Quality Textual Corpora and New South Slavic Language Models

Scientific paper DOI: 10.18485/judig.2025.1.ch19

Mihailo Škorić[1], *0000-0003-4811-8692*
Saša Petalinkar[2], *0009-0007-9664-3594*

Abstract

This paper presents a newly published quality corpus, *Znanje*, which builds on the existing S.T.A.R.S. corpus of Serbian academic research, but extends its range to more South Slavic languages, primarily Croatian and Slovenian. The new corpus is compiled using texts from the S.T.A.R.S corpus, as well as Croatian academic writings collected from the DABAR repository, and Slovenian academic writings collected from several online repositories under the Open Science Slovenia initiative. Secondly, the paper presents two new encoding language models for Serbo-Croatian, trained using this dataset, as well as the *Kišobran* umbrella web-corpus of Serbian and Croatian, and Wikipedia and Wiki source dumps for Serbo-Croatian, Serbian, Croatian and Bosnian. Finally, it discusses the evaluation of these new models on several previously established text vectorization tasks (for Serbian and Serbo-Croatian), their comparison with the similar models previously trained in this language domain and the outlook of preparing large language models for South Slavic languages in the future.

Key words: language models, corpora, Serbian, Croatian, parameter-efficient training.

[1] Language Resources and Technologies Society (JeRTeh), mihailo@jerteh.rs
[2] Language Resources and Technologies Society (JeRTeh), sasa5@jerteh.rs

1. Introduction

The performance and reliability of language models are fundamentally tied to the quality of the textual data they are trained on. High-quality corpora ensure that models learn accurate linguistic patterns, semantic relationships, and contextual nuances, which are essential for tasks such as text classification, translation, summarization, and semantic search. In low-resource languages or dialects, such as South Slavic languages, the scarcity of corpora in general, but also especially well-curated corpora, has historically limited the development of robust models.

Recent research underscores that not only the size but also the diversity, cleanliness, and representativeness of the corpus significantly affect model generalization and downstream task performance. Poorly curated datasets can introduce noise, bias, and inconsistencies, leading to models that misinterpret or misrepresent linguistic phenomena (*Yang et al., 2025*). For multilingual models, the imbalance in corpus quality across languages can further exacerbate bias and reduce cross-lingual transfer efficiency (*Xu et al., 2024*).

Therefore, constructing a high-quality academic corpus is a critical step toward advancing language technologies for South Slavic languages. It not only addresses the data scarcity issue but also lays the groundwork for training fairer and more effective language models in this linguistic domain.

This research builds upon (*Škorić & Janković, 2024)*, where several new corpora for Serbian and Serbo-Croatian were introduced, including a doctoral dissertation corpus, S.T.A.R.S. and a parallel academic abstract corpus, *PaSaž*. Doctoral dissertations represent a highly valuable resource for building academic corpora due to their rigorous standards in methodology, linguistic quality, and peer review. These texts offer a rich and diverse thematic range across scientific disciplines and are often timely in addressing contemporary research questions. In Serbia, the NaRDuS [3]repository provides open access to thousands of doctoral theses, supported by legal mandates and standardized metadata. This makes it an ideal foundation for constructing a large, high-quality scientific corpus in Serbian.

To compile the corpus, metadata for all dissertations was programmatically retrieved and enriched with additional fields to assess document accessibility, language, OCR requirements, and copyright status. Dissertations were filtered to include only those written in Serbian, available in full-text PDF format, not requiring OCR, and not under restrictive copyright. This process yielded 11,624 dissertations or approximately 87.5% of the repository. The final corpus extracted using

[3] https://nardus.mpn.gov.rs

PyMuPDF[4] library and cleaned to retain only paragraph text, comprised over 560 million words, making it one of the largest and most reliable academic corpora in Serbian.

Since the initial compilation of the doctoral dissertation corpus, S.T.A.R.S has been further expanded by incorporating research papers (peer-reviewed journal articles, conference proceedings, and other scholarly publications) from multiple academic online repositories in Serbia, i.e. active institutional repositories of higher education institutions, where researchers and peers store the published papers (*Otašević, 2023*). This expansion increased the size of the corpus to over 700 million words.

That work established the methodological framework for corpus compilation, which this paper extends to a broader South Slavic context by incorporating Croatian and Slovenian academic texts.

2. New Corpus

To further enrich the corpus and extend its linguistic and thematic coverage, additional academic texts were gathered from repositories across several countries of former Yugoslavia. For each country, prominent open-access academic repositories were identified, and custom scraping scripts were developed to extract relevant documents. Beyond the full-text PDFs, the collection process prioritized key metadata fields essential for corpus structuring and future filtering: title, creator, source, subject (where available), PDF URL, and language. In cases where language metadata was missing, particularly in smaller repositories, manual annotation was performed to ensure consistency and usability across the dataset.

For the Croatian portion of the corpus, academic texts were sourced from DABAR (Digital Academic Archives and Repositories)[5], the national infrastructure for institutional repositories in Croatia. DABAR hosts a wide range of scholarly outputs, including theses, dissertations, and research papers from Croatian universities and research institutions. Its standardized metadata and open-access structure made it an ideal candidate for large-scale corpus construction. Despite this, DABAR had not previously offered a publicly available, structured corpus suitable for language modelling. Using a custom scraping pipeline tailored to DABAR's architecture, a total of 108,786 documents were collected, comprising approximately 53.4 million sentences and over 1.2 billion words. This substantial addition significantly enhances the corpus's linguistic diversity and supports the development of more robust language models for Croatian.

[4] https://pymupdf.io

[5] https://dabar.srce.hr

For Bosnia and Herzegovina, the corpus was expanded using academic texts from a single well-structured and accessible repository: the institutional repository of the University of East Sarajevo[6]. This repository provided a consistent metadata schema and open access to a range of academic documents, making it suitable for automated collection. A tailored scraping script was developed to extract both full-text PDFs and essential metadata. In total, 8 documents were collected, containing 21,156 sentences and approximately 510,000 words. While modest in scale compared to other national segments, this addition contributes valuable linguistic and thematic diversity to the overall corpus, as the texts were written in Ijekavian Serbian. The repository of the University of Zenica[7] was also considered, but eventually the effort was scrapped, since these texts were mostly written in English, and metadata was scarce and .

For Montenegro, academic texts were collected from the institutional repository of the University of Montenegro[8], the country's only public university. This repository provided a centralized and structured source of scholarly documents, making it suitable for targeted scraping. A custom script was developed to extract both full-text PDFs and essential metadata: title, author, source, and language. In total, 315 documents were collected, comprising approximately 605,119 sentences and 14.5 million words. This contribution marks the first large-scale academic corpus from Montenegro and adds valuable linguistic representation to the South Slavic dataset.

For Slovenia, academic repositories were defined through the Open Science Slovenia initiative[9], which aggregates institutional repositories from major Slovenian universities and research centers. The repositories of the University of Ljubljana[10], University of Maribor[11], University of Primorska[12], and University of Nova Gorica[13] were included, along with specialized repositories such as DiRROS (Digital Repository of Research Outputs in Slovenia)[14] and REVIS (Repository of Educational and Scientific Works)[15]. These platforms provided structured metadata and open access to a wide range of academic documents. It should be noted that unlike other South Slavic countries, Slovenia had already previously

[6] https://repozitorijum.ues.rs.ba
[7] https://www.epub.unze.ba
[8] https://eteze.ucg.ac.me
[9] https://www.openscience.si
[10] https://repozitorij.uni-lj.si
[11] https://dk.um.si
[12] https://repozitorij.upr.si
[13] https://repozitorij.ung.si
[14] https://dirros.openscience.si
[15] https://revis.openscience.si

established a large-scale academic corpus from Open Science Slovenia, *OSS 1.0*, which reports over 3.2 billion words from more than 150,000 scientific texts (*Erjavec, Fišer & Ljubešić, 2021*). During this effort, however, for the Serbo-Croatian portion, 176 documents were collected, totaling 113,339 sentences and 2.7 million words. For Slovenian-language texts, a significantly larger dataset was compiled: 148,158 documents, comprising 104.9 million sentences and over 2.29 billion words. This makes the Slovenian segment the largest national contribution to the corpus and a major resource for future Slovenian language modeling efforts.

For North Macedonia, no suitable academic repository was identified that met the criteria for inclusion in the corpus. The institutional repositories of Goce Delčev University[16] and St. Clement of Ohrid University[17] were reviewed, but both presented challenges in terms of metadata accessibility and document structure. Manual inspection revealed that only a small number of documents were written in South Slavic languages, with many texts either lacking language metadata or being authored in English, Albanian and Greek. Due to these limitations, Macedonia was not included in the current version of the corpus, though future expansions may revisit this if repository infrastructure improves.

Table 1 presents contributions to the Serbo-Croatian portion of the final corpus by repository source or sources. The total word count of this portion is nearly 2 billion, making it the largest compiled academic corpora for these languages. This, however, still falls short of the Slovenian portion, which has a word count of nearly 2.3 billion.

Table 1:Total contributions to the Znanje corpus from each source repository (Serbo-Croatian macro-language portion that was used for model training).

Source	Doc. count	Sent. count	Word count	Share
NARDUS	11,432	22,779,252	574,600,000	30%
Institutional repositories in Serbia	10,889	4,192,656	109,400,000	5.70%
University of Montenegro	315	605,119	14,500,000	0.80%
University of East Sarajevo	8	21,156	510,000	< 0.1%
DABAR	108,786	53,369,657	1,214,000,000	63.40%
Open Science Slovenia	176	113,339	2,694,331	0.10%
Total	**131,606**	**81,060,023**	**1,915,704,331**	**100%**

[16] https://eprints.ugd.edu.mk
[17] https://repository.ukim.mk

3. New Models

Building on the Znanje corpus[18], we trained two new transformer-based language models using the *XLM-R* architecture (Conneau et al., 2020). These models are specifically designed for the Serbo-Croatian macro-language, and thus only Serbo-Croatian portion of the Znanje corpus was used. To further enrich the training data, we also incorporated the umbrella web corpus Kišobran[19], which encompasses all published web-crawled datasets of Serbian, Croatian, Bosnian and Montenegrin texts, as well as compiled Wikipedia dumps for Serbian[20], Croatian, and Serbo-Croatian, and Wikisource dumps for Serbian, Croatian, and Bosnian[21]. These sources provided a diverse and balanced representation of contemporary and formal language use across the macro-language spectrum. In addition to these modern sources, the training data included the SrpELTeC corpus[22], a collection of digitized Serbian literary texts from the 19th and early 20th centuries (*Stanković, Krstev, Todorović & Škorić, 2021*). This corpus, consisting of historical novels and prose, added valuable diachronic depth to the language model and helped capture stylistic and lexical variation over time. Altogether, the combined training dataset exceeded 20 billion words, making it the largest corpus ever used for training South Slavic language models. The resulting models were published on Hugging Face[23], ensuring open access for further research and application in NLP tasks involving South Slavic languages.

To preserve the multilingual capabilities of the original *XLM-R* architecture while minimizing computational costs, we adopted parameter-efficient fine-tuning using LoRA (Low-Rank Adaptation) methods (*Hu et al., 2021*). LoRA enables selective updating of a small subset of model parameters, allowing the model to adapt to new data without full retraining. This approach was particularly suitable given the scale of the training corpus, which was over 20 billion words, and the goal of maintaining compatibility with the broader multilingual XLM-R framework. By using LoRA, we achieved efficient specialization for Serbo-Croatian while retraining the model's generalization capacity across related languages.

To accommodate different use cases and computational environments, two model variants were developed: *XLMali*[24] and

[18] https://huggingface.co/datasets/procesaur/znanje
[19] https://huggingface.co/datasets/procesaur/kisobran
[20] https://huggingface.co/datasets/procesaur/Vikipedija
[21] https://huggingface.co/datasets/procesaur/Vikizvornik
[22] https://huggingface.co/datasets/jerteh/SrpELTeC
[23] https://huggingface.co
[24] https://huggingface.co/te-sla/XLMali

TeslaXLM[25]. *XLMali* is a compact model based on the *XLM-R-base* architecture, with 279 million parameters, optimized for faster inference and lower resource consumption, ideal for real-time applications and deployment on limited hardware. *TeslaXLM*, on the other hand, is built on the *XLM-R-large* architecture, with 561 million parameters, offering superior performance on complex NLP tasks due to its greater representational capacity. This dual-model approach ensures flexibility, allowing researchers and developers to choose between speed and precision depending on their needs. Due to usage of LoRA, models were trained updating only 0.03% of the model parameters, significantly reducing computational demands. The training was conducted on a setup of four NVIDIA A4000 GPUs, with *XLMali* completing in 52 hours and *TeslaXLM* in 186 hours.

4. Evaluation and Results

To assess the performance of the newly trained models (XLMali and TeslaXLM) alongside three existing baselines: *XLM-R-base* and *XLM-R-large* (*Conneau et al., 2020*)., as well as *XLM-R-BERTić* (*Ljubešić et al., 2024*). We conducted evaluations on two previously defined upstream tasks (*Škorić, 2024*). The first task was masked token prediction in literary texts: *The Adolescent* by Fyodor Mikhailovich Dostoevsky represented by two different Serbian translations[26], and *Around the World in 80 Days* by Jules Verne in both Serbian and Croatian[27]. All models use the same tokenizer, which was also employed to mask the input texts, ensuring consistency across evaluations. This task tested the models' ability to recover contextually appropriate words in literary prose, a challenging domain due to its stylistic and lexical complexity.

The second task focused on sentence pair recognition, evaluating semantic similarity and alignment. This was performed on parallelized versions of the same novels: *The Adolescent* for Serbian–Serbian pair recognition, and *Around the World in 80 Days* for Serbian–Croatian pair recognition. These evaluations measured how well each model could distinguish between semantically aligned and misaligned sentence pairs while also reflecting their capacity for cross-lingual understanding within the Serbo-Croatian macro-language. The results of both tasks are presented in Figures 1 and 2, showing comparative performance across all five models.

[25] https://huggingface.co/te-sla/TeslaXLM

[26] Mladić, Beograd (1972), Dečko, Beograd (1988)

[27] Put oko sveta za 80 dana, Beograd (1962), Put oko svijeta u osamdeset dana, Zagreb (1961)

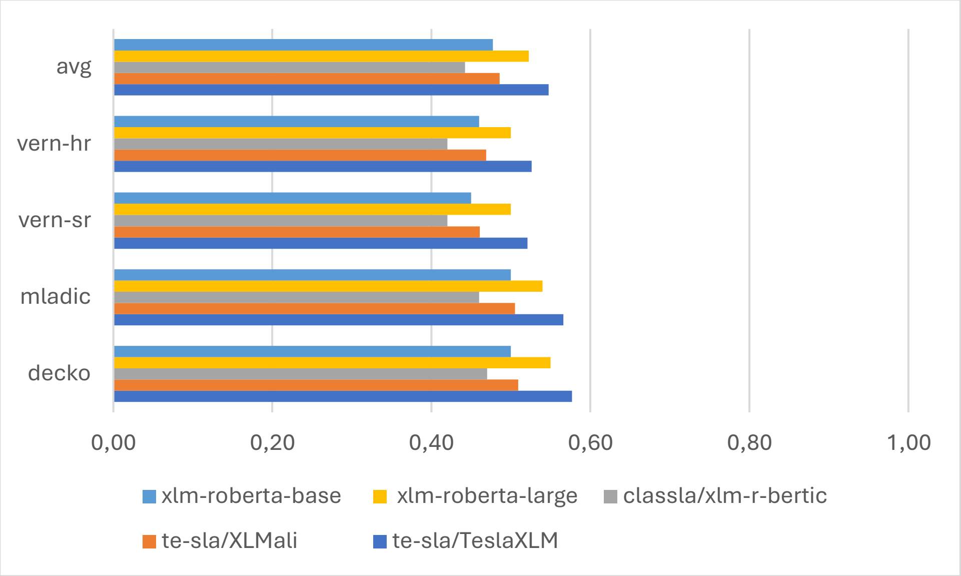

Figure 1: Masked token prediction accuracy across five models on literary texts in Serbian and Croatian. Evaluation performed on masked excerpts from Dostoevsky's The Adolescent and Verne's Around the World in 80 Days.

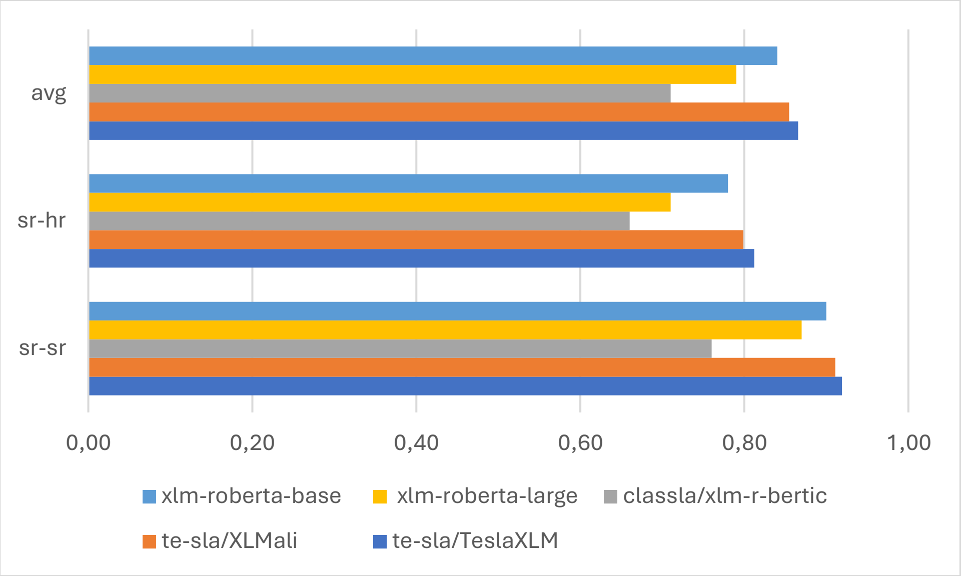

Figure 2: Sentence pair recognition performance for Serbian–Serbian and Serbian–Croatian parallel texts. Models evaluated on an aligned and misaligned sentence pairs from The Adolescent and Around the World in 80 Days, measuring semantic similarity and cross-lingual understanding.

On the first task *TeslaXLM* performed best in all categories, likely due to its larger parameter count and domain-specific training. *XLM-R-large* also performed well, showing that size and multilingual pretraining help. *XLMali*, while smaller, still outperformed *XLM-R-BERTić* and *XLM-R-base*, validating its efficiency-focused design. On the second task, *TeslaXLM* again leads, showing strong cross-lingual semantic understanding. Interestingly, *XLMali* slightly outperforms *XLM-R-base* and *large*, suggesting that targeted fine-tuning on Serbo-Croatian data yields better results than general multilingual pretraining.

In addition to upstream evaluations, all five models were tested on a suite of five downstream NLP tasks to assess their practical utility and generalization capabilities. These tasks included part-of-speech (POS) tagging (17 classes) using SrpKor4Tagging dataset[28], named entity recognition (NER, 5 classes) using SrpELTeC-gold-NER[29], sentiment classification (3 classes) (*Stanković et al., 2022*), as well as emotion classification (8 classes), and moral classification (5 classes)[30]. Each task was selected to reflect a different aspect of linguistic understanding, from syntactic structure to affective and ethical interpretation.

The results of these evaluations are presented in Figure 3, which compares model performance across all five tasks. These benchmarks provide insight into how well each model adapts to real-world applications in South Slavic language processing. Notably, *TeslaXLM* consistently outperformed other models in tasks requiring deeper semantic understanding such as emotion and moral classification, while in the other categories results were similar. *XLMali* also offered competitive results with significantly faster inference times, but on average fell shortly out of *XLM-R-base* and *XLM-R-Bertić* on downstream tasks.

[28] https://huggingface.co/datasets/jerteh/SrpKor4Tagging

[29] https://huggingface.co/datasets/jerteh/SrpELTeC-gold-NER

[30] Unpublished annotated datasets

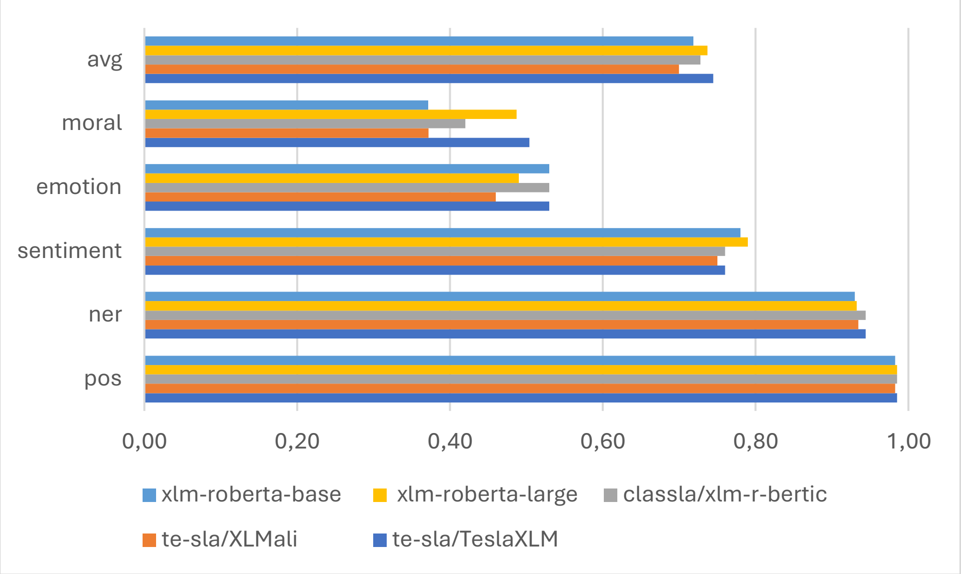

Figure 3:Downstream task performance comparison across five models. Tasks include part-of-speech tagging (17 classes), named entity recognition (5 classes), sentiment classification (3 classes), emotion classification (8 classes), and moral classification (5 classes).

5. Conclusion

This research presents a significant advancement in the development of language resources and models for South Slavic languages. By compiling the *Znanje* corpus, which integrates academic texts from Serbia, Croatia, Bosnia and Herzegovina, Montenegro, and Slovenia, we have created one of the most comprehensive and linguistically diverse corpora in the region. The inclusion of over 20 billion words from curated sources such as doctoral dissertations, national repositories, Wikipedia, and historical literature marks a milestone in corpus construction for low-resource languages. This effort not only fills a critical gap in available training data but also lays the groundwork for future multilingual and cross-lingual NLP research in the South Slavic domain.

The development of two new Serbo-Croatian language models, XLMali and TeslaXLM, demonstrates the effectiveness of parameter-efficient fine-tuning using LoRA, enabling high performance with reduced computational cost. Evaluation across upstream and downstream tasks shows that these models outperform or match existing multilingual baselines, particularly in semantic understanding and classification tasks.

By publishing the models on Hugging Face, we ensure open access and reproducibility, supporting further research and practical applications.

While the Znanje corpus and the newly trained models represent a major step forward for South Slavic NLP, several avenues remain open for future research. Expanding the corpus to include underrepresented languages such as Macedonian and minority dialects within the Serbo-Croatian continuum would further enhance linguistic coverage.

References

[1] Xu, Y., Hu, L., Zhao, J., Qiu, Z., Ye, Y., & Gu, H. (2024). *A Survey on Multilingual Large Language Models: Corpora, Alignment, and Bias*. arXiv. https://arxiv.org/pdf/2404.00929

[2] Yang, S., Jing, M., Wang, S., Kou, J., Shi, M., Xing, W., Hu, Y., & Zhu, Z. (2025). *Exploring Large Language Models in Healthcare: Insights into Corpora Sources, Customization Strategies, and Evaluation Metrics*. arXiv. https://arxiv.org/abs/2502.11861

[3] Škorić, M., & Janković, N. (2024). *New Textual Corpora for Serbian Language Modeling*. Infotheca. - Vol. 24, No. 1 (2024), p. 71–96.

[4] Otaševic, V. *Transfer of Metadata into the National Information System of Scientific Research Activities with Automatic Authorship Association*. Infotheca. - Vol. 23, No. 2 (2024), p. 27–48.

[5] Erjavec, T., Fišer, D., & Ljubešić, N. (2021). *The KAS corpus of Slovenian academic writing*. Language Resources and Evaluation, 55(2), 551-583.

[6] Conneau, A. et al. (2019). *Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116.*

[7] Stanković, R., Krstev, C., Todorović, B. Š., & Škorić, M. (2021). *Annotation of the serbian eltec collection*. Infotheca–Journal for Digital Humanities, 21(2), 43-59.

[8] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2022). *Lora: Low-rank adaptation of large language models*. *ICLR*, *1*(2), 3.

[9] Ljubešić, N., Suchomel, V., Rupnik, P., Kuzman, T., & van Noord, R. (2024). *Language Models on a Diet: Cost-Efficient Development of Encoders for Closely-Related Languages via Additional Pretraining*. arXiv preprint arXiv:2404.05428.

[10] Škorić, M. (2024). *New Language Models for Serbian*. Infotheca. - Vol. 24, No. 1 (2024), p. 7–28.

[11] Stanković, R., Košprdić, M., Nešić, M. I., & Radović, T. (2022, June). Sentiment analysis of Serbian old novels. In *Proceedings of the 2nd Workshop on Sentiment Analysis and Linguistic Linked Data* (pp. 31-38).

Квалитетни текстуални корпуси и нови јужнословенски језички модели

Михаило Шкорић, Саша Петалинкар

Овај рад представља значајан допринос развоју језичких ресурса и модела за јужнословенске језике. Описује конструисање новообјављеног корпуса *Знање*, који обухвата академске текстове из Србије, Хрватске, Босне и Херцеговине, Црне Горе и Словеније. Корпус је изграђен комбиновањем више извора: докторских дисертација из НаРДУС-а, научних радова из националних репозиторијума као што су ДАБАР (Хрватска) и *Open Science Slovenia*, као и институционалних репозиторијума широм Србије, универзитета у Источном Сарајеву и Универзитета Црне Горе. Корпус се може грубо поделити на део који чине текстови на српскохрватском макро-језику (нешто мање од две милијарде речи) и део на словеначком језику (нешто мање од 2,3 милијарде речи).

На основу српскохрватском дела овог корпуса, и неколико других (Кишобран веб-корпус, Википедија, Викизворник и СрпЕЛТеК) обучена су два нова језичка модела за векторизацију текста: *XLMali* (279 милиона параметара) и *TeslaXLM* (561 милион параметара), као адаптације *XLM-R* архитектуре. Модели су обучени су уз помоћ LoRA (Low-Rank Adaptation) технике и циљањем свега 0.3% параметара, што је омогућило ефикасну обуку уз минималне рачунарске ресурсе. Модели су тестирани на низу задатака: попуњавање маскираних речи у књижевним текстовима, препознавање парова реченица, као и пет задатака обраде природног језика — обележавање врстом речи, препознавање именованих ентитета, класификација сентимената, емоција и моралних вредности.

Резултати показују да *TeslaXLM* постиже најбоље резултате у већини задатака, док *XLMali* нуди одличан баланс између брзине и прецизности. Упоредна анализа са постојећим моделима као што су *XLM-R-base, XLM-R-large* и *XLM-R BERTić* потврђује предности доменски специфичне обуке. Модели су објављени на платформи *HuggingFace* и доступни су за даљу употребу и истраживање. Рад отвара простор за будући развој језичких технологија за мање заступљене језике и дијалекте, као и за проширење корпуса и модела на македонски и друге варијетете.

Кључне речи: Језички модели, корпуси, српски језик, хрватски језик, параметарски ефикасно учење.