, Wentao Han, Minlie Huang, Qin Jin, Yanyan Lan, Yang Liu, Zhiyuan Liu, Zhiwu Lu, Xipeng Qiu, Ruihua Song, Jie Tang, Ji-Rong Wen, Jinhui Yuan, Wayne Xin Zhao, and Jun Zhu. 2021. [Pre-trained models: Past, present and future](#). *AI Open*, 2:225–250.
- Fan Huang, Haewoon Kwak, and Jisun An. 2023. [Is ChatGPT better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech](#). In *Companion Proceedings of the ACM Web Conference 2023, WWW ’23*. ACM.
- Anthony Kenny. 2003. *Action, Emotion and Will*, 2nd edition. Routledge.
- Jan Ole Krugmann and Jochen Hartmann. 2024. [Sentiment Analysis in the Age of Generative AI. Customer Needs and Solutions](#), 11:1–19.
- Taja Kuzman, Igor Mozetič, and Nikola Ljubešić. 2023. [ChatGPT: Beginning of an End of Manual Linguistic Data Annotation? Use Case of Automatic Genre Identification](#). *ArXiv*.
- Xiaofeng Liao and Zhiming Zhao. 2019. [Unsupervised Approaches for Textual Semantic Annotation, A Survey](#). *ACM Computing Surveys*, 52(4):66:1–66:45.
- Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. [Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing](#). *ACM Computing Surveys*, 55(9).

- Kyle Mahowald. 2023. *A Discerning Several Thousand Judgments: GPT-3 Rates the Article + Adjective + Numeral + Noun Construction*. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 265–273, Dubrovnik, Croatia. Association for Computational Linguistics.
- George Miller, R. Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. *Introduction to WordNet: An on-line lexical database*. *International Journal of Lexicography*, 3:235–244.
- Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. *Large Language Models: A Survey*. *ArXiv*.
- OpenAI. 2023. *GPT-4 Technical Report*. *ArXiv*.
- Letian Peng, Zilong Wang, Feng Yao, Zihan Wang, and Jingbo Shang. 2024. *MetalE: Distilling a Meta Model from LLM for All Kinds of Information Extraction Tasks*. *ArXiv*.
- Christopher Piñón. 1997. *Achievements in an event semantics*. *Semantics and Linguistic Theory*, 7:276–293.
- Susan Rothstein. 2004. *Structuring Events: A Study in the Semantics of Lexical Aspect*. Blackwell, Oxford.
- Lei Shi and Rada Mihalcea. 2005. *Putting pieces together: combining framenet, verbnet and wordnet for robust semantic parsing*. In *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing, CICLing’05*, page 100–111, Berlin, Heidelberg. Springer-Verlag.
- Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. *Text Classification via Large Language Models*. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8990–9005, Singapore. Association for Computational Linguistics.
- Tiago Timponi Torrent, Thomas Hoffmann, Arthur Lorenzi Almeida, and Mark Turner. 2024. *Copilots for Linguists: AI, Constructions, and Frames*. Elements in Construction Grammar. Cambridge University Press.
- Satoru Uchida. 2024. *Using early LLMs for corpus linguistics: Examining ChatGPT’s potential and limitations*. *Applied Corpus Linguistics*, 4(1).
- Zeno Vendler. 1957. *Verbs and times*. *The Philosophical Review*, 66:143–160.
- Henk J. Verkuyl. 1972. *On the Compositional Nature of the Aspects*. Foundations of Language Supplementary Series 15. D. Reidel Publishing Company, Dordrecht, Netherlands.
- Henk J. Verkuyl. 1993. *A Theory of Aspectuality: the Interaction between Temporal and Atemporal Structure*. Cambridge Studies in Linguistics; 64. University Press, Cambridge.
- Eric Wallace, Kai Xiao, Reimar Leike, Lilian Weng, Johannes Heidecke, and Alex Beutel. 2024. *The Instruction Hierarchy: Training LLMs to Prioritize Privileged Instructions*. *ArXiv*.
- Haifeng Wang, Jiwei Li, Hua Wu, Eduard Hovy, and Yu Sun. 2022. *Pre-Trained Language Models and Their Applications*. *Engineering*, 25:51 – 65.
- Xiao Wang, Guangyao Chen, Guangwu Qian, Pengcheng Gao, Xiao-Yong Wei, Yaowei Wang, Yonghong Tian, and Wen Gao. 2023. *Large-scale Multi-Modal Pre-trained Models: A Comprehensive Survey*. *ArXiv*.
- Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, Yong Jiang, and Wenjuan Han. 2023. *Zero-Shot Information Extraction via Chatting with ChatGPT*. *ArXiv*.
- Zheng Xin Yong and Tiago Timponi Torrent. 2020. *Semi-supervised Deep Embedded Clustering with Anomaly Detection for Semantic Frame Induction*. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3509–3519, Marseille, France. European Language Resources Association.
- Danni Yu, Luyang Li, Hang Su, and Matteo Fuoli. 2024. *Assessing the potential of AI-assisted pragmatic annotation: The case of apologies*. *ArXiv*.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. *A Survey of Large Language Models*. *ArXiv*.
- Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, Hao Peng, Jianxin Li, Jia Wu, Ziwei Liu, Pengtao Xie, Caiming Xiong, Jian Pei, Philip S. Yu, and Lichao Sun. 2023. *A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT*. *ArXiv*.
- Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. *Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis*. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

## Appendix A Example of a prompt for augmenting FrameNet semantic frames with lexical units and annotations

**Arriving** frame in FrameNet has the definition: “An object **Theme** moves in the direction of a **Goal**. The **Goal** may be expressed or it may be understood from context, but its is always implied by the verb itself.” The core frame elements are: **Goal** defined as “**Goal** is any expression that tells where the **Theme** ends up, or would end up, as a result of the motion” and **Theme** defined as “**Theme** is the object that moves. It may be an entity that moves under its own power, but it need not be”.

There is a list of verbs:

*catch up* (Definition: to reach the point where one should be after a delay)

*scale* (Definition: to reach the highest point of access)

*surmount* (Definition: to reach the highest point of access)

*get at* (Definition: to reach or gain access to)

Which of the verbs in the list evoke the semantic frame “**Arriving**” and which do not; give three examples of each verb that evokes the frame and annotate the core frame elements in the examples. Use this pattern:

“**verb**” (Definition: “**definition**”):

Example 1: “example”. (“core frame element name”: “annotation”, “core frame element name”: “annotation”)

## Appendix B Example of a prompt for classification of FrameNet lexical units into semantic classes

There is a list of verbs:

*approach* (Definition: come near or nearer to in distance)

*arrive* (Definition: reach a destination)

*near* (Definition: come near or nearer to in distance)

*drive in* (Definition: arrive by motorcar)

*draw near* (Definition: come near or nearer to in distance)

*draw close* (Definition: come near or nearer to in distance)

*enter* (Definition: come or go into)

*come on* (Definition: come near or nearer to in distance)

*go up* (Definition: come near or nearer to in distance)

*come near* (Definition: come near or nearer to in distance)

*bear down on* (Definition: sail towards another vessel, of a ship)

*bear down upon* (Definition: sail towards another vessel, of a ship)

*close* (Definition: draw near)

*get* (Definition: reach a destination)

*come* (Definition: reach a destination)

*appear to* (Definition: suddenly arrive)

*find* (Definition: to reach a destination)

*hit* (Definition: to reach a destination or location)

*reach* (Definition: to arrive at)

*return* (Definition: come or go back to a place)

Give examples with each of the verbs from the list in the following constructions only if the examples are grammatically correct.

“Theme” is verb-*ing* “Goal” *for an hour*.

“Theme” *finish* verb-*ing* “Goal”.

“Theme” *stop* verb-*ing* “Goal”.



## LIST OF AUTHORS

|                                  |  |
|----------------------------------|--|
| <b>Atanasov, Atanas</b>          | Sofia University <i>St. Kliment Ohridski</i> , Bulgaria  |
| <b>Badea, Valentin</b>           | Research Institute for AI, Romanian Academy, Romania   |
| <b>Barbu Mititelu, Verginica</b> | Research Institute for Artificial Intelligence <i>Mihai Drăgănescu</i> , Romanian Academy, Romania                 |
| <b>Birtić, Matea</b>             | Institute for the Croatian Language, Croatia   |
| <b>Bojadziev, Andrej</b>         | Sofia University <i>St. Kliment Ohridski</i> , Bulgaria  |
| <b>Brač, Ivana</b>               | Institute for the Croatian Language, Croatia   |
| <b>Bruinsma, Bastiaan</b>        | Chalmers University of Technology, Sweden  |
| <b>Bucur, Ana-Maria</b>          | Interdisciplinary School of Doctoral Studies, University of Bucharest, Romania                                     |
| <b>Chitez, Madalina</b>          | West University of Timisoara, Romania  |
| <b>Dankova, Natalia</b>          | Université du Québec en Outaouais (UQO), Canada  |
| <b>Deda, Albana</b>              | Faculty of History and Philology, University of Tirana, Albania  |
| <b>Derzhanski, Ivan</b>          | Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria                                  |
| <b>Di Bratto, Martina</b>        | Department of Humanities, University of Naples <i>Federico II</i> , Logogramma s.r.l, Italy                        |
| <b>Di Maro, Maria</b>            | Department of Electrical Engineering and Information Technologies, University of Naples <i>Federico II</i> , Italy |
| <b>Dimitrova, Tsvetana</b>       | Institute for Bulgarian Language, Bulgarian Academy of Sciences, Bulgaria  |
| <b>Dinca, Andreea</b>            | West University of Timisoara, Romania  |
| <b>Ferro, Marcello</b>           | National Research Council, Institute for Computational Linguistics, Pisa, Italy                                    |
| <b>Gaftandzhieva, Silvia</b>     | University of Plovdiv <i>Paisii Hilendarski</i> , Bulgaria   |
| <b>Goliakova, Ekaterina</b>      | Université de Lorraine, France   |
| <b>Haddad Haddad, Amal</b>       | Universidad de Granada, Spain  |
| <b>Haveriku, Alba</b>            | Polytechnic University of Tirana, Albania  |
| <b>Ion, Radu</b>                 | Research Institute for Artificial Intelligence <i>Mihai Drăgănescu</i> , Romanian Academy, Romania                 |
| <b>Irimia, Elena</b>             | Research Institute for Artificial Intelligence <i>Mihai Drăgănescu</i> , Romanian Academy, Romania                 |
| <b>Karev, Lyuboslav</b>          | Sofia University <i>St. Kliment Ohridski</i> , Bulgaria  |

|                                 |   |
|---------------------------------|---|
| <b>Khokhlova, Maria</b>         | St. Petersburg State University, Russia   |
| <b>Klints, Agute</b>            | Institute of Mathematics and Computer Science, University of Latvia, Latvia   |
| <b>Koeva, Svetla</b>            | Institute for Bulgarian Language, Bulgarian Academy of Sciences, Bulgaria   |
| <b>Koryshev, Mikhail</b>        | St. Petersburg State University, Russia   |
| <b>Kote, Nelda</b>              | Polytechnic University of Tirana, Albania   |
| <b>Koychev, Ivan</b>            | Faculty of Mathematics and Informatics, Sofia University <i>St. Kliment Ohridski</i> , Bulgaria                             |
| <b>Krlev, Jordan</b>            | Technical University of Sofia, Bulgaria   |
| <b>Kukova, Hristina</b>         | Institute for Bulgarian Language, Bulgarian Academy of Sciences, Bulgaria   |
| <b>Kurshumova, Denitza</b>      | University of Plovdiv <i>Paisii Hilendarski</i> , Bulgaria  |
| <b>Langlois, David</b>          | LORIA, Université de Lorraine, France   |
| <b>Lavigne, Camille</b>         | Université de Nancy, France   |
| <b>Lejeune, Gaël</b>            | Sorbonne University, France   |
| <b>Lento, Alessandro</b>        | Biomedical Campus University, Rome, Italy   |
| <b>Leseva, Svetlozara</b>       | Institute for Bulgarian Language, Bulgarian Academy of Sciences, Bulgaria   |
| <b>Lokmane, Ilze</b>            | Institute of Mathematics and Computer Science, University of Latvia, Latvia   |
| <b>Maion, Fabio</b>             | University of Innsbruck, Austria  |
| <b>Margova, Ruslana</b>         | Sofia University <i>St. Kliment Ohridski</i> , GATE Institute, Bulgaria   |
| <b>Marzi, Claudia</b>           | National Research Council, Institute for Computational Linguistics, Pisa, Italy   |
| <b>Mece, Elinda Kajo</b>        | Faculty of Information Technology, Polytechnic University of Tirana, Albania  |
| <b>Mennella, Sabrina</b>        | Department of Humanities, University of Catania; URBAN/ECO Research Center, University of Naples <i>Federico II</i> , Italy |
| <b>Morita, Junya</b>            | Kinjo Gakuin University, Japan  |
| <b>Nadalini, Andrea</b>         | National Research Council, Institute for Computational Linguistics, Pisa, Italy   |
| <b>Nešić, Milica Ikonić</b>     | Faculty of Philology, University of Belgrade, Serbia  |
| <b>Nešpore-Bērzkalne, Gunta</b> | Institute of Mathematics and Computer Science, University of Latvia, Latvia   |
| <b>Nikolova-Stoupak, Iglīka</b> | Sorbonne University, France   |

|                                 |  |
|---------------------------------|--|
| <b>Nivre, Joakim</b>            | Uppsala University and RISE, Sweden  |
| <b>Occhipinti, Laura</b>        | University of Bologna, Italy   |
| <b>Osenova, Petya</b>           | Sofia University <i>St. Kliment Ohridski</i> and IICT, Bulgarian Academy of Sciences, Bulgaria     |
| <b>Paikens, Peteris</b>         | University of Latvia, Latvia   |
| <b>Pashev, Georgi</b>           | University of Plovdiv <i>Paisii Hilendarski</i> , Bulgaria   |
| <b>Penkov, Stanislav</b>        | Faculty of Slavic Studies, Sofia University <i>St. Kliment Ohridski</i> , Bulgaria                 |
| <b>Petalinkar, Saša</b>         | University of Belgrade, Serbia   |
| <b>Pirrelli, Vito</b>           | National Research Council, Institute for Computational Linguistics, Pisa, Italy                    |
| <b>Premasiri, Damith</b>        | Lancaster University, United Kingdom   |
| <b>Păiș, Vasile</b>             | Research Institute for Artificial Intelligence <i>Mihai Drăgănescu</i> , Romanian Academy, Romania |
| <b>Rakiplari, Elsa Skënderi</b> | Faculty of History and Philology, University of Tirana, Albania                                    |
| <b>Rituma, Laura</b>            | Institute of Mathematics and Computer Science, University of Latvia, Latvia                        |
| <b>Rogobete, Roxana</b>         | West University of Timisoara, Romania  |
| <b>Rujević, Biljana</b>         | Faculty of Mining and Geology, University of Belgrade, Serbia                                      |
| <b>Rushiti, Rozana</b>          | Faculty of History and Philology, University of Tirana, Albania                                    |
| <b>Schaeffer-Lacroix, Eva</b>   | Sorbonne University, France  |
| <b>Siruk, Olena</b>             | Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria                  |
| <b>Stade, Madara</b>            | Faculty of Humanities, University of Latvia, Latvia  |
| <b>Stanković, Ranka</b>         | Faculty of Mining and Geology, University of Belgrade, Serbia                                      |
| <b>Stasica, Alex</b>            | Utrecht University, Netherlands  |
| <b>Stefanova, Valentina</b>     | Institute for Bulgarian Language, Bulgarian Academy of Sciences, Bulgaria                          |
| <b>Stoyanov, Veselin</b>        | Tome AI, USA   |
| <b>Stoyanova, Ivelina</b>       | Institute for Bulgarian Language, Bulgarian Academy of Sciences, Bulgaria                          |
| <b>Tarpomanova, Ekaterina</b>   | Sofia University <i>St. Kliment Ohridski</i> , Bulgaria  |
| <b>Todorova, Maria</b>          | Institute for Bulgarian Language, Bulgarian Academy of Sciences, Bulgaria                          |

|                             |   |
|-----------------------------|---|
| <b>Todorova, Maria</b>      | Sofia University <i>St. Kliment Ohridski</i> , Bulgaria   |
| <b>Trandafili, Evis</b>     | Polytechnic University of Tirana, Albania   |
| <b>Tsonkov, Todor</b>       | Faculty of Mathematics and Informatics, Sofia University <i>St. Kliment Ohridski</i> , Bulgaria |
| <b>Voicu, Tudor</b>         | Tudor Vianu High School, Romania  |
| <b>Xhanari, Lindita</b>     | Faculty of History and Philology, University of Tirana, Albania                                 |
| <b>Zmiycharov, Valentin</b> | Faculty of Mathematics and Informatics, Sofia University <i>St. Kliment Ohridski</i> , Bulgaria |
| <b>Çepani, Anila</b>        | Faculty of History and Philology, University of Tirana, Albania                                 |
| <b>Škorić, Mihailo</b>      | Faculty of Mining and Geology, University of Belgrade, Serbia                                   |



## Organised by:



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

## With the support of:



NATIONAL  
SCIENCE  
FUND



SoftUni



GATE  
big data for smart society



**ISSN: 2367-5675**