Institute for Bulgarian Language

BULGARIAN ACADEMY OF SCIENCES

Department of Computational Linguistics

# Computational Linguistics in Bulgaria

**Volume** 1, **Issue** 1 • 2025

**Editor-in-Chief:** Prof. Dr. Svetla Koeva

Sofia, Bulgaria

**PUBLICATION AND CATALOGUING INFORMATION**

| | |
|---|---|
| Journal title: | *Computational Linguistics in Bulgaria* |
| Volume and issue: | Volume 1, Issue 1, 2025 |
| ISSN: | 3033-1382 (print) / 3033-2397 (online) |
| Published and distributed by: | Institute for Bulgarian Language<br>Bulgarian Academy of Sciences<br>Department of Computational Linguistics |
| Editorial address: | Institute for Bulgarian Language<br>Bulgarian Academy of Sciences<br>Department of Computational Linguistics<br>52 Shipchenski Prohod Blvd., Bldg. 17<br>Sofia 1113, Bulgaria<br>Phone/Fax: +359 2/ 872 23 02 |
| Copyright: | The copyright of each article published in the journal belongs to its author(s).<br><br>All content is licensed under a Creative Commons Attribution 4.0 International Licence (CC BY 4.0).<br><br>License details:<br>http://creativecommons.org/licenses/by/4.0<br><br>Cover artwork by Bozhidar Chemshirov.<br><br>© 2025 Institute for Bulgarian Language, Bulgarian Academy of Sciences, Department of Computational Linguistics |

## Table of contents

# Computational Linguistics in Bulgaria

Svetla Koeva
Department of Computational
Linguistics,
Institute for Bulgarian Language,
Bulgarian Academy of Sciences
52 Shipchenski prohod, Bldg. 17,
1113 Sofia, Bulgaria
svetlal@dcl.bas.bg

*The article introduces the journal **Computational Linguistics in Bulgaria**, an annual open access peer-reviewed journal published by the Department of Computational Linguistics at the Institute for Bulgarian Language of the Bulgarian Academy of Sciences. The relationship between the terms computational linguistics, natural language processing and artificial intelligence is briefly commented on in order to clarify the concept behind the journal's name. The focus is then placed on the Bulgarian language and the Bulgarian research community, emphasising the importance of international contributions for the development of scientific cooperation and progress.*

*The scope of the journal Computational Linguistics in Bulgaria is presented: It publishes articles on all areas of theoretical computational linguistics as well as on existing language resources, datasets and technologies for natural language processing and artificial intelligence. The journal promotes new approaches and methods, especially those aimed at applying language technologies to small and still resource-poor languages such as Bulgarian.*

**Keywords**: *computational linguistics, natural language processing, artificial intelligence, Computational Linguistics in Bulgaria*

## 1. Introduction

The first issue of the ***Computational Linguistics in Bulgaria* journal** (JCLIB),[1] an annual open-access peer-reviewed journal, is published by the Department of Computational Linguistics at the Institute for Bulgarian Language of the Bulgarian Academy of Sciences. The editorial policy of the *Computational Linguistics in Bulgaria* journal includes the publication of articles from all areas of **theoretical computational linguistic**s in combination with **available language resources, datasets and technologies for natural language processing and artificial intelligence**. The focus is on new approaches and methods, especially with

1 https://jclib.dcl.bas.bg/

https://doi.org/10.47810/JCLIB.1.2025.01

regard to their application to **small and resource-poor languages such as Bulgarian**, in order to bridge the gap between large and small languages in terms of language technologies.

The idea for the journal is not new, it was born together with the idea for the biennial conference *Computational Linguistics in Bulgaria*, organised by the Department of Computational Linguistics at the Institute for Bulgarian Language of the Bulgarian Academy of Sciences, which started in 2014. However, there were many objective and subjective reasons that prevented the publication of the journal. The Scientific Council of the Institute for Bulgarian Language of the Bulgarian Academy of Sciences has decided to launch the journal on **7 November 2024**.

The aim of the publisher and the editorial board is to make the journal a recognised forum for the publication of scientific research in the field of computational linguistics, natural language processing and artificial intelligence, with a focus on the Bulgarian language, which is either the direct subject of the studies and applications or whose research could be significantly influenced in the future by a variety of innovative developments. It goes without saying that both young and established researchers from Bulgaria and abroad, as well as outstanding researchers who can contribute to significant advances in the field of computational linguistics, natural language processing and artificial intelligence, are welcome as authors.

Since the terms computational linguistics, natural language processing and artificial intelligence appear several times in this text, it is worth clarifying what we mean by each of them, outlining their specific area of application and explaining how they relate to each other and overlap.

## 2. Computational Linguistics, Natural Language Processing and Artificial Intelligence

Several terms (and corresponding concepts) are used to describe related areas of research and development: computational linguistics (CL), natural language processing (NLP), language engineering (LE), human language technology (HLT), language technology (LT), artificial intelligence (AI), etc. However, we will only focus on some of them.

If we compare the terms computational linguistics and natural language processing, the word *linguistics* in the first term refers to the scientific discipline and, with its modifier, forms a term for a new specific field of research, namely computational linguistics. In contrast, the word *processing* from the term *natural language processing* refers to the processing of certain data, in this case natural language.

There are many definitions for both terms, computational linguistics and natural language processing, which show a different understanding of their content: In different interpretations, the terms can overlap, subsume or refer to related but nevertheless different concepts.

For example, there is a narrow understanding of computational linguistics that implies that computational linguistics provides sophisticated methods for linguistic research. In the modern development of technologies, this is a more appropriate understanding of theoretical linguistics itself, which uses various language data analyses to prove or reject theoretical linguistic hypotheses.

Another, more widespread view of computational linguistics is that computational linguistics seeks to define how humans compute and produce language by formulating formal grammars and probabilistic models and developing efficient algorithms for machine learning, language generation and understanding that appear suitable for capturing the range of phenomena in human languages.

Almost 16 years ago, the following definition was made (Calzolari 2009):

> *The term CL includes the disciplines dealing with models, methods, technologies, systems and applications concerning the automatic processing of a language, both spoken and written. CL therefore includes both Speech Processing (or processing of the spoken word, and Natural Language Processing (NLP) or text processing. SP and NLP have closely linked objectives such as human-machine vocal interaction and human language understanding, to be used in many applications, such as machine translation, speech-to-speech translation, information retrieval, and so on.*

Another definition of computational linguistics, published in 2020, states that (Schubert 2020):

> *Computational linguistics is the scientific and engineering discipline concerned with understanding written and spoken language from a computational perspective, and building artifacts that usefully process and produce language, either in bulk or in a dialogue setting. To the extent that language is a mirror of mind, a computational understanding of language also provides insight into thinking.*

Some authors emphasise that it is difficult to distinguish between computational linguistics and natural language processing (Hirschberg and Manning 2015, 261):

> ***Computational linguistics, also known as natural language processing (NLP)****, is the subfield of computer science concerned with using computational techniques to learn, understand, and produce human language content. Computational linguistic systems can have multiple purposes: The goal can be aiding human-human communication, such as in machine translation (MT); aiding human-machine communication, such as with conversational agents; or benefiting both humans and machines by analyzing and learning from the enormous quantity of human language content that is now available online.*

Many more definitions could be given, but we can summarise that computational linguistics is concerned with the theoretical modelling and formal description of language, while natural language processing applies theoretical investigations to solve real language problems in the context of interaction with computers. The terms are often used interchangeably because theory and application are inherently interdependent – neither can exist without the other.

To complicate things further, let us briefly examine how computational linguistics and natural language processing relate to artificial intelligence. In many views, the first two (or at least natural language processing) are areas of artificial intelligence (Navigli 2018, 5697):

> *Natural Language Processing (NLP) is a challenging field of Artificial Intelligence which is aimed at addressing the issue of automatically processing human language, called natural language, in written form. This is to be achieved by way of the automatic analysis, understanding and generation of language.*

Cole Stryker and Jim Holdsworth have posted similar thoughts:[2]

> *Natural language processing (NLP) is a subfield of computer science and artificial intelligence (AI) that uses machine learning to enable computers to understand and communicate with human language. NLP enables computers and digital devices to recognize, understand and generate text and speech by combining computational linguistics, the rule-based modeling of human language together with statistical modeling, machine learning and deep learning. NLP research has helped enable the era of generative AI, from the communication skills of large language models (LLMs) to the ability of image generation models to understand requests.*

One of the most frequently cited definitions of artificial intelligence is that of John MacCarthy, which was given several decades ago and revised in 2007 (McCarthy 2007, 2):

> Artificial intelligence *is the science and engineering of making intelligent machines, especially intelligent computer programs.*

(Human) intelligence is not just about a person's ability to learn and use language. There is ample evidence from various fields where machines successfully (in some cases even better than humans) perform tasks that require human intelligence. In road transport, for example, self-driving cars are no longer experiments: Waymo, one of the largest providers in the US, offers more than 150,000 autonomous rides every week, while Baidu's affordable Apollo Go robotaxi fleet now serves numerous cities in China (Maslej et al. 2025, 156).

When we turn to the question of the relationship between computational linguistics, natural language processing and artificial intelligence, it may be helpful to use the categorisation of artificial intelligence systems into four basic design principles: acting humanly, thinking humanly, thinking rationally and acting rationally (Russell and Norvig 2022, 19-20):

> *Historically, researchers have pursued several different versions of AI. Some have defined intelligence in terms of fidelity to human performance, while others prefer an abstract, formal definition of intelligence called rationality – loosely speaking, doing the "right thing". The subject matter itself also varies: some consider intelligence to be a property of internal thought processes and reasoning, while others focus on intelligent behavior, an external characterization. From these two dimensions – human vs. rational and thought vs. behavior – there are four possible combinations, and there have been adherents and research programs for all four.*

2 https://www.ibm.com/think/topics/natural-language-processing

It is also pointed out that the rationalist approach to artificial intelligence involves a combination of mathematics and engineering and is associated with statistics and control theory (Russell and Norvig 2022, 20). However, when it comes to human language, it has not been possible to avoid (computational) linguistics.

Even if we restrict ourselves to the fact that artificial intelligence is currently mainly concerned with acting rationally, the activities related to human language are a subset of the field artificial intelligence, so we should agree that computational linguistics (natural language processing) is part of artificial intelligence understood in this way.

### 3. Computational Linguistics in Bulgaria

The connection between the journal ***Computational Linguistics in Bulgaria*** and the conference ***Computational Linguistics in Bulgaria*** is close, although the conference publishes its own proceedings. The *Computational Linguistics in Bulgaria* (CLIB)[3] conference is an international event with the aim of exploring new approaches and methods in computational linguistics and natural language processing, especially with regard to their application to small and less well-resourced languages such as Bulgarian, and bridging the gap between "large" and "small" languages in terms of language technologies.

Original contributions on the following topics are expected at the CLIB conference (and also in the journal): computer-assisted language learning, training and education; information retrieval; information extraction; text mining and knowledge graph inference; linguistic foundations for computer vision and robotics; language modelling; language theories and cognitive modelling for NLP; large language models and NLP evaluation methods; language resources and benchmarking for large language models; language resource construction and annotation; machine learning for NLP; machine translation; multilingualism; translation aids; morphology and segmentation; natural language generation, understanding, summarisation and simplification; ontologies, terminology and knowledge representation; sentiment analysis; authorship analysis; opinion and argumentation analysis; speech recognition, synthesis and understanding of spoken language; tagging, chunking, syntax and parsing; and other related topics.

The part "in Bulgaria" in the names of the journal and the conference means several things:

- The journal is published in **Bulgaria**, the conference takes place in **Bulgaria**.
- The focus of the journal (as well as the conference) is on the **Bulgarian language** in the broadest sense: computational linguistic research in Bulgarian, but also datasets, models, technologies for other languages that can be newly implemented or adopted for Bulgarian. Of course, outstanding achievements in the field of computational linguistics, which can influence not only computational linguistic research in Bulgarian, but also the whole scientific field

3 https://dcl.bas.bg/clib/

of computational linguistics and natural language processing, are extremely important and of particular interest to the journal and the conference.

- The aim of the journal and the conference is to connect **Bulgarian researchers** at home and abroad and to promote the exchange of ideas, resources and successes. Both forums welcome research contributions from all over the world and promote scientific communication, collaboration and mutual support in the knowledge that research thrives on exchange and joint endeavour.

## 4. Journal *Computational Linguistics in Bulgaria*

It has already been mentioned that the journal ***Computational Linguistics in Bulgaria*** publishes research papers from all areas of theoretical computational linguistics as well as studies on existing resources, datasets and technologies for natural language processing and artificial intelligence. The journal emphasises innovative approaches and methods, especially those aimed at the application of language technologies to small and resource-poor languages such as Bulgarian, with the overarching goal of narrowing the gap between large and small languages in the development and accessibility of language technologies.

The journal accepts submissions of original research with a particular focus on Bulgarian and other related languages, but also welcomes contributions on new theories, datasets and technologies applicable to a wide range of languages.

Finally, we must mention some formal features of the journal *Computational Linguistics in Bulgaria:*

It is an **open access** journal, which means that the full text of all articles is freely available on the Internet so that users can read, download, copy, distribute, print, search, link or index the content and use it as data in software or for any other lawful purpose – without financial, legal or technical barriers beyond Internet access.[4]

The research results published in the journal are in the public domain and may be used under the terms of a Creative Commons Attribution 4.0 International Public Licence (**CC-BY-4.0**).[5] Anyone is free to share – copy, distribute and transmit, remix – adapt the work, under the condition of attribution – the original authors must be credited.

The journal *Computational Linguistics in Bulgaria* adheres to the ethical guidelines for journal publications of the **Committee on Publication Ethics** (COPE).[6]

To summarise, by combining research on Bulgarian and related languages with global perspectives, we aim to create a space where ideas thrive, innovation flourishes and language technologies bridge the gap between resource-rich and resource-poor languages.

4 https://www.budapestopenaccessinitiative.org/
5 https://creativecommons.org/licenses/by/4.0/
6 https://publicationethics.org/

## References

Calzolari, Nicoletta. 2009. *Encyclopedia of Life Support Systems (EOLSS)*, volume 1, chapter Linguistics. Computational Linguistics. EOLSS Publishers, Paris, France. Developed under the Auspices of the UNESCO.

Hirschberg, Julia and Christopher D. Manning. 2015. Advances in natural language processing. *Science*, 349(6245):261–266.

Maslej, Nestor, Loredana Fattorini, Raymond Perrault, Yolanda Gil, Vanessa Parli, Njenga Kariuki, Emily Capstick, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, Toby Walsh, Armin Hamrah, Lapo Santarlasci, Julia Betts Lotufo, Alexandra Rome, Andrew Shi, and Sukrut Oak. 2025. *The AI Index 2025 Annual Report.* AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA. https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf.

McCarthy, John. 2007. *What is artificial intelligence?* Computer Science Department, Stanford University.

Russell, Stuart J. and Peter Norvig. 2022. *Artificial Intelligence: A Modern Approach*, 4 edition. Pearson.

Schubert, Lenhart. 2020. Computational Linguistics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Spring 2020 edition. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2020/entries/computational-linguistics/.

# Does ChatGPT Adapt Itself to the Language Used and the Audience It Implies?


Iglika Nikolova Stoupak
Sens Texte Informatique Histoire
Sorbonne Université, Paris, France
`iglika.nikolova-stoupak@etu.sorbonne-universite.fr`

Gaël Lejeune
Sens Texte Informatique Histoire
Sorbonne Université, Paris, France
`gael.lejeune@sorbonne-universite.fr`

Eva Schaeffer-Lacroix
Sens Texte Informatique Histoire
Sorbonne Université, Paris, France
`eva.lacroix@inspe-paris.fr`



*This paper seeks to quantify and analyse the progress that ChatGPT has made from its GPT-3.5 (2022) to its GPT-4.5 (2025) version when it comes to answering prompts in a selection of differently-resourced languages: English, Bulgarian, Greek, French, Hebrew, Japanese and Russian. Factual correctness, textual quality and an answer's linguistic and cultural independence from an English baseline are evaluated in the process. Each response is marked positively or negatively for each of the three metrics based on a set of defined criteria and careful human-based analysis. In addition, three categories of questions are experimented with: general (e.g. communication assistance or request for jokes), perception-related (e.g. creative writing or explanation of physical processes) and geography-/culture-sensitive (questions in a specific language that address a particular, slightly sensitive topic related to the implied audience e.g. 'Why do French people eat snails?'). As hypothesised, the recent GPT-4.5 version demonstrates significant progress in all evaluated categories, thereby resolving past issues such as decreased textual quality of low-resourced languages and, notably, very limited variety in answers to the same question across languages. The metric 'Independence from the (English) Baseline' receives 80.95% of positive marks in the GPT-4.5 version as opposed to 26.19% for GPT-3.5. Lingering problems include ChatGPT's incomplete ability to generate relevant and culturally-sensitive jokes and poems.*



**Keywords**: *ChatGPT, GPT-3.5, GPT-4.5, multilingualism, bias*


## 1. Introduction

OpenAI's ChatGPT barely needs introduction as of today. Appreciated by AI specialists and the general public alike, the chatbot assists internet users from all around the world in a number of tasks from social advice to academic writing and coding. It is based on the GPT (Generative Pre-trained Transformer) model, which achieves state-of-the-art performance


https://doi.org/10.47810/JCLIB.1.2025.02

in a number of natural language processing (NLP) tasks. ChatGPT is user-friendly with its strong zero-shot capabilities and, due to the incorporation of Reinforcement learning from Human Feedback in its training stages, its output highly resembles human language. On the other hand, ChatGPT's limitations as expressed in research since its release in 2022 include lower functionality in low-resourced languages as well as the presence of sometimes concerning biases in output. Since the system's conception, serious steps have been taken to eliminate these problems as well as to increase ChatGPT's overall quality.

The GPT-4 family, trained on a larger dataset than the initial GPT-3.5, comes with an increased general ability to understand and generate text (Caramancion 2023). It has been noted to provide improved answers to math- and logic-related questions (Plevris, Papazafeiropoulos, and Rios 2023). Whilst the GPT-3 family is, in general, competitive with state-of-the-art language models fine-tuned for specific tasks, its GPT-4 descendant outperforms most state-of-the-art systems (Brown et al. 2020; OpenAI et al. 2024a). Similarly, the latter's scores on a simulated bar exam fall within the top 10% for human test takers, compared to the bottom 10% for GPT-3.5 (OpenAI et al. 2024a).

In light of this raising of the bar, the time might be right to move beyond the initial concerns of textual errors in low-resourced languages and outright harmful content and extend the focus onto the system's more subtle malleability. Language implies underlying culture, geography and history and, ideally, ChatGPT should be expected to respond in accordance with this user background. For instance, lengthy and English-based formulas of politeness might not sound natural to a Japanese speaker, passive constructions may impede reader comprehension in Slavic languages, and a translation or explanation of an already French-based culinary term may be rather unnecessary for a French audience.

In an attempt to test ChatGPT's success and progress in adapting to its multilingual audience, a selection of languages with significantly different resource availability was opted for. See Table 1 for a rough overview of each selected language's online prevalence as based on the existent number of Wikipedia pages written in it[1]. A variety of geographical and cultural characteristics related to the languages' native speakers was also sought. Two Slavic languages were deliberately opted for: Russian and Bulgarian. The fact that the former is significantly more highly-resourced than the latter allows for the detection of possible interference as well as for conclusions to be drawn about the relative importance of language resourcedness versus language similarity within the proposed experiments.

In order to evaluate ChatGPT's full up-to-date progress, we will compare the performance of its first and most recent versions; respectively, GPT-3.5 and GPT-4.5. As per its system card, GPT-4.5 builds upon the previous model, GPT-4o, while being more general-purpose in nature. Its underlined strenghts include 'alignment with user intent' and 'improved emotional intelligence' (OpenAI 2025). The model is currently limited to Plus and Pro subscription plans and a quota of around 50 messages per week is attributed in the former.

Whilst the focus of the current study is specifically ChatGPT's ability to adapt to its user as implied by the prompt's language, we are using the opportunity to coincidentally

1 as per `https://meta.wikimedia.org/wiki/List_of_Wikipedias`

| Language Code | Number of Pages |
|---|---:|
| EN | 62,907,668 |
| FR | 13,440,180 |
| RU | 8,246,967 |
| JA | 4,279,143 |
| HE | 1,579,325 |
| EL | 716,967 |
| BG | 684,829 |

**Table 1**
Number of Wikipedia pages per utilised language

evaluate the model's performance and progress in more traditional aspects, such as factual correctness and textual quality, especially given evidence that low-resourced languages have been associated with output of lower quality due to an involved process of translation (Zhang et al. 2023).

Based on ChatGPT's specifications as per OpenAI and relevant academic research, it is hypothesised that ChatGPT's recent version will perform significantly better across languages and across the different question categories. The following specific research questions are brought forward in relation to the experiments' results:

1. Is subordinate multilingualism (characterised with prior translation) less prominent in the case of GPT-4.5?

2. Does ChatGPT's GPT-4.5 version provide a higher proportion of language-/culture-sensitive answers? If yes, to what extent?

3. Does ChatGPT provide a different degree of malleability with regard to the different defined question categories?

4. Are there discernible tendencies in relation to specific languages (e.g. as characterised by language family or script)? In particular, is there interference from the higher- to the lower-resourced Slavic language involved?

## 2. Background

Although to our knowledge, no extensive research has addressed the specific ability of ChatGPT to adapt itself to its audience as demonstrated by the prompt language, it is worth examining the chatbot's performance in two directly related aspects: biases and multilingual output.

### 2.1 ChatGPT and Biases

Since the launch of ChatGPT, ethical issues have been raised in relation to detected biases based on gender, race, religion, occupation, etc., likely caused by existing biases in the utilised training data. For instance, Plevris et al (2023) discover that the word 'Black' is associated with a consistently low sentiment in the model's output. Deshpande et al (2023) note the model's increased tendency to incorporate stereotypes when prompted to act as a specific persona. In their extensive study on ChatGPT's robustness and ethics, Vidhya et al (2023) examine the phenomenon of 'jailbreak' i.e. the model's tendency to bypass ethical norms as a result of specific prompt engineering techniques. Possibly controversial output is detected in several languages, for instance in relation to contested territory (Vidhya et al. 2023). Rozado discusses political and demographic biases as displayed by ChatGPT and other large language models (LLMs), such as their failure to always flag negative content accordingly (Rozado 2023).

The OpenAI team set out to work towards removing the existing biases 'in a holistic manner' (Brown et al. 2020). Their measures, manifested in the GPT-4 model family, include alignment to human preferences during post-training (OpenAI et al. 2024b), training the model for refusals, and the assembly of a red team of experts that monitors the model's activity (OpenAI et al. 2024a). However, biases and the bypassing of ethical norms continue to represent a problem. For example, in an instance of jailbreak, ChatGPT is led to propose an antisemitic comment that would not trigger flagging on Twitter (OpenAI et al. 2024a). To go further, the activity of the red team itself, which is mostly made up of English speakers from Western countries, may lead to the favouring of specific opinions and worldviews (OpenAI et al. 2024a).

In their detailed study, Puttaparthi et al (2023) research the effect of multilingual wrapping of prompts on the probability of ChatGPT jailbreak. 5.07% of questions asked in a single language lead to jailbreak (none of them being in English). Out of the remaining questions, a further 3.21% cause the model to fall into the trap when reformulated in a mixture of languages. In a final experiment, the yet remaining questions are repeated, instructing ChatGPT to provide output in a language not present in the prompt, leading to 1.61% of jailbreak. Furthermore, the phenomenon's probability increases significantly when prompt injections[2] are used. Puttaparthi et al also note that the process of multilingual wrapping tends to impede the model's comprehension and leads to output of limited quality.

### 2.2 ChatGPT and Multilingualism

ChatGPT's multilingual abilities and the limitations therein also come as a frequent subject of discussion. The GPT-3.5 model is trained almost exclusively on high-quality English language data from the Common Crawl, text in other languages accounting for only 7% of the data (Brown et al. 2020). Still, ChatGPT has been noted to outperform earlier LLMs such as T5 and BERT across a variety of NLP tasks when diverse languages, including low-resourced ones, are concerned (Lai et al. 2023). However, the model's performance sharply deteriorates in the presence of extremely low-resourced languages, such as Buginese. It has also been noted that while ChatGPT exhibits comprehension in some rare languages, it struggles to

2 a cybersecurity exploit that seeks to confuse a model by providing it with both a legitimate prompt and a request to ignore it and offer specific, different output

identify the language itself (Bang et al. 2023). Peng et al (2023) discover a higher number of hallucinations in machine translation (MT) output when low-resourced languages as well as distant language pairs are involved. The model's limitations in multilingual performance also impact tasks that are tightly associated with ethics. For instance, Das et al (2023) evaluate ChatGPT's ability to detect hate speech within input in 11 discrete languages and detect weaknesses in the model's identification of irony as well as the differentiation of protected target groups within non-English languages.

As is the case with the majority of the model's functions, ChatGPT demonstrates significantly increased multilingual performance in its GPT-4 version. Manakhimova et al (2023) test the model's performance in MT in the following language pairs: German-English, English-German and English-Russian. Whilst the pair that involves Russian (a language linguistically remote from the other two) performs worst and issues such as the interpretation of idioms are still detected, the results are comparable to those of state-of-the-art systems. Jiao et al (2023) research ChatGPT's translation ability in 101 languages based on both its GPT-3.5 and GPT-4 versions. When it comes to the former, the gap between high- and low-resourced languages is claimed to be drastic, although it can be partially mitigated through techniques such as pivot-prompting (i.e. explicitly asking the model to translate into a higher-resourced language as an additional step). In contrast, the more recent GPT version is associated with strikingly higher BLUE scores, and its performance is deemed to be of sufficiently high quality even when low-resourced languages are concerned. With GPT-4, languages such as Latvian, Welsh and Swahili show strong performance on the MMLU benchmark of multiple-choice questions on 57 subjects (OpenAI et al. 2024a). In turn, GPT-4o demonstrates improved reading comprehension and reasoning abilities in historically underrepresented languages, significantly narrowing their gap with English (OpenAI et al. 2024b).

A specific study that is worth mentioning in view of the current research is Zhang et al's (2023) comprehensive investigation of the type of bilingualism exhibited by ChatGPT when it provides non-English output (assuming that English is the model's 'native language'). They bring forward the following terms as drafted by Marcos (1976): coordinate bilingualism (wherein one's lexicons for each language are associated with discrete mental images) versus subordinate bilingualism (wherein translation occurs into one's main language prior to textual production). They experiment with three prompt categories which imply different degrees of impact of the language involved: Reasoning, Knowledge Access, and Articulation. The last category is defined as 'translation variant'; that is to say, different output in terms of content is to be expected in different languages[3]. Produced answers in non-English languages in this category are discovered to be very similar to their English counterparts, as per the cosine distance of their BERT embeddings (following initial automatic translation of the non-English text). Overall, ChatGPT is concluded to exhibit a mixture of coordinate and subordinate bilingualism. The authors also note that as a result of underlying translation, 'an errorprone process', the accuracy of output achieved through subordinate bilingualism suffers deterioration (Zhang et al. 2023). The question remains of whether later versions

3 An example is the composition of a cover letter, which is affected by linguistic and societal norms.

of ChatGPT have led to a shift or even a qualitative change in the described multilingual abilities.

## 3. Methods

ChatGPT as per its GPT-3.5 and GPT-4.5 versions was asked to provide answers to a series of prompts in a zero-shot setting. A new session was started following each output in order to avoid the system gaining knowledge of the fact that an academic experiment was being performed and/or that its interlocutor was using multiple languages. 'Incognito' mode was not made use of for the following reason: experiments for the current research commenced in 2023, when ChatGPT did not typically demonstrate any knowledge of the user's background or prior conversations once a new conversation was started. The system was accessed via its web-based direct chat interface, through a 'Plus' subscription plan.

### 3.1 Question Types

Different types of questions were employed (see Table 3.1), calling for different degrees of sensitivity to the language and culture at hand within the output. Coincidentally, different tasks and language registers were also implied.

Question type 1 involves general, frequently encountered questions, linked to help with writing and planning, social interaction and entertainment. Social norms and accepted levels of politeness are dependent upon historical phenomena, such as a country's relationships with industrialisation and occupational self-direction (Schooler 1996). Humour in turn has been proven to be both universal and culture-specific. For instance, Jiang et al (2019) note that Eastern cultures are less receptive to aggressive humour compared to Western ones and that, in particular, humour is often used as a coping social device in Japan.

Type 2, broadly named 'perception-related', touches on aspects of language that are directly related to one's perception of the world, including spacial orientation, climatic specificity, agency, colour perception and gender representation. Various theories exist on the question of whether or to what extent language determines thought. As per the Sapir-Whorf hypothesis, language at the very least mediates or influences perception (Fulga 2012). For instance, colours interrelate with metaphorical extensions (van Leeuwen 2010), and the direction of writing in one's native language affects the response to and production of movement (Boroditsky 2011). Furthermore, the linguistic representation of time has been claimed to strongly affect one's thought process (Fulga 2012).

Finally, question type 3 includes topics and concerns that are specifically relevant to an audience from a particular country or countries. The topics were selected to be emotionally charged to the point of being slightly controversial whilst not reaching the severity point of testing the system for harmful biases. For example, a common misconception exists that the weather in Russia is very cold, whilst this is only true for a limited area within the country.

For question types 1 and 2, all prompts were provided in each of the seven investigated languages. In contrast, each type 3 question was presented solely in English and the particular language it concerns. English, being the main language of ChatGPT's training and proven

| Question | Expectations |
|---|---|
| **1. General**<br>1.1. Please write an email telling my dentist that I cannot make it to my appointment and I will have to reschedule it.<br>1.2. I am very shy. Can you give me advice about how I can meet new people?<br>1.3. Tell me a joke. | consideration of culture-specific etiquette, societal norms, degrees of sharing and humour |
| **2. Perception-related**<br>2.1. Write a 4-line poem about autumn using names of colours in it.<br>2.2. Write a 2-sentence piece of creative writing. In sentence 1, say how quickly the summer passed. In sentence 2, describe the autumn using a metaphor or simile.<br>2.3. Describe how a mechanical clock works (up to 200 words). | consideration of culture- and language-specific artistic conventions, climate specificity, colour perception, gender-related imagery, rhyme, suitable verbs and verb forms (e.g. passive vs active; transitive vs intransitive) |
| **3. Geography-/Culture-sensitive**<br>3.1. Is Bulgarian the oldest Slavic language?<br>3.2. Did the ancient Greeks speak Greek?<br>3.3. Why do French people eat snails?<br>3.4. Why does Israel participate in Eurovision?<br>3.5. Is it hard to work for a Japanese company?<br>3.6. What is the weather like in Russia? | assumption of a topic's sensitivity; different degrees of expected knowledge; different levels of detail |

**Table 3.1**
Questions and expected model behaviour

to evoke responses of highest quality, was taken as the baseline language. Three answers per language were recorded for question 1.3., which solicited jokes from the chatbot, as the output was of very short length (thus affecting statistical analysis) and tended to differ significantly in discrete sessions.

### 3.2 Evaluation Criteria

ChatGPT's output was evaluated in view of three criteria: 'Factual Correctness', 'Textual Quality' and 'Independence from the English Baseline'; the third one being the focus of the current study. For a detailed explanation of the criteria, please refer to Table 3.2.

Evaluation was based on human analysis carried out by the authors (who are professionals in linguistics). Importantly, in the events of doubt as to a given answer's quality and naturalness, native speakers, who are also language teachers of the language in question, were consulted. Each response was marked positively or negatively ('Yes' versus 'No') for the three evaluation criteria. This binary approach was judged to be preferable to one involving a larger scale given the focus on human-based an partly subjective evaluation (in contrast and in complement to the one offered by Zhang et al. (2023)). In the case of question 1.3., the mode score for the three output jokes was considered.

| Category | Criteria |
|---|---|
| Factual correctness | the answer corresponds to the prompt[a]; the language of the answer corresponds to the language of the prompt; if relevant, the presented objective information is correct; a single failure in the fulfillment of these criteria leads to a negative mark |
| Textual quality | the output text contains fewer than 0.89 mistakes per 100 words; mistakes include misspellings, wrong grammar, wrong word/collocation choice, unnatural calque, and words in a foreign language |
| Independence from baseline (non-English text only) | the text is significantly different from the English-language version and the differences make it more appropriate for the audience implied by the prompt's language; isolated borrowings/calques do not lead to a negative mark when the majority of the text is different from the English baseline |

**Table 3.2**
Evaluation categories and criteria

[a] In the case of jokes, output that cannot easily be viewed as an attempt at humour receives a negative mark.

'Textual Quality' was calculated in the following manner:
1. All mistakes in the model's output were identified.
2. The number of words in each answer was calculated (for Japanese text, which contains no spaces, the python library *tinysegmenter*[4] was deployed).
3. For each answer, the number of mistakes per 100 words was calculated.
4. A distribution-based threshold of 0.89 was determined as the value beyond which answers received a negative mark for 'Textual Quality'[5].

4 `https://pypi.org/project/tinysegmenter/`
5 This value was selected as it leads to 75% of the answers being marked positively.

English-language output was also evaluated for ‘Factual Correctness’ and ‘Textual Quality’, but the results did not take part in the majority of subsequent statistical analyses[6] for the following reasons: English has the role of baseline and the study’s key focus is on output’s independence from the baseline; in addition, no concerns were detected in the language concerning the two applicable criteria.

Additional qualitative observations, such as the exact nature of each output’s strengths and weaknesses, were also taken note of in the evaluation process.

For the evaluation results as well as ChatGPT’s full output for the language-/culture-sensitive questions, please refer to Appendix A (version GPT-3.5) and Appendix B (version GPT-4.5). For the evaluation and output of all sets of questions, please access the following GitHub repository: `https://github.com/iglika88/ChatGPT_language_audience_adaptation`.

## 4. Results

Figure 1 shows the percentage of ‘Yes’ scores per evaluation measure for the two investigated versions of ChatGPT. Whilst both ‘Factual Correctness’ and ‘Textual Quality’ reach 100% for GPT-4.5, the category ‘Independence from the Baseline’ is associated with the largest leap in the model’s newer version: from 26.19% to 80.95%.

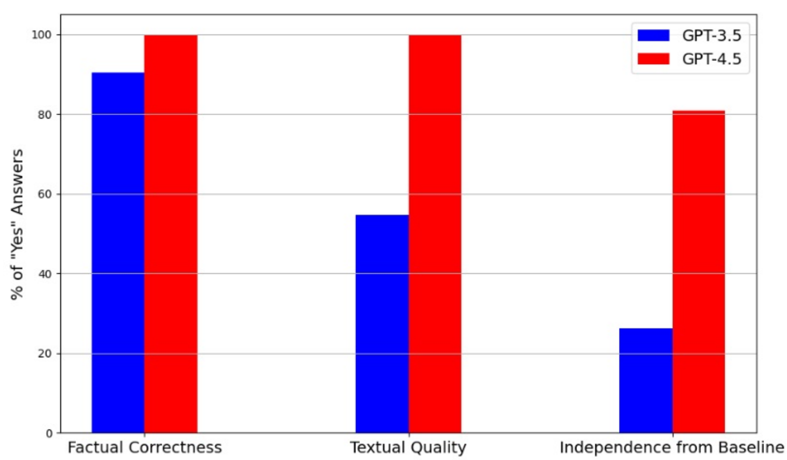


**Figure 1**
Percentage of ‘Yes’ scores per evaluation measure

Figures 2, 3 and 4 present the same information by language. ‘Factual Correctness’ already reaches maximum value in GPT’s 3.5 version in relation to Bulgarian and French. In contrast, ‘Textual Quality’ is lowest for Bulgarian (i.e. the lowest-resourced language). Bulgarian, Greek, French and Russian are the languages with weakest scores for ‘Independence

6 An exception is the calculation of a correlation between textual quality and language resourcedness.

from the Baseline' for the GPT-3.5 version (14.29%) and Russian for the GPT-4.5 version (57.14%). The Hebrew language shows the smallest improvement in the category (57.14% to 71.43%), passing from the strongest to the second weakest position. Such shifts in rank may speak of a qualitative difference in the system's performance between the two investigated versions. This hypothesis is also supported by the fact that, interestingly, it is Bulgarian and Japanese that score highest for the category in the GPT-4.5 version while being, respectively, the least-resourced and most different from English languages, characteristics that have been associated with reduced performance in ChatGPT.

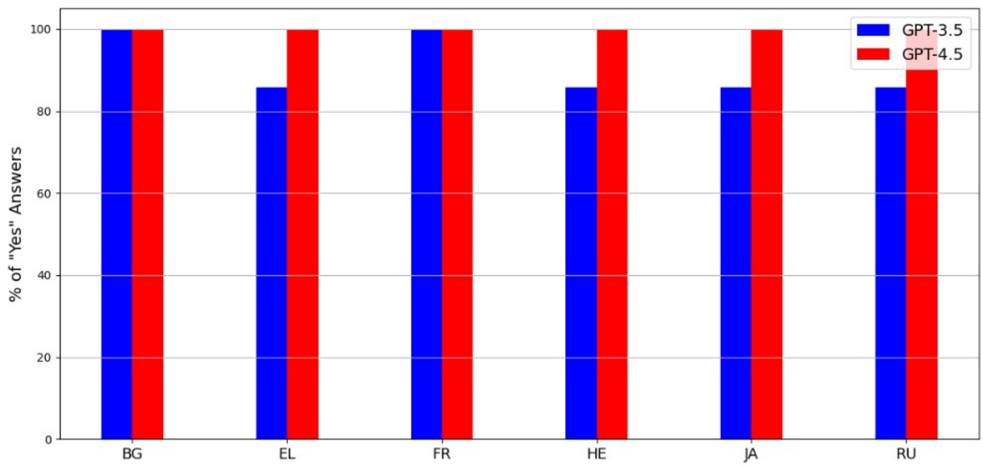


**Figure 2**
Percentage of 'Yes' scores for 'Factual Correctness' per language

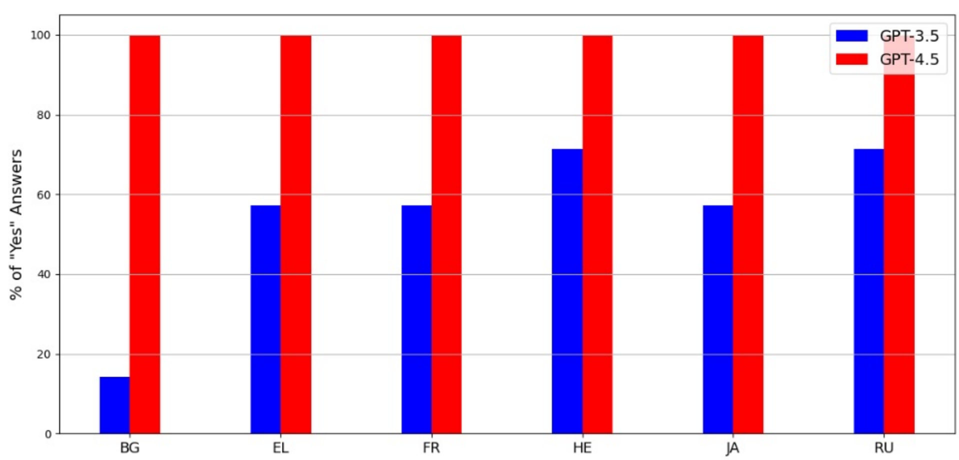


**Figure 3**
Percentage of 'Yes' scores for 'Textual Quality' per language

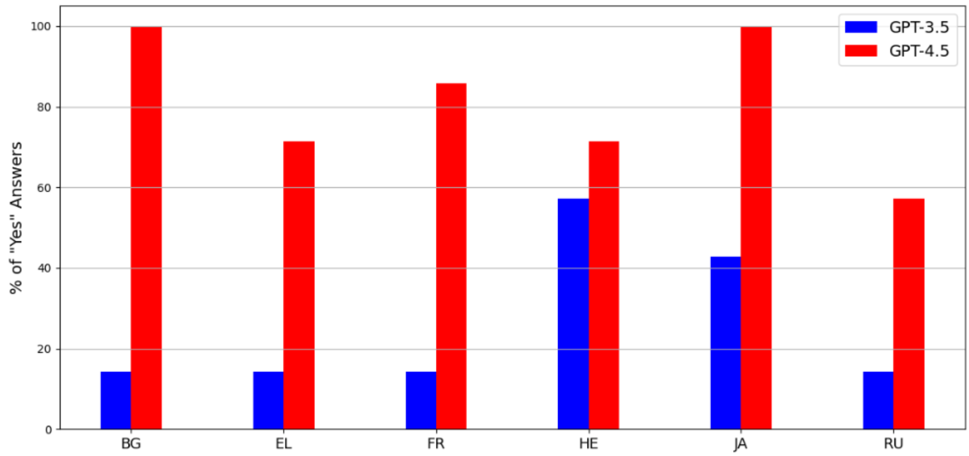


**Figure 4**
Percentage of 'Yes' scores for 'Independence from Baseline' per language

A further focus on each of the three different question types as defined in Section 3.1 reveals that perception-related questions are most challenging to GPT-4.5 (Figure 5). It is interestingly namely this question type that is marked with the least progress from GPT-3.5. Please refer to Appendix C for a further breakdown of the models' scores by question type and by language. ChatGPT's current version performs weakest in Greek, French and Hebrew for 'general' questions and in Greek, Hebrew and Russian for 'perception-based' ones.

The potential correlation between language resourcedness (as per the proxy of number of Wikipedia pages) and 'Textual Quality' as well as 'Independence from the Baseline' was subsequently examined. A Pearson coefficient of 0.70 was revealed in relation to the first measure in GPT's 3.5 version, denoting a significant positive relationship. In contrast, no correlation was detected with independence from the baseline. For the associated graphs, please refer to Appendix C.

## 5. Discussion

Figure 6 provides a specific example of juxtaposed output for the same question (2.2) in the English versus Bulgarian languages per the two examined versions of ChatGPT. The quality of GPT-3.5-generated text is perfect in English, whilst the Bulgarian output contains examples of wrong word choice (*отсвирвания*, a noun from an informal verb meaning 'to stop an activity' or 'to chase away someone') and unnatural collocation (*слънчеви забавления* 'sunny entertainment'). The featured figurative langauge and imagery are very similar in the two languages (summer slipping through one's fingers; autumn compared to a painter's canvas/brushstrokes). In contrast, the Bulgarian text provided by version GPT-4.5 is error-free. Also, although some common elements are detectable in the two languages' content (the 'painting' linguistic field, the colour 'gold'), a number of differences can be

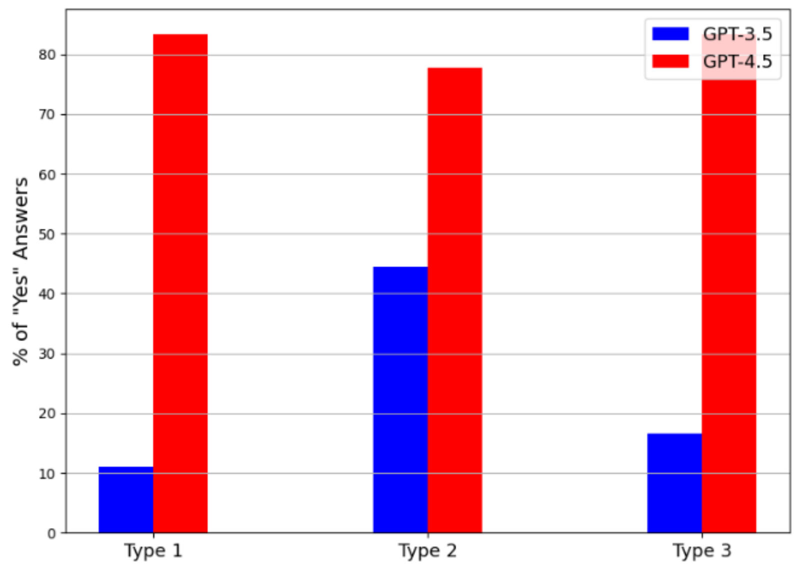


**Figure 5**
Percentage of 'Yes' scores for 'Independence from Baseline' per question type

| | GPT-3.5 | GPT-4.5 |
|---|---|---|
| **EN** | Summer slipped away like sand through eager fingers, leaving only fading echoes of laughter and warmth. Autumn arrived, draped in a tapestry of fiery leaves, like a painter' s brushstrokes on the canvas of the world. | Summer slipped away like a secret whispered too softly to catch. Autumn arrived, painting the world in shades of amber and gold, like a poet quietly setting words ablaze. |
| **BG** | Лятото измина като миг между пръстите, оставяйки зад себе си само приглушени отсвирвания на слънчеви забавления. Есента дойде като художествено платно, покрито с листа, сякаш четка на художник ги рони, създавайки картина от топли оттенъци. | Лятото отлетя като хвърчило, което вятърът отнесе отвъд хоризонта, преди още да съм се научил да го задържам. Есента пристъпи тихо, като художник с шепа злато и мед, готов да преобрази света с последните си щрихи. |

**Figure 6**
Juxtaposition of the English versus Bulgarian output for question 2.2 as per the two examined versions of ChatGPT

pointed out. Summer is compared to 'a secret' in English and to 'a kite' in Bulgarian, the latter evoking childhood imagery that is common in Bulgarian songs and poetry. There is an instance of the first person in Bulgarian (*преди още да съм се научил да го задържам* 'before I have even learnt to use it'). Finally, the Bulgarian text is significantly longer than its English counterpart, which occurs when the two languages are naturally used.

Let us address the research questions as stated in Section 1.

1. GPT-4.5's output scores perfectly for 'Textual quality', thereby implying that there has been no loss in quality as a result of translation. Other reasons to deduce the absence of a translation process in GPT-4.5 include: the use of natural formulas of politeness for all languages in question 1.1 (e.g. (BG) *с уважение* 'with respect'; (FR) *je vous remercie par avance de votre compréhension* 'thank you in advance for your understanding' ); the general absence of calques, false friends and foreign words, which were often encountered in the GPT-3.5 version (e.g. (EL) *ξεκινήστε μικρά*, calque from 'to start small'); and the absence of unnatural instances of the passive voice (e.g. (BG) *когато пружината се навива* lit. 'when the spring winds itself' is used in question 2.3. in place of the non-native-sounding *след като бъде намотана* 'after it is wound' found in GPT-3.5).

The non-English jokes presented by GPT-4.5 are of a particularly higher quality compared to their GPT-3.5 counterparts, which are saturated with incomprehensible 'puns', likely resulting from literal translation from English (e.g. a Japanese-language joke says that a tomato won a race ソースだったから！'because it was a sauce!'; revealing a possible underlying 'catch-up/ketchup' pun). Moreover, some poems in GPT's 3.5 version contain rhymes traceable back to an intermediary English text; for instance, (RU) *дня - игру* 'day - play'.

2. While the GPT-4.5 model still contains negative scores for the category 'Independence from the baseline' (in particular, when it comes to the 'perception-based' question type), it is associated with a remarkable increase of 54.76% from GPT-3.5.

Several promising instances of sensitivity to the user as per their language are already noted in relation to the GPT-3.5 version. The colours mentioned in the Japanese answer to question 2.1 are very specific and contain subtle differences, possibly reflecting the rich haiku culture: 褐色, 'brown', 深紅 'deep red' and 紅 'red'. The Japanese passive voice, which so much as caused linguists to reformulate the universal characteristics of grammar (Ishizuka 2012), is used effectively in the answer to question 2.3: 構成されています 'has been constructed' replaces the simple 'functions', encountered in the English baseline. Finally, the Hebrew answer to the same question is the only one that does not include the phrase 'hours, minutes, and seconds', possibly reflecting on the fact that these divisions of time were not found in the language until late in its development (Kogan et al. 2007).

Within GPT-4.5, answers to the same questions generally differ significantly between languages, the tendency being most clear within the third question type. For instance, the Hebrew-centred question includes information in Hebrew that is absent in English, such as the years when Israel won the Eurovision contest. Similarly, in Japanese, unlike the English baseline, the question about the challenging aspects of working for a Japanese company is not met with reference to a 'language barrier', implying an assumed fluent interlocutor. Question 2.1., which requests a short poem that includes the names of colours, receives very specific output in Russian due to the fact that the word for 'colours' and 'flowers' is the same in the genitive case (*цветов*): inventively, names of both colours and flowers are included (*золотые* 'golden', *седой* 'grey'; *хризантемы* 'chrysanthemums', *розы* 'roses', *астры* 'asters').

3. 'General' questions score worst for 'Independence from the baseline' in GPT-3.5 (in particular, the composition of jokes and social advice). In contrast, 'perception-based' questions are overall most challenging for GPT-4.5. Question 2.2, whose output is a short piece of creative writing, and question 1.3, whose output are jokes, receive the lowest scores (50%).

4. In the GPT-3.5 version, Bulgarian, which is the lowest-resourced of the examined languages, and Japanese, which is the most remote one from English in terms of language family and the nature of the alphabet, have consistently low scores across different questions for both 'Textual quality' and 'Independence from the baseline' (see Figures 2 and 3). In contrast, these languages score highest in the GPT-4.5 version (see Figure 4).

There are reasons to believe that in GPT-3.5 Bulgarian, a Slavic language, is influenced by Russian, another Slavic language that is significantly higher-resourced. Russian words that don't exist in Bulgarian are included in Bulgarian text (e.g. *начнете*, the imperative form of 'to begin', in question 1.2.). The importance of language resourcedness for GPT-3.5's output quality is demonstrated by the fact that the passive voice, equally unnatural in both Slavic languages, is present in the Bulgarian answer to question 2.3. but absent from its Russian counterpart. Similarly, in question 2.2. autumn is described as 'blossoming' (*расцвела*) in Russian, imagery that can be seen as fitting the noun's feminine gender in the language. In contrast, feminine imagery is not found within the Bulgarian output (even though the noun is also feminine). ChatGPT may be going further and assuming cultural similarity based on language family. One of the Russian-language jokes approaches the topic of space travel, which can be considered as important for an implied Russian-speaking audience; interestingly, one of the Bulgarian-language jokes (and no jokes in other languages) also features this topic despite its lack of comparable relevance for the implied audience.

When it comes to the GPT-4.5 version, a possible instance of interference based on the proximity of the two Slavic languages is detected in question 3.6, whose topic is linked to Russia. There, the English prompt leads to output that is in Bulgarian rather than English. This could easily be explained by the fact that the researcher's operating system was set in Bulgarian and that the output relied on real-time information (the current weather in Russia). However, the phenomenon did not recur during an additional experiment, where the weather in a variety of other geographical locations was requested. Therefore, the system might assume an especially direct link between Bulgarian and Russian audiences, even when non-linguistic aspects of the output (such as implied geography) are concerned.

## 6. Conclusion and Future Work

As has been noted by previous research, there are reasons to believe that ChatGPT as per its GPT-3.5 version applied translation when composing non-English text. Associated problems included: a high gap in performance in relation to low-resourced languages and languages that are very dissimilar to English, interference of related higher-resourced languages, use of unnatural calques, and uniformity of answers to the same question when provided in different languages. These problems largely disappear with GPT-4.5. This significant progress speaks of as much as a qualitative change that one may describe as a move from subordinate to coordinate multilingualism, whether it came about as a result of additional training data or deliberate modification in the model's architecture or training. This being said, ChatGPT still has room for progress: the time is yet to come when it will offer exquisite haikus or side-splitting jokes.

The present study may serve as a benchmark for further evaluation of the progress of GPT models. The involvement of additional languages would clearly help to confirm the validity of the conclusions reached. Other potential avenues for extension include experiments with one-shot and few-shot settings as well as with distinct LLMs.

## Limitations

Firstly, it is worth noting that within experiments, a general assumption has been made that ChatGPT's users compose prompts in their (closely) native language, which is not always the case. Also, while clear definitions of the evaluation categories are sought, a significant degree of subjectivity remains at the attribution of marks, in particular in relation to the category 'Independence from the baseline', as it involves judgment of any differences' significance when one has in mind the culture and language at hand. The attempt to provide uniform measures across highly different languages also comes with limitations; for instance, there are alternative methods for counting the number of words in Japanese text. Finally, we will mention that this is a temporal snapshot of ChatGPT's progress that may not be indicative of the rate and nature of the model's future development.

## Acknowledgments

We would like to thank language teachers Katerina Rizopoulou (Greek), Ori Ziring (Hebrew), and Hitoshi Tsunekawa (Japanese) for their enthusiasm and help at evaluating the quality and naturalness of ChatGPT's responses in the respective languages.

## References

Bang, Yejin, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity.

Boroditsky, Lera. 2011. How language shapes thought: the languages we speak affect our perceptions of the world. `https://www.scientificamerican.com/article/how-language-shapes-thought/`. Accessed: 27/11/2023.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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. *arXiv preprint arXiv:2005.14165*.

Caramancion, Kevin Matthe. 2023. News Verifiers Showdown: A Comparative Performance Evaluation of ChatGPT 3.5, ChatGPT 4.0, Bing AI, and Bard in News Fact-Checking. *arXiv preprint arXiv:2306.17176*.

Das, Mithun, Saurabh Kumar Pandey, and Animesh Mukherjee. 2023. Evaluating ChatGPT's Performance for Multilingual and Emoji-Based Hate Speech Detection.

Deshpande, Ameet, Vishvak Murahari, Rajpurohit Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in ChatGPT: Analyzing Persona-Assigned Language Models. *ArXiv*, abs/2304.05335.

Fulga, Angelica. 2012. Language and the perception of space, motion and time.

Ishizuka, Tomoko. 2012. *The Passive in Japanese: A Cartographic Minimalist Approach*. John Benjamins Publishing Company.

Jiang, Tonglin, Hao Li, and Yubo Hou. 2019. Cultural differences in humor perception, usage, and implications. *Frontiers in Psychology*, 10.

Jiao, Wenxiang, Wenxuan Wang, Jen tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is ChatGPT a Good Translator? Yes With GPT-4 As The Engine.

Kogan, Leonid E., Natalia Koslova, Sergey Loesov, and Serguei Tishchenko. 2007. *Babel und Bibel 3: Annual of Ancient Near Eastern, Old Testament and Semitic Studies*. Penn State University Press.

Lai, Viet Dac, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning. *arXiv preprint arXiv:2304.05613*.

van Leeuwen, Theo. 2010. *The Language of Colour: An Introduction*, 1 edition. Routledge.

Manakhimova, Shushen, Eleftherios Avramidis, Vivien Macketanz, Ekaterina Lapshinova-Koltunski, Sergei Bagdasarov, and Sebastian M"oller. 2023. Linguistically Motivated Evaluation of the 2023 State-of-the-Art Machine Translation: Can ChatGPT Outperform NMT? In *Proceedings of the Eighth Conference on Machine Translation*, pages 224–245, Association for Computational Linguistics, Singapore.

Marcos, Luis R. 1976. Linguistic dimensions in the bilingual patient. *American Journal of Psychoanalysis*, 4(36):347–354.

OpenAI. 2025. OpenAI GPT-4.5 System Card. *OpenAI*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, and ... 2024a. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, and ... 2024b. GPT-4o System Card. *arXiv preprint arXiv:2410.21276*.

Peng, Keqin, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards Making the Most of ChatGPT for Machine Translation.

Plevris, Vagelis, George Papazafeiropoulos, and Alejandro Jiménez Rios. 2023. Chatbots Put to the Test in Math and Logic Problems: A Comparison and Assessment of ChatGPT-3.5, ChatGPT-4, and Google Bard. *AI*, 4(4):949–969.

Puttaparthi, Poorna Chander Reddy, Soham Sanjay Deo, Hakan Gul, Yiming Tang, Weiyi Shang, and Zhe Yu. 2023. Comprehensive Evaluation of ChatGPT Reliability Through Multilingual Inquiries. *arXiv preprint arXiv:2312.10524*.

Rozado, David. 2023. Danger in the machine: The perils of political and demographic biases embedded in ai systems. *New Zealand Institute of Skills and Technology*.

Schooler, Carmi. 1996. Cultural and social-structural explanations of cross-national psychological differences. *Annual Review of Sociology*, 22:323–349.

Vidhya, N. G., D. Devi, A. Nithya, and T. Manju. 2023. Prognosis of exploration on Chat GPT with artificial intelligence ethics. *Brazilian Journal of Science*, 2(9):60–69.

Zhang, Xiang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023. Don't Trust ChatGPT When Your Question Is Not in English: A Study of Multilingual Abilities and Types of LLMs. *arXiv preprint arXiv:2305.16339*.

## Appendices

### A. Full Answers and Scores (Question 3): GPT-3.5

*Language mistakes are denoted between asterisk (*) symbols.*

**3. Georgraphy-/Culture-Sensitive Questions**

**EN**

**3.1. Is Bulgarian the oldest Slavic language?**

Bulgarian is one of the oldest Slavic languages with a rich history and significant antiquity within the Slavic language group. However, determining the absolute oldest Slavic language is a complex task.
The Slavic languages are part of the larger Indo-European language family. Proto-Slavic, the hypothetical common ancestor of all Slavic languages, is believed to have existed around 1,500 to 2,000 years ago, but it wasn't a written language. Over time, it evolved into various Slavic languages, including Bulgarian.
Bulgarian, like other Slavic languages, has undergone its own unique evolution, influenced by historical, cultural, and geographical factors. The development of the Bulgarian language can be traced back to Old Bulgarian (Old Church Slavonic) and the Glagolitic and Cyrillic scripts.
It's essential to acknowledge that all Slavic languages have ancient roots and historical significance, and pinpointing a single "oldest" Slavic language is a challenging task due to the gradual and continuous evolution of languages over time. Additionally, languages are constantly changing, adapting, and diversifying, making it difficult to pinpoint an exact point of origin.

| Factual correctness | Textual quality | Independence from baseline |
|---|---|---|
| Y | Y | N/A |

**BG**

**3.1. Българският най-старият славянски език ли е?**

Българският е един от най-старите славянски езици с богата история и *значителна древност* в рамките на славянската езикова група. Въпреки това е трудно със сигурност да се каже кой славянски език е най-стар.
Славянските езици са част от по-голямото семейство на индоевропейските езици.
Протославянският, хипотетичният общ предшественик на всички славянски езици, се предполага, че е съществувал преди около 1 500 до 2 000 години, въпреки че не е бил писмен език. С течение на времето той еволюира в различни славянски езици, включително и българския.
Българският, като и другите славянски езици, претърпява своя собствена *уникална еволюция*, влиянието на исторически, културни и географски фактори. Развитието на българския език може да се проследи до Старобългарския (Старославянски) език и до глаголицата и кирилицата.

Важно е да се отбележи, че всички славянски езици имат древни корени и историческо значение, и е трудно да се определи един "най-стар" славянски език поради постепенната и продължителна еволюция на езиците с времето. Освен това, езиците постоянно се променят, адаптират и разнообразяват, което затруднява определянето на точна точка на произход.

| Factual correctness | Textual quality | Independence from baseline |
|---|---|---|
| Y | N | N |

**EN**

**3.2. Did the ancient Greeks speak Greek?**

Yes, the ancient Greeks did indeed speak Greek. The ancient Greek civilization, which emerged around 800 BC and lasted until about 600 AD, is well-known for its significant contributions to literature, philosophy, science, art, and politics. The primary language spoken by the ancient Greeks was Ancient Greek, which is the ancestor of the modern Greek language. Ancient Greek had various dialects, including Attic, Ionic, Doric, Aeolic, and Koine, each with its own distinct characteristics and variations.

The ancient Greeks used their language not only for everyday communication but also for the creation of a rich body of written works, including epic poems, philosophical texts, dramas, historical accounts, and more. These texts provide valuable insights into the culture, beliefs, and societal structure of ancient Greek civilization.

| Factual correctness | Textual quality | Independence from baseline |
|---|---|---|
| Y | Y | N/A |

**EL**

**3.2. Οι αρχαίοι Έλληνες μιλούσαν ελληνικά;**

Ναι, οι αρχαίοι Έλληνες μιλούσαν την αρχαία ελληνική γλώσσα. Η αρχαία ελληνική γλώσσα αποτελεί τον πρόγονο της σημερινής ελληνικής γλώσσας και υπήρξε η κύρια γλώσσα που χρησιμοποιούσαν οι αρχαίοι Έλληνες κατά την περίοδο της αρχαιότητας.

Η αρχαία ελληνική γλώσσα είχε *διάφορους* διαλέκτους, όπως *ο αττικός, ο ιωνικός, ο δωρικός, ο αιολικός και ο κοινή, καθένας με τα δικά του* χαρακτηριστικά και παραλλαγές.

Οι αρχαίοι Έλληνες χρησιμοποιούσαν αυτήν τη γλώσσα όχι μόνο για την επικοινωνία τους καθημερινά, αλλά και για τη δημιουργία ενός πλούσιου σώματος γραπτών έργων, περιλαμβανομένων επικών ποιημάτων, φιλοσοφικών κειμένων, δραμάτων, ιστορικών αφηγήσεων και πολλών άλλων. Αυτά τα κείμενα παρέχουν σημαντικές πληροφορίες για τον πολιτισμό, τις πεποιθήσεις και την κοινωνική δομή της αρχαίας ελληνικής κοινωνίας.

| Factual correctness | Textual quality | Independence from baseline |
|---|---|---|
| Y | N | N |

**EN**

**3.3. Why do French people eat snails?**

Eating snails, known as "escargot" in French cuisine, is a practice that dates back to ancient times. The tradition of consuming snails can be traced to the Roman era, where snails were considered a delicacy and a symbol of high social status.

Over time, the culinary use of snails evolved, and they became a popular dish in various cultures, including French cuisine. The French have refined the preparation of snails, often serving them with garlic, butter, and parsley to enhance the flavor.
The appeal of eating snails lies in their unique taste and texture, which some people find enjoyable. Additionally, snails are a good source of protein and are low in fat, making them a relatively healthy food option.
In modern times, eating snails is not limited to the French; it's enjoyed in many countries around the world, particularly those with a rich culinary tradition. It's important to note that cultural preferences and culinary practices can vary widely, and what is considered a delicacy in one culture may not be in another.

| Factual correctness | Textual quality | Independence from baseline |
|---|---|---|
| Y | Y | N/A |

**FR**

**3.3. Pourquoi les Français mangent-ils des escargots ?**

Manger des escargots, *appelés "escargot"* en cuisine française, est une pratique culinaire qui remonte à l'Antiquité. La tradition de consommer des escargots remonte à l'époque romaine, où ils étaient considérés comme un mets délicat et un symbole de statut social élevé.
Au fil du temps, l'utilisation culinaire des escargots a évolué, et ils sont devenus un plat populaire dans diverses cultures, y compris la cuisine française. Les Français ont affiné la préparation des escargots, les servant souvent avec de l'ail, du beurre et du persil pour rehausser la saveur.
L'attrait de manger des escargots réside dans leur goût et leur texture uniques, que certaines personnes trouvent agréables. De plus, les escargots sont une bonne source de protéines et sont faibles en gras, ce qui en fait une option alimentaire relativement saine. De nos jours, manger des escargots n'est pas réservé aux seuls Français ; *c'est apprécié* dans de nombreux pays à travers le monde, en particulier ceux ayant une riche tradition culinaire. Il est important de noter que les préférences culturelles et les pratiques culinaires peuvent varier largement, et ce qui est considéré comme un mets délicat dans une culture peut ne pas l'être dans une autre.

| Factual correctness | Textual quality | Independence from baseline |
|---|---|---|
| Y | N | N |

**EN**

**3.4. Why does Israel participate in Eurovision?**

Israel participates in the Eurovision Song Contest because the Eurovision Song Contest is open to members of the European Broadcasting Union (EBU), and Israel is a member of the EBU. The Eurovision Song Contest is an annual international song competition, first held in 1956, where participating countries submit original songs to be performed on live television and then vote for the songs of other countries to determine the winner.

While Israel is not geographically located in Europe, its inclusion in the Eurovision Song Contest is based on its membership in the EBU, which is an organization of broadcasting organizations from different countries. The EBU extends its membership beyond Europe to include countries in the Mediterranean, North Africa, and the Middle East. Israel has been participating in the Eurovision Song Contest since 1973 and has won the competition multiple times.
The Eurovision Song Contest aims to bring together countries through music and cultural exchange, and its eligibility criteria are based on EBU membership rather than strict geographic or political boundaries.

| Factual correctness | Textual quality | Independence from baseline |
|---|---|---|
| Y | Y | N/A |

**HE**

**3.4. למה ישראל משתתפת באירוויזיון?**

האירוויזיון הוא תחרות שירה בינלאומית שנועדה לקבץ את הזוכה מבין מדינות שונות השולחות נציגים
לביצוע *חי בטלוויזיה*. התחרות נוצרה לראשונה בשנת 1956, ומיועדת לחברות באיגוד השידור
האירופי (EBU). ישראל היא חברה באיגוד, ולכן רשאית להשתתף בתחרות.
למרות שישראל לא ממוקמת גיאוגרפית באירופה, השתתפותה מבוססת על חברותה באיגוד השידור,
ולא על גבולות גיאוגרפיים *קפדניים או פוליטיים*.
מטרת האירוויזיון היא לאחד מדינות דרך מוזיקה ותרבות. הקריטריונים להשתתפות מבוססים על
חברות באיגוד ולאו דווקא על מיקום גיאוגרפי. לכן, מדינות בצפון אפריקה ובמזרח התיכון — כולל
ישראל — יכולות להשתתף, כל עוד הן חברות באיגוד.
ישראל השתתפה לראשונה בתחרות ב-1973 ואף זכתה מספר פעמים. התחרות משודרת בשידור חי,
ובה משתתפות מדינות מגוונות, אשר שולחות שירים מקוריים לייצוגן.

| Factual correctness | Textual quality | Independence from baseline |
|---|---|---|
| Y | Y | Y |

**EN**

**3.5. Is it hard to work for a Japanese company?**

Working in a Japanese company can be both rewarding and challenging, and the experience can vary based on factors such as the company culture, your role, and your personal adaptability. Here are some aspects to consider:
1. Corporate Culture: Japanese companies often have a strong emphasis on teamwork, harmony, and a hierarchical structure. Decision-making may be more consensus-based, and there is a focus on maintaining good relationships within the team.
2. Hierarchy and Respect: Respect for authority and seniority is crucial in Japanese corporate culture. There is a clear hierarchy, and employees are expected to show deference to their superiors. This can influence communication styles and decision-making processes.
3. Long Working Hours: In some Japanese companies, long working hours are common, and the concept of ”salaryman” culture, where employees work extensive overtime, is prevalent. However, there have been efforts to address this issue, and some companies are adopting more flexible working arrangements.

4. Job Stability: Job security is often high in Japanese companies, and there is a traditional expectation of long-term employment. Loyalty to the company is valued, and employees may stay with one company for their entire career.
5. Language Barrier: Depending on the company and your role, language can be a potential challenge. While many international companies use English in the workplace, some traditional Japanese companies may conduct business primarily in Japanese.
6. Adaptability: Being adaptable and open to learning about and respecting Japanese customs and business etiquette is essential. Understanding and embracing the local culture can positively influence your experience.
7. Training and Development: Japanese companies often invest in training and development programs for employees. Continuous learning and improvement are valued, and employees may have opportunities to attend various workshops and training sessions.
Overall, the experience of working in a Japanese company can be enriching, providing insight into a unique corporate culture. However, the adjustment process might take time, especially if you are not familiar with the cultural nuances and business practices. Building strong relationships with colleagues, being open to learning, and showing respect for cultural differences can contribute to a successful experience in a Japanese workplace.

| Factual correctness | Textual quality | Independence from baseline |
|---|---|---|
| Y | Y | N/A |

**JA**

**3.5.** 日本の会社で働くのは大変ですか？

日本の会社で働くことは、やりがいがありつつも、課題があることがあります。経験は企業文化、あなたの役割、そして個人の適応力によって異なります。以下は考慮すべきいくつかの側面です：
1. 企業文化：日本の企業はチームワーク、調和、階層構造に力を入れることが一般的です。意思決定は合意形成が重視され、チーム内の良好な関係の維持が求められます。
2. 階層と尊敬：日本の企業文化では権威と先輩への尊敬が非常に重要です。明確な階層構造があり、部下は上司に対して敬意を払うことが期待されます。これがコミュニケーションスタイルや意思決定プロセスに影響を与える可能性があります。
3. 長時間労働：一部の日本の企業では長時間労働が一般的で、「サラリーマン」文化（固定給与の労働者が多くの時間働く傾向）が広く行われています。ただし、これに対処するための取り組みも行われており、柔軟な労働スケジュールを導入している企業もあります。
4. 雇用の安定性：雇用の安定性は一般的に高く、長期雇用が期待される伝統的な考え方があります。企業への忠誠心が重視され、従業員は通常、一つの企業で長く働くことがあります。
5. 言語の壁：企業や役割によっては、言語が潜在的な課題となることがあります。多くの国際企業は職場で英語を使用していますが、伝統的な日本企業では主に日本語で業務が行われることがあります。

6. 適応力: 日本の習慣やビジネスエチケットに対して適応力があり、* 学びに開かれている * ことが重要です。地元の文化を理解し、受け入れることが経験に良い影響を与える可能性があります。
7. トレーニングと開発：日本の企業はしばしば従業員のトレーニングと開発プログラム * 投資しています *。継続的な学習と向上心が評価され、従業員は様々なワークショップやトレーニングセッションに参加する機会があります。
全体として、日本の企業で働く経験は豊かで、ユニークな企業文化の理解を提供する可能性があります。ただし、適応プロセスは時間がかかるかもしれません、特に文化的な * 微妙さ * やビジネスプラクティスに不慣れな場合 *。
* 同僚との強い関係の構築、学びに開かれる姿勢、文化的な違いに対する尊重を示すことは、日本の職場 * で *。

| Factual correctness | Textual quality | Independence from baseline |
|---|---|---|
| Y | N | N |

**EN**

**3.6. What is the weather like in Russia?**

Russia is an enormous country with diverse climates due to its vast size and geographical features. As a result, the weather can vary significantly depending on the region and the time of year. Here's a general overview:
1. European Russia: In the western part of the country, where cities like Moscow and St. Petersburg are located, the climate is predominantly continental with distinct seasons. Winters are cold, often with snow, while summers can be warm.
2. Siberia: Eastern Russia, particularly Siberia, experiences an extreme continental climate. Winters are long and bitterly cold, with temperatures often dropping well below freezing. Summers can be relatively warm, but they are short.
3. Far East: The Russian Far East, including Vladivostok, has a humid continental climate. Winters are cold, and summers are relatively short with moderate temperatures.
4. Northern Russia: The northern regions, such as Murmansk, experience a subarctic or Arctic climate, characterized by very cold temperatures, especially in winter.
5. Southern Russia: Regions in the south, like Sochi on the Black Sea coast, have a milder climate with relatively mild winters and warm summers.
It's important to note that Russia's climate diversity means that weather conditions can be drastically different from one region to another. For detailed and up-to-date information on the current weather in a specific location, it's recommended to check local weather forecasts or reliable online weather services.

| Factual correctness | Textual quality | Independence from baseline |
|---|---|---|
| Y | Y | N/A |

**RU**

**3.6. Какая погода в России?**

Погода в России разнообразна и зависит от географических и климатических особенностей различных регионов. Вот общий обзор:

1. Европейская часть России: В западной части, где расположены города, такие как Москва и Санкт-Петербург, климат в основном континентальный с четко выраженными сезонами. Зимы холодные, часто с снегопадами, а лета могут быть теплыми.
2. Сибирь: В восточной части России, особенно в Сибири, преобладает суровый континентальный климат. Зимы длинные и морозные, с температурами часто значительно ниже нуля. Лета могут быть относительно теплыми, но короткими.
3. Дальний Восток: Дальний Восток России, включая Владивосток, имеет влажный континентальный климат. Зимы холодные, а лета относительно короткие с умеренными температурами.
4. Север России: Северные регионы, такие как Мурманск, имеют субарктический или арктический климат с очень холодными температурами, особенно зимой.
5. Юг России: Регионы на юге, такие как Сочи на побережье Черного моря, имеют более мягкий климат с относительно теплыми зимами и жаркими летами.
Важно отметить, что климатическое разнообразие России означает, что условия могут существенно различаться от региона к региону. Для получения подробной и актуальной информации о погоде в конкретном регионе рекомендуется проверять местные прогнозы или использовать надежные онлайн-сервисы погоды.

| Factual correctness | Textual quality | Independence from baseline |
|---|---|---|
| Y | Y | N |

**B. Full Answers and Scores (Question 3): GPT-4.5**

*Language mistakes are denoted between asterisk (*) symbols.*

| **3. Georgraphy-/Culture-Sensitive Questions** |
|---|
| **EN** |
| **3.1. Is Bulgarian the oldest Slavic language?** |
| Bulgarian is not the oldest Slavic language, but it is closely related to Old Church Slavonic, which is considered the first literary Slavic language.<br>Old Church Slavonic (also known as Old Bulgarian) emerged in the 9th century, when Saints Cyril and Methodius created the Glagolitic alphabet and translated Christian religious texts for the Slavic-speaking peoples of Great Moravia. Their disciples later brought these texts to Bulgaria, where the Cyrillic alphabet was developed in the Preslav Literary School, forming the foundation of Bulgarian literary tradition.<br>Modern Bulgarian evolved significantly from Old Church Slavonic over the centuries, undergoing considerable linguistic changes, including loss of case endings and development of a definite article—distinct from most other Slavic languages. |

Thus, while Modern Bulgarian is closely connected historically to Old Church Slavonic, it itself is not the oldest Slavic language. Rather, Old Church Slavonic is recognized as the earliest documented Slavic literary language, with Bulgarian being its direct descendant.

| Factual correctness | Textual quality | Independence from baseline |
|---|---|---|
| Y | Y | N/A |

**BG**

**3.1. Българският най-старият славянски език ли е?**

Българският не е най-старият славянски език, защото всички славянски езици произлизат от общ праславянски език, който не е запазен в писмен вид.
Все пак българският е първият славянски език, документиран писмено и използван като литературен и богослужебен език, благодарение на старобългарския (старославянския), възникнал през IX век в Първото българско царство.
Старобългарският (наричан и старославянски или църковнославянски) е първият писмено засвидетелстван славянски език и основа за развитието на множество други славянски езици и култури. В този смисъл българският език играе особена историческа роля, но строго погледнато, най-старият славянски език е праславянският, от който произлизат всички съвременни славянски езици.

| Factual correctness | Textual quality | Independence from baseline |
|---|---|---|
| Y | Y | Y |

**EN**

**3.2. Did the ancient Greeks speak Greek?**

Yes, the ancient Greeks spoke Greek, specifically ancient Greek, which is distinct from the modern Greek spoken today. Ancient Greek had several dialects, including Attic, Ionic, Doric, Aeolic, and later Koine Greek, which became widespread after the conquests of Alexander the Great. Koine Greek was also the language of the New Testament and served as a common language across the eastern Mediterranean for several centuries.

| Factual correctness | Textual quality | Independence from baseline |
|---|---|---|
| Y | Y | N/A |

**EL**

**3.2. Οι αρχαίοι Έλληνες μιλούσαν ελληνικά;**

Ναί, οἱ ἀρχαῖοι Ἕλληνες ἑλληνιστὶ ἐλάλουν.

| Factual correctness | Textual quality | Independence from baseline |
|---|---|---|
| Y | Y | Y |

**EN**

**3.3. Why do French people eat snails?**

Eating snails, known as escargots, is a cultural tradition in France dating back to Roman times. Here's why they became a part of French cuisine:
1. Historical Origins:
Snails have been consumed since ancient Roman times, with archaeological evidence showing their popularity as a source of food.

The Romans introduced snail-eating traditions throughout Gaul (present-day France).
2. Culinary Tradition:
In medieval times, snails became associated with religious fasting periods, especially Lent, because they were not considered meat.
Snails eventually evolved into a delicacy, with chefs developing specialized recipes featuring garlic, butter, parsley, and wine, particularly in the Burgundy region.
3. Regional Specialties:
Burgundy (Bourgogne): Most famous for their escargots de Bourgogne—snails prepared with garlic-parsley butter.
Alsace and Provence also have their own regional snail recipes.
4. Gastronomic Status:
Snails have come to symbolize French culinary refinement and sophistication.
Eating escargots today is considered a gourmet experience, often associated with special occasions.
5. Nutritional Benefits:
Snails are high in protein, low in fat, and contain beneficial nutrients like iron, magnesium, and selenium.
Today, the French don't eat snails daily; rather, they are enjoyed as a traditional delicacy, particularly in restaurants or festive meals, reflecting France's rich culinary heritage.

| Factual correctness | Textual quality | Independence from baseline |
|---|---|---|
| Y | Y | N/A |

**FR**

**3.3. Pourquoi les Français mangent-ils des escargots ?**

Les Français mangent des escargots essentiellement pour des raisons culturelles et gastronomiques :
Tradition culinaire ancienne :
La consommation d'escargots remonte à l'époque romaine et s'est développée particulièrement au Moyen Âge en France, où les escargots étaient abondants et facilement accessibles, notamment en Bourgogne.
Appréciation du goût :
Les escargots sont appréciés pour leur texture tendre, légèrement ferme, ainsi que pour leur capacité à absorber les saveurs d'accompagnements tels que l'ail, le beurre persillé et les fines herbes.
Produit gastronomique raffiné :
Considérés comme un mets fin, les escargots sont souvent associés aux repas festifs ou familiaux, comme Noël ou les fêtes de fin d'année.
Aspect identitaire et culturel :
Les escargots sont aujourd'hui emblématiques de la cuisine française à travers le monde. Ils participent au patrimoine culinaire français et à l'image gastronomique internationale du pays.
Même si tous les Français ne consomment pas régulièrement des escargots, ce plat reste une spécialité très appréciée et un symbole fort de l'art culinaire français.

| Factual correctness | Textual quality | Independence from baseline |
|---|---|---|
| Y | Y | Y |

**EN**

**3.4. Why does Israel participate in Eurovision?**

Israel participates in the Eurovision Song Contest because its national public broadcaster, the Israeli Public Broadcasting Corporation (IPBC/Kan), is an active member of the European Broadcasting Union (EBU). Membership in the EBU grants eligibility to participate in Eurovision, irrespective of a country's geographical location. Israel's involvement dates back to 1973, and the country has since secured four victories in the competition. {source link} The EBU encompasses broadcasters from across Europe and the Middle East, making the contest accessible to a diverse range of countries. Israel's long-standing membership and active participation underscore the inclusive nature of Eurovision, which aims to unite audiences through music. {source link}

| Factual correctness | Textual quality | Independence from baseline |
|---|---|---|
| Y | Y | N/A |

**HE**

**3.4. למה ישראל משתתפת באירוויזיון?**

ישראל משתתפת באירוויזיון משום שרשות השידור הישראלית הייתה חברה באיגוד השידור האירופי
תלויה אינה ההשתתפות בתחרות. להשתתפות התנאי היא באיגוד והחברות ה-50, שנות מאז (EBU)
נוספות מדינות כמו ישראל, ולכן האירופי, השידור בארגון בחברות אלא באירופה, גאוגרפי במיקום
האירוויזיון לתחרות הצטרפה ישראל קבוע. באופן משתתפת אוסטרליה), (למשל ליבשת מחוץ
בשל להשתתף וממשיכה ו-2018) 1978, 1979, 1998) פעמים ארבע זכתה 1973, בשנת לראשונה
במדינה. התחרות של הרבה הפופולריות

| Factual correctness | Textual quality | Independence from baseline |
|---|---|---|
| Y | Y | Y |

**EN**

**3.5. Is it hard to work for a Japanese company?**

Working for a Japanese company can be challenging for many people, but the difficulty depends greatly on the individual's cultural background, work style, and expectations.
Here are some common factors that can make it challenging:
1. Work Culture and Hours
Japanese companies traditionally expect long working hours, including overtime.
Commitment, loyalty, and company dedication are strongly emphasized.
Taking frequent vacations or leaving work precisely at the end of your scheduled hours might be viewed negatively.
2. Hierarchy and Communication
Hierarchical structures are prevalent, with clearly defined roles and seniority.
Decision-making can be slower, with multiple layers of approvals required.
Indirect communication is common; clear criticism or direct feedback may be avoided, making it harder for foreigners accustomed to straightforward discussions.
3. Language Barrier

Japanese proficiency is often crucial, even if the company's official business language is English.
Many internal documents, meetings, and day-to-day communications will likely occur in Japanese.
4. Group Harmony (Wa 和)
Harmony within the group is prioritized, sometimes above individual needs.
Expressing disagreement openly can be considered disruptive, so employees often need to navigate social cues carefully.
5. Expectations and Pressure
High expectations regarding quality, punctuality, attention to detail, and consistency.
Performance pressure can be significant, with regular evaluations and feedback.
Potential Benefits and Positive Aspects
On the other hand, there are positive aspects to working in a Japanese company:
Stability: Many Japanese companies provide job security, stable income, and comprehensive benefits.
Professional Growth: Exposure to disciplined, meticulous business practices that emphasize quality and efficiency.
Skill Development: Opportunity to gain proficiency in Japanese and develop intercultural competencies.
Team Orientation: Emphasis on teamwork and collective achievement, which can create a strong sense of belonging and cooperation.
Recommendations:
Develop at least intermediate Japanese language proficiency.
Familiarize yourself with Japanese business etiquette and social customs.
Be open-minded, patient, and willing to adapt to a culturally different working environment.
In short, working at a Japanese company can be rewarding, but it requires patience, flexibility, and a willingness to adapt to cultural differences.

| Factual correctness | Textual quality | Independence from baseline |
|---|---|---|
| Y | Y | N/A |

**JA**

**3.5.** 日本の会社で働くのは大変ですか？

日本の会社で働くのは一般的に大変だと言われます。業界や会社、職種によっても違いますが、多くの場合以下のような特徴が挙げられます。
大変と感じる主な理由：
労働時間が長い
日本の会社では残業が比較的多く、「定時退社」が難しい職場もまだまだあります。
上下関係やマナーが厳しい
上司や先輩に対して丁寧な敬語を使い、組織の中での立場を強く意識しなければなりません。
有給休暇が取りにくい
日本では有給休暇を自由に取りづらい雰囲気が残っている企業も多くあります。

集団行動が求められる
チームや組織内の協調性を重視するため、個人主義的な働き方が難しい場合もあります。
暗黙のルールが多い
明文化されていないルールやマナーが多く、外国人や新入社員にとっては馴染むまで苦労することがあります。
一方で、改善の傾向も：
最近では働き方改革やワークライフバランスを重視する企業も増え、残業規制やテレワーク、柔軟な働き方を取り入れる企業も出てきています。
IT や外資系、ベンチャー企業などでは比較的自由な働き方が可能な場合もあります。
結論：
日本の会社で働くことは、文化や労働慣習の違いから特に外国人にとっては大変だと感じる人が多いですが、業界・企業によって差があります。職場選びを慎重に行えば、ストレスを軽減することも可能です。

| Factual correctness | Textual quality | Independence from baseline |
|---|---|---|
| Y | Y | Y |

**EN**

**3.6. What is the weather like in Russia?**

Russia is the largest country in the world, spanning multiple climate zones, so weather conditions can vary significantly across different regions. Here's an overview of current weather conditions and forecasts for several key cities:
В момента е 10° ·Ясно
Москва, Русия
Днес 14° 3° Слънчево
петък 16° 1° Ветровито
събота 6° -3° По-студено
неделя 0° -3° Сняг
понеделник 2° -5° Няколко превалявания от сняг
вторник 4° -4° Слънчево
сряда 3° -3° Облачно
В момента е 9° ·Ясно
Сейнт Питърсбърг, Русия
Днес 12° 3° Облаци и слънце
петък 8° -1° Ветровито
събота 2° -3° Малко снеговалеж
неделя 3° -2° Слънчево
понеделник 5° -1° Няколко превалявания от сняг
вторник 7° 0° Облаци и слънце
сряда 6° -2° Няколко превалявания, предимно в по-късните часове
В момента е 6° ·Облачно
Новосибирск, Русия
Днес 9° 5° Ветровито
петък 17° 1° По-топло

събота 6° 3° Значително по-студено
неделя 12° 2° Облачно
понеделник 16° 5° Сгъстяваща се облачност
вторник 15° 3° Облачно
сряда 14° 2° Частично слънчево
В момента е 3° ·Дъжд
Екатеринбург, Русия
Днес 7° 0° Дъжд
петък 4° 1° Ниски облаци
събота 9° -1° Възможно преваляване
неделя 10° 1° Облаци и слънце
понеделник 5° 0° Слаб валеж от дъжд
вторник 8° 1° Предимно облачно
сряда 8° 1° Слаб валеж от дъжд
As seen, western cities like Moscow and Saint Petersburg are experiencing relatively mild temperatures with a mix of sun and clouds, while eastern cities like Novosibirsk and Yekaterinburg are cooler with chances of precipitation. It's advisable to check the specific forecasts for your destination, as conditions can vary widely across Russia.

| Factual correctness | Textual quality | Independence from baseline |
|---|---|---|
| N | Y | N/A |

**RU**

**3.6. Какая погода в России?**

Россия —страна с обширной территорией, поэтому погодные условия значительно различаются в разных регионах. Ниже представлены текущие погодные условия и прогнозы для некоторых крупных городов России:
Москва:
В настоящее время 10° ·Ясно
Москва, Россия
Сегодня 14° 3° Солнечно
пятница 16° 1° Ветрено
суббота 6° -3° Холоднее
воскресенье 0° -3° Снег
понедельник 2° -5° Кратковременный снегопад
вторник 4° -4° Солнечно
среда 3° -3° Облачно
Санкт-Петербург:
В настоящее время 9° ·Ясно
Санкт-Петербург, Россия
Сегодня 12° 3° Переменная облачность
пятница 8° -1° Ветрено
суббота 2° -3° Небольшой снег
воскресенье 3° -2° Солнечно
понедельник 5° -1° Кратковременный снегопад
вторник 7° 0° Переменная облачность

среда 6° -2° Кратковременные ливни, преимущественно в позднее время
Новосибирск:
В настоящее время 6° ·Облачно
Новосибирск, Россия
Сегодня 9° 5° Ветрено
пятница 17° 1° Теплее
суббота 6° 3° Намного холоднее
воскресенье 12° 2° Облачно
понедельник 16° 5° Растущая облачность
вторник 15° 3° Облачно
среда 14° 2° Небольшая облачность
Для получения более точной информации о погоде в вашем регионе рекомендуется воспользоваться местными метеорологическими службами или онлайн-сервисами.

| Factual correctness | Textual quality | Independence from baseline |
|---|---|---|
| Y | Y | N |

## C. Statistical Analysis

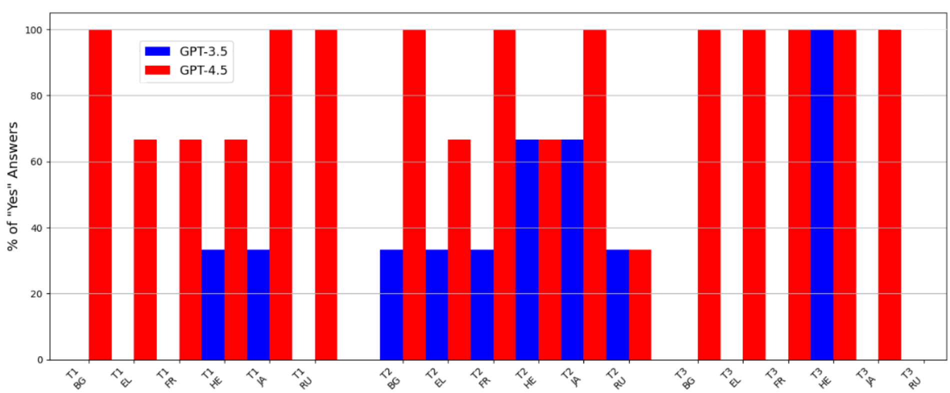


**Figure 7**
Percentage of 'Yes' scores for 'Independence from Baseline' per question type and per language

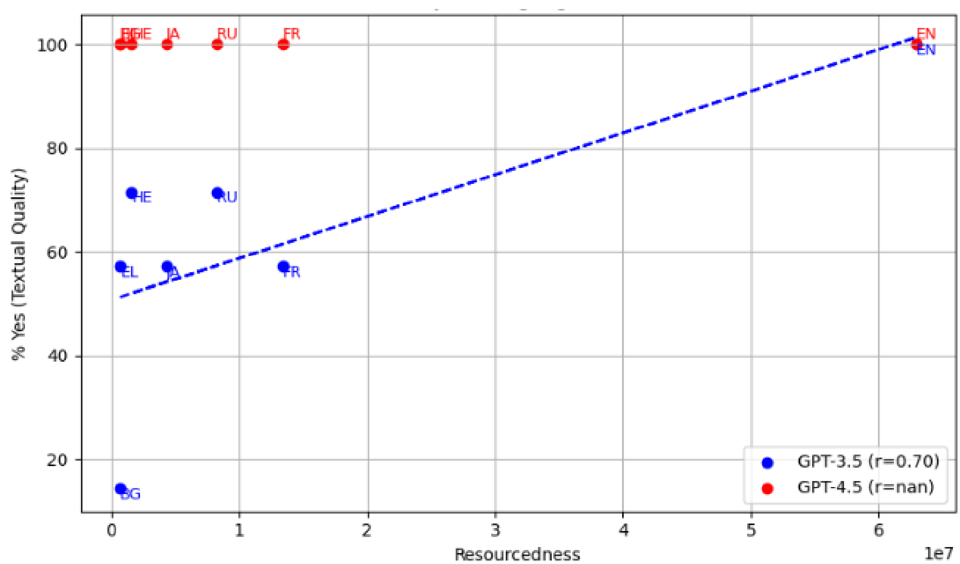


**Figure 8**
Resourcedness vs 'Textual' Quality
* The number of Wikipedia pages in the language is taken as a proxy for resourcedness.

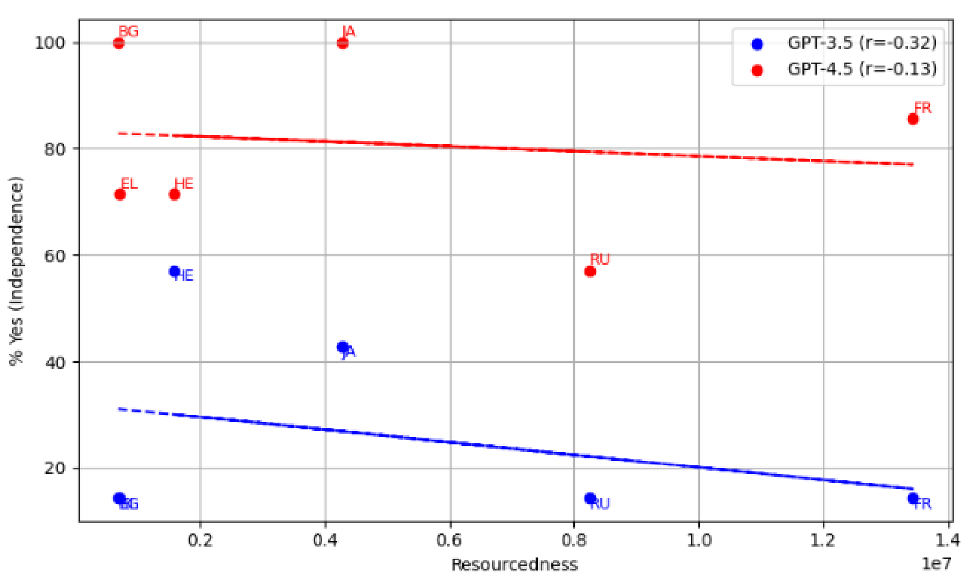


**Figure 9**
Resourcedness vs 'Independence from Baseline'
* The number of Wikipedia pages in the language is taken as a proxy for resourcedness.

# Light Verb Constructions in ELEXIS-WSD – Annotation, Comparisons and Issues

Cvetana Krstev
Association for Language Resources and Technologies
Belgrade, Serbia
cvetana@jerteh.rs

Ranka Stanković
University of Belgrade
F. of Mining and Geology
Belgrade, Serbia
ranka@rgf.rs

Aleksandra Marković
Institute for the Serbian Language SASA
Belgrade, Serbia
malexa39@gmail.com

*This paper deals with light verb constructions and their annotation in ELEXIS-SR, the Serbian extension of the ELEXIS-WSD corpus. In Section 1, general introductory remarks are given about these constructions, the notion of light verbs, and their treatment and further classification in the PARSEME annotation guidelines (subtypes LVC.full and LVC.cause). Section 2 offers an insight into ELEXIS-WSD corpus, annotated with VMWEs for several languages, with a remark that these VMWEs were not further subcategorised into finer classes. For this paper, we classified them ourselves to facilitate comparisons of the LVCs annotated in ELEXIS-SR. Tools and resources used for the automatic annotation of ELEXIS-SR are presented in Section 3, as well as the results of manual checking. In Section 4, we offer a comparison of LVCs in four ELEXIS-WSD sub-collections: Serbian, Bulgarian, Slovene, and English. We use Serbian as a starting point for this comparison, as it has been thoroughly annotated with MWEs (and NEs). We present the results of the comparison of all the occurrences of LVCs in the Serbian extension with their occurrences and annotation both in ELEXIS-WSD and Parseme sub-corpora for other languages.*

*An important conclusion is that the most equivalents among LVCs are between Serbian and Bulgarian, closely related Slavic languages (a total of 34 equivalents), while between Serbian and Slovene, also Slavic, there are 11 equivalents, as between Serbian and English. It seems that this could be explained by the number of VMWES and LVCs annotated, or by the strategy used by different annotators.*

**Keywords**: *Light verb constructions, annotation, ELEXIS-WSD, Serbian, Bulgarian, Slovene, English*

https://doi.org/10.47810/JCLIB.1.2025.03

## 1. Light verb constructions – LVCs

### 1.1 General

This study deals with the identification and annotation of light verb constructions (LVCs) in the Serbian extension of ELEXIS-WSD corpus. One of our aims is to take a comparative look at these constructions in Serbian and their different translations in the ELEXIS corpora for some other languages. Our study could also serve as a contribution to the study of LVCs in the Serbian language.

Light verbs, first mentioned in (Jespersen 1965),[1] was a frequent research subject (a rather detailed overview of related work can be found in (Stoyanova, Leseva, and Todorova 2016). While the term 'light verb construction' is used in English, in other languages, the terminology for this notion is not uniform (for example, as (Wittenberg 2016) notes in German, or as we will demonstrate below in Serbian).

The first mention of the phenomenon we are dealing with in this paper in Serbian was in (Radovanović 1990). The author names it 'predicate decomposition' and treats it as a language universal and a part of global nominalization processes in language. The author also mentions that syntactic models with predicate decomposition show a considerable extent of phraseologisation and that their lexical components exhibit a kind of collocation. Relevant to our research is that, as the author observes, nominalization processes and predicate decomposition are frequently represented in certain functional styles, namely those that favour abstract and intellectual, general language use, like those of official documents, scientific prose, and publicistic literature. This fact is relevant because the nature of the ELEXIS corpus we are working on (see Section 2) supports this observation: it lacks the language of belles-lettres, as well as that of everyday use. Similar observations are found in (Samardžić 2007), where the author uses the term 'light verb constructions' for the first time for the Serbian language and says it is a widespread linguistic phenomenon, which, although very productive, exhibits some collocational properties, and because of this requires special attention in translation, second language learning and teaching, as well as in lexicography. Based on the analysis of a small sample of parallel sentences, the conclusion showed that LVCs in English cannot be translated into Serbian word-for-word and are always translated with a single perfective verb. Besides these, there was a bundle of research dedicated to this and related problems (periphrastic predicates in Slavic languages (Topolińska 1982), and in Serbian (Lazić-Konjik 2006); on predicate decomposition (Ivić 1988), to mention some of them).

In (Wittenberg 2016), the author mentions what she considers the essential characteristics of light verb constructions. Namely, they are complex predicates composed of the light or semantically bleached verb and its nominal part, the event nominal. As this author says, most of the predicative meaning of LVCs comes from the event nominal, which assigns semantic roles to the subject, besides the light verb itself. In these constructions,

1 Otto Jespersen was the first author who used the term *light verb*. However, it seems that he used it conditionally, since the word *light* is in quotation marks.

a phenomenon known as argument sharing occurs: a subject is not an agent only of the light verb but also of the event nominal. There is a suspension of the canonical one-to-one correspondence between syntactic positions and semantic roles. On the syntactic level, LVCs behave similarly to their non-light counterparts (for example, in the sentences *The woman is giving the man a kiss* and *The woman is giving the man a book* both predicates are ditransitive). But LVCs exhibit a mismatch between semantic and syntactic representation levels: "The linguistic structure of light verb constructions looks syntactically like in non-light constructions, but semantically like in base verbs." (Wittenberg 2016). This author also thinks that the specific semantics and structure of LVCs, as well as an undefined repertoire of LVCs in languages, make their recognition and annotation an important task.

### 1.2 LVC in Parseme corpora

PARSEME corpus version 1.3 represents a multilingual corpus comprising 26 languages (including Serbian, English, Bulgarian and Slovene) that is annotated with verbal multiword expressions (VMWE) (Savary et al. 2023). PARSEME annotation guidelines[2] define multi-word expressions (MWE) as continuous or discontinuous sequences of words that show some degree of orthographic, morphological, syntactic, or semantic idiosyncrasy with respect to what is considered general grammar rules of a language. The component words of such a sequence have to include a headword and at least one other syntactically related word and at least two of its components have to be lexicalized. The most salient property of MWEs is semantic non-compositionality, that is, it is often impossible to deduce the meaning of the whole unit from the meanings of its parts and from its syntactic structure.

A verbal MWE is a multiword expression whose neutral form[3] is such that it has a distribution of a verb, a verbal phrase or a verbal clause, and its syntactic head is a verb.

Parseme distinguishes the following categories of VMWEs:

- universal categories:
  - *light verb constructions* (LVC);
  - *verbal idioms* (VID);
- quasi-universal categories:
  - *inherently reflexive verbs* (IRV);
  - *idiomatic verb-particle constructions* (IVPC);
  - *multiverb constructions* (MVC);
- language-specific categories, defined for some particular languages;
- an optional experimental category, *inherently adpositional verbs* (IAV).

*Light verb constructions* have the following two characteristics:

- They are formed by a verb $v$ and a single or compound noun $n$, which either directly depends on $v$ or is introduced by a preposition.

2 Parseme annotation Guidelines
3 In Parseme, a neutral form of a MWE is its least syntactically marked form which preserves its meaning.

- The noun $n$ is predicative and refers to an event (e.g. decision, visit) or a state (e.g. fear, courage). Predicative nouns are nouns that have semantic arguments, that is, they express predicates whose meaning is only fully specified by their semantic arguments.

Two sub-categories of VMWEs are recognized that define two different categories of LVCs:

- Type LVC.full: The verb $v$ is “light” in that it contributes to the meaning of the whole only by bearing morphological features: person, number, tense, mood and morphological aspect. This implies that $v$’s syntactic subject is $n$’s semantic argument. Examples:[4]
  - EN *to make a presentation* – a semantic argument of a noun (*presentation*) is a syntactic subject, a presenter;
  - BG *давам изявление* lit. to give a statement, ‘to make a statement’ – a semantic argument of a noun (*изявление* – statement) is a syntactic subject;
  - SL *imeti predavanje* lit. to have a lecture, ‘to give a lecture’ – a semantic argument of a noun (*predavanje* – lecture) is a syntactic subject, a lecturer;
  - SR *донети одлуку* lit. to bring a decision, ‘to make a decision’ – a semantic argument of a noun (*одлука* – decision) is a syntactic subject.
- Type LVC.cause: The verb $v$ is “causative” in that it indicates that the subject of $v$ is the cause or source of the event/state expressed by $n$. The noun $n$ has semantic arguments expressed as non-subject elements in the sentence, and the subject of the verb brings additional information, indicating the cause or source of the event/state. Examples:
  - EN *to grant rights* – $X$ has the right to $Y$, the granter is not a semantic argument of *rights*, but it causes $X$ to have the right to do $Y$;
  - BG *давам възможност* ‘to give an opportunity’ – $X$ has an opportunity to do $Y$, the giver is not a semantic argument of *възможност* ‘opportunity’, but it causes $X$ to have the opportunity for $Y$;
  - SL *narediti konec <nečemu>* lit. to make an end <to something> ‘to end <something>’, $X$ has an end, the syntactic subject is not a semantic argument of *konec* ‘end’, but it causes $X$ to ends;
  - SR *задати главобољу* lit. to cause headache ‘to give a headache’, $X$ has a headache, the syntactic subject is not a semantic argument of *главобоља* ‘headache’, but it causes $X$ to have it.

The Guidelines themselves contain tests that allow for the distinction of VMWEs from other MWEs, and then the distinction between various types of VMWEs.[5] We will not present these tests in detail here, but will mention briefly the tests used for LVCs. These tests

4 All examples in this subsection are taken from the Parseme Guidelines 1.3 website.
5 Specific tests for categorizing verbal MWEs

will be applied if the generic decision tree for verbal MWE candidates has determined that the candidate contains a unique verb $v$ as functional syntactic head of the whole, that this verb has a unique dependent, has no lexicalized subject and the morphosyntactic category of the dependent is an extended nominal phrase $n$.

- LVC.0 – *Noun is abstract*: Is the noun $n$ (single or compound) abstract? A "no" answer rejects the candidate as an LVC.
- LVC.1 – *Noun is predicative*: Does the noun $n$ have at least one semantic argument, implying that it is a predicative noun? A "no" answer rejects the candidate as an LVC.
- LVC.2 – *Verb's subject is noun's semantic argument*: Is the subject of the verb a semantic argument of the noun $n$? The answer "no" leads to test LVC.5.
- LVC.3 – *Verb with light semantics*: Is $v$ semantically light, that is, is the semantics that $v$ adds to $n$ restricted to: (i) what stems from its morphological features (e.g. future, plural, perfective aspect, etc.), (ii) pointing at the semantic role of $n$ played by $v$'s subject? A "no" answer rejects the candidate as an LVC. A "yes" answer or "unsure" leads to the next test.
- LVC.4 – *Verb reduction*: Is it possible to build an NP without the verb, in which $v$'s subject s becomes $n$'s dependent. A "yes" answer means that it is an LVC.full, a "no" answer rejects the candidate.
- LVC.5 – *Verb's subject is noun's cause*: Is the subject of the verb expressing the cause of the predicate expressed by the noun? A "yes" answer means that it is an LVC.cause, a "no" answer rejects the candidate.

### 1.3 LVCs in the Parseme: SR, EN, SL, BG

All four languages that we are dealing with in this paper are represented in the PARSEME corpus version 1.3. The size of corresponding sub-corpora measured in tokens, as well as the types of annotated VMWEs and their number differ significantly (see Table 1).

| Lng. | Tokens | VID | IRV | LVC | | VPC | | IAV | MVC |
|---|---|---|---|---|---|---|---|---|---|
| | | | | full | cause | full | semi | | |
| BG | 480,413 | 1,260 | 3,223 | 1,909 | 222 | 0 | 0 | 90 | 0 |
| EN | 124,203 | 187 | 0 | 333 | 51 | 368 | 53 | 71 | 51 |
| SL | 586,187 | 724 | 1626 | 239 | 64 | 0 | 0 | 710 | 0 |
| SR | 87,367 | 269 | 564 | 402 | 69 | 0 | 0 | 0 | 0 |

**Table 1**
Number of occurrences of VMWEs of different type in the Parseme corpus 1.3 for BG, EN, SL and SR.

We will briefly report on some research conducted in connection with the Parseme corpus.

Authors in (Gantar et al. 2019a) report on the structural and semantic classification of VMWEs in Slovene. Quantitative analyses of 3,364 sentences annotated with VMWEs showed that the least frequent category by type was LVC.cause (2%), LVC.full being right after with 7%, and the most frequent was IRV, 48%. The distinction between these two types of LVCs in the NLP context is explained in the next subsection. Interesting is their finding that LVC.full and LVC.cause are the least diverse categories (when talking about different VMWEs, which indicates that they form a closed class, with a limited list of their lexical components). Qualitative analysis showed: that combinations of a verb and a PP are more typical for the LVC.cause category; a relatively limited set of nouns is found in the annotated examples, some of which occur exclusively in LVC.cause (EN: 'effect', 'influence', 'help'), while others are characteristic for LVC.full type (en 'possibility', 'role', 'opinion'). The latter are more diverse, speaking of semantic classes. Among the most frequent verbs in LVCs are 'to have', 'to be', and 'to give'. This observation follows our results (see Section 3).

In (Gantar et al. 2019b), authors mention that LVCs are among MWEs (just like some verb + particle combinations, and some compounds, like *bus driver*) which can be included in dictionaries as lexical units, although semantically transparent.[6] LVCs appear with different degrees of idiomaticity, verbs in these constructions are sometimes void of meaning and can be paraphrased with the verbal form of the noun complement (*take a walk* vs. *walk*). Authors mention that morphology and syntax of LVCs can be unpredictable (e.g., there is only a limited number of nouns light verbs can combine with). The place of these constructions in the dictionary micro- and macro-structure varies; sometimes the LVCs are given as separate entries, sometimes under particular senses (lexical units), and sometimes among other MWEs, in the phrase section.

The semi-automatic compilation of the Dictionary of Bulgarian MWEs, among which nominal and verbal ones were predominant, was described in (Koeva et al. 2016). Since Bulgarian, like Serbian, is a morphologically rich language, many issues need to be addressed in the appropriate description of MWEs, including LVCs. One characteristic of LVCs is that they often take modifiers (BG: *vzemam (trudno/vazhno) reshenie* 'to make a (difficult/important) decision').

In (Leseva et al. 2024), the authors focus on developing a uniform approach to the description of MWEs, intending to create an electronic bilingual lexicon (BG-RO) of MWEs. The lexicon is derived from Bulgarian and Romanian wordnets, and verbal MWEs are being covered so far. The work offers the following description levels: lexical, derivational, morphological, syntactic, semantic, contextual, and stylistic. The authors conclude that among VMWEs, LVC and VID cases pose several challenges for proper description and analysis (on the other hand, IRVs have regular structure, word order and syntactic properties). As for the internal syntactic structure of LVCs, for the LVC.full type, most expressions exhibit the V+obj structure (EN: *give check*), while LVC.cause type displays two internal structure types, V+xcomp (EN: *make public, make equal*), and V+[case+obl] (EN: *put into circulation*). As for the external syntactic structure of LVCs, LVC.cause type has a valence frame with

6 "Many word combinations are very frequent and fixed in structure, but in terms of predictability of meaning, they dwell in the gray area between free combinations and MWEs (e.g. *dark chocolate*), and their classification and inclusion in dictionaries is rather arbitrary" (Gantar et al. 2019b, 7).

a subject and an obligatory object, e.g. “Bank puts money into circulation''. The LVC.full valence frame can contain only the subject, or the subject and a nominal, or a clause, etc.

## 2. ELEXIS-WSD corpus

ELEXIS-WSD is a parallel sense-annotated corpus in which content words (nouns, adjectives, verbs, and adverbs) have been assigned senses for 10 languages: Bulgarian (BG), Danish (DA), English (EN), Spanish (ES), Estonian (ET), Hungarian (HU), Italian (IT), Dutch (NL), Portuguese (PT), and Slovene (SL).[7] The list of sense inventories is based on WordNets for DA (Pedersen et al. 2023), EN, IT, NL, Wiktionary is used for ES, and national digital dictionaries are used for BG, ET, HU, PT, and SL (Martelli et al. 2021).

All corpora were morpho-syntactically tagged, and to a certain extent, multi-word expressions (MWE) and named entities (NE) were also annotated. The number of different MWEs and NEs annotated per language and the number of different senses associated with them is represented in Table 2. We can observe that the number of annotated MWEs and NEs differs significantly per language; e.g., 7 MWEs for Hungarian compared to 440 for Danish. It should also be noted that the different types of MWEs and NEs were not distinguished. Also, for some languages, for example, Slovene, NEs were annotated as MWEs.

| | MWE | | NE | |
|---|---|---|---|---|
| **Lang.** | **lemma** | **sense** | **lemma** | **sense** |
| Bulgarian | 299 | 465 | 2 | 2 |
| Danish | 440 | 477 | 440 | 459 |
| English | 179 | 309 | 1 | 1 |
| Spanish | 36 | 40 | 4 | 8 |
| Estonian | 177 | 217 | 112 | 145 |
| Hungarian | 7 | 7 | 6 | 6 |
| Italian | 41 | 42 | 0 | 0 |
| Dutch | 33 | 37 | 27 | 27 |
| Portuguese | 113 | 115 | 14 | 15 |
| Slovenian | 385 | 451 | 0 | 0 |
| **Total** | 1,710 | 2,160 | 606 | 663 |

**Table 2**
Number of MWEs and NEs in the repository; the second and fourth columns present the number of unique lemmas in the WSD, while the third and fifth columns present the number of unique senses.

Since this paper deals with comparing verbal multi-word expressions in Serbian and their usage in Bulgarian, Slovene, and English, we analysed the types of VMWEs annotated in the ELEXIS-WSD corpus for these three languages. As we already explained, annotated MWEs in ELEXIS-WSD were not classified into finer categories, so we classified them ourselves based on the morphosyntactic tagging, VMWE syntactic structure and information

7 Parallel sense-annotated corpus ELEXIS-WSD 1.1

obtained from the Parseme corpus (for VMWEs that occur in it, that is, that are annotated as VMWEs in it). The results are presented in Table 3. We can observe that in Bulgarian and Slovene corpora, most annotated VMWEs are reflexive verbs, while in English, verb-particle constructions prevail.

| | IRV | | VPC | | LVC.full | | VID | | VERB | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Lang.** | **L** | **S** | **L** | **S** | **L** | **S** | **L** | **S** | **L** | **S** |
| BG | 141 | 306 | / | / | 5 | 5 | 9 | 10 | 17 | 17 |
| EN | / | / | 34 | 156 | 2 | 2 | 3 | 3 | 2 | 3 |
| SL | 75 | 130 | / | / | / | / | 1 | 1 | 2 | 3 |

**Table 3**
Number of VMWEs in the repository for BG, EN, and SL, per lemmas (L) and senses (S); VERB refers to VMWEs that could not be categorized.

The Serbian part of the ELEXIS-WSD corpus, dubbed ELEXIS-SR, is being prepared in the scope of Working Group 2 of the UniDive COST action.[8] The translation from the English corpus of 2,024 sentences has been completed. Care was taken to ensure that the translation was in idiomatic, natural Serbian language while maintaining at the same time the meaning of the sentences in English. Tokenization, lemmatization, and POS-tagging were done automatically (Stanković et al. 2020; Stanković, Škorić, and Šandrih Todorović 2022), and controlled by at least three evaluators. It has been annotated with MWEs (see Section 3), and NEs that have also been linked to Wikidata knowledge database (Nešić et al. 2024). The repository of senses has been prepared on the basis of the Serbian WordNet (Krstev et al. 2025; Stanković et al. 2018), while the mapping of words (including MWEs) to senses is work in progress.

### 3. Annotation of MWE in the ELEXIS-SR

Automatic annotation of the Serbian set of 2,024 sentences with MWEs was done using different resources and tools. Resources consisted of morphological e-dictionary of Serbian simple and multi-word units, while tools were based on systems of finite-state automata that rely on these e-dictionaries (Krstev 2008).

- The e-dictionary of non-verbal MWEs (nominal, adjectival and functional) was used to annotate this type of MWEs. This dictionary contains all inflected forms of MWEs, if the words are subject to inflection, and associates them with the part-of-speech, lemma, and morphosyntactic category codes (Krstev et al. 2013). Among them were 444 nominal MWE occurrences, 80 preposition occurrences, 44 adverb occurrences, 35 conjunction occurrences, and 2 adjective occurrences.

8 COST Action CA21167 UniDive – universality, diversity and idiosyncrasy in language technology

| Type | Tot. | TP | FN | FP | Precision | Recall | F1 |
|---|---|---|---|---|---|---|---|
| NID | 724 | 413 | 311 | 31 | 0.93 | 0.57 | 0.71 |
| AdjID | 5 | 2 | 3 | 0 | 1.00 | 0.40 | 0.57 |
| AdpID | 78 | 73 | 5 | 7 | 0.91 | 0.94 | 0.92 |
| AdvID | 83 | 38 | 45 | 6 | 0.86 | 0.46 | 0.60 |
| ConjID | 52 | 34 | 18 | 1 | 0.97 | 0.65 | 0.78 |
| IRV | 290 | 195 | 95 | 10 | 0.95 | 0.67 | 0.79 |
| LVC.full | 82 | 37 | 45 | 3 | 0.93 | 0.45 | 0.61 |
| LVC.cause | 5 | 1 | 4 | 9 | 0.10 | 0.20 | 0.13 |
| VID | 51 | 11 | 40 | 0 | 1.00 | 0.22 | 0.35 |
| PronID | 7 | - | 7 | - | - | - | - |
| PartID | 6 | - | 6 | - | - | - | - |
| NV.LVCfull | 2 | - | 2 | - | - | - | |
| Total | 1385 | 804 | 581 | 67 | 0.92 | 0.58 | 0.71 |

**Table 4**
Total number of MWEs per type; true positives, false negatives, false positives; precision, recall and F1 measure. True positives are correctly recognised MWEs, their scope and their type. For the explanation of types, see Parseme Annotation Guidelines 2.0.

- A system for the recognition of verbal MWEs based on e-dictionaries, rules, and the repertoire of VMWEs annotated in the Serbian part of the PARSEME Corpus Release 1.3 (Savary et al. 2023) retrieved 266 occurrences, distributed by type: IRV – 205, LVC.full – 40, VID – 11, and LVC.cause – 10. This system was developed for Unitex[9], a program that uses electronic dictionaries and finite state transducers (FST) for corpus analysis. For this purpose, a collection of FSTs was developed that recognises and annotates various types of VMWEs. In simplified terms, the rule for verbs of type LVC.full, which use the verb *dati* and its imperfective counterpart *davati* 'to give', would be:

```
((<dati.V>|<davati.V>) <WORD>{0,n} (primer|ocenu|mišljenje|...))|
((primer|ocenu|mišljenje|...) <WORD>{0,n} (<dati.V>|<davati.V>))
```

  In this expression `<dati.V>` and `<davati.V>` recognize all inflective forms of corresponding verbs, `<WORD>{0,n}` recognizes occurrences of 0 to $n$ arbitrary word forms, and `primer, ocenu, mišljenje,...` are some of the predicative nouns used with these verbs ('example', 'assessment', 'opinion',...) in the expected inflected forms.

- A system for the recognition of adjectival and verbal similes described in (Krstev, Jaćimović, and Vitas 2020; Krstev, Stanković, and Marković 2023) did not retrieve a single simile in this set of sentences, which could be expected given the factual genre of sentences.

9 Lexicon-based Corpus Processing Suite

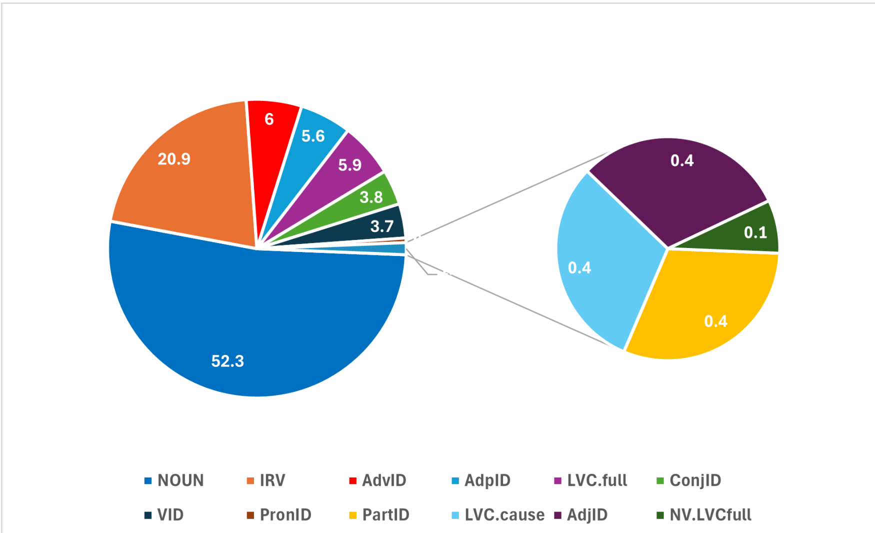


**Figure 1**
The distribution of MWE types in the manually verified dataset.

All automatically annotated MWEs were manually checked, and missing annotations were identified. Parseme guidelines were used to check the VMWE annotations, that is, decision trees presented in Section 1.2. Guidelines were not used for other MWE annotations (nominal and functional) because they were still in preparation at the time of this experiment, and we relied on our resources for the Serbian language. These resources record more than 20,000 MWEs of these types obtained from traditional sources and corpora research (Krstev et al. 2013). Of 871 automatically recognized MWEs, 817 were confirmed by manual check as MWEs, giving a precision of 0.94. 1385 MWEs were finally identified in ELEXIS-sr, which gives a recall of 0.59.

In some cases, the recognition was incorrect because the noun phrase was not associated with a correct verb. For instance, in sentence 38: *Ovo iskustvo imalo je snažan uticaj na njegov kasniji rad.* 'This experience had a strong influence on his later work.', system recognised as a VMWE *imati iskustvo* 'have experience' instead of *imati uticaj* 'have influence'. Such cases are a consequence of the fact that the system works locally on text that was not syntactically parsed.

Among these 817 correctly identified MWEs, only 13 were incorrectly classified. In 7 cases, the incorrect classification was due to the confusion between adverbs and prepositions, e.g. *u toku*_ADP 'during' and *u toku*_ADV 'ongoing'. In the remaining 6 cases, verbal MWEs were not appropriately classified as LVC.full, LVC.cause or VID. When we take classification into account, 804 MWEs were correctly classified, giving a precision 0.92. The annotation and check results per MWE type are presented in Table 4.

Nominal idioms (NID) are the most numerous in the entire set (52.63%), followed by inherently reflexive verbs (IRV) (20.9%) (see Figure 1). Except for pronominal and particle idioms,[10] that were not recognised by the automatic procedures, the most omissions, relative to the total number, occurred among verbal idioms (VID, $R = 0.22$), while the fewest occurred among prepositions (AdpID, $R = 0.94$). The most incorrect recognitions occurred among causative light verb constructions (LVC.cause, $P = 0.1$), while none occurred among adjectives (AdjID, $P = 1.00$), and very few among conjunctions (ConjID, $P = 0.97$) (see Figure 2).

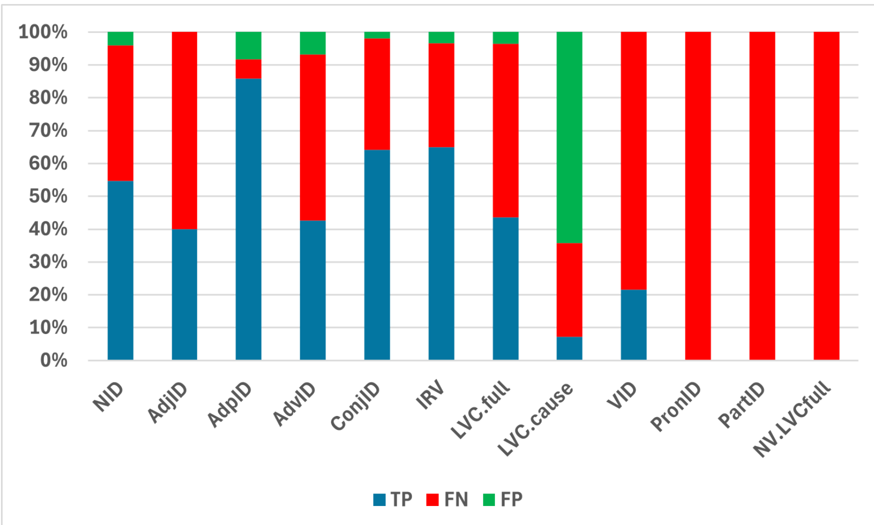


**Figure 2**
The percentage of true positives, false negatives and false positives per MWE type.

Five different causative LVCs were annotated in the corpus. The verb *izazvati* was used twice – *izazvati reakciju* 'arouse reaction' and *izazvati protivljenje* 'draw opposition'. The other three verbs are: *doneti* (*slavu*) 'bring (glory)', *praviti* (*problem*) 'cause (trouble)', and *napraviti* (*revoluciju*) 'make (revolution)'.

Among 82 VMWEs annotated as LVC.full, there are 56 different. The most frequent is *dobiti ime* 'lit. to get a name' with 11 occurrences, and *imati uticaj* 'have influence' with 5 occurrences. In these 56 VMWEs appear 33 different verbs. The most frequently used verb in this category is *imati* 'to have' and its negation *nemati*, occurring in 11 different VMWEs with the frequency 22. They are followed by verbs *dobiti* and *dati* 'to give' occurring in 5 different VMWEs (frequency 15) and 6 different VMWEs (frequency 6), respectively.

The noun *podrška* 'support' occurs in three different LVCs: *imati podršku* 'to have support', *dobiti podršku* 'to get support', *pružati podršku* 'to provide support'. Nouns occurring in two different LVCs are: *ime* 'name' (*dati ime* 'to give name', *dobiti ime* 'to get name'), *mišljenje* 'opinion' (*imati (visoko) mišljenje* 'to have (high) opinion', *izneti mišljenje* 'to express

10 Particle MWEs are not included in the current version of the Parseme Guidelines v2.0. MWEs like *na primer* 'for example' and *sve u svemu* 'all in all' were annotated as PartID.

opinion'), *nagrada* 'award' (*dobiti nagradu* 'to get award', *dodeliti/dodeljivati nagradu* 'to award prize'), *posao* 'job' (*dobiti posao* 'to get job', *obavljati posao* 'to do job'), *pravo* 'right' (*imati/nemati pravo* 'to have right', *dati pravo* 'to grant right'), *problem* 'problem' (*imati problem* 'to have problem', *rešiti problem* 'to solve problem'), *uslov* 'condition' (*zadovoljavati uslov* 'to satisfy condition', *ispunjavati uslov* 'to fulfill condition').

Some verbs occur in pairs: perfective and imperfective verbs. Those are *dodeliti* and *dodeljivati* 'to award', both used with the same noun *nagrada* 'award', and *izneti* and *iznositi* 'to state' used with different but semantically related nouns: *mišljenje*, *stav* and *viđenje* 'opinion'.

Two MWEs are annotated in this corpus as deverbal nouns derived from VMWEs, specifically LVC.full (NV.LVCfull).[11] These are *postizanje cilja* 'achieving a goal' connected to *postizati cilj* 'to achieve a goal' and *pružanje podrške* 'providing support' connected to *pružati podršku* 'to provide support'. However, the mentioned VMWEs do not occur in our corpus.

Light verb constructions are not systematically identified in Serbian monolingual dictionaries (either under verbal senses, among phrases, usage examples or collocations). The check was done for three of the most frequently used verbs in our corpus.[12] None of the 11 VMWEs that occur in our corpus using the verb *imati* 'to have' is mentioned in its dictionary entry listing 12 senses. Two LVCs evidenced in our corpus for the verb *dobiti* 'to get' are listed as usage examples in the verb's entry: *dobiti ime* 'to get a name', illustrating the sense *steći neku oznaku* 'to acquire a mark, label' and *dobiti posao* 'to get a job' related to the sense *biti postavljen na neku dužnost, položaj, biti proizveden, unapređen u neko zvanje* 'to be appointed to a certain position, to be promoted to a certain rank'. One LVC with the verb *dati*, namely *dati pravo* 'to grant right', is mentioned as a usage example of this verb in the sense *priznati, dodeliti* 'to recognize, to grant', as a part of the phrase *dati pravo građanstva* 'to grant citizenship'. The entry of this verb also states that "a verb with a noun as an object has the meaning of a verb derived from that noun." It holds for four annotated LVCs using this verb: *dati donaciju* 'to make a donation' ↔ *donirati* 'to donate', *dati doprinos* 'to make a contribute' ↔ *doprineti* 'to contribute', *dati ime* 'to give a name' ↔ 'imenovati', 'to name', *dati ocenu* 'to give a rating' ↔ *oceniti* 'to rate'.

## 4. Comparision of light verb constructions in the ELEXIS-WSD: SR, EN, SL, BG

We compared the occurrences of LVCs in four ELEXIS-WSD sub-collections: Serbian, English, Slovene, and Bulgarian. We have taken the Serbian sub-collection as a starting point for the comparison, since it has been carefully annotated with MWEs and NEs, as explained in Section 3, while in the currently available sub-collections for other languages, the annotation of MWEs was not given priority, as discussed in Section 2. The numeric results of the comparison are presented in Table 5. The results are presented for all the occurrences of LVCs in the ELEXIS-SR, since a LVC's equivalent in another language can in one case be

11 This category is introduced in the version 2.0 of Parseme guidelines.
12 For this check, we used *Речник српскога језика*. 2. изд. Нови Сад: Матица српска. 2018.

a LVC of the same type, while in the other case it can be a VMWE of another type, or not a VMWE at all. For instance, the equivalents of the LVC *imati pravo* are in ELEXIS-EN a LVC 'to have right' (twice) and an adjective 'eligible' (once), while in ELEXIS-SL they are a LVC 'imeti pravico' (twice) and an adverb 'lahko' (once).

| | LVC.full | | | | LVC.cause | | | |
|---|---|---|---|---|---|---|---|---|
| Lang. | LVC candidate | ELEXIS MWE | Parseme same | Parseme diff. | LVC candidate | ELEXIS MWE | Parseme same | Parseme diff. |
| SR | | 82 | 24 | 3 | | 5 | 1 | 0 |
| EN | 54 | 1 | 8 | 3 | 4 | 0 | 1 | 0 |
| SL | 40 | 2 | 6 | 5 | 3 | 0 | 0 | 0 |
| BG | 57 | 6 | 25 | 9 | 4 | 0 | 0 | 1 |

**Table 5**
Comparison of LVCs in SR, EN, SL and BG in ELEXIS and Parseme corpora.

In Table 5, results are presented separately for the light word constructions of type LVC.full and LVC.cause. The number of phrases in one of the analysed languages that correspond to LVCs in Serbian and can be candidates for LVC are given in column "LVC candidate". We treated as LVC candidates phrases that have a unique verb as the functional syntactic head, one nominal phrase as the syntactic dependent that is not the subject of the phrase, and that besides this satisfy tests LVC.0 and LVC.1 (see Subsection 1.2). This means that the referent of the nominal phrase has abstract and predicative meaning. The column "ELEXIS MWE" gives the number of LVC candidates that were annotated as MWE in the ELEXIS corpus of the corresponding language. The column "Parseme" gives the number of LVCs that were retrieved in the Parseme corpus (Savary et al. 2023): the number of LVCs that were always assigned the same type as in ELEXIS-SR is given in column "same", while the number of VMWEs (both LVC and VID) that were always or in some cases assigned a different type are given in column "diff." When retrieving VMWEs from Parseme corpus, we overlooked differences in conjugation or declension. We also treated as a match those cases in which the used verb has a different aspect, for instance in Serbian *ispuniti obavezu* (perfective) and *ispunjavati obavezu* (imperfective) 'fulfill an obligation' or in Bulgarian *изпълня ангажимент* (perfective) and *изпълнявам ангажимент* (imperfective). For instance:

- A LVC.full *iskoristiti prednost* used in sentence 588 (+1 in column 'ELEXIS MWE' for SR) has as an equivalent in EN a LVC candidate *take advantage* (+1 in column 'LVC candidate' for EN) that was annotated as MWE in ELEXIS-EN (+1 in column 'ELEXIS MWE' for EN). This LVC occurs in Parseme-EN three times: twice it is annotated as LVC.full and once as VID (+1 in column 'Parseme diff.' for EN)
- A LVC.full *dati donaciju* used in sentence 1950 (+1 in column 'ELEXIS MWE' for SR) has as an equivalent in EN a LVC candidate *make donation* (+1 in column 'LVC candidate' for EN), it was not annotated as MWE in ELEXIS-EN

(+0 in column 'ELEXIS MWE' for EN), but it occurs in Parseme corpus where it is annotated as LVC.full (+1 in column 'Parseme same' for EN).[13]

The phrase *imati pravo*, which occurs three times in ELEXIS-SR and is annotated as LVC.full, is the only phrase for which in other languages the same equivalent phrase was used and annotated as LVC.full.[14] One example is (sentence number 1790):

SR Narod *ima pravo*$_{\text{LVC.full}}$ da od ombudsmana traži intervenciju.

EN The people *have right*$_{\text{LVC.full}}$ to ask the ombudsman for intervention.

SL Ljudje *imajo pravico*$_{\text{LVC.full}}$ zaprositi varuha človekovih pravic za posredovanje.

BG Хората *имат право*$_{\text{LVC.full}}$ да помолят омбудсмана да се намеси.

The phrase *imati za cilj* 'to have aim', which also occurs three times in ELEXIS-SR and is annotated as LVC.full never has as the equivalent in other languages the LVC.full phrase nor any VMWE phrase of other type.[15] One example is (sentence number 178):

SR Evropska komisija *ima za cilj*$_{\text{LVC.full}}$ da nagradama podstakne prekograničnu cirkulaciju popularnog muzičkog repertoara…

EN With the awards, the European Commission *aims* to stimulate the cross-border circulation of popular music repertoire…

SL Z nagrado *želi* Evropska komisija spodbuditi čezmejno kroženje repertoarja popularne glasbe…

BG С тези награди ЕК *цели* да се стимулира разпространението на популярна музика в различните държави…

The phrase *imati/nemati uticaj* 'to have (no) effect', which occurs six times in ELEXIS-SR and is annotated as LVC.full, has in one case as the equivalent in other languages the VMWE phrase, but its category varies.[16] The example is (sentence number 1551):

SR Vaspitni stil roditelja izgleda *nema* veliki *uticaj*$_{\text{LVC.full}}$,…

EN Parenting style seems to *have* no major *effect*$_{\text{LVC.full}}$,…

SL Zdi se, da slog starševstva *nima* veliko *vpliva*$_{\text{LVC.cause}}$,…

BG Изглежда родителският стил *няма* голям *ефект*$_{\text{LVC.full/VID}}$,…[17]

13 This LVC.full occurs in Parseme-EN once.

14 The similar phrase is used in other languages and annotated as LVC.full in the Parseme corpus, for example: HR *imati pravo*, FR *avoir droit*, PL *mieć prawa*.

15 The similar phrase is used in other languages and annotated as LVC.full in the Parseme corpus, for example: HR*imati cilj*, FR *avoir but*, PL *mieć cel*, SL *imeti cilj* (although it was not used in ELEXIS-SL.).

16 The similar phrase used in other languages is annotated as LVC.full in the Parseme corpus, for example: HR*imati utjecaj*, FR *avoir influence*, PL *mieć wpływ*.

17 In Bulgarian Parseme corpus this VMWE is twice annotated as LVC.full, once as VID, and one occurrence is skipped.

An interesting issue arises with this phrase, which has equivalents of different verbal types. We conclude that this particular LVC causes problems in that its meaning is causative, but its type is LVC.full. This is similar to the cases which contain typical causative verbs, but their type is LVC.full. There is a subtle distinction between the cause subject as a canonical argument (LVC.full) and its "external", non-canonical use (LVC.cause).

As mentioned in Section 3, the verb *imati/nemati* is used in the largest number of LVCs (11). In cases when an LVC is also used in another language, it is always *to have* in English, *imeti* in Slovene, and *имам/нямам* and *окажа/оказвам* in Bulgarian. The verb *dati* used in six LVCs, has equivalents in English with verbs 'to give' and 'to make', and in Bulgarian with verbs 'имам', 'дам/давам' and 'подавам'. The verb *dobiti* 'to get', used in five different LVCs, has no equivalents in English and Slovene, but two equivalents in Bulgarian: *dobiti ime* 'to give name' ↔ 'получавам име' and *dobiti posao* 'to get job' ↔ 'наема на работа'. In the last case, there is no true equivalence, since the Bulgarian phrase is an exact translation of English 'to hire for a job'. Serbian light verb constructions with *imati* and *dati* that have equivalents in other languages are represented in Table 6.

| Serbian LVC | | Equivalent |
|---|---|---|
| *imati/nemati pravo* | EN | to have (no) right |
| | SL | imeti pravico |
| | BG | имам право |
| *imati/nemati uticaj* | EN | to have (no) effect |
| | SL | imeti vpliv |
| | BG | окаже влияние, имам влияние, окажа/оказвам въздействие |
| *imati/nemati problem* | EN | have (no) problem |
| | SL | imeti težave |
| *imati/nemati dejstvo* | EN | to have (no) effect |
| | BG | оказвам влияние |
| *imati/nemati šansu* | SL | imeti možnost |
| *imati pristup* | BG | имам достъп |
| *dati pravo* | EN | to give right |
| | BG | давам право |
| *dati donaciju* | EN | to make donation |
| *dati ocenu* | BG | дам оценка |
| *dati doprinos* | BG | имам принос |
| *dati ostavku* | BG | подавам оставка |

**Table 6**
Serbian LVCs with verbs *imati* and *dati* and their equivalents in EN, SL, BG.

The data in Table 5 indicate that a significant number of candidates were not marked as LVCs in the Parseme corpus. However, we cannot conclude that these candidates cannot be considered LVCs in the analysed languages, because our analysis did not include checking whether these candidates occurred in that corpus at all. One such example is (sentence 1699):

SR *Dobila*$_{\text{LVC.full}}$ je *podršku* mnogih bogatih i uticajnih sponzora.

EN It *won*$_{\text{LVCcandidate}}$ the *support* of many wealthy and influential backers.

SL *Uživalo*$_{\text{LVCcandidate}}$ je *podporo* številnih bogatih in vplivnih podpornikov.

BG Беше *спечелена*$_{\text{LVCcandidate}}$ *подкрепата* на много заможни и влиятелни поддръжници.

However, the equivalent LVC.full is annotated in the Polish part of the Parseme corpus: *otrzymywać wsparcie*, in the Bulgarian part (using a different verb): *получа/получавам подкрепа*, as well as in the French part: *recevoir soutien*. It should also be noted that the phrases *win / get support* do not occur in the Parseme-EN, and neither do *uživati podporo* in Parseme-SL (however, the near synonym *pridobiti podporo* occurs once in Parseme-SL and is not annotated as VMWE).

It should be stated here that some of the annotated LVCs in the ELEXIS-SR raise doubt. We can take as an example LVC.full *izneti/iznositi mišljenje/stav/viđenje* 'express/present opinion/view'. These phrases pass test LVC.3 (the subject of the verb is a semantic argument of the noun), but can we say that the paraphrase 'subject's opinion/view' expresses the same meaning? Our position here was that 'opinion/view' has to be expressed much as a lecture has to be given. We see in Table 7 that various similar phrases were used across languages, none of which was annotated as LVC.full (or other type of VMWE) either in Parseme or in the ELEXIS corpus. Corpus search using the GrewMatch tool[18] (Guillaume 2021) reveals that *express/present/put forth opinion/view* is not used in Parseme-EN, *predstaviti/izraziti stališće/mnenje* is used four times in Parseme-SL and never annotated as VMWE, while *представя/изразя-(се)/изкажа поглед/гледен точка/мнение* occurs 14 times in Parseme-BG in the form *изразя-(се) мнение* and it is 10 times annotated as LVC.full and 7 times skipped, while the form *изкажа мнение* occurs twice and is both times skipped. Finally, we should add that the annotation of Parseme-SR was not more consistent: *izneti/iznositi mišljenje/stav/viđenje* occurs 6 times: three occurrences of *izneti stav* were annotated as LVC.full, one is skipped, while both single occurrences of *izneti mišljenje* and *izneti viđenje* were skipped.

| Sent. | ELEXIS-SR | ELEXIS-EN | ELEXIS-SL | ELEXIS-BG |
|---|---|---|---|---|
| 23 | iznositi viđenje | present view | predstaviti sališće | представя поглед |
| 156 | izneti stav | express view | izraziti stališče | изразя гледен точка |
| 1731 | izneti mišljenje | put forth opinion | biti mnjenja | изкажа мнение |

**Table 7**
LVC.full *izneti/iznositi mišljenje/stav/viđenje* and its equivalents in ELEXIS-EN/SL/BG.

The data from Table 5 further show that the most equivalents among LVCs are between Serbian and Bulgarian (a total of 34), while between Serbian and English, as well as Serbian and Slovene, there are a total of 11 equivalents. This does not seem to depend on the size of

18 Grew-match

the Parseme corpus, as the Slovenian part is the largest but has significantly fewer LVCs than the Bulgarian, which is shorter in length (see Table 1). The Serbian corpus is the smallest of all but has a number of LVCs comparable to Slovene and English. It may be that the type of text plays a role in this or even the strategy that the annotators applied when annotating their corpora.

## 5. Conclusion

This paper deals with light verb constructions and their annotation in ELEXIS-sr, the Serbian extension of the ELEXIS-WSD corpus. We made general introductory remarks about these constructions, the notion of light verbs, and their treatment and further classification in the PARSEME annotation guidelines (subtypes LVC.full and LVC.cause).

Section 2 offers an insight into the ELEXIS-WSD corpus, annotated with VMWEs for several languages, with a remark that these VMWEs weren't further sub-categorized into smaller classes. For this paper, we classified them ourselves to be able to make comparisons of the LVCs annotated in ELEXIS-sr.

In Section 3 we presented tools and resources used in the automatic annotation with MWEs of ELEXIS-sr, as well as the results of manual checking. There are 5 VMWEs annotated as LVC.cause (all different) and 82 annotated as LVC.full (among them 56 different), with 33 different verbs, the most frequent being *imati* 'to have', followed by *dobiti* 'to get' and *dati* 'to give'. As for the nominal part of LVCs, the noun *podrška* 'support' appears in three different LVCs, and nouns *ime* 'name', *mišljenje* 'opinion', *nagrada* 'award', *posao* 'job', *pravo* 'right', *problem* 'problem', *uslov* 'condition' occur in two different LVCs each. We checked whether these and other LVCs are identified and represented in the macro- or microstructure of a monolingual Serbian dictionary and conclude that, as a rule, LVCs are usually not identified (with a few exceptions, which occur in different parts of the dictionary structure).

In Section 4, we offer a comparison of LVCs in four ELEXIS-WSD sub-collections: Serbian, Bulgarian, Slovene, and English. We use Serbian as a starting point for this comparison, as it has been manually checked for annotation with MWEs (and NEs). We present the results for all the occurrences of LVCs in the Serbian extension. Sometimes LVC has an equivalent LVC in another language, sometimes VMWE of another type, and sometimes a single word.

We took LVC candidates from the ELEXIS subcollections of other languages, as well as from the Parseme corpus, separately for LVC.cause and LVC. full types. These comparisons gave us the following insight: Only one phrase, *imati pravo*, annotated in ELEXIS-sr as LVC.full, has as its equivalent the same phrase, annotated as LVC.full in other sub-collections. On the contrary, sr *imati za cilj*, 'to have aim', never has either LVC.full phrase as the equivalent, or any other VMWE type.

An important conclusion is that the most equivalents among LVCs are between Serbian and Bulgarian, closely related Slavic languages (a total of 34 equivalents), while between Serbian and Slovene, also Slavic, there are 11 equivalents, as between Serbian and English. It seems that this could be explained by the number of VMWEs and LVCs annotated, or by the strategy used by different annotators.

This research showed that although LVCs are a universal phenomenon, their repertoire in each language has to be established separately, and their translation from one language to another has an idiomatic character. This is the reason why facts collected through lexicological inventorying facilitate the work on LVCs.

The results obtained from the presented research suggest that its extension to other languages and to other types of VMWEs could yield interesting results. Considering that the annotation of a multilingual corpus with other types of top-level MWEs (nominal, adjectival, adpositional, etc.) is being prepared within the framework of the UniDive COST action, the research could be extended in that direction as well.

## Acknowledgments

This research was supported by the COST ACTION CA21167 – UniDive, the Science Fund of the Republic of Serbia, #7276, Text Embeddings – Serbian Language Applications - TESLA, and the Ministry of Science, Republic of Serbia #GRANTS 451-03-136/2025-03/200174 and 451-03-136/2025-03/200126.

## References

Gantar, Polona, Špela Arhar Holdt, Jaka Čibej, and Taja Kuzman. 2019a. Structural and Semantic Classification of Verbal Multi-Word Expressions in Slovene. *Contributions to Contemporary History*, 59(1):99–119.

Gantar, Polona, Lut Colman, Carla Parra Escartín, and Héctor Martínez Alonso. 2019b. Multiword expressions: between lexicography and NLP. *International Journal of Lexicography*, 32(2):138–162.

Guillaume, Bruno. 2021. Graph Matching and Graph Rewriting: GREW tools for corpus exploration, maintenance and conversion. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 168–175, Association for Computational Linguistics, Online.

Ivić, Milka. 1988. Još o dekomponovanju predikata. *Južnoslovenski filolog*, 44:1–5.

Jespersen, Otto. 1965. A modern English grammar on historical principles.

Koeva, Svetla, Ivelina Stoyanova, Maria Todorova, and Svetlozara Leseva. 2016. Semi-automatic compilation of the dictionary of Bulgarian multiword expressions. In *Proceedings of the Workshop on Lexicographic Resources for Human Language Technology*, pages 86–95.

Krstev, Cvetana. 2008. *Processing of Serbian – Automata, Texts and Electronic Dictionaries.* University of Belgrade, Faculty of Philology, Belgrade.

Krstev, Cvetana, Jelena Jaćimović, and Duško Vitas. 2020. Analysis of Similes in Serbian Literary Texts (1840-1920) Using Computational Methods. In *Proceedings of the Fourth International Conference Computational Linguistics in Bulgaria (CLIB 2020)*, pages 31–41, Institute for Bulgarian Language "Prof. Lyubomir Andreychin", Bulgarian Academy of Sciences, Sofia, Bulgaria.

Krstev, Cvetana, Ivan Obradović, Ranka Stanković, and Duško Vitas. 2013. An approach to efficient processing of multi-word units. *Computational Linguistics: Applications*, pages 109–129.

Krstev, Cvetana, Ranka Stanković, and Aleksandra Marković. 2023. Multiword Expressions – Comparative Analysis Based on Aligned Corpora. In *Book of Abstracts of the UniDive 1st general meeting, 16-17 March 2023, Paris-Saclay University, France.*

Krstev, Cvetana, Ranka Stanković, Aleksandra Marković, and Milica Ikonić Nešić. 2025. Progress in SR-ELEXIS Semantic Annotation: Focusing on Multiword Expressions, Named Entities, and Sense Repository. In *Book of Abstracts of the UniDive 3rd general meeting, 29-30 January 2025, Budapest, Hungary.*

Lazić-Konjik, Ivana. 2006. Struktura, funkcija i leksikografska obrada perifrastičnih predikata (na primerima iz dnevne štampe). *Zbornik Matice srpske za filologiju i lingvistiku*, 49(1):219–304.

Leseva, Svetlozara, Verginica Barbu Mititelu, Ivelina Stoyanova, and Mihaela Cristescu. 2024. A uniform multilingual approach to the description of multiword expressions. *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, pages 73–116.

Martelli, Federico, Roberto Navigli, Simon Krek, Jelena Kallas, Polona Gantar, Svetla Koeva, Sanni Nimb, Sussi Olsen, Margit Langemets, Kristina Koppel, et al. 2021. Designing the ELEXIS Parallel Sense-Annotated Dataset in 10 European Languages. In *Proceedings of the eLex 2021 conference*, pages 377–395, Lexical Computing.

Nešić, Milica Ikonić, Saša Petalinkar, Ranka Stanković, Miloš Utvić, and Olivera Kitanović. 2024. SrpCNNeL: Serbian Model for Named Entity Linking. In *2024 19th Conference on Computer Science and Intelligence Systems (FedCSIS)*, pages 465–473, IEEE.

Pedersen, Bolette, Sanni Nimb, Sussi Olsen, Thomas Troelsgård, Ida Flörke, Jonas Jensen, and Henrik Lorentzen. 2023. The DA-ELEXIS corpus - a sense-annotated corpus for Danish with parallel annotations for nine European languages. In *Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023)*, pages 11–18, Association for Computational Linguistics, Tórshavn, the Faroe Islands.

Radovanović, Milorad. 1990. *Spisi iz sintakse i semantike.* Izdavačka knjižarnica Zorana Stojanovića.

Samardžić, Tanja. 2007. Light verb constructions in English and Serbian. *English Language and Literature Studies–Structures across Cultures*, pages 59–73.

Savary, Agata, Cherifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev, Chaya Liebeskind, Hongzhi Xu, Sara Stymne, Tunga Güngör, Thomas Pickard, Bruno Guillaume, Eduard Bejček, Archna Bhatia, Marie Candito, Polona Gantar, Uxoa Iñurrieta, Albert Gatt, Jolanta Kovalevskaite, Timm Lichte, Nikola Ljubešić, Johanna Monti, Carla Parra Escartín, Mehrnoush Shamsfard, Ivelina Stoyanova, Veronika Vincze, and Abigail Walsh. 2023. PARSEME corpus release 1.3. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35, Association for Computational Linguistics, Dubrovnik, Croatia.

Stanković, Ranka, Miljana Mladenović, Ivan Obradović, Marko Vitas, and Cvetana Krstev. 2018. Resource-based WordNet Augmentation and Enrichment. In *Proceedings of the Third International Conference Computational Linguistics in Bulgaria (CLIB 2018)*, pages 104–114, Institute for Bulgarian Language "Prof. Lyubomir Andreychin", Bulgarian Academy of Sciences, Sofia, Bulgaria.

Stanković, Ranka, Branislava Šandrih, Cvetana Krstev, Miloš Utvić, and Mihailo Škorić. 2020. Machine Learning and Deep Neural Network-Based Lemmatization and Morphosyntactic Tagging for Serbian. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3954–3962, European Language Resources Association, Marseille, France.

Stanković, Ranka, Mihailo Škorić, and Branislava Šandrih Todorović. 2022. Parallel Bidirectionally Pretrained Taggers as Feature Generators. *Applied Sciences*, 12(10).

Stoyanova, Ivelina, Svetlozara Leseva, and Maria Todorova. 2016. Towards the automatic identification of light verb constructions in Bulgarian. In *Proceedings of the Second International Conference on Computational Linguistics in Bulgaria (CLIB 2016)*, pages 28–37, Department of Computational Linguistics, Institute for Bulgarian Language, Bulgarian Academy of Sciences, Sofia, Bulgaria.

Topolińska, Zuzanna. 1982. Perifrastični predikatski izrazi na međuslovenskim relacijama. *Južnoslovenski filolog*, 38:35–49.

Wittenberg, Eva. 2016. *With light verb constructions from syntax to concepts*, volume 7. Universitätsverlag Potsdam.

# Automatic Detection of the Bulgarian Evidential Renarrative

Irina Temnikova
Big Data for Smart Society Institute (GATE)
5, James Boucher St., Sofia, 1164, Bulgaria
irina.temnikova@gate-ai.eu

Ruslana Margova
Big Data for Smart Society Institute (GATE)
5, James Boucher St., Sofia, 1164, Bulgaria
ruslana.margova@gate-ai.eu

Stefan Minkov
Big Data for Smart Society Institute (GATE)
5, James Boucher St., Sofia, 1164, Bulgaria
stefan.minkov@gate-ai.eu

Tsvetelina Stefanova
Big Data for Smart Society Institute (GATE)
5, James Boucher St., Sofia, 1164, Bulgaria
tsvetelina.stefanova@gate-ai.eu

Nevena Grigorova
Big Data for Smart Society Institute (GATE)
5, James Boucher St., Sofia, 1164, Bulgaria
freelancer.neva@gmail.com

Silvia Gargova
Big Data for Smart Society Institute (GATE)
5, James Boucher St., Sofia, 1164, Bulgaria
silvia.gargova@gate-ai.eu

Venelin Kovatchev
School of Computer Science, University of Birmingham
University Rd W, Edgbaston, Birmingham, United Kingdom
v.o.kovatchev@bham.ac.uk

*Manual and automatic verification of the trustworthiness of information is an important task. Knowing whether the author of a statement was an eyewitness to the reported event(s) is a useful clue. In linguistics, such information is expressed through "evidentiality". Evidentials are especially important in Bulgarian, as Bulgarian journalists often use a specific type of evidential ("renarrative") to report events that they did not directly observe, nor verify. Unfortunately, there are no automatic tools to detect Bulgarian renarrative. This article presents the first two automatic solutions for this task. Specifically - a fine-tuned BERT classifier (renarrative BERT detector, BGRenBERT), achieving 0.98 Accuracy on the test split, and a renarrative rule-based detector (BGRenRules), created with regular expressions, matching a parser's output.*

https://doi.org/10.47810/JCLIB.1.2025.04

*Both solutions detect Bulgarian texts containing the most frequently encountered forms of renarrative. Additionally, we compare the results of the two detectors with the manual annotation of subsets of two Bulgarian fake text datasets. BGRenRules obtains substantially higher results than BGRenBERT. The error analysis shows that the errors from BGRenRules most frequently correspond to cases in which humans also have doubts. The training dataset (BgRenData), the annotated dataset subsets, and the two detectors are made publicly accessible on Zenodo, GitHub, and HuggingFace. We expect that these new resources will be of invaluable assistance to 1) Bulgarian-language researchers, 2) researchers of other languages with similar phenomena, especially those working on verifying information.*

**Keywords:** *evidentiality, Bulgarian, renarrative, fine-tuned BERT classifier, Python, annotation*

## 1. Introduction

Verifying the trustworthiness of information and automatically detecting factually incorrect information has become a topic that in the past few years attracted more attention (Guo et al. 2020; Shu et al. 2017; Das et al. 2023). We define "factually incorrect information" as information which contradicts the facts. It is considered that there are different types of factually incorrect information. "Misinformation" refers to factually incorrect information that is not intended to cause harm, while "disinformation" is factually incorrect information that is spread with the intention to deceive, and to cause harm[1]; "malinformation" is called information that stems from the truth, but is often exaggerated in a misleading way (Newman 2021; Wardle et al. 2018). In this article, we address all these three types of factually incorrect information and focus on the information's trustworthiness or reliability.

There are various automatic methods for verifying the trustworthiness of textual information. Depending on where it appears (for example in news media or social media), it may be verified by one or a combination of more than one from the following methods: using specific linguistic features (Dinkov, Koychev, and Nakov 2019; Atanasova et al. 2019; Zhou and Zafarani 2020), matching statements to databases of fact-checked claims (Panchendrarajan and Zubiaga 2024; Hangloo and Arora 2023), using social network-specific features (Rani, Das, and Bhardwaj 2022; Santhosh, Cheriyan, and Nair 2022) such as the popularity of the author, the number of likes, reposts, and comments.

Knowing whether the author of a news article or a social media post has been a witness to the information shared in the text, can help to verify the trustworthiness of the reported information (Grieve and Woodfield 2023). This knowledge is usually expressed linguistically by "evidentiality". The linguistic expressions of evidentiality are called "evidentials". Evidentials have been already used in natural language processing (NLP) for text trustworthiness detection (Su, Huang, and Chen 2010; Su, Chen, and Huang 2011) for other languages, but not for Bulgarian. The evidentials in Bulgarian language

1 https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2020%3A790%3AFIN&qid=1607079662423

are four - indicative, conclusive, renarrative, and dubitative. When using renarrative, the speaker reports something which he did not witness, without assessing whether the source is reliable or not (Nitzolova 2008).

Margova (2023) found that the form of Bulgarian evidential, called "ренаратив" (in English "renarrative") is most frequently used by Bulgarian journalists in one of two situations: 1) when journalists are unsure of the reliability of the information and want to put distance between themselves and their statement; and 2) in news articles that have been flagged as misleading by independent fact-checkers. In fact Renarrative is used in Bulgarian when the speaker/author wants to transmit information provided by a third party without expressing the knowledge and views of the speaker/author. In this way, the speaker/author does not express doubt, nor guarantee the truthfulness of the reported information. Additionally, he/she shows that the responsibility for the reliability of the reported information belongs to its source and depends on the interpretation of the recipient of the information (Nitzolova 2008; Margova 2023).

Due to these useful characteristics, the automatic detection of the Bulgarian "renarrative" can be an important additional linguistic cue for verifying the trustworthiness of information. We hypothesized that since renarrative is expressed morphologically through specific verb forms, it should be detectable by syntactic parsers or Part-of-Speech (POS) taggers. For this reason, we reviewed several publicly available Bulgarian Natural Language Processing (NLP) tools (see Section 2 for details). Unfortunately, we discovered that none of them was explicitly designed for detecting evidentials.

In this article, we propose the first automatic solutions to detect renarrative in Bulgarian texts. Our solutions include a fine-tuned BERT classifier (the renarrative BERT detector, BGRenBERT) and a renarrative rule-based detector (BGRenRules). Both detectors recognize when a short text or a sentence contains at least one form of renarrative. As the forms of renarrative often match the forms of other Bulgarian verb forms, our solutions detect only the renarrative forms which are pointed by an expert to be the most frequent ones, and also those which are the easiest to be automatically distinguished. The article describes the methods followed for fine-tuning the model and adapting the rule-based detector to reach the best results. We also report the performance results and an error analysis of comparing the performance of the two detectors on manually annotated subsets of two Bulgarian fake text datasets (Hardalov, Koychev, and Nakov 2016; Nakov et al. 2022).

In summary, our contributions are the following:

- a new machine learning (ML) text classifier and a rule-based detector for detecting texts with at least one of the most frequent forms of Bulgarian renarrative;
- a new text dataset used for fine-tuning BGRenBERT and adjusting BGRenRules to recognize the correct renarrative forms (BGRenData);
- manually annotated subsets of two Bulgarian fake texts datasets;
- linguistic insights about the presence of renarrative in these datasets.

All these resources are publically shared on Zenodo, GitHub, and HuggingFace to increase their visibility to the research community. We believe that our work will be very useful to Bulgarian linguists and other researchers working on similar topics.

The next sections include: Section 2 presents the linguistic theory about evidentiality and relevant related work, Section 3 explains which forms of renarrative we detect, Section 4 describe all the datasets used, Section 5 provides details about the two detectors, Section 6 contains the results of the comparison of the two methods on the manually annotated dataset subsets and an error analysis. Section 9 contains our planned future work and the conclusions, Section 7 discusses the ethical and legal considerations, Section 8 presents this work's limitations, and finally, Section 10 lists the authors' contributions and the acknowledgements.

## 2. Linguistic Theory and Related Work

In this section, we provide a short theoretical overview of evidentiality and present the work most relevant to ours. We consider the works most similar to ours those which automatically detect Bulgarian evidentials, and those which use evidentials for verifying texts' trustworthiness.

### 2.1 Evidentiality

**Evidentiality** is the grammatical expression of the information source, and usually assists in distinguishing whether the speaker witnessed the reported information, was told about it by somebody else, or, for example, inferred it based on common sense (Aikhenvald 2004, 2015). Evidentiality can express **the source of the knowledge** and **the subjective certainty of the speaker about the truthfulness of the statement** (Ifantidou 2001; Mushin 2000). The following two examples taken from (Su, Huang, and Chen 2010) show some evidentiality differences in English:

- *It must be raining.*
- *I can see it raining.*

According to researchers of this phenomenon (Aikhenvald 2015; Jackobson 1957), evidentiality exists in every language, but it can be expressed differently (for example morphologically or lexically) (Su, Huang, and Chen 2010).

### 2.2 Evidentiality in Bulgarian

In Bulgarian (Nitzolova 2008) evidentiality expresses the four combinations of the presence or absence of two aspects:

- a subjective view toward the reported information;
- transmitting somebody else's information.

The Bulgarian evidential system includes the following evidentials:

- **Indicative** (the main evidential - when the speaker is a witness of what he or she reports as information);
- **Conclusive** (when the speaker makes conclusions, based on the information);
- **Renarrative** (when the speaker re-transmits the information without saying if it is true or false);
- **Dubitative** (when the speaker expresses doubt about the information);
- **Admirative** (when the speaker admires the information) (Karagjosova 2021). *Some authors don't see admirative as an evidential, but here we accept it as a part of the evidential system in Bulgarian.

Bulgarian evidentials are expressed morphologically through the forms of verb tenses. Bulgarian language has nine tenses. See below examples of the verb "пиша", with their translations into English ("to write"):

Present tense (praesens) "пиша" ("I am writing, I write"), aorist "аз писах" ( I wrote), imperfect past tense (imperfectum) "пишеше" ( "I was writing"), perfect (perfectum) "писал съм" ( "I wrote"), pluperfect (plusquamperfectum) "писал бях" ( "I have written"), future tense (futurum) "ще пише" ("I will write"), future perfect tense (futurum exactum) "ще е/бъде писал", ("will have written" and "will have been writing"), future in the past tense (futurum praeteriti) "щях да пиша" ("I was going to write", "I would have written"), future perfect in the past (futurum exactum praeteriti) "щях да съм / да бъда писал" ("I would have written", "I would have been writing", and "I would have had written").

The most important feature of the indicative in modern Bulgarian is the fact that the past tenses (aorist, imperfectum, plusquamperfectum, futurum praeteriti, futurum exactum praeteriti) implicitly indicate that the speaker witnessed the reported event(s). These tenses in indicative can only be used if the speaker is a witness or if the speaker presents himself as a witness (Aleksova 2024).

### 2.3 Automatic Methods

It was already shown that **evidentiality improves automatic text trustworthiness detection** when used as additional features to an English-language machine learning model (Su, Huang, and Chen 2010).

According to our knowledge, **there are no works which automatically detect Bulgarian evidentials**. As already mentioned in Section 1, we ran an extensive analysis of automatic tools for Bulgarian and none of them was detecting evidentials. The summary of our analysis can be seen in Table 1.

Table 1: Overview of Existing Bulgarian POS Taggers and Parsers and their Support of Evidentials

| **Name** | **Supports Evidentials** | **Comments** |
|---|---|---|
| SketchEngine (Kilgarriff et al. 2014) | No | Evidentials not included in the tagset |
| AMontgomerie / bulgarian-nlp (Montgomery 2023) | No | Relying on Universal Dependencies, only Part-of-Speech (POS), no grammatical tags, but supports Named Entity Recognition (NER), sentiment analysis, etc. |
| polyglot (Al-Rfou 2015) | No | Relying on Universal Dependencies, only POS, no grammatical tags, but supports NER |
| UDPipe, Universal Dependencies 2.15, Deep Universal Dependencies 2.8, Universal Dependencies 2.5 Models for UDPipe, Universal Dependencies 2.4 Models for UDPipe (Straka 2018) | No | `https://universaldependencies.org/ext-feat-index.html`, `https://ufal.mff.cuni.cz/udpipe` |
| iarfmoose/roberta-small-bulgarian-pos (Montgomerie 2021b) | No | Relying on Universal Dependencies |
| iarfmoose/roberta-base-bulgarian-pos (Montgomerie 2021a) | No | Relying on Universal Dependencies |
| CLASSLA Fork of the Official Stanford NLP Python Library for Many Human Languages (CLASSLA 2021; Ljubešić, Terčon, and Dobrovoljc 2024) | No | Relying on Universal Dependencies |
| GATE: Universal Dependencies POS Tagger for bg / Bulgarian (Roberts 2020) | No | Relying on Universal Dependencies |
| LIMA - Libre Multilingual Analyzer (LIMA 2021) | No | Not supporting morphological features |

*Continued on next page*

Table 1 – *Continued from previous page*

| Name | Supports Evidentials | Comments |
|---|---|---|
| NLP Cube (Boroş, Dumitrescu, and Burtica 2018) | No | Relying on Universal Dependencies |
| NooJ (Silberztein 2005) | No | Evidentials not included in the tagset |
| RNNTagger (Schmid 2019) | No | Only Linux supported, only POS tagger |
| spaCy (Honnibal et al. 2020) | No | No information in the documentation about Bulgarian, but claimed support of Macedonian |
| Sparv (Borin et al. 2016) | No | TreeTagger integrated for Bulgarian (see row 18) (relying on Universal Dependencies), working with Stanza for POS and lemmatization, best for Swedish and English |
| Stanza (Qi et al. 2020) | No | Relying on UniversalDependencies |
| Text Tonsorium (Jongejan 2020) | No | Relying on Universal Dependencies |
| TreeTagger - a part-of-speech tagger for many languages (Schmid 1994) | No | Based on the BulTreeBank tagset, only finite indicative recognized (`http://bultreebank.org/wp-content/uploads/2017/06/BTB-TR03.pdf`) |
| UniMorph - Schema and datasets for universal morphological annotation (Sylak-Glassman 2016) | No information[2] | Supporting over 212 features, including evidentiality, need to check for Bulgarian, might rely on UniversalDependencies |
| melaniab / spacy-pipeline-bg (Berbatova and Ivanov 2023) | No | Supports morphological features, relying on UniversalDependencies |

*Continued on next page*

2 There is no information in the documentation whether UniMorph supports evidentials for Bulgarian language

Table 1 – *Continued from previous page*

| **Name** | **Supports Evidentials** | **Comments** |
|---|---|---|
| Bulgarian NLP pipeline in CLaRK System (BTB-Pipe) (BulTreeBank) | No | Based on the BulTreeBank tagset, only finite indicative recognized (`http://bultreebank.org/wp-content/uploads/2017/06/BTB-TR03.pdf`) |
| The CLASSLA-Stanza model for morphosyntactic annotation of standard Bulgarian 2.1 (Terčon et al. 2023) | No | A version of Classla from 2023 based on the MULTEXT-East tagset |
| wietsedv/xlm-roberta-base-ft-udpos28-bg (de Vries 2021; de Vries, Wieling, and Nissim 2022) | No | Relying on UniversalDependencies |
| KoichiYasuoka/bert-base-slavic-cyrillic-upos (Yasuoka 2021a) | No | Relying on UniversalDependencies |
| KoichiYasuoka/bert-large-slavic-cyrillic-upos (Yasuoka 2021b) | No | Relying on UniversalDependencies |
| rmihaylov/bert-base-pos-theseus-bg (Mihaylov 2023) | No | Only POS tagger |
| bgnlp (Fauzi 2021) | No information[3] | Supports lemmatization, NER, POS and morphological features tagging, keyword extraction and commatization; the POS and morphological features model is trained on the Wiki1000+ dataset |
| AzBuki.ml (Cholakov) | No | Evidentiality not included in the tagset |

3 There is no information in the documentation whether bgnlp supports evidentials for Bulgarian language.

## 3. Detected Forms of Bulgarian Renarrative

The renarrative in Bulgarian has the peculiarity of having homonymous forms with other evidentials and tenses. Due to this, its recognition often depends only on the context and the interpretation. The present and imperfect forms of renarrative match the forms of the conclusive imperfect, with a difference in the third person. The forms of present and imperfect have the same form as a specific form of the indicative perfect - when the third person form of the auxiliary verb *be* is omitted. The forms of the dubitative aorist are homonymous with renarrative perfectum/plusquamperfectum. The present tense form of the admirative is the same as the present form of the renarrative.

Due to this homonymy, **we focus on the forms of renarrative which can be recognised more easily**, namely those of **the third-person singular and the third-person plural**. To achieve that, some of our datasets are from journalistic headlines where renarrative is usually in the third person. We did not work with the forms of the auxiliary verb *be* (in Bulgarian съм), which has high homonymy with other evidentials.

Table 2 shows the forms which we detect for the Bulgarian verb "нося" (English translation: "to bring, to carry").

| Tense | Forms |
|---|---|
| **Present/Imperfect** | носел,-а,-о **съм**<br>носел,-а,-о [empty] |
| **Perfect/Pluperfect** | бил **съм** носил,-а,-о<br>бил [empty] носил,-а,-о |
| **Aorist** | носил,-а,-о,**съм**<br>[empty] носил,-а,-о |
| **Renarrative futurum/futurum praeteriti** | **щял,-а,-о съм** да нося<br>щял,-а,-о [empty] да носи |
| **Futurum exactum/Futurum exactum praeteriti** | щял,-а,-о съм да **съм** носил,-а,-о<br>щял,-а,-о да е носил,-а,-о |

**Table 2**
Automatically detected forms of renarrative (examples, taken from Nitzolova (2008).)

Finally, we made sure that the forms which we aimed to detect are important in the Bulgarian world of news media. We assured this by consulting one of this article's co-authors, who has a long experience in Bulgarian news media.

## 4. Datasets Used

In this section, we first (in Section 4.1) introduce the text dataset which we built for training BGRenBERT and testing and modifying the BGRenRules. Next, we describe the datasets (Section 4.2) on which we compared the results of the two methods with manual annotation.

### 4.1 BgRenData - Dataset Used for Preparing the Automatic Solutions

As our work is connected to detecting factually incorrect information, we focus only on the two most frequently considered types of texts in this domain – namely, news articles and social media texts.

For training BGRenBERT and testing and modifying the BGRenRules, we have created a special dataset (BgRenData) of 2891 short texts, half of which contained renarrative, and half did not contain renarrative. The texts were a mixture of news article titles, short social media posts, and ChatGPT-generated sentences. Table 3 shows the contents of this dataset. The news article titles came from two sources: a random selection of funny/fake titles with a variety of topics from Hardalov (Hardalov, Koychev, and Nakov 2016)'s Credible News dataset; and a compilation of news article titles with renarrative taken from (Margova 2023). Originally, the Credible News dataset contained credible and fake news. We selected only the fake news subsets for our analysis. Specifically, they came from:

1. The Bulgarian website with humorous news Ne!Novinite[4] (translation into English: "No!News"), containing topics such as politics, sports, culture, world news, horoscopes, interviews, and user-written articles;
2. The blog website Bazikileaks[5], containing fictitious blogs;
3. The Bulgarian news media bTV humorous (Duplex) Lifestyle subsection[6].

The social media posts were also a random selection of posts with various topics from Temnikova et al. (2023)'s datasets. These datasets were originally selected to contain keywords related to manipulation, lies, and similar topics.

One of our linguists used ChatGPT 3.5 to generate a selection of sentences, containing 3rd person singular and plural forms of renarrative from those in Table 2. We restricted the generation to 3rd person singular and plural because these are the most frequent forms of renarrative in news, according to our colleague Ruslana Margova.

Each text in the dataset was selected in a way to contain at least one form of renarrative from those in Table 2. We did not count their numbers per text, as the task of both our solutions was only to identify the texts which contained at least one renarrative form. BGRenData is available in Zenodo[7], Github[8], and HuggingFace[9] with the license Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0), due to containing tweets and as required by Temnikova's datasets (Temnikova et al. 2023).

4 https://www.nenovinite.com/
5 https://neverojatno.wordpress.com/
6 https://www.btv.bg/lifestyle/all/
7 https://zenodo.org/records/15871397
8 https://github.com/silviavg/bg-renarrative
9 https://huggingface.co/datasets/gate-institute/BGRenData

| Class | Sources and Counts | | | | | |
|---|---|---|---|---|---|---|
| | All Sources | News Articles Titles | | ChatGPT | | Social Media |
| | Counts | Type | Counts | Type | Counts | Counts |
| Renarr. | 1445 | Margova (2023) | 377 | All | 998 | 70 |
| | | | | Sg. | 843 | |
| | | | | Pl. | 155 | |
| No Renarr. | 1446 | Hardalov, Koychev, and Nakov (2016) | 375 | All | 1036 | 35 |
| | | | | Sg. | 845 | |
| | | | | Pl. | 191 | |
| Both | 2891 | | | | | |

**Table 3**
BGRenData - the dataset used for training and testing BGRenBERT and BGRenRules.

## 4.2 Analyzed Existing Datasets

To compare BGRenBERT with BGRenRules, we used subsets of two publicly accessible datasets with Bulgarian texts, used in NLP methods for fake news detection. The datasets contained the two types of texts of our interest: news article titles and Twitter posts. Specifically, we used the fake news titles subsets of Credible News (Hardalov, Koychev, and Nakov 2016) (already introduced in Section 4.1) and the Twitter posts dataset, provided for the 2022 Conference and Labs of the Evaluation Forum (CLEF) CHECK-THAT! Lab[10], TASK 1D. CLEF 2022 Check-That! Lab Task 1 (Nakov et al. 2022) included Bulgarian and required to predict which Twitter posts are worth fact-checking, with topics related to COVID-19 and politics. We decided to include in the analysis only the tweets from Task 1D "Attention-worthy tweet detection": Given a tweet, predict whether it should get the attention of policymakers and why. The tweets from Task 1D were annotated with 5 classes: *harmful* (333 tweets), *yes_discusses_cure* (79 tweets), *yes_blame_authorities* (51 tweets), *yes_discusses_action_taken* (25 tweets), and *no_not_interesting* (3186 tweets). To limit manual annotation efforts, we excluded the last category. We refer to the obtained in this way dataset as ***CT1D***.

Before starting manual annotation, we first preprocessed Hardalov's datasets by:

- removing the titles, used in the training dataset, described in Section 4.1.
- after removing the training titles we were left with only 6 titles from bTV Lifestyle. As they were too few for a meaningful comparison with the Bazikileaks and Ne!Novinite, we removed all bTV titles.
- leaving only one example from two sets of titles, which were almost completely identical, but differing by a date and/or a person's name. See Examples 1 and 2 below:

10 https://checkthat.gitlab.io/clef2022/

**Example 1** (Highly similar titles 1)
“Петъчен оптимизъм с Росен Плевнелиев - 30.08”;
“Петъчен оптимизъм с Александър Томов - 04.04”
**Translations into English**:
“Friday optimism with Rosen Plevnevliev - 30.08”;
“Friday optimism with Alexander Tomov - 04.04”

**Example 2** (Highly similar titles 2)
“Седмичен не!хороскоп: 21.01-27.01”;
“Седмичен не!хороскоп: 07.01-13.01”
**Translations into English**:
“Weekly no-horoscope: 21.01-27.01”;
“Weekly no-horoscope: 07.01-13.01”

We refer to the resulting dataset as ***CNClean***, to its Ne!Novinite’s subset as ***CNCleanN*** and to the Bazikileaks subset as ***CNCleanB***.

Next, we manually annotated all the datasets for the presence of the forms of renarrative in Table 2. Annotation was done by two Bulgarian linguists (co-authors of this article). The annotators had to assign the category “renarrative” if the text contained at least one of the forms in Table 2, and “No renarrative” if it did not contain any of these forms.

The cases in which they did not agree or had doubts about were resolved in a discussion with a third Bulgarian linguist (also co-author of this article).

Table 4 summarizes the number of texts per dataset and how many of them were manually annotated as *Renarrative* and *No renarrative*.

As it can be seen, CT1D contains fewer tweets with renarrative (17 texts or 3.48% from 488 texts) than the CNClean (447 texts, or 8.55% from 5230 texts in total). Specifically, CNCleanN contain the highest number of titles with renarratives (438 texts or 9.57% from the total of 4579 texts). The class “yes discusses action taken” from CT1D, instead, contains 0 texts with renarratives.

| **Datasets** | **Subsets** | **Classes and Counts** | | |
|---|---|---|---|---|
| | | Ren. | No Ren. | All |
| **CNClean** | CNCleanB | 9 | 642 | 651 |
| | CNCleanN | 438 | 4141 | 4579 |
| | all the above | 447 | 4783 | 5230 |
| **CT1D** | “harmful” | 9 | 324 | 333 |
| | “yes discusses cure” | 2 | 77 | 79 |
| | “yes blames authorities” | 6 | 45 | 51 |
| | “yes discusses action taken” | 0 | 25 | 25 |
| | all the above | 17 | 471 | 488 |

**Table 4**
Counts of all items in the CNCleanB, CNCleanN, and CT1D datasets, including counts of the texts annotated as containing “renarrative” and “no renarrative”.

| Class | Acc. | Recall | Precision | F1-Score |
|---|---|---|---|---|
| Dev Split Results | | | | |
| All | 0.98 | | | |
| Renarr. | | 0.99 | 0.97 | 0.98 |
| No Renarr. | | 0.97 | 0.99 | 0.98 |
| Test Split Results | | | | |
| All | 0.98 | | | |
| Renarr. | | 1.00 | 0.97 | 0.98 |
| No Renarr. | | 0.97 | 1.00 | 0.98 |

**Table 5**
Results of BGRenBERT on BGRenData splits.

CNCleanB, CNCleanN, and CT1D are available in Zenodo[11]. The datasets are shared with the same license as the original Hardalov's datasets (Hardalov, Koychev, and Nakov 2016), and due to the fact that they contain tweets (Creative Commons Attribution-NonCommercial 4.0 International – CC BY-NC 4.0).

## 5. Automatic Detection Methods

### 5.1 BGRenBERT

To create BGRenBERT, we fine-tuned BERT-WEB-BG-CASED[12] with the BGRenData dataset, described in Section 4.1. We split BGRenData into train, dev, and test sections with the following proportions: 80, 10, 10. The number of training epochs was 5. For the rest of the hyperparameters, we used their default values. We made sure that all targeted forms appeared in all the splits, however they did not have equal distributions. All three splits contained 50% of texts with one or more forms of renarrative and 50% of texts with no renarratives.

Table 5 shows the results of the classifier. As the dataset is balanced for the presence of renarrative or not, but not balanced regarding the number of forms, we report both Accuracy and F1-scores.

### 5.2 BGRenRules

We compared BGRenBERT with a BGRenRules which used regular expressions on the top of *Classla*'s output (BGRenRules). The regular expressions covered all renarrative forms in Table 2 and were built after consultations with all three Bulgarian linguists, who are co-authors of this article. The dataset used for training the classifier (BGRenData) was also

11 `https://zenodo.org/records/15882529`
12 `https://huggingface.co/usmiva/bert-web-bg-cased`.

| Dataset | Prec. | Rec. | Acc. | F1-Score |
|---|---|---|---|---|
| **BGRenBERT** | | | | |
| **CNClean** | 0.714 | 0.846 | 0.958 | 0.774 |
| **CNCleanB** | 0.127 | 1.000 | 0.905 | 0.225 |
| **CNCleanN** | 0.806 | 0.842 | 0.965 | 0.824 |
| **CT1D** | 0.292 | 0.824 | 0.924 | 0.431 |
| **BGRenRules** | | | | |
| **CNClean** | 0.958 | 0.928 | 0.990 | **0.943** |
| **CNCleanB** | 0.583 | 0.778 | 0.989 | 0.667 |
| **CNCleanN** | 0.969 | 0.931 | 0.991 | 0.949 |
| **CT1D** | 0.500 | 0.882 | 0.965 | **0.638** |

**Table 6**
Results of BGRenBERT and BGRenRules on CNClean, CNCleanB, CNCleanN, and CT1D.

used to recursively test and refine BGRenRules. For BGRenRules we used all three splits of BGRenData. We expected that BGRenRules would give better results as it was based on regular expressions, matching the well-structured Bulgarian grammar, compared with the probabilities-based BGRenBERT. Both BGRenRules and BGRenBERT are shared on Zenodo[13], HuggingFace[14] and Github[15] with Creative Commons Attribution 4.0 International license.

## 6. Comparison of the Methods on the Annotated Subsets of Datasets

To compare the two methods on the same texts, we applied them to CNClean subsets and CT1D. Exactly like human annotators, the methods were assigning the label "renarrative" if the text contained at least 1 form of the renarratives in Table 2 and "no renarrative" if none. We then compared the success of both methods in matching the manual annotations.

Table 6 shows the results of the two methods on these datasets. We would like to clarify that Table 6 shows the evaluation of the two methods on different texts than those used for creating BGRenBERT and BGRenRules.

Due to the higher number of items, we report the results for CNClean, and separately for CNCleanB and CNCleanN. We report only the results for CT1D as a whole, due to the too low numbers of renarratives in its subsets. We report Precision, Recall, Accuracy, and F1-Scores, compared to the manual annotations. As the datasets are not balanced, more attention should be given to the F1-Scores, but we report both. It is clearly visible that the BGRenRules outperforms BGRenBERT. The CT1D results are lower for both methods, probably because of the low number of texts with renarrative.

13 https://zenodo.org/records/15802264
14 https://huggingface.co/gate-institute/BGRenBERT
15 https://github.com/silviavg/bg-renarrative

Figures 1 and 2 show the percentages of False Positives (FP) and False Negatives (FN) from all errors of the two methods in all datasets. It is clear that BGRenBERT makes more FP errors, while BGRenRules – more FN errors. Given this, if a user is interested in having fewer FP (i.e. fewer texts not containing renarrative, but automatically identified as containing renarratives), they should use BgRenRules. If the user, instead, would like to have fewer FN (i.e. texts with renarrative, wrongly recognised to not contain any of its forms) – then BGRenBERT would be a better option.

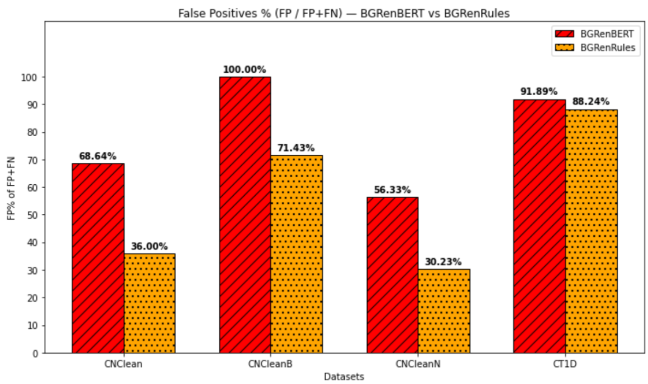


**Figure 1**
Percentages of False Positives (FP) from all errors of the two methods in all datasets.

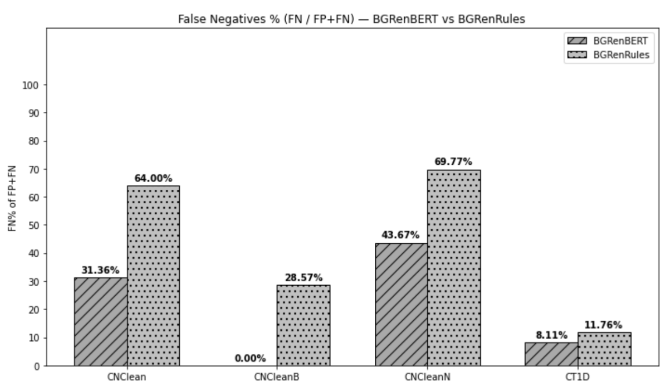


**Figure 2**
Percentages of False Negatives (FN) from all errors of the two methods in all datasets.

To see a more detailed picture, we manually analysed the False Positives and False Negatives for both datasets and the two methods, comparing them with the manual annotations. The error analysis is presented below.

### 6.1 Error Analysis of CNClean

As CNCleanB and CNCleanN differ in terms of style and contain enough items, allowing us to build a picture of each dataset's subset, we analyzed them separately.

#### 6.1.1 Analysis of CNCleanB

CNCleanB contains 651 titles, from which 9 (1.38%) were manually annotated as containing renarrative. BGRenBERT made 62 errors, all False Positives (cases in which the model decided that a title contains a renarrative, but there was none). Errors could be grouped as:

- Words ending on '-л'. For example, the model shows the word 'капитал' – in English: 'capital' as a renarrative - obviously because of the suffix '**-ал**';
- Grammatically mistaken absences of comma + past participles ('Доживотен затвор за дядо заклал прасето си!' (in English: 'Life imprisonment for an elderly man who slaughtered his pig!')
- Some quotations headlines, recognised as containing renarrative, like "Марин Райков: 'Бизнесът и медиите трябва да са солидарни с управляващите, а не с народа"', (in English: "Marin Raykov: 'The Business and the media should be in solidarity with the government, not with people."). The possible explanation is that the function of renarrative is to retell a story; thus, the quotation marks are a sign of retelling.
- Linked to the previous case, a lot of punctuation tricks BGRenBERT into deciding that a form is a renarrative, while it actually is not. Only two cases of renarrative were not recognized as such by the model - one is a typical renarrative "Менделеев починал след консумация на български колбас!" (in English: "Mendeleev died after eating Bulgarian sausage!") and the second is difficult to recognise even by humans: "Циганите изпищяха, купували им гласовете с фалшиви банкноти. ЦИК заплаши да касира изборите." (in English: "The gypsies cried out, their votes were bought with fake banknotes. The CEC (the Central Electoral Commission) threatened to cancel the elections.") The second case could be understood either as the evidential conclusive, or as indicative of a statal perfect.

All BGRenRules mistakes instead corresponded to cases in which all human annotators were not sure.

#### 6.1.2 Analysis of CNCleanN

The dataset contains 4579 titles, from which the manually annotated ones with renarrative are 438 (or 9.56%). BGRenBERT tagged 89 titles as containing renarrative, while there was no renarrative. The types of these errors were:

- Words ending in "-л". Examples: "МВР погна похитители на играчки" - translation into English: "The Ministry of Internal Affairs chased kidnappers of toys"; "Протестър №1 роди пудел от Саня Армутлиева" - translation into English: "Protester No. 1 gave birth to a poodle by Sanya Armutlieva"; participle finishing with "-л".
- Present tense, indicative, third person, singular. Examples: "изнасили" (translation into English: "raped"), "уцели" (translation into English: "hit").
- There were also 45 other cases which could not be grouped.

There are also 69 cases of titles, containing renarrative, but not recognised as such by BGRenBERT:

- 12 of them are in a subordinate clause (in the remaining similar cases, the renarrative is correctly recognised). For example: ( "Монтират топломери на хората, също излъчвали топлина, приспадат я от сметките." (in English: "They are installing heat meters on people too, since they emit heat — it's deducted from the bills.").
- Elliptical phrases without a personal pronoun (for example: "Ердоган се отказа от мол в Истанбул, прицелил се в Народното събрание в София" (in English: "Erdogan gave up a mall in Istanbul, now he is targeting the Parliament in Sofia").
- There was a case in which the construction was revealed, even though the whole headline was misspelt.

BGRenRules gave better results: it made only 43 errors. From these, the main ones were such in which the renarrative was not recognized:

- Set phrases like "***каквито и да било***" (in English "***any***"), for example, in this already translated sentence: "CSKA insists that the derby with Levski should not be officiated by ***any*** referees"'.
- Renarrative of the auxilary verb "съм" (in English "to be"). For example in "ИЗВЪНРЕДНО!111! РУСИЯ НИ ОБЯВИ ВОЙНА, УДАРЪТ ПО ВОЛЕН БИЛ УДАР ПО ТЯХНАТА ТЕРИТОРИЯ" (translation into English: "BREAKING NEWS!111! RUSSIA DECLARED WAR ON US, THE STRIKE ON VOLEN WAS A STRIKE ON THEIR TERRITORY").
- A reflexive verb, like in this example: "Най-богатият българин във Facebook си платил, за да го следят" (translation into English: "The richest Bulgarian on Facebook paid to be followed.").

### 6.2 Analysis of the CT1D

CT1D contains small numbers of the different classes, as well as of manually annotated renarratives, and for this reason, we run an error analysis of all classes together. The dataset contains 488 items, out of which 17 were manually annotated as containing renarrative -

the renarrative is 3.48% of all the items. BGRenBERT tagged 34 cases as containing forms of renarrative, but they did not. The errors are related to the following issues:

- Long sentences.
- Insertion of Latin transcription (COVID-19).
- Insertion of special symbols such as #.
- Atypical errors.

BGRenRules worked better than BGRenBERT, as in the previous cases. Specifically, it flagged wrongly as containing renarratives 15 cases and also failed to recognise 2 cases with actual renarratives. All of its errors were hard to resolve even for humans. We give as an example a case which is in general difficult to be recognised as a renarrative: "Три учителки от детска градина в София са с Covid-19, ходили до Сърница" (translation into English: "Three kindergarten teachers in Sofia have Covid-19: they went to Sarnitsa" or "it was said that they went to Sarnitsa"). It is possible that the subordinated clause is in perfect tense, indicative, or that it is in renarrative.

### 6.3 Overall Observations

We observed that BgRenRules generally made mistakes only in the cases which were doubtful also for humans. This is most probably due to the fact that BGRenBERT is based on probabilities, which may cause errors, while BGRenRules is using regular expressions which closely match the Bulgarian grammar. Such observations, in fact, confirm our initial expectations.

Additionally, we observed that the journalists' poor punctuation skills confuses BGRenBERT. This brings up another factor not previously accounted for in attempts at automatic renarrative detection: punctuation issues. Based on the current analysis, when punctuation is correctly used, BGRenBERT works fine. However, since the lack of commas is a frequent writing mistake, this limitation should be considered in the future. Another main problem for the automatic detection of renarrative is the confusion with past perfect/imperfect participle, as well as with some nouns, finishing with "-л", and verbs in the present tense finishing with "-ели".

## 7. Ethical and Legal Considerations

The annotators were among the authors of this article. Their annotation work was part of their regular salaries and was decently paid for Bulgaria. The annotated datasets are shared per the requirements and licenses of the datasets the texts were originally part of. We are making BGRenBERT, BGRenRules, and all the annotated datasets open-access with specifically written legal disclaimers. In addition to the license, the disclaimers state that: 1) it should be taken into account that the automatic detection of texts with "renarrative" generates some errors; 2) the presence of a form of "renarrative" should not be taken as a sole indication of the lack of trustworthiness of a statement or a text.

## 8. Work Limitations

Our work makes an important contribution by presenting the first automatic solutions for detecting Bulgarian texts containing renarrative. However, it also has some limitations:

- We are considering only the most frequent forms of renarrative and do not offer a definitive solution for cases in which humans would also doubt. It would be a better solution to cover all forms of renarrative, including identifying correctly those forms for which humans are also unsure.
- We covered only the 3rd person singular and plural, as these are the most common ones in news articles.
- We recognise only Bulgarian texts, written in Cyrillic alphabet. Social media and forums' posts and answers in Bulgarian are often written with different Latin transliterations.
- Our solutions allow detecting short texts, containing these main forms of renarrative, but not the forms themselves. Future work could include creating a solution which detects the actual beginning and the end of a form of renarrative in the text.

Such limitations could be resolved in our future work or by other researchers. It is also preferable if the automatic detection of renarrative is not a stand-alone solution, but possibly a functionality offered by a Bulgarian syntactic parser.

## 9. Conclusions and Future Work

In this article, we presented the first automatic solutions for detecting Bulgarian texts or sentences, containing the most frequent forms of the evidential renarrative. After applying them to manually annotated subsets of two datasets with fake Bulgarian texts, we obtained better results from our rule-based solution, BGRenRules. The fine-tuned BGRenBERT, instead, had a worse performance. Such results were justified by the principles on which the classifier and BGRenRules were based. We provided an extensive error analysis. All resources, including the two automatic solutions, the training dataset and the manually annotated subsets of the third-party text datasets, are made publicly available to assist other researchers. In future work, we plan to create a detector which recognizes the exact span of the forms of renarrative, as well as to cover more forms. We also consider adding the indicative present perfect forms as they are often used instead of the renarrative ones. Important questions to consider are whether: 1) the presence of renarrative is statistically meaningful and so frequent; 2) we need automatic methods to protect society from reading fake news containing renarratives. Additionally, while renarrative and evidentiality are important for truthfulness, their combination with other linguistic markers is still not analysed. Finally, a diachronic analysis of the use of renarrative could also be done.

## 10. Acknowledgments and Authors' Contributions

We would like to thank the "Computational Linguistics in Bulgaria" journal editors and the anonymous reviewers for their suggestions and for signalling imperfections in the article.

This research is part of the GATE project funded by the Horizon 2020 WIDESPREAD-2018-2020 TEAMING Phase 2 Programme under grant agreement no. 857155, the Programme "Research, Innovation and Digitalisation for Smart Transformation" 2021-2027 (PRIDST) under grant agreement no. BG16RFPR002-1.014-0010-C01 and, the project BROD (Bulgarian-Romanian Observatory of Digital Media) funded by the Digital Europe Programme of the European Union under contract number 101083730.

The authors made the following contributions: ***Irina Temnikova*** initiated the research, contributed to most of the datasets collection, and experiments design, calculated the comparative analysis results, generated the graphics, worked on fine-tuning BERT-WEB-BG-Cased, and wrote almost all the article; the focus on renarrative followed ***Ruslana Margova***'s observations during her work as an international news editor in Bulgarian newspapers. She also decided which are the main renarrative forms to detect, supplied part of BGRenData, manually annotated parts of CNCLean and CT1D, contributed to writing the sections on Evidentiality, and on the Detected forms of Bulgarian renarrative, conducted and wrote the manual error analysis, and made changes to the article regarding the linguistic comments of the reviewers. ***Stefan Minkov*** created BGRenRules and provided comments on the final version of the article. ***Tsvetelina Stefanova*** gave linguistic advice, participated in collecting examples with and without renarrative, manually annotated CT1D, checked whether the existing Bulgarian POS taggers and parsers detected evidentials, created Table 1, and made small changes to the final version of the article. ***Nevena Grigorova*** contributed with linguistic advices, generated examples for BGRenData, and gave suggestions for the final version of the paper. ***Silvia Gargova*** created the BGRenData splits for BGRenBERT, fine-tuned one of its previous versions, uploaded BGRenBERT, BGRenRules, and BGRenData to the respective platforms, and made suggestions for the final version of the article. Finally, ***Venelin Kovatchev*** gave suggestions on answering the reviewers' requests and on how to improve the final version of the article, including about the contents of Figures 1 and 2.

## References

Aikhenvald, Alexandra Y. 2004. *Evidentiality*. OUP Oxford.

Aikhenvald, Alexandra Y. 2015. Evidentials: Their links with other grammatical categories. *Linguistic Typology*, 19(2):239–277.

Al-Rfou, Rami. 2015. *Polyglot: A massive multilingual natural language processing pipeline*. Ph.D. thesis, State University of New York at Stony Brook.

Aleksova, Krasimira. 2024. Kategoriata evidentsialnost na balgarskiya glagol v obuchenieto po balgarski v srednoto uchilishte [The evidentiality category of Bulgarian verbs in Bulgarian language education at secondary school]. *Balgarski ezik i literatura [Bulgarian Language and Literature]*, 66(1):30–47.

Atanasova, Pepa, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, Georgi Karadzhov, Tsvetomila Mihaylova, Mitra Mohtarami, and James Glass. 2019. Automatic fact-checking using context and discourse information. *Journal of Data and Information Quality (JDIQ)*, 11(3):1–27.

Berbatova, Melania and Filip Ivanov. 2023. An improved Bulgarian natural language processing pipeline. *Annual of Sofia University St. Kliment Ohridski. Faculty of Mathematics and Informatics*, 110:37–50.

Borin, Lars, Markus Forsberg, Martin Hammarstedt, Dan Rosén, Roland Schäfer, and Anne Schumacher. 2016. Sparv: Språkbanken's corpus annotation pipeline infrastructure. In *The Sixth Swedish Language Technology Conference (SLTC), Umeå University*, pages 17–18.

Boroş, Tiberiu, Ştefan 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.

BulTreeBank. Bulgarian NLP Pipeline in CLARK System. `http://bultreebank.org/clark/bulgarian-nlp-pipeline-in-clark-system/`. BulTreeBank Project website.

Cholakov, Radostin Lozanov. AzBuki.ML - Machine learning platform for natural language processing, implementing recurrent & convolutional neural networks and linguistic algorithms with applications for the Slavic Languages.

CLASSLA. 2021. CLASSLA Fork of the Official Stanford NLP Python Library for Many Human Languages. `https://live.european-language-grid.eu/catalogue/tool-service/15639`. Software (Tool/Service).

Das, Anubrata, Houjiang Liu, Venelin Kovatchev, and Matthew Lease. 2023. The state of human-centered NLP technology for fact-checking. *Information Processing & Management*, 60(2):103219.

Dinkov, Yoan, Ivan Koychev, and Preslav Nakov. 2019. Detecting Toxicity in News Articles: Application to Bulgarian. *arXiv preprint arXiv:1908.09785*.

Fauzi, Adam. 2021. POS-BERT-bg: BERT model for Bulgarian Part-of-Speech Tagging. `https://huggingface.co/auhide/pos-bert-bg`. Hugging Face model repository.

Grieve, Jack and Helena Woodfield. 2023. *The Language of Fake News*. Cambridge University Press.

Guo, Bin, Yasan Ding, Lina Yao, Yunji Liang, and Zhiwen Yu. 2020. The future of false information detection on social media: New perspectives and trends. *ACM Computing Surveys (CSUR)*, 53(4):1–36.

Hangloo, Sakshini and Bhavna Arora. 2023. Evidence-aware fake news detection: A review. In *2023 International Conference on Advanced Computing & Communication Technologies (ICACCTech)*, pages 81–86, IEEE.

Hardalov, Momchil, Ivan Koychev, and Preslav Nakov. 2016. In search of credible news. In *Artificial Intelligence: Methodology, Systems, and Applications: 17th International Conference, AIMSA 2016, Varna, Bulgaria, September 7-10, 2016, Proceedings 17*, pages 172–180, Springer.

Honnibal, Matthew, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Ifantidou, Elly. 2001. *Evidentials and Relevance*. John Benjamins Publishing Company.

Jackobson, Roman. 1957. *Shifters, Verbal Categories and the Russian Verb*. Department of Slavic Languages and Literatures, Harvard University.

Jongejan, Bart. 2020. The CLARIN-DK Text Tonsorium. In *CLARIN Annual Conference*, pages 111–121.

Karagjosova, Elena. 2021. Mirativity and the bulgarian evidential system. In *Advances in formal Slavic linguistics 2018*. Language Science Press, pages 133–167.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlỳ, and Vít Suchomel. 2014. The Sketch engine. *Lexicography*, 1(1):7–36.

LIMA. 2021. LIMA - Libre Multilingual Analyzer. `https://live.european-language-grid.eu/catalogue/tool-service/15344`. Software (Tool/Service).

Ljubešić, Nikola, Luka Terčon, and Kaja Dobrovoljc. 2024. CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages. In *Conference on Language Technologies and Digital Humanities (JT-DH-2024)*, Institute of Contemporary History, Ljubljana, Slovenia.

Margova, R. 2023. *Linguistic Features of Fake News*. Phd thesis, Sofia University.

Mihaylov, Rumen. 2023. BERT-base POS model with Theseus training for Bulgarian. `https://huggingface.co/rmihaylov/bert-base-pos-theseus-bg`. Hugging Face model repository.

Montgomerie, Adam. 2021a. RoBERTa Base Bulgarian POS. `https://huggingface.co/iarfmoose/roberta-base-bulgarian-pos`. Hugging Face model repository.

Montgomerie, Adam. 2021b. RoBERTa Small Bulgarian POS. `https://huggingface.co/iarfmoose/roberta-small-bulgarian-pos`. Hugging Face model repository.

Montgomery, Adam. 2023. Bulgarian NLP. `https://github.com/AMontgomerie/bulgarian-nlp`. GitHub repository.

Mushin, Ilana. 2000. Evidentiality and deixis in retelling. *Journal of Pragmatics*, 32:927–957.

Nakov, Preslav, Alberto Barrón-Cedeño, Giovanni Da San Martino, Firoj Alam, Rubén Míguez, Tommaso Caselli, Mucahid Kutlu, Wajdi Zaghouani, Chengkai Li, Shaden Shaar, et al. 2022. Overview of the CLEF-2022 CheckThat! lab task 1 on identifying relevant claims in tweets. In *2022 Conference and Labs of the Evaluation Forum, CLEF 2022*, pages 368–392, CEUR Workshop Proceedings (CEUR-WS. org).

Newman, Hadley. 2021. Understanding the differences between disinformation, misinformation, malinformation and information: Presenting the dmmi matrix. *Draft Online Safety Bill (Joint Committee).*

Nitzolova, R. 2008. *Balgarska gramatika. Morfologiya [Bulgarian Grammar: Morphology].* UI "St. Kliment Ohridski".

Panchendrarajan, Rrubaa and Arkaitz Zubiaga. 2024. Claim detection for automated fact-checking: A survey on monolingual, multilingual and cross-lingual research. *Natural Language Processing Journal*, 7:100066.

Qi, Peng, 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*.

Rani, Neetu, Prasenjit Das, and Amit Kumar Bhardwaj. 2022. Rumor, misinformation among web: A contemporary review of rumor detection techniques during different web waves. *Concurrency and Computation: Practice and Experience*, 34(1):e6479.

Roberts, Ian. 2020. GATE: Universal Dependencies POS Tagger for bg / Bulgarian. Software (Tool/Service).

Santhosh, Nikita Mariam, Jo Cheriyan, and Lekshmi S Nair. 2022. A multi-model intelligent approach for rumor detection in social networks. In *2022 International Conference on Computing, Communication, Security and Intelligent Systems (IC3SIS)*, pages 1–5, IEEE.

Schmid, Helmut. 1994. Treetagger-a part-of-speech tagger for many languages. *Ludwig-Maximilians-Universität Munich*.

Schmid, Helmut. 2019. Deep learning-based morphological taggers and lemmatizers for annotating historical texts. In *Proceedings of the 3rd international conference on digital access to textual cultural heritage*, pages 133–137.

Shu, Kai, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36.

Silberztein, Max. 2005. Nooj: a linguistic annotation system for corpus processing. In *Proceedings of HLT/EMNLP 2005 Interactive Demonstrations*, pages 10–11.

Straka, Milan. 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.

Su, Qi, Helen Kai-Yun Chen, and Chu-Ren Huang. 2011. A research on the text reliability based on evidentiality. *International Journal of Computer Processing of Languages*, 23(02):201–214.

Su, Qi, Chu-Ren Huang, and Helen Kaiyun Chen. 2010. Evidentiality for text trustworthiness detection. In *Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground*, pages 10–17.

Sylak-Glassman, John. 2016. The composition and use of the universal morphological feature schema (unimorph schema). *Johns Hopkins University*, page 6.

Temnikova, Irina, Silvia Gargova, Ruslana Margova, Veneta Kireva, Ivo Dzhumerov, Tsvetelina Stefanova, and Hristiana Krasteva. 2023. New bulgarian resources for studying deception and detecting disinformation. In *Proceedings of the 10th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, Poznań, Poland.

Terčon, Luka, Nikola Ljubešić, Petya Osenova, Kiril Simov, and Luka Krsnik. 2023. The CLASSLA-Stanza model for morphosyntactic annotation of standard Bulgarian 2.1. Slovenian language resource repository CLARIN.SI.

de Vries, Wietse. 2021. XLM-RoBERTa base fine-tuned on Bulgarian Universal Dependencies POS. https://huggingface.co/wietsedv/xlm-roberta-base-ft-udpos28-bg. Hugging Face model repository.

de Vries, Wietse, Martijn Wieling, and Malvina Nissim. 2022. Make the best of cross-lingual transfer: Evidence from POS tagging with over 100 languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7676–7685, Association for Computational Linguistics, Dublin, Ireland.

Wardle, Claire et al. 2018. Information disorder: The essential glossary. *Harvard, MA: Shorenstein Center on Media, Politics, and Public Policy, Harvard Kennedy School.*

Yasuoka, Koichi. 2021a. BERT-base Slavic Cyrillic model fine-tuned for UPOS. https://huggingface.co/KoichiYasuoka/bert-base-slavic-cyrillic-upos. Hugging Face model repository.

Yasuoka, Koichi. 2021b. BERT-large Slavic Cyrillic model fine-tuned for UPOS. https://huggingface.co/KoichiYasuoka/bert-large-slavic-cyrillic-upos. Hugging Face model repository.

Zhou, Xinyi and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40.

# The (Possible) Use of AI Tools for Processing Texts in Journalism in Bulgarian

Ruslana Margova
Big Data for Smart Society Institute
(GATE)
Sofia University
5, James Boucher St., Sofia, 1164,
Bulgaria
ruslana.margova@gate-ai.eu

*This study examines the technological gaps in text-processing AI tools available to Bulgarian-language journalists and how these tools might better support journalistic practices. Through a systematic analysis of current technologies across three key domains—monitoring and information gathering, content production, and content dissemination—the research reveals significant disparities between international standards and local capabilities. While some resources exist for Bulgarian journalism, including news aggregators, translation services, these tools often lack transparency, update infrequently, or provide insufficient functionality for professional journalistic needs. Large language models (LLMs) offer promising possibilities but remain underutilised in Bulgarian newsrooms. The article provides a case study about the practical use of AI. The study recommends strategic investment in language-specific AI, targeted training, transparency standards, and ethical frameworks to improve journalistic capacity and information quality in Bulgaria, as trustworthy journalism must reach wider audiences to drown out disinformation.*

**Keywords**: *artificial intelligence, Bulgarian, journalism, text processing, computational linguistics*

## 1. Introduction

Journalism operates under severe time pressures, demanding quick decision-making and substantial output from professionals who often rely on intuition honed through experience. The recent rise of generative artificial intelligence (AI) has intensified debates around misinformation proliferation and the potential displacement of human journalists. Despite these emerging challenges, journalism remains an inherently human-centred field, though its practices continue to evolve with technological advances.

This study investigates a critical research question: What technological gaps exist in text-processing tools available to Bulgarian-language journalists, and how might these tools be enhanced to better support journalistic needs in content creation and quality reporting? Our analysis concentrates on current text-processing technologies in Bulgarian, deliberately

https://doi.org/10.47810/JCLIB.1.2025.05

excluding OSINT[1] and visual tools to maintain focus, while examining AI's dual role as both collaborative assistant and subject of ethical consideration.

The research examines three primary applications of AI in journalistic practice:

1. Monitoring and information gathering
2. Content production
3. Dissemination

Our simple framework for evaluating AI tools for journalism includes language support, transparency and usability of the tools in Bulgarian. Additionally, we conduct a discourse analysis of selected articles where artificial intelligence is explicitly acknowledged as a co-author of journalistic content, exploring the implications of this emerging practice and raising pertinent ethical questions.

Identifying and addressing technological gaps faced by Bulgarian journalists is crucial for enhancing newsroom efficiency, accuracy, and adaptability in today's rapidly evolving digital media landscape. Many journalists currently lack access to advanced technologies due to financial constraints, insufficient training opportunities, or organizational resistance to change – factors that significantly impair their capacity to gather, verify, and present information effectively. By first addressing text-processing AI tools (visual applications to be explored in future research), this study aims to provide a foundation for Bulgarian journalists to leverage innovative technologies for data analysis, data gathering and content creation. Such technological integration ultimately enhances journalistic quality and credibility and would lead to the sustainability of the Bulgarian media industry.

Drawing upon years of journalistic experience, the author uses his own experience, also incorporating insights gleaned from informal discussions with fellow journalists. This blend of direct professional involvement and peer perspectives provides a nuanced understanding of the issues at hand within the journalistic community.

## 2. Context and Related Work

Research on technological integration in journalism has evolved significantly over time. In 2017, the first comprehensive survey on newsroom technologies revealed a concerning lack of technical expertise, with merely 5% of newsroom staff possessing technology-related degrees, only 2% of newsrooms employing technologists, and a mere 1% having analytics editors on staff (Betts et al. 2017). The survey further revealed that 82% of newsroom positions remained in traditional roles such as reporters, editors, and editorial writers, while journalists generally demonstrated limited proficiency in digital skills. By 2024, however, the International Center for Journalists (ICFJ)[2] predicted substantial transformations in both social and traditional media, asserting that ”there will be major changes in social and traditional media, AI disruption will be everywhere, funding for traditional media will evolve”(Newman et al. 2024).

1 OSINT is an abbreviation of Open Source Intelligence - a practice of gathering and analyzing publicly available information to check it for disinformation

2 https://www.icfj.org/news/3-predictions-media-2024, last accessed 9.4.2025

The discourse surrounding artificial intelligence in journalism gained significant momentum in 2023, when the World Press Institute and various partners convened the ”New Horizons in Journalism” conference in Sofia. This event facilitated critical exchanges between journalists and media professionals regarding AI-related risks, with particular emphasis on the interplay between journalistic practice and artificial intelligence technologies. While participants acknowledged AI's dual potential in both propagating and detecting misinformation, as well as its applications in investigative journalism and content personalization, they consistently emphasized the enduring importance of human editorial judgment amid challenges posed by algorithmic biases and deepfakes.[3] Notably, journalism, historically characterized by slow adaptation to technological innovations, has begun proactively embracing large language models (LLMs) and generative AI, recognizing their capabilities. The Open Society Foundation's AI in Journalism Challenge (AIJC) exemplifies this shift, providing funding, mentorship, and educational resources to 12 innovative, digital-first newsrooms globally to develop pragmatic AI projects.[4] Preliminary observations from the AIJC indicate that hands-on engagement with AI tools rapidly enhances teams' understanding and enthusiasm, while suggesting that journalistic expertise frequently proves more valuable than technical knowledge in determining project success.

The latest Reuters Institute Digital News Report (Suárez et al. 2024) documents significant shifts in news consumption patterns, with increasing preference for private messaging and video platforms, while maintaining access via social media, search engines, and content aggregators. Only approximately 22% of consumers primarily utilise news websites or applications, representing a notable decline from 32% in 2018. The report identifies several key trends, including the growing influence of content creators, the increasing presence of AI in public discourse, and the introduction of ”new layers to news production and distribution” with which journalists and audiences continue to grapple. Particularly concerning is the ”proliferation of AI 'slop', low-quality, mass-produced content designed for clicks,” occurring alongside democratic backsliding, deteriorating working conditions for journalists, and a rapidly changing digital environment (ibid.). The same report indicates that investigative journalism outlets worldwide are leveraging AI to enhance workflows and expand audience reach through diverse initiatives – from automating social media content creation and summarising articles to analysing complex documents and detecting audio deepfakes – ultimately allowing journalists to focus on in-depth reporting (ibid.). In this context, ”platformization becomes increasingly important for how people communicate and access information, including news” (Nielsen and Fletcher 2023). As media organisations increasingly establish agreements with AI companies, it is noteworthy that ”most of the media managers didn't expect that money would be shared equally across publishers” (Newman et al. 2024). Trust in traditional media has declined (Newman et al. 2024), particularly in countries experiencing diminished television news consumption and increased social media news usage.

3 https://aej-bulgaria.org/new-horizons-in-journalism-2023-summary/, last accessed 9.4.2025

4 https://www.journalismfestival.com/programme/2024/applying-ai-in-small-newsrooms.-lessons-from-the-ai-in-journalism-challenge, last accessed 9.4.2025

Integrating AI into journalistic practice will catalyse numerous transformations, including automated content analysis, novel journalistic formats, evolving platform-media relationships, implications for privacy and transparency, regulatory developments, ethical considerations, emerging professional roles, and shifts in journalist training (Veleva 2024). Researchers have critically examined issues of transparency, accountability, and bias in AI systems, emphasising the necessity for ethical standards in AI-driven journalism (Broussard et al. 2019),(Opdahl et al. 2023), (Verma 2024). While AI will not supplant human journalists or journalistic intuition, "the new technology brings science to storytelling, helping newsrooms scale their production" (Marconi 2020). Core journalistic competencies – such as editorial judgment, understanding audience needs, identifying and verifying stories, and communicating effectively – will remain essential, requiring new skills in abstract thinking and analytical audience understanding (Veleva 2024). Some researchers have demonstrated that so-called Constitutional AI offers potential solutions to existing challenges in journalism by customizing LLMs to address misinformation and rebuild reader trust (Cheng 2025).

The literature also addresses AI's transformative role in journalism and the emergence of new paradigms such as open data journalism, big data journalism, blockchain journalism, and cloud journalism (Hassan and Albayari 2022). Scholars have investigated public perception of AI-generated news content and humans' ability to distinguish between articles authored by humans versus those created by artificial intelligence (Moravec et al. 2024). Research on AI's capabilities in composing various journalistic texts – both informative and opinion-based – reveals that common applications include automatic content generation, data analysis, documentation, and text translation, while usage decreases for creative tasks such as headline writing or advanced editorial functions (Fernández-Barrero, López-Redondo, and Aramburú-Moncada 2024). The term "algorithmic journalism" has emerged to describe technological transformations in the field, defined as "a process of using software or algorithms to automatically generate news stories without human intervention" (Graefe 2016). Some scholars conceptualise algorithmic journalism more broadly, encompassing automated content production, data mining, news dissemination, and content optimisation (Kotenidis and Veglis 2021).

Although not the primary focus of this research, natural language processing for fact-checking represents a critical area for future investigation. Such research should "include collaboration with fact-checkers, as well as incorporation of human-centred design practices in model development, to further guide technology development for human use and practical adoption"(Das et al. 2023).

As AI solutions continue to demonstrate potential for transforming newsrooms by enhancing efficiency, accuracy, and accessibility in news production, scholars have raised important questions regarding tool ownership and transparency. Research indicates that "the lack of transparency is a significant concern, particularly with regard to the transparency of AI tools utilized for fact-checking information in journalism: of the 100 AI tools identified, 23 included AI fact-checking services, and of these 23, only five (21%) could be classified as adequately transparent. 13 of them, or over 56% of the total, are considered not transparent" (Martin 2024). The advantages of AI could lead to slow journalism, including generating and enhancing content, reducing workloads, and consequently giving journalists more time for creative tasks (Albizu-Rivas, Parratt-Fernández, and Mera-Fernández 2024).

## 3. Three Types of Possible AI Tools and Their Uses in Bulgarian

### 3.1 General Observation for Bulgaria

Shi and Sun summarise three main applications of AI in journalism: information gathering; content production; customization and dissemination (Shi and Sun 2024). First, monitoring and gathering of news and the analysis of data can forecast trends and prepare journalists for emerging events (like Reuters News Tracer filtering Twitter for breaking news), addressing journalists' time and space constraints by providing real-time information monitoring, and being trained by journalists to align with editorial standards (ibid.). Second, AI can assists journalists in content creation by generating text that matches news organizations' style and tone, managing tasks as facilitating transcription, translation, and improving readability, producing multilingual content, expanding audience reach – demonstrated by Le Monde's AI-assisted translation of approximately 30 stories daily for its English edition, and News-GPT (launched March 2023), which analyzes data from multiple sources to create reports using AI anchors. AI supports content production in two key ways: first, by creating summaries and headlines and converting between text, images, audio, and video formats; second, by diversifying writing styles—adapting stories for different publications and audiences, as shown Claude's ability to narrate news in styles mimicking various major publications (ibid.). Third, customisation and dissemination, where the traditional news aims at a specific audience with previously analysed preferences (ibid.).

When conducting a comparative analysis of technological trends in Bulgarian media landscapes versus international standards, several significant disparities emerge. The Bulgarian media ecosystem demonstrates notable deficiencies in technology across multiple dimensions of journalistic practice. Developing a mini-framework for evaluating AI tools for journalism, we analyse language support, transparency and usability of the tools.

The Bulgarian media landscape is characterised by a persistent lack of transparency in both ownership structures and editorial practices, despite the existence of legal frameworks intended to promote openness (Bleyer-Simon et al. 2024). Within the journalistic community itself, there is a notable division regarding professional standards and responses to disinformation, particularly as digital transformations and debates over misinformation dominate the public discourse. Actors from across the political spectrum actively engage in these debates, each asserting the legitimacy of their own narratives. Many websites, which name themselves as news websites, use copy-paste journalism (see below the journalistic investigation about the so-called mushroom news websites).

Every journalist, who attends an event, must create text. Again, depending on the subject matter, journalists use various open public data for verification, although in Bulgaria, so-called data journalism is not well developed. When dealing with non-international news, journalists work in direct connection with the main "subjects" of their news texts. They can directly verify certain information from the primary source and conduct direct journalistic investigations. Despite the existence of various tools advertised as facilitating journalists' work, journalists in Bulgaria still rely on their own capabilities and journalistic intuition. There are several reasons for that – one is the lack of understanding in many newsrooms about the need to introduce technology and the lack of sufficient information about new technologies. Journalists are largely self-taught in terms of technological advances. On the other hand, each journalist develops their verification channels and personal methodology for checking specific cases. In this sense, every journalist, in the classical meaning of the word journalist, possesses both a solid background in a given field as well as contacts and their personal strategy. In Bulgaria, in the recent past, there was an entire specific team of librarians in every newspaper editorial office, called Documentation, where journalists were assisted, verifying data and conducting comprehensive research about specific historical events and individuals. These departments have now been closed, and for data verification, journalists rely, as already mentioned, on their own resources – knowledge, experience, and instinct.

### 3.2 Monitoring and Gathering

Primary media organizations in Bulgaria predominantly lack automated monitoring systems, relying instead on manual monitoring of the news flow. The technological tools widely utilized for English-language content processing have minimal Bulgarian-language equivalents, creating a substantial capability gap for journalists working in this linguistic context.

Creating a journalistic article depends on the field to which it is dedicated. As is well known, journalists typically specialize in different fields and develop in-depth knowledge of their subject matter. However, with the emergence of large online platforms and social media, journalism has become increasingly dynamic, requiring much faster reactions.

We will skip the deep analysis of the various most popular browsers on the internet, since the results for the Bulgarian language in terms of news are close, and each journalist chooses which one to work with, as may be the editorial policy. In this regard, we could expect that some AI-powered tools could be useful, which could synthesise information from the web and providing concise, well-sourced responses in the Bulgarian language. Some of these tools are checked in Bulgarian and provide efficient research, fact-checking, and quick access to background information, as well as summarising articles, generating topic overviews, and suggesting related content streamlines (ex. Perplexity).[5]

In Bulgaria, journalists monitor their own information – they follow the agenda of the parliament, the council of ministers, the work of their colleagues. Each editorial office has a subscription to certain agencies and channels, as well as to news exchanges. A huge number of Bulgarian technology companies offer media monitoring, press clipping, internet clipping, and media analysis, but these possibilities are used by specific actors, not by journalists daily, and they monitor the media flow in an "analog" way – checking other news websites, monitoring national televisions (here we are referring to transmedia journalism, in which news from one type of media is transferred to another type of media), monitoring the Bulgarian Telegraph Agency (which often does not get ahead of the news flow), and certain social media groups and accounts.

5 https://www.perplexity.ai/

However, there are some free options for monitoring information in Bulgarian, although these technological tools are much slower than journalists manually reviewing websites, and they are news aggregators, which collect and republish online already published news from various sources. One of them is Google's news aggregator[6], created back in 2002, which also works with Bulgarian. The user can explore the aggregator by their Google account, and as stated in the announcement of company, the: news articles are ranked based on their quality, originality of content, freshness of content, and where permitted based on settings and previous activity and purchases within Google News, and activity in other Google products. Google may have a license agreement with some of these publishers, but it has no impact on the ranking of results. Even having that disclaimer, it is not transparent how the algorithm works. That lack of transparency of the algorithm appears in all news aggregators.

The advantage of this aggregator is that the user can choose the sources to follow. Thus, having enough experience with reliable and unreliable sources, journalists choose which source to follow. However, this can be a problem when the journalist wants to catch disinformation and oppose it before it even takes on gigantic proportions. Another advantage of this aggregator is that the topics that the journalist follows can be indicated, and at the same time, get an overview of the entire flow, following Top Stories. A third advantage is that the aggregator does not offer summarised news, but original articles and the editor can choose where to focus on. The aggregator has a search field, where the journalist could search by keywords, Boolean search, and period. However, unlike the same aggregator in English, where there is also a fact-checking section, the Bulgarian tool lacks this functionality.

Over the years, there have been various aggregators only in Bulgarian, for which it was not clear who created them, and most of them are no longer available. The Topnovini[7] aggregator is of this type. It collects various news, distributed by topics. Through the settings, the user can choose from 17 websites that work in Bulgarian. There is no field for searching by keywords, nor for a certain period. There is a ranking of the most-read news in the different categories. These functionalities are not available with all browsers. The aggregator is updated every 10 minutes. Despite the transparency of the way that news was collected, the ownership of the website is not clear, and a check at https://hostingchecker.com/ gives results that the site is hosted by Hetzner Online GmbH, Helsinki, Finland.

Another similar Bulgarian news aggregator is radar.bg. [8] It offers 20 different subtopics from which the user can choose which ones to follow. It seems that the news stream is updated every hour, which is not particularly convenient for the dynamic news environment. The channel allows keyword search. The aggregator provides a link to the articles, as well as the first 500 characters of each article, without summarising them. There is a lack of transparency about what news websites it is fed to, as well as who is behind it. A check at https://hostingchecker.com/ shows that it is hosted by: MAIL.BG Ssc, Sofia, Bulgaria.

6 https://news.google.com/home?, last accessed 8.3.2025
7 https://www.topnovini.com/, last accessed 8.3.2025
8 https://radar.bg/news

The tool NewsGPT[9] is not relevant for the Bulgarian environment. The news related to Bulgaria is from months ago and cannot work as a tool for monitoring and gathering news. The slogan "Tomorrow news today", implementing the AI predictions of news is not working in Bulgarian and it could be used only for amusement, based on the logo *News by AI. Share the unhuman truth.* The Google Trends tool also works for the Bulgarian language. It is convenient to get a very general picture of individual searches in Google, providing statistical data on searches both on the territory of Bulgaria and by regions of Bulgaria. As the company says, it shows what's trending across Google Search, Google News and YouTube.

Many of the world's news agencies have tools for personal notification of breaking news, but the content is in English and is of interest mainly to international news editors.

### 3.3 Content Creation

The identified technological deficiencies have substantive implications for journalistic practice across multiple dimensions. Limited technological capacity compromises information verification processes, particularly in environments characterized by high information velocity and sophisticated misinformation. Content creation and content processing are the most interesting from a linguistic point of view. Without underestimating the data collection itself, which is sometimes a real journalistic investigation, the creation of content itself is perceived as a creative process.

There have been multiple developments in speech and language technology, and Bulgarian is part of some multilingual systems for machine translation, speech analysis and recognition (Koeva 2023). A Bulgarian General Language Understanding Evaluation Benchmark - bgGLUE, was also created, for evaluating language models on natural language understanding tasks in Bulgarian, targeting a variety of NLP problems (e.g., natural language inference, fact-checking, named entity recognition, sentiment analysis, question answering, etc.) and machine learning tasks (sequence labeling, document-level classification, and regression) (Hardalov et al. 2023). In recent years, many NLP scholars have been working specifically on the topic of disinformation in Bulgarian (Hardalov, Koychev, and Nakov 2016), (Koeva 2021), (Nakov et al. 2021), (Osenova and Simov 2024), (Temnikova et al. 2023), (Margova 2023). Disinformation is not the main focus here, however, when talking about content, fact-checking in journalistic content is a basic norm. In the autumn of 2024, the Association of European Journalists Bulgaria (AEJ) and the lisenced by European Fact-Checking Standards Network (EFSCN)[10] fact-checking organisation FactCheck.bg are launching a new partnership to combat election disinformation as a part of Google News Initiative, with a main goal of fighting misinformation and monitoring electoral fraud.[11] In 2018, Google launched its News Initiative to scale the work with journalists, publishers, and

9 https://newsgpt.ai/
10 https://efcsn.com/
11 https://aej-bulgaria.org/en/google-news-initiative-aej-bulgaria-launch-a-new-partnership-against-election-fraud-and-disinformation/, last accessed 3.8.2025

industry leaders to help build a resilient future for news. Thus, out of the entire package of different Google News Initiative capabilities, the fact-checking part has a Bulgarian version and is powered by Bulgarian verified facts.

In the context of fact-checking and relevant debunkings across languages, the Database of Known Fakes (DBKF)[12] allows users to check whether a claim, image or video has already been verified by trusted sources, when and how, using AI-powered technologies that go way beyond a keyword search. This could be particularly useful for fact-checkers and journalists as part of their verification workflow. The Bulgarian part consists of more than 2000 debunked articles. The database is organized by claims and articles.

Another similar database at the European level, containing identical articles in Bulgarian, like that of DBKF, is the Truly media[13] platform, which is, however, paid. In Bulgarian, a number of media outlets are engaged in fact-checking, with the licensed ones being under the umbrella of the BROD project – Bulgarian National Television Bnt[14] and AFP[15], and Factcheck.bg[16]. These details are important because, as noted above, too often, the disseminators of disinformation accuse others of such actions.

Another important tool that can be used in fact-checking is the InVid[17] plugin for Google's Chrome browser. This toolkit is a "Swiss army knife" helping journalists save time and be more efficient in their fact-checking and debunking tasks on social networks especially when verifying videos and images. It is important that the tool also works in Bulgarian and it is able to analyse the intensity of emotions in a given text. Nowadays, new automated fact-checker tools[18] have appeared and some of them performed quite well in Bulgarian, but it is a fast-developing field and many improvements are needed. Special evaluator system is created (Wang et al. 2024).

Unfortunately, the other possibilities available to journalists within the framework of the Google News Initiative are not accessible in the Bulgarian language. However, it should be noted that the programme offers a number of courses for journalists related to the use of AI, as well as machine learning, LLMs, and offers a Pinpoint workspace in which files can be easily transcribed, but these options are not working in Bulgarian.

Microsoft also has experience in creating healthy news ecosystems, explaining how technology played a role in the disruption of news, but can also be an important part of the rebuilding effort. Thus, Microsoft, and its Democracy Forward programme, provide a host of tools[19] and services to help journalists to rebuild capacity, restore trust, and reduce risk. As in the case with Google, these tools are not available in Bulgarian.

12 Developed in the frame of the European projects WeVerify and BROD project by Ontotext https://brodhub.eu/en/fact-checking/ontotext-dbkf/

13 https://www.truly.media/

14 https://bntnews.bg/bnt-provereno-108929tag.html

15 https://brodhub.eu/bg

16 https://factcheck.bg/

17 https://www.invid-project.eu/tools-and-services/invid-verification-plugin/

18 https://www.longshot.ai/fact-check-free, https://www.factiverse.ai/, https://originality.ai

19 https://www.microsoft.com/en-us/corporate-responsibility/journalism-hub

A huge number of initiatives like Partnership on AI (PAI)[20] for example tries to connect academic with civil society, industry, and media organizations to create sustainable development for journalists in the age of AI, but they give general recommendations, and not the exact tools, especially in Bulgarian.

Several specific tools work for the Bulgarian language and are somehow helpful for journalists. Various spell-checking programmes are now taken for granted, even for the Bulgarian language. However, one big issue remains in word processing – the so-called transcribing of sound files, as writing journalists continue to work with them. Unfortunately, most free tools do not support the work of journalists, as their use is limited and they end up having to manually download the content of the sound files. The only possible option is the Turboscribe[21] tool, which has a free version and allows processing up to 3 audio and video files in 24 hours, where a GPU-powered transcription engine converts audio and video to text, to be exported as DOCX, PDF, TXT, captions, and subtitles (SRT, VTT). However, the result is full of errors and must be manually checked by the journalist in questionable places, and editing takes time.

Another major need is related to translations, especially when it comes to covering international news. Journalists use various free online translators, which in recent years have been getting better and better at handling the Bulgarian language. Naturally, Google Translate[22] and Deepl[23] are mainly used. They are convenient for short texts. Deepl handles the Bulgarian language especially well, as it also allows for additional editing, by containing different suggestions. Even not very popular, Gourmet[24] is other tool which can be used for translation from and to Bulgarian, with language pairs or models for it. Although international news and the translation of news articles are not a priority, translator tools provide great opportunities when it comes to languages unfamiliar to the journalist and for conducting journalistic investigations. However, the main shortage of journalists during investigations remains time. Translation tools could also be used for creating content in English to make news websites more visible outside Bulgaria.

The main game changer nowadays – large language models (LLMs) and AI tools, such as bots, could automate and enhance news production. To date, no research has been conducted on whether and to what extent LLMs and chatbots have been used for content creation in Bulgarian media, and for what kind of tasks. Only sporadic freelancers share their news-generating experiences and one media shares such a rubric.[25] Tools like BGGPT[26], ChatGPT[27], Claude[28], Deepseek[29], Gemini[30] could assist content creation, research, data analysis, summarising translations, fact-checking, and editorial decision-making. These

20 https://partnershiponai.org/resources/
21 https://turboscribe.ai/
22 https://translate.google.com/?hl=bg
23 https://www.deepl.com/en/translator
24 https://gourmet-project.eu/project-output/gourmet-translate-tool/
25 https://karamanev.me/ and https://clubz.bg/151538
26 https://bggpt.ai
27 https://openai.com/index/chatgpt/
28 https://claude.ai/
29 https://chat.deepseek.com/
30 https://gemini.google.com

agents could also personalise news delivery by analysing reader preferences, suggesting related topics. All the outputs require human oversight to ensure accuracy and uphold journalistic standards. Claude has a similar application as ChatGPT, but is aligned with ethics, minimising harmful bias, while Deepseek is blamed for propaganda dependencies.[31] Generally speaking, drafting articles, summarising reports, and translating are among the main possibilities that AI tools offer.

### 3.4 Dissemination

Bulgarian media demonstrates limited implementation of artificial intelligence systems for audience segmentation and content personalisation. While media organisations in many countries employ sophisticated algorithms to tailor content delivery according to user preferences and behavioural patterns, Bulgarian outlets have not substantively integrated these capabilities into their operational frameworks.

Editors of news websites are usually also involved in the distribution of content on various social platforms. All major media outlets have their accounts on the various social networks popular in our country. Publications are made manually, although there are also many content optimisation experts who can assist journalists. Again, it comes down to the issue of enriching experience and knowledge in harmony with new technologies.

In Bulgaria, dissemination is a priority for malicious actors in the field of journalism because it's directly linked to monetisation. Some individuals without journalistic education can still influence social life within the country. For example, the Bulgarian journalist Georgi Angelov[32], made an investigative reportage, in which he became part of the so-called disinformation machine - part of "mushroom cites". Angelov managed to become involved in a scheme that disseminated disinformation, as well as specific articles about certain politicians designed to shape public opinion in Bulgaria. He manages to expose the system of disinformation and to show how it works. The creation of "mushroom sites" that operate in favour of foreign political interests and threaten a country's national security should not remain solely within serious online media outlets. The reason for highlighting this case is the possibility for any individual to become a journalist by joining such a mushroom site machine and monetising their work directly—that is, although it constitutes political propaganda, disinformation, and actions that are dishonest from society's perspective, this represents a scheme in which technologies play a role and, with human assistance, create false misleading content that contradicts the law. This raises the question of ethics in using various tools, which is essential and must be considered in journalism.

31 https://www.politico.eu/article/we-asked-deepseek-about-geopolitics-chinese-government-propaganda-artificial-intelligence/

32 https://www.svobodnaevropa.bg/a/saytove-gabi–mashina–dezinformatsiya-rusiya/32889599.html, last accessed 1.4.2025

### 3.5 Analysis

The existing tools for monitoring and collecting information in Bulgarian on the Bulgarian Internet can provide a fairly general picture of what is happening in the country, but they would not adequately serve journalists, who are certainly ahead of the events being covered. In addition to the delay, which is critically important in the profession, there is also a lack of transparency of the algorithms behind the various aggregators, as well as their owners. This can be a problem in times of disinformation and impaired information integrity. At the same time, as already mentioned, many Bulgarian technology companies are engaged in media clipping and monitoring, and it is possible to assist journalists with free, simplified versions of their tools in the name of a better journalistic environment that is more stable against disinformation attacks.

Bulgarian transcription tools are still a limitation for journalists. Large language models are limited in time (the data on which they are trained is limited to a certain date) and cannot be used to check today's news. However, they can provide ideas. A main problem when working with AI tools is that the content they generate must be checked, and in the hectic journalistic everyday life, time is one of the main resources. Although our emphasis is not on fact-checking, this is a field that is developing, and Bulgarian journalists can use it. The same applies to new search engines. It is important to note that many of the opportunities that exist in the field of journalist education are not offered in Bulgarian, and are not particularly well-known among media managers. Training journalists is an opportunity to improve the media environment as a whole. Despite the limited time of journalists, newsrooms can invest in their knowledge and upgrade their qualifications regarding AI, which will contribute to the overall efforts against disinformation. Since the so-called data journalism is not developed in our country, the opportunities provided by chatbots and AI tools can contribute to the development of this data journalism, as they can easily work with large data sets. Many automated fact-checker tools have appeared and need to be deeply analysed for Bulgarian.

Bulgarian media exhibits a notable lag in adopting AI for audience segmentation and content personalisation, contrasting with global trends. This limited integration hinders audience engagement, advertising effectiveness, and the ability to understand evolving user needs. The manual distribution of content on social platforms, despite the availability of optimization expertise, further underscores this technological gap.

Compounding this challenge is the exploitation of these very technological limitations by malicious actors. The case of "mushroom sites" demonstrates how the ease of online content creation and dissemination, coupled with a focus on monetisation, enables the spread of disinformation and politically motivated propaganda. This not only undermines journalistic integrity but also poses a threat to national security. The ethical considerations surrounding the use of technology in journalism, particularly in creating and amplifying misleading content, become crucial. It is paramount that trustworthy media outlets, committed to genuine journalism, achieve greater reach to effectively counter the noise generated by disinformation sources. In an era saturated with information, discerning credible news from falsehoods is increasingly challenging for the public. Amplifying the voices of reputable journalists and news organisations is crucial for fostering an informed citizenry capable of making sound decisions.

These trusted sources adhere to journalistic ethics, prioritise factual accuracy, and provide well-researched analysis, offering a clear contrast to the often emotional and fabricated content propagated by disinformation networks. By strategically leveraging digital platforms and innovative engagement strategies, credible media can penetrate echo chambers and connect with wider audiences. Investing in media literacy initiatives can further empower individuals to critically evaluate information and identify manipulation tactics. Ultimately, ensuring the prominence of authentic journalism is vital for safeguarding democratic values and fostering a society grounded in truth and informed discourse, effectively drowning out the cacophony of misinformation.

## 4. Ethical Questions

Ethics in journalism is a main question. The integration of artificial intelligence tools in Bulgarian journalism raises significant ethical questions that require careful consideration. While technological advancements offer numerous benefits for news production and distribution, they simultaneously introduce complex ethical challenges that must be addressed through robust frameworks and standards.

### 4.1 Transparency and Accountability

Without diving deeper, here we will mention the need to develop ethical standards for the use of AI. European Union made a big step in that direction, developing the Digital Service Act, where the Code of Practice on Disinformation is recognised as a robust set of commitments for Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs) to constitute strong mitigation measures against online disinformation, as demonetisation: cutting financial incentives for purveyors of disinformation; transparency of political advertising: more efficient labeling for users to recognize political advertising; ensuring the integrity of services: reducing fake accounts, bot-driven amplification, malicious deep fakes, and other manipulative behavior used to spread disinformation; and empowering users, researchers, and the fact-checking community with better tools for users to identify disinformation, wider access to data, and fact-checking coverage across the EU. This framework encompasses crucial measures such as demonetisation of disinformation, transparency in political advertising, service integrity maintenance, and empowerment of users, researchers, and fact-checkers. However, these broad European standards require adaptation and specific implementation for the Bulgarian media landscape, taking into account local journalistic practices and challenges.

### 4.2 Attribution and Authorship

The emergence of AI-generated or AI-assisted content creates fundamental questions about proper attribution and intellectual responsibility. When Bulgarian journalists utilise large language models or other AI tools to produce content, how should readers be informed about AI involvement in content creation? This issue becomes particularly salient in the context of the case study analysis conducted in this research, examining articles where AI is explicitly identified as a co-author of journalistic content.

Without clear standards for attributing AI contributions, it potentially creates confusion among readers, searching for original and reliable information and may erode public trust in media institutions already struggling with credibility challenges in Bulgaria's complex information environment.

### 4.3 AI Slop and Journalistic Standards

AI tools could be efficient but may compromise journalistic quality if implemented without appropriate oversight mechanisms. In the resource-constrained environment of Bulgarian newsrooms, where time pressures are significant and technological expertise is limited, AI tools might be deployed without sufficient quality control processes. The "AI slop" phenomenon identified in the Reuters Institute Digital News Report (Suárez et al. 2024) – low-quality, mass-produced content designed primarily for click generation – represents a concerning potential outcome when AI implementation prioritises quantity over quality. The best practice is to maintain journalistic standards, creating clear editorial policies about AI usage, including protocols for human oversight, fact verification, and editorial judgment.

### 4.4 Data Privacy and Consent

The work with a large amount of data is not eas, and Bulgarian journalists generally do it manually. When journalists utilise AI tools for content analysis, audience segmentation, or personalisation, they inevitably engage with significant amounts of user data, the questions about privacy protection raised, particularly in the context of Bulgaria's implementation of the General Data Protection Regulation (GDPR). Media organisations must ensure that data collection and processing practices comply with legal requirements, respecting audience privacy. Bulgarian journalists must navigate these boundaries thoughtfully, considering both legal compliance and ethical responsibility.

### 4.5 Misinformation and AI Detection

As AI-generated content becomes increasingly sophisticated, distinguishing between human-authored and AI-generated texts presents growing challenges for Bulgarian journalists and media consumers. The potential for malicious actors to employ AI tools for creating and disseminating misinformation, as illustrated in the "mushroom sites" investigation discussed earlier, highlights the dual nature of AI technologies in the information ecosystem.

Developing effective strategies for detecting AI-generated content and countering misinformation requires specialised knowledge and tools that many Bulgarian media organisations currently lack. This technological gap creates vulnerabilities in the information environment that could be exploited to undermine public discourse and democratic processes.

## 5. Case Study: *AI Reporting Live from the Scene*

In July 2024, the Bulgarian online media outlet ClubZ announced that it was starting a project in which editors would test and develop ethical use of artificial intelligence in journalism under the name "AI reporting live from the scene". [33] From July 2024 to March 2025, 172 articles were published in this section, co-authored by artificial intelligence. Thus, an average of 20 articles per month are created by AI. Of all the titles, only 16 are related to Bulgaria, all the rest concern foreign policy events or are related to artificial intelligence and its implementations. Those related to Bulgaria are distributed by topic, respectively: most are related to the Ministry of Defence, a few political articles related to the formation of a government, and a few for consumers. In all articles, the AI is indicated as the author, without having a name of journalists, and at the end of the text there is a disclaimer reflecting the co-authorship of artificial intelligence: „This material was written with the help of artificial intelligence under the control and editing of at least two journalists from Club Z. The material is part of the project "AI reporting live from the scene". The rest of the articles demonstrate a thematic preponderance of AI, with Elon Musk's name appearing across discussions related to technological advancements, legal frameworks, and innovative methodologies.

Geopolitical tensions and cybersecurity concerns, particularly those involving China, Russia, and the United States, constitute another prominent trend. Furthermore, a substantial portion of the articles addresses the legal and ethical implications of AI's societal impact. While less frequent, articles dedicated to pure technological innovation, without political or legal framing, remain a smaller but still visible topic.

In addition to manual discourse analysis, we used ChatGPT for a quick analysis of the topics in the articles. The check shows that the disclaimer containing AI reporting live from the scene is perceived as a permanent topic for artificial intelligence, and according to ChatGPT, all articles contain this topic. However, this is not the case, because the texts related to Bulgaria are not aimed at artificial intelligence, but at completely different topics. The tool's error proves that although very good for big data analysis, tools like Chat GPT should be used carefully and checked. Chat GPT does not recognise topics related to Bulgaria as a different topic at all. According to the tool, the next main topic is geopolitics and security, followed by the work of technology giants, the actions of Elon Musk, Social & Legal Issues, and Innovation & Future Tech. The experiment they are doing at Club Z is commendable for several reasons – the editorial team dares to use AI, tries to introduce ethical norms into this use, and publicises it. Proportionally, based on this experiment, it becomes clear that the use of AI is currently more applicable to international news and topics related to AI. One possible explanation is that Bulgarian topics require editorial intervention and scrutiny, especially when it comes to politics, defence, and consumer interests. Nevertheless, the experiment is worth it. So far, only freelancers have been doing such experiments with text generation. As a potential recommendation, we would suggest greater transparency regarding the type of AI employed and the specifics of its application. Clearly outlining how AI tools

33 https://clubz.bg/151538

are integrated into journalistic processes can foster public trust and understanding. This openness can help distinguish ethical AI usage from potentially manipulative applications, contributing to a more informed media landscape.

### 6. Conclusions and Recommendations

This research underscores that while significant technological gaps exist in the Bulgarian journalistic landscape, targeted investment, training, and ethical framework development could substantially enhance the sector's capacity for rigorous reporting and strengthen its institutional sustainability in an increasingly complex information environment.

Bulgarian media organisations demonstrate notable deficiencies in technological adoption across multiple dimensions of journalistic practice, particularly in comparison to international counterparts. Primary media organisations in Bulgaria predominantly rely on manual monitoring rather than automated systems. Available news aggregators (Google News, Topnovini, Radar.bg) offer limited functionality, lack transparency in their algorithms, and update too infrequently for the dynamic news environment. Perplexity AI represents a promising but underutilised resource for Bulgarian journalists. Bulgarian journalism faces significant challenges in content processing, with limited AI-powered tools specifically developed for the language. While spell-checking programmes exist, transcription tools for audio files remain inadequate. Translation tools like DeepL show promise but have limitations for journalistic investigations. Large language models (BGGPT, ChatGPT, Claude, Deepseek, Gemini) offer potential assistance with content creation but require further evaluation and integration into newsroom workflows.

Several fact-checking initiatives operate in Bulgarian, including those under the BROD project (Bulgarian National Television, AFP) and Factcheck.bg. Tools like InVid and databases such as the Database of Known Fakes provide some support for verification processes, though automated fact-checkers for Bulgarian require further development. Bulgarian media demonstrates limited implementation of artificial intelligence systems for audience segmentation and content personalisation. Trustworthy journalism must reach wider audiences to drown out disinformation. In today's information overload, distinguishing fact from fiction is a growing challenge. Amplifying reputable news sources, committed to ethics and accuracy, is vital for an informed public. Strategic use of digital platforms and audience engagement can help credible media break through echo chambers.

While the European Union's Digital Service Act (DSA) provides a broad framework for addressing disinformation, more specific ethical standards for AI use in Bulgarian journalism are needed to ensure responsible implementation of the DSA. Media organisations should prioritise investment in AI tools specifically developed or adapted for Bulgarian language processing, particularly in areas of transcription, automated monitoring, and content analysis. New comprehensive training initiatives for journalists on effectively utilising AI tools in their workflow, with special attention to language-specific capabilities and limitations, are needed. AI tools could be used for trustworthy data journalism practices in Bulgaria, an underdeveloped area with significant potential for enhancing reporting quality and depth.

It is necessary to create specific standards for ethical AI use in Bulgarian journalism, addressing the unique challenges of the local media landscape while aligning with broader European frameworks. Media organisations should collaborate with technology companies to develop more effective Bulgarian-language AI tools tailored to journalistic needs. Addressing ethical challenges requires an approach involving media organisations, technology developers, regulatory bodies, academic institutions, and civil society organisations.

Specific actions include: developing ethical guidelines for AI implementation in Bulgarian journalism, adapted to local conditions; establishing transparency requirements for AI tools used in newsrooms; creating standard practices for AI-assisted or AI-generated content to ensure audience understanding of content provenance; implementing robust quality control mechanisms to prevent the proliferation of AI slop; providing specialised ethical training for journalists and editors on responsible AI usage; fostering cross-sector collaboration between media and technology companies to develop AI tools that support journalistic values. These standards could contribute to the ethical integration of AI in journalism, supporting democratic discourse and enhancing media credibility.

Implementing the recommendations above, Bulgarian journalism would be a tech-empowered, ethically grounded media ecosystem, where AI tools tailored to the Bulgarian language would support the journalistic process. Receiving training, helping human editorial judgment with the precision of machine assistance, and approving ethical standards would rebuild public trust in the media, and data journalism would make disinformation fade, leading to information integrity.

## Acknowledgments

The author thanks the reviewers of the journal for their remarks. The author is also grateful to his colleagues Irina Temnikova and Venelin Kovatchev. This research has been supported by funding from the GATE project funded by the Horizon 2020 WIDESPREAD-2018-2020 TEAMING Phase 2 Programme under grant agreement no. 857155, the Programme "Research, Innovation and Digitalisation for Smart Transformation" 2021-2027 (PRIDST) under grant agreement no. BG16RFPR002-1.014-0010-C01, and the project BROD (Bulgarian-Romanian Observatory of Digital Media) funded by the Digital Europe Programme of the European Union under contract number 101083730.

## References

Albizu-Rivas, I., S. Parratt-Fernández, and M. Mera-Fernández. 2024. Artificial intelligence in slow journalism: Journalists' uses, perceptions, and attitudes. *Journalism and Media*, 5(4):1836–1850.

Betts, T., D. Bowers, J. Cohen, L. Evans, C. Gicheru, M. Jenkins, R. Khursheed, A. Mhanna, H. Owusu, M. Peralta, S. Vucinic, and M. Zayani. 2017. Icfj survey: The state of technology in global newsrooms. Technical report, ICFJ Survey.

Bleyer-Simon, K., E. Brogi, R. Carlini, D. Da Costa Leite Borges, J. Kermer, I. Nenadic, M. Palmer, P. Parcu, R. Reviglio, M. Trevisan, S. Verza, and M. Žuffová. 2024. Monitoring media pluralism in the digital era: Application of the media pluralism monitor in the european member states and in candidate countries in 2023. Technical report, Centre for Media Pluralism and Media Freedom, Robert Schuman Centre.

Broussard, M., N. Diakopoulos, A. L. Guzman, R. Abebe, M. Dupagne, and C. H. Chuan. 2019. Artificial intelligence and journalism. *Journalism & Mass Communication Quarterly*, 96(3):673–695.

Cheng, S. 2025. When journalism meets AI: Risk or opportunity? *Digital Government: Research and Practice*, 6(1):1–12.

Das, Anubrata, Houjiang Liu, Venelin Kovatchev, and Matthew Lease. 2023. The state of human-centered nlp technology for fact-checking. *Information Processing & Management*, 60(2).

Fernández-Barrero, M. Á., I. López-Redondo, and L. G. Aramburú-Moncada. 2024. Possibilities and challenges of artificial intelligence in the teaching and learning process of journalism writing: The experience in spanish universities. *Communication & Society*, pages 241–256.

Graefe, A. 2016. Guide to automated journalism. Technical report, Tow Center for Digital Journalism.

Hardalov, M., I. Koychev, and P. Nakov. 2016. In search of credible news. In *Artificial Intelligence: Methodology, Systems, and Applications: 17th International Conference, AIMSA 2016, Varna, Bulgaria, September 7–10, 2016, Proceedings*, volume 17, pages 172–180, Springer International Publishing.

Hardalov, Momchil, Pepa Atanasova, Todor Mihaylov, Galia Angelova, Kiril Simov, Petya Osenova, Veselin Stoyanov, Ivan Koychev, Preslav Nakov, and Dragomir Radev. 2023. bgGLUE: A Bulgarian general language understanding evaluation benchmark. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8733–8759, Association for Computational Linguistics, Toronto, Canada.

Hassan, A. and A. Albayari. 2022. The usage of artificial intelligence in journalism. In *Future of Organizations and Work after the 4th Industrial Revolution: The Role of Artificial Intelligence, Big Data, Automation, and Robotics*. Springer International Publishing, Cham, pages 175–197.

Koeva, S. 2021. The Bulgarian WordNet: Structure and specific features. *Bulgarian Academy of Sciences*, 8(1):47–70.

Koeva, Svetla. 2023. Language Report Bulgarian. In *European Language Equality: A Strategic Agenda for Digital Language Equality*. Springer, pages 103–106.

Kotenidis, E. and A. Veglis. 2021. Algorithmic journalism—current applications and future perspectives. *Journalism and Media*, 2(2):244–257.

Marconi, F. 2020. *Newsmakers: Artificial intelligence and the future of journalism*. Columbia University Press.

Margova, R. 2023. *Linguistic markers of Fake News*. Ph.D. thesis, Sofia University St. Kliment Ohridski. https://bpos.bg/publication/57700.

Martin, S. 2024. Control beyond code: Ownership and financial insights of AI used in journalism. Technical report, Media and Journalism Research Center (MJRC), Tallinn/London/Santiago de Compostela.

Moravec, V., N. Hynek, M. Skare, B. Gavurova, and M. Kubak. 2024. Human or machine? The perception of artificial intelligence in journalism, its socio-economic conditions, and technological developments toward the digital future. *Technological Forecasting and Social Change*, 200:123162.

Nakov, P., D. Corney, M. Hasanain, F. Alam, T. Elsayed, A. Barrón-Cedeño, and G. D. S. Martino. 2021. Automated fact-checking for assisting human fact-checkers. https://arxiv.org/abs/2103.07769. ArXiv preprint arXiv:2103.07769.

Newman, N., R. Fletcher, C. T. Robertson, A. Ross Arguedas, and R. K. Nielsen. 2024. Reuters Institute Digital News Report 2024. Technical report, Reuters Institute for the Study of Journalism.

Nielsen, R. K. and R. Fletcher. 2023. Comparing the platformization of news media systems: A cross-country analysis. *European Journal of Communication*, 38(5):484–499.

Opdahl, A. L., B. Tessem, D. T. Dang-Nguyen, E. Motta, V. Setty, E. Throndsen, and C. Trattner. 2023. Trustworthy journalism through AI. *Data & Knowledge Engineering*, 146:102182.

Osenova, P. and K. Simov. 2024. Representation of multiword expressions in the Bulgarian integrated lexicon for language technology. In *Multiword Expressions in Lexical Resources: Linguistic, Lexicographic, and Computational Perspectives*. pages 117–146.

Shi, Y. and L. Sun. 2024. How generative AI is transforming journalism: Development, application and ethics. *Journalism and Media*, 5(2):582–594.

Suárez, E., M. Adami, M. Leake, and G. Kahn. 2024. How 2024 shaped journalism: Insights from the Reuters Institute's work. Technical report, Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/news/how-2024-shaped-journalism-insights-reuters-institutes-work, last accessed 8.4.2025.

Temnikova, I., I. Marinova, S. Gargova, R. Margova, and I. Koychev. 2023. Looking for traces of textual deepfakes in Bulgarian on social media. In *Proceedings of the 14th International Conference on Recent*

*Advances in Natural Language Processing*, pages 1151–1161.

Veleva, M. 2024. Навлизането на изкуствения интелект в медиите – предимства и рискове [The introduction of artificial intelligence in the media – advantages and risks]. *Медии и обществени комуникации [Media and Public Communications]*, 56.

Verma, D. 2024. Impact of artificial intelligence on journalism: A comprehensive review of AI in journalism. *Journal of Communication and Management*, 3(02):150–156.

Wang, Y., M. Wang, H. Iqbal, G. Georgiev, J. Geng, and P. Nakov. 2024. Openfactcheck: A unified framework for factuality evaluation of LLMs. https://arxiv.org/abs/2405.05583. ArXiv preprint arXiv:2405.05583.

*Computational Linguistics in Bulgaria*
Volume 1, Issue 1, 2025

Editor-in-Chief: Svetla Koeva
Cover artwork by Bozhidar Chemshirov

Institute for Bulgarian Language
Bulgarian Academy of Sciences
Department of Computational Linguistics
https://ibl.bas.bg/

ISSN 3033-1382 (print) | 3033-2397 (online)