
Linguistic Annotation of South Slavic Languages with Large Language Models

Katerina Zdravkova¹[\[0000-0002-9674-3081\]](#)

¹ Ss. Cyril and Methodius University in Skopje, Faculty of Computer Science and Engineering, North Macedonia, katerina.zdravkova@finki.ukim.mk

Scientific paper

DOI: 10.18485/judig.2026.2.ch18

Abstract

This study investigates the applicability of ChatGPT 5 Plus to linguistic annotation in four South Slavic languages: Bulgarian, Macedonian, Serbian, and Slovene. While recent studies have demonstrated the potential of large language models (LLMs) for linguistic annotation, their ability to learn detailed language-specific morphosyntactic annotation conventions remains insufficiently understood. Using the corresponding translations of the first 1235 tokens of Orwell's 1984 together with their MULTEXT-East annotations, we evaluated the model for lemmatization, part-of-speech (PoS) tagging, and morphosyntactic description (MSD) annotation. The evaluation consisted of zero-shot annotation, one-shot calibration, iterative calibration through corrective human feedback, and an independent blind evaluation. Zero-shot annotation demonstrated high lemmatization accuracy and reliable PoS tagging, while iterative calibration substantially improved adherence to language-specific annotation conventions, particularly for MSD annotation. Blind evaluation on previously unseen sentences confirmed high lemmatization (94.95–99.22%) and PoS-tagging accuracy (94.69–96.61%) across all four languages, with MSD accuracy ranging from 75.16% to 89.21%. Most observed errors reflected the model's tendency to rely on own linguistic knowledge rather than consistently adopting the language-specific annotation conventions introduced during calibration. Regardless of their relatively frequent presence, especially during MSD annotation, the current results show that this chatbot can effectively support morphosyntactic annotation using iterative calibration based on the human-in-the-loop concept and highlight the importance of expert feedback for adapting large language models to corpus-specific annotation standards.

Keywords: lemmatization, LLMs, MSD annotation, PoS tagging

1. Introduction

Computational linguists regard tokenization, lemma extraction, part-of-speech (PoS) tagging, and morphosyntactic description (MSD) annotation as interrelated layers of linguistic annotation that transform unstructured text into data suitable for automatic natural language processing (Ide, 2017; Straka and Straková, 2017). Accurate linguistic annotation forms the foundation of many contemporary NLP pipelines, as errors introduced during these preprocessing stages inevitably propagate to downstream applications. The resulting annotations support information retrieval, machine translation, sentiment analysis, and, more recently, large language models (LLMs), stimulating the development of specialized systems for their automation (Ginn and Palmer, 2025). This progress is particularly evident for high-resource languages supported by extensive annotated corpora and mature language technologies, where annotation systems achieve high levels of accuracy (Kondratyuk et al., 2018).

Although a substantial gap remains between high- and low-resource languages, recent studies have demonstrated that LLMs can become a credible and effective solution for linguistic annotation tasks, functioning as valuable assistants to human annotators (Vidal-Gorène et al., 2026). Existing studies have primarily evaluated LLMs on individual low-resource languages or historical corpora (Vidal-Gorène et al., 2026).

The time-consuming nature of manual linguistic annotation, together with the promising results reported by Vidal-Gorène et al. (2026), motivated this study, which evaluates the capabilities of ChatGPT 5 Plus for lemmatization, PoS tagging, and morphosyntactic annotation in Bulgarian, Macedonian, Slovene, and Serbian. ChatGPT 5 Plus was selected because it represents one of the most capable commercially available LLMs and supports extended interactive calibration through long conversational sessions, making it particularly suitable for studying human-guided annotation refinement.

The four languages were selected because they are supported by the European CLARIN research infrastructure, which provides annotated corpora, lexical resources, and NLP tools for South Slavic languages (CLARIN, 2026). At the same time, these languages remain considerably less represented in contemporary NLP research than major world languages.

The CLARIN registry includes numerous systems for lemmatization, PoS tagging, and related linguistic tasks. Nevertheless, these tools typically address individual annotation layers or individual languages, and no single system currently performs lemmatization, PoS tagging, and MSD annotation according to a common morphosyntactic standard across all four investigated languages. To the best of our knowledge, no previous study has systematically compared ChatGPT across four contemporary South Slavic languages using a common morphosyntactic annotation standard together with an iterative human-in-the-loop calibration methodology. This motivated the comparative evaluation presented in this paper.

The study employs the MULTEXT-East framework [Erjavec \(2012\)](#), which was the predominant standard for morphosyntactic annotation of Slavic and several other Central and Eastern European languages before the widespread adoption of Universal Dependencies ([de Marneffe et al., 2021](#)). MULTEXT-East provides detailed and well-documented morphosyntactic specifications, making it a reliable framework for comparative evaluation.

The objective of this study is to evaluate the capability of ChatGPT 5 Plus to perform morphosyntactic annotation of four South Slavic languages using the MULTEXT-East annotation framework. The study addresses the following research questions:

The study addresses the following research questions:

- RQ1.** How accurately can ChatGPT perform lemmatization, PoS tagging, and MSD annotation without prior calibration?
- RQ2.** To what extent does iterative human-in-the-loop calibration improve annotation quality?
- RQ3.** Which linguistic categories remain challenging after calibration, and are these errors primarily grammatical or annotation-convention related?

2. Language resources and methodology of the project

In this project, MULTEXT-East version 4 was employed for Bulgarian, Serbian, and Slovene ([Erjavec, 2010](#)). The corresponding manually annotated corpora were obtained from the CLARIN repository ([CLARIN.SI Repository, 2010](#)), while the MULTEXT-

East version 4 annotation specifications were obtained from the official MULTEXT-East resource repository (Erjavec et al., 2010).

The Macedonian corpus, which is not yet publicly available, was developed in accordance with the MULTEXT-East version 6 specifications (Zdravkova, 2020). Since the available resources were created at different times, the investigated languages follow different versions of the MULTEXT-East framework. The portion of the Macedonian corpus used in this study was subsequently reviewed by an experienced linguist and revised according to her recommendations.

Considering the differences between the available annotation standards, Orwell’s 1984 was selected because manually annotated MULTEXT-East corpora are available for all four investigated languages, enabling direct comparison with gold-standard annotations. The corresponding translations were transformed into 4-tuples consisting of a token and three annotation layers: lemma, PoS tag, and MSD tags, which encode detailed grammatical information including gender, number, case, tense, person, and other morphosyntactic features. Tokenization was performed implicitly during this transformation.

The experiment began with a zero-shot evaluation based on the first three manually annotated sentences. Their corresponding gold-standard annotations were subsequently presented to ChatGPT and discussed interactively, forming the basis of the one-shot calibration. The effectiveness of this initial calibration was then evaluated on the following five previously unseen sentences. Calibration subsequently continued through four iterative rounds, each consisting of three additional unseen sentences, depending on sentence alignment. In every round, ChatGPT first annotated the new sentences, after which the corresponding gold-standard annotations were revealed and discussed. The identified disagreement was used as corrective feedback before proceeding to the next round. This progressive calibration strategy was designed to evaluate whether repeated expert feedback enabled ChatGPT to generalize language-specific annotation conventions beyond the calibration examples. All remaining manually annotated sentences were reserved for an independent blind evaluation and were not used during any stage of the calibration process. This final evaluation assessed the model’s ability to generalize the acquired annotation conventions to previously unseen text.

3. Experimental results

Annotation accuracy was computed as the proportion of correctly predicted annotations divided by the total number of evaluated tokens for each annotation layer (lemma, PoS, and MSD). Because MSD tags are PoS-dependent, an incorrect PoS prediction automatically results in an incorrect MSD annotation. Consequently, the reported MSD accuracy combines errors arising from PoS misclassification with errors in assigning fine-grained morphosyntactic features to correctly identified parts of speech. This dependency should be considered when interpreting the reported MSD accuracies. The results of the zero-shot evaluation are presented in Table 1.

Table 1: Annotation accuracy of zero-shot linguistic annotation.

<i>Language</i>	<i>Bulgarian</i>	<i>Macedonian</i>	<i>Slovene</i>	<i>Serbian</i>
Tokens	52	57	54	54
Lemma accuracy	92.31%	98.25%	98.15%	96.30%
PoS accuracy	92.31%	91.23%	100.00%	100.00%
MSD accuracy	52.77%	15.79%	79.63%	38.89%

The results of the one-shot calibration are presented in Table 2. Although MSD accuracy improved substantially for Macedonian and moderately for Bulgarian and Serbian, lemmatization and PoS tagging did not consistently benefit from a single round of corrective feedback.

Table 2: Accuracy of one-shot calibration.

<i>Language</i>	<i>Bulgarian</i>	<i>Macedonian</i>	<i>Slovene</i>	<i>Serbian</i>
Tokens	67	83	80	89
Lemma accuracy	91.04%	91.57%	92.50%	98.83%
PoS accuracy	77.61%	89.16%	95.00%	96.63%
MSD accuracy	64.18%	79.52%	75.00%	65.17%

After iterative calibration, the accuracy of all annotation layers improved (Table 3). The largest gains were observed for MSD annotation, indicating that repeated corrective feedback progressively aligned the model with the language-specific MULTEXT-East annotation conventions (Erjavec, 2012).

Table 3: Accuracy of linguistic annotation after iterative calibration.

<i>Language</i>	<i>Bulgarian</i>	<i>Macedonian</i>	<i>Slovene</i>	<i>Serbian</i>
Tokens	225	236	236	262
Lemma accuracy	92.04%	96.61%	97.03%	96.95%
PoS accuracy	94.69%	94.92%	97.03%	96.56%
MSD accuracy	78.76%	79.66%	83.90%	82.82%

The results of the independent blind evaluation are presented in Table 4. The model maintained consistently high lemmatization and PoS-tagging accuracy across all four languages, with lemma accuracy ranging from 94.95% to 99.22% and PoS accuracy from 94.69% to 96.61%.

Table 4: Accuracy of linguistic annotation after blind evaluation.

<i>Language</i>	<i>Bulgarian</i>	<i>Macedonian</i>	<i>Slovene</i>	<i>Serbian</i>
Tokens	694	765	753	681
Lemma accuracy	94.96%	99.22%	94.95%	95.59%
PoS accuracy	94.96%	96.61%	94.69%	95.74%
MSD accuracy	89.21%	75.16%	80.38%	78.85%

As in the previous evaluation phases, MSD annotation remained the most challenging task because of the language-specific complexity of the MULTEXT-East annotation scheme. Nevertheless, the model achieved MSD accuracies ranging from 75.16% to 89.21%, demonstrating that the annotation conventions acquired during calibration generalized successfully to previously unseen data. Bulgarian achieved a higher MSD accuracy in the blind evaluation than after iterative calibration, whereas Macedonian, Slovene, and Serbian exhibited decreases despite the substantially larger and completely unseen evaluation corpus. These findings indicate that iterative human-guided calibration primarily improved adherence to language-specific annotation conventions, while ChatGPT 5 Plus consistently demonstrated strong grammatical knowledge through high lemmatization and PoS-tagging accuracy across all evaluation phases.

The following section analyses the annotation behaviour of ChatGPT 5 Plus for each language individually and examines the principal annotation disagreements that remained after calibration.

4. Discussions

The results presented in the previous section reveal distinct learning patterns across the four South Slavic languages and demonstrate the impact of human-guided calibration on annotation quality. Even in the zero-shot setting, ChatGPT 5 Plus already exhibited strong grammatical competence, as reflected by consistently high lemmatization and PoS tagging accuracy across all four languages. In contrast, MSD annotation proved considerably more challenging, highlighting that the principal difficulty was not grammatical analysis itself but adherence to the language-specific conventions of the MULTEXT-East framework.

One-shot calibration produced only partial improvements, indicating that a single round of corrective feedback was insufficient for reliable adaptation to the language-specific annotation conventions. In contrast, iterative calibration progressively aligned the model with these conventions, demonstrating that repeated expert feedback is substantially more effective than isolated corrections.

The independent blind evaluation demonstrated that the knowledge acquired during calibration generalized successfully to previously unseen data. While the extent of the improvements varied across languages, the consistently high lemmatization and PoS-tagging accuracy indicates that ChatGPT 5 Plus already possessed substantial grammatical knowledge in all languages, whereas the remaining disagreements primarily reflected language-specific annotation conventions.

A detailed examination of the annotation disagreements revealed that most were concentrated in a limited number of linguistic categories. Although the specific phenomena differed across languages, the majority originated from differences in annotation conventions, lemma-selection strategies, the interpretation of morphosyntactic features defined by the MULTEXT-East framework and according to authors' impression, model's tendency to preserve its pre-existing linguistic generalizations instead of fully adapting to the language-specific annotation conventions established during calibration. The following subsections examine the annotation behaviour of ChatGPT 5 Plus for each of the four South Slavic languages individually, highlighting their principal annotation disagreements and the effects of iterative calibration before identifying the common cross-linguistic patterns observed throughout the study.

4.1. Bulgarian

The Bulgarian results demonstrate that ChatGPT 5 Plus possessed satisfactory annotation skills even before calibration, as evidenced by the high zero-shot accuracy for both lemmatization and PoS tagging. Although MSD annotation was considerably less accurate in the initial evaluation, iterative calibration progressively aligned the model with the MULTEXT-East annotation conventions (Pavlov et al., 2009), resulting in the highest blind-evaluation MSD accuracy (89.21%) among the four investigated languages. This finding indicates that the Bulgarian annotation scheme was comparatively well suited to interactive adaptation through corrective feedback.

A detailed analysis of the remaining disagreements showed that most errors were associated with annotation conventions rather than deficiencies in grammatical analysis. The most persistent problem concerned the treatment of verbal adjectives. Throughout all evaluation phases, ChatGPT repeatedly analysed forms functioning as adjectives as verbal participles, preferring a morphologically motivated verbal interpretation instead of the adjectival analysis prescribed by the Bulgarian MULTEXT-East specification. Similar behaviour was observed in the other South Slavic languages, suggesting that this reflects a broader tendency of the model rather than a language-specific weakness.

Another recurrent source of disagreement involved the annotation of the preposition *kamo*, which was assigned an incorrect morphosyntactic category throughout the calibration process despite repeated corrective feedback. Unlike the errors involving participial forms, this inconsistency cannot be attributed to grammatical ambiguity alone but rather indicates persistent difficulty in reproducing language specific annotation conventions for highly frequent functional words.

The remaining MSD disagreements were primarily related to the encoding of fine-grained grammatical features, including definiteness, gender, number, and other language-specific morphosyntactic distinctions. Overall, the Bulgarian results suggest that ChatGPT 5 Plus acquired the underlying grammatical structure of the language with relatively little difficulty, while the remaining disagreement largely reflected the challenge of reproducing the highly spe-

cific annotation conventions defined by the MULTEXT-East framework (Pavlov et al., 2009).

4.2. Macedonian

Macedonian exhibited the most pronounced learning effect among the four investigated languages. While zero-shot lemmatization and PoS tagging already demonstrated substantial grammatical competence, MSD annotation achieved only 15.79% accuracy, indicating that ChatGPT 5 Plus was initially unfamiliar with the highly specific conventions of the Macedonian MULTEXT-East version 6 annotation scheme. The subsequent calibration process produced the largest improvement observed in this study, confirming that repeated corrective feedback effectively aligned the model with the language-specific annotation framework. Blind evaluation further demonstrated that the acquired knowledge generalized successfully to previously unseen data, resulting in high lemma accuracy and reliable PoS tagging, while MSD accuracy increased to 75.16%.

A detailed examination of the remaining disagreement revealed that most errors were related to the interpretation of MULTEXT-East conventions rather than to deficiencies in grammatical analysis. The most persistent difficulties involved the classification of verbal adjectives, adjective subcategories, auxiliary constructions, and fine-grained morphosyntactic distinctions encoded in the MSD tags. During the calibration process, numerous discussions focused on the distinction between adjectival and participial analyses, the treatment of auxiliary verbs, and the interpretation of language-specific morphological features that are represented differently from the corresponding grammatical descriptions. Many of these inconsistencies disappeared after iterative calibration, demonstrating that the underlying linguistic analysis was frequently correct, while the principal challenge consisted of mapping grammatical interpretations onto the exact MULTEXT-East specifications.

Unlike Bulgarian, where only a limited number of annotation categories required repeated correction, Macedonian demanded substantially more extensive calibration due to greater complexity of its language-specific annotation scheme and the relatively limited availability of annotated language resources. Nevertheless, the final results indicate that ChatGPT 5 Plus successfully acquired most corpus-specific annotation conventions through iterative human guid-

ance. The remaining annotation disagreements were primarily associated with fine-grained MSD distinctions rather than deficiencies in lexical analysis or morphosyntactic interpretation.

4.3. Serbian

Serbian demonstrated strong grammatical competence throughout the experiment and required only moderate adaptation during calibration. Although MSD annotation remained more challenging than lemmatization and PoS tagging, iterative calibration substantially improved the model’s ability to reproduce the Serbian MULTEXT-East annotation conventions (Krstev and Vitas, 2010). The stable performance observed during the blind evaluation indicates that the acquired annotation knowledge generalized successfully to previously unseen data.

A detailed examination of the remaining disagreement showed that similarly to other South Slavic languages most errors were associated with language-specific annotation conventions. The inconsistencies involved lemma selection and the treatment of verbal and participial forms. Recurrent disagreements concerned forms of the auxiliary verb *jesam* and *biti*, active participles (*bio*, *radio*, *imao*, *gledao*), passive participles, and adjectivized participles.

In many cases, ChatGPT selected linguistically plausible lemmas and grammatical analyses that differed from the conventions adopted in the gold-standard corpus. For example, forms such as *prikačen* and *otkinut* were repeatedly interpreted as verbal participles instead of adjectives.

Similar tendencies were observed in the other South Slavic languages, suggesting that the distinction between verbal and adjectival interpretations represents one of the most persistent challenges for LLM-based morphosyntactic annotation.

Additional disagreements were observed in the classification of pronouns, numerals, and multifunctional lexical items. For example, the frequent form *kao* continued to receive an incorrect lexical classification despite repeated corrective feedback, while fine-grained MSD distinctions for pronouns, numerals, and auxiliary forms remained difficult to reproduce consistently. These observations indicate that the remaining Serbian errors were concentrated in highly specific annotation decisions defined by the MULTEXT-East framework rather than in the model’s understanding of Serbian grammar

itself. Unfortunately, in many cases, ChatGPT 5 Plus continued to rely on linguistic generalizations acquired during pre-training instead of fully adopting the language-specific annotation conventions introduced through iterative calibration.

4.4. Slovene

Slovene exhibited the strongest overall performance throughout the study and required the least adaptation during calibration. Unlike the other investigated languages, the model was already closely aligned with the Slovene MULTEXT-East annotation conventions (Erjavec and Krek, 2010), so iterative calibration served primarily as refinement rather than correction. The stable performance observed during the blind evaluation further indicates that this alignment generalized successfully to previously unseen data.

A detailed examination of the remaining annotation disagreements revealed that most were associated with MSD distinctions rather than lexical categorization or grammatical analysis. The model generally selected the correct lemmas and PoS categories but occasionally assigned incorrect values for individual morphosyntactic features, particularly within complex adjectival and participial paradigms.

Like other three South Slavic languages, occasional disagreements were observed in the treatment of participial and adjectival forms, although these inconsistencies typically reflected differences in annotation conventions rather than incorrect linguistic interpretation. Other recurrent disagreements involved highly specific MSD attributes, including adjective subcategories and inflectional features, where the predicted analyses remained linguistically plausible but did not exactly match the MULTEXT-East gold standard.

Compared with the other three South Slavic languages, Slovene required the least corrective feedback to achieve stable annotation behaviour. This suggests that ChatGPT 5 Plus already possessed internal linguistic representations that were closely aligned with the Slovene MULTEXT-East framework. Consequently, the remaining annotation disagreements primarily reflected the challenge of reproducing highly detailed annotation conventions rather than limitations in grammatical competence.

4.5. Cross-linguistic patterns

Across all four languages, the disagreements between ChatGPT-generated and human annotations demonstrated consistent patterns despite differences in the MULTEXT-East annotation specifications. Iterative calibration reduced many of these disagreements, although several challenges remained unresolved, indicating that the linguistic knowledge acquired during pre-training frequently prevailed over the language-specific annotation conventions introduced during calibration. Despite language-specific differences, the remaining annotation disagreements can be grouped into following four categories:

Lemma selection. The model frequently produced linguistically valid lemmas that differed from those adopted in human annotation. This phenomenon was particularly evident for auxiliary verbs, participles, and derived adjectival forms, where several alternative lemma-selection strategies coexist across the South Slavic languages.

Participial constructions. Active, passive, and adjectivized participles represented the single most persistent source of disagreement. While Macedonian and Slovene predominantly treated passive participles as adjectives, Bulgarian and Serbian more frequently preserved their verbal origin. Similar inconsistencies were observed across all four languages, suggesting that participial constructions constitute one of the most demanding aspects of LLM-based morphosyntactic annotation.

MSD convention mismatches. Many disagreements resulted from incomplete adherence to language-specific MULTEXT-East conventions, including feature ordering, category selection, adjective subcategories, and the encoding of fine-grained morphological attributes. In most cases, the predicted analyses were linguistically plausible but did not exactly correspond to the required annotation specifications.

Fine-grained functional categories. Pronouns, numerals, determiners, auxiliary verbs, and highly frequent multifunctional words generated a disproportionate number of disagreements despite constituting only a relatively small portion of the corpus. Frequent forms such as *kamo* in Bulgarian and *kao* in Serbian remained particularly resistant to corrective feedback, illustrating that corpus-specific lexical classification conventions can be more difficult to acquire than broader grammatical categories.

Across all four languages, iterative human-guided calibration substantially improved adherence to these conventions, while the remaining disagreement were largely confined to language-specific morphosyntactic distinctions and corpus-specific annotation practices.

These findings suggest that the adoption of more standardized annotation frameworks, such as Universal Dependencies (de Marneffe et al., 2021), may facilitate more consistent annotation and enable large language models to achieve even higher levels of agreement with human annotators with less calibration effort.

5. Limitations

This study has several limitations that should be considered when interpreting the results. The evaluation was conducted on a relatively small corpus consisting of the opening section of Orwell’s 1984. Although the blind evaluation contained substantially more previously unseen material than the calibration phases, larger and more diverse corpora are needed to determine whether the reported findings generalize across different text genres and linguistic styles.

Another limitation is that the study investigated only one commercially available large language model, ChatGPT 5 Plus. Consequently, the observed learning behaviour and annotation performance may differ for other LLMs or future model versions.

The evaluation was further restricted to the MULTEXT-East annotation framework (Erjavec, 2012), whose detailed language-specific conventions represent only one approach to morphosyntactic annotation. As a result, the findings cannot be directly generalized to alternative frameworks such as Universal Dependencies (de Marneffe et al., 2021).

An additional methodological limitation concerns the multilingual calibration procedure. Since the four languages were calibrated within a single continuous interaction, language-specific annotation conventions occasionally influenced subsequent annotations in related languages. Although this behaviour provided valuable insights into cross-linguistic transfer, future studies should also investigate independently calibrated language-specific sessions to isolate these effects more systematically.

Finally, the calibration process relied on interactive expert feedback provided by a single researcher. Although this ensured methodological consistency throughout the experiment, future stud-

ies should investigate the influence of multiple annotators, larger calibration datasets, and complementary evaluation measures, including precision, recall, F1-scores, and confusion matrix analyses.

6. Conclusions and future work

The three research questions posed in the Introduction can now be answered. RQ1 showed that ChatGPT 5 Plus achieved high zero-shot performance in lemmatization and PoS tagging across all four South Slavic languages, whereas MSD annotation proved considerably more challenging due to the complexity of the language-specific MULTEXT-East framework. RQ2 proved that iterative human-in-the-loop calibration improved annotation quality, particularly for MSD annotation, and enabled the acquired annotation conventions to generalize successfully to previously unseen data, as confirmed by the blind evaluation. Finally, RQ3 indicated that the remaining disagreements were concentrated in a limited number of linguistic categories, including participial constructions, auxiliary verbs, lemma selection, and fine-grained morphosyntactic distinctions.

Taken together, these findings demonstrate that the main challenge for LLM-based morphosyntactic annotation is no longer grammatical analysis itself, but the reliable acquisition of corpus-specific annotation conventions while preserving the valuable linguistic knowledge acquired during pre-training. The results of this pilot study show that ChatGPT 5 Plus can successfully acquire these conventions through iterative human-in-the-loop calibration and corrective feedback, although the adaptation remains incomplete when pre-trained linguistic knowledge prevails over corpus-specific annotation standards. Consequently, the findings support the use of iterative human-in-the-loop calibration for morphosyntactic annotation, particularly in morphologically rich languages (Popović and Arčan, 2015) employing detailed annotation frameworks such as MULTEXT-East (Erjavec, 2012).

One of the main challenges observed during the experiment was the tendency of the model to transfer annotation patterns across related languages. For example, after learning that many Macedonian passive participles were annotated as adjectives, the model initially applied a similar strategy to Bulgarian. Subsequent corrections successfully reduced these disagreements but occasionally introduced new inconsistencies, particularly in Serbian, where adjectival

and verbal interpretations of participial forms follow different annotation conventions. Cross-linguistic transfer therefore appears to be a double-edged phenomenon, simultaneously facilitating learning while introducing language-specific annotation errors.

Another persistent challenge involved the lemmatization and classification of auxiliary verb forms. Despite repeated calibration, the model occasionally assigned linguistically plausible lemmas that differed from those adopted in the gold-standard corpora. These observations suggest that, in some cases, ChatGPT 5 Plus continued to rely on linguistic generalizations acquired during pre-training instead of fully adopting the language-specific annotation conventions introduced through iterative calibration. These findings further suggest that more effective adaptation strategies, such as retrieval-augmented generation (RAG), (Lewis et al., 2020) or extended human-in-the-loop calibration, may be required to achieve full compliance with corpus-specific annotation standards.

Future work will extend this pilot study by replacing the relatively small calibration sample with substantially larger manually annotated corpora. Particular attention will be devoted to evaluating whether larger calibration datasets further improve annotation quality and reduce cross-linguistic interference effects. Forthcoming experiments will also calibrate each language independently until annotation performance stabilizes before extending the methodology to the next language, thereby minimizing cross-linguistic interference between annotation conventions and enabling more reliable multilingual comparisons.

In addition to overall annotation accuracy, future evaluations will incorporate precision, recall, F1-scores, and confusion matrix analyses to provide a more comprehensive assessment of annotation performance and language-specific error patterns.

Another promising direction is the investigation of RAG (Lewis et al., 2020), where language-specific MULTEXT-East specifications are dynamically retrieved and supplied during annotation. Unlike iterative human-in-the-loop calibration, which gradually adapts the model through corrective feedback, RAG provides the relevant annotation conventions at inference time and may reduce the model’s tendency to rely on linguistic knowledge acquired during pre-training.

Finally, MULTEXT-East annotations will be mapped to Universal Dependencies (UD), enabling comparison with existing NLP resources through the CoNLL-U format (Straka et al., 2016). Together, these developments are expected to improve annotation consistency, facilitate interoperability with existing NLP tools, and reduce the amount of manual calibration required.

Ultimately, the present study suggests that the future of LLM-assisted morphosyntactic annotation lies not in replacing linguistic expertise, but in combining the broad grammatical knowledge acquired during pre-training with systematic human guidance and language-specific annotation resources.

We invite researchers and language experts working on South Slavic languages to join this effort. Broader expert participation will facilitate the development of larger manually annotated corpora, improve annotation consistency, and help transform the traditionally labour-intensive process of manual morphosyntactic annotation into a human-guided refinement workflow supported by large language models. We hope that this study will encourage closer collaboration between computational linguists and AI researchers toward the development of reliable multilingual annotation workflows for South Slavic and other morphologically rich languages.

Acknowledgements

This work is partially financed by the Ministry of Education and Science of the Republic of North Macedonia through the project *Utilising AI and National Large Language Models to Advance Macedonian Language Capabilities*.

Language Resources

CLARIN. 2026. Part-of-speech taggers and lemmatisers. <https://www.clarin.eu/resource-families/tools-part-speech-tagging-and-lemmatisation>.

CLARIN.SI Repository. 2010. MULTEXT-East “1984” Annotated Corpus 4.0. <https://www.clarin.si/repository/xmlui/handle/11356/1041>.

- Tomaž Erjavec. 2010. MULTEXT-East Resources, Version 4 “MondiLex”. <https://nl.ijs.si/ME/Vault/V4/>.
- Tomaž Erjavec and Simon Krek. 2010. MULTEXT-East Morphosyntactic Specifications, Version 4, 3.6. Slovene Specifications. <https://nl.ijs.si/ME/Vault/V4/msd/html/msd-sl.html>.
- Tomaž Erjavec et al. 2010. MULTEXT-East “1984” Annotated Corpus 4.0. <http://hdl.handle.net/11356/1043>.
- Cvetana Krstev and Duško Vitas. 2010. MULTEXT-East Morphosyntactic Specifications, Version 4, 3.9. Serbian Specifications. <https://nl.ijs.si/ME/Vault/V4/msd/html/msd-sr.html>.
- Radoslav Pavlov, Ludmila Dimitrova, Ludmila Sinapova, and Kiril Simov. 2009. MULTEXT-East Morphosyntactic Specifications, Version 4, 3.13. Bulgarian Specifications. <https://nl.ijs.si/ME/Vault/V4/msd/html/msd-bg.html>.
- Katerina Zdravkova. 2020. MULTEXT-East Morphosyntactic Specifications, Version 6, 3.12. Macedonian Specifications. <https://nl.ijs.si/ME/V6/msd/html/msd-mk.html>.

References

- Marie-Catherine de Marneffe et al. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.
- Tomaž Erjavec. 2012. MULTEXT-East: Morphosyntactic Resources for Central and Eastern European Languages. *Language Resources and Evaluation*, 46(1):131–142.
- Michael Ginn and Alexis Palmer. 2025. LLM Dependency Parsing with In-Context Rules. In *Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025)*, pages 186–196.
- Nancy Ide. 2017. Introduction. In *Handbook of Linguistic Annotation*, pages 1–18. Springer Netherlands, Dordrecht.
- Daniel Kondratyuk et al. 2018. LemmaTag: Joint Tagging and Lemmatization for Morphologically Rich Languages with BRNNs. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4921–4928.

- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, et al. 2020. [Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.
- Maja Popović and Mihael Arčan. 2015. [Identifying Main Obstacles for Statistical Machine Translation of Morphologically Rich South Slavic Languages](#). In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, pages 97–104.
- Milan Straka, Jan Hajič, and Jana Straková. 2016. [UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 4290–4297.
- Milan Straka and Jana Straková. 2017. [Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe](#). In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99.
- Chahan Vidal-Gorène, Bastien Kindt, and Florian Cafiero. 2026. [Under-Resourced Studies of Under-Resourced Languages: Lemmatization and POS-Tagging with LLM Annotators for Historical Armenian, Georgian, Greek and Syriac](#). In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026)*, pages 324–334.

Lingvistička anotacija južnoslovenskih jezika pomoću velikih jezičkih modela

Katerina Zdravkova

Sažetak

Ova studija istražuje primenljivost ChatGPT 5 Plus za lingvističku anotaciju četiri južnoslovenska jezika: bugarskog, makedonskog, srpskog i slovenačkog. Iako su nedavna istraživanja pokazala potencijal velikih jezičkih modela za lingvističku anotaciju, njihova sposobnost da usvoje detaljne jezički specifične konvencije morfosintaksičke anotacije još uvek nije dovoljno istražena. Koristeći odgovarajuće prevode prvih 1235 tokena iz Orvelovog romana „1984“, zajedno sa njihovim MULTEXT-East anotacijama, ocenili smo model za lematizaciju, označavanje vrsta reči (PoS) i anotaciju morfosintaksičkog opisa (MSD). Eksperiment je obuhvatio četiri faze: anotaciju bez ijednog pokušaja, kalibraciju sa jednokratnim učenjem zasnovanom na ručno anotiranim primerima, iterativnu kalibraciju putem korektivnih povratnih informacija i nezavisnu slepu evaluaciju. Anotacija bez ijednog pokušaja je već pokazala visoku tačnost lematizacije i pouzdano PoS označavanje, dok je iterativna kalibracija značajno poboljšala usklađenost sa jezički specifičnim konvencijama anotacije, naročito za MSD anotaciju. Slepa evaluacija na prethodno neviđenom skupu podataka potvrdila je visoku tačnost lematizacije (94.95–99.22%) i PoS označavanja (94.69–96.61%) u sva četiri jezika, dok se tačnost MSD anotacije kretala od 75.16% do 89.21%. Većina primećenih grešaka odražavala je tendenciju modela da se oslanja na sopstveno lingvističko znanje, umesto da dosledno usvaja jezičke specifične konvencije anotacije uvedene tokom kalibracije. Bez obzira na njihovu relativno čestu pojavu, naročito tokom MSD anotacije, dobijeni rezultati pokazuju da ovaj četbot može efikasno podržati morfosintaksičku anotaciju primenom iterativne kalibracije zasnovane na konceptu čovek u petlji i ističu značaj stručnih povratnih informacija za prilagođavanje velikih jezičkih modela korpusno specifičnim standardima anotacije.

Ključne reči: lematizacija, veliki jezički modeli, morfosintaksička anotacija, označavanje vrste reči

Works are licensed under the terms and conditions of the Creative Commons Attribution 4.0 International License.

<https://creativecommons.org/licenses/by/4.0/>

