Prompting Large Language Models for Church Slavic Translation

Edoardo Signoroni and Pavel Rychlý

NLP Centre, Faculty of Informatics, Masaryk University
Botanická 68a, 60200 Brno, Czech Republic
e.signoroni@mail.muni.cz, pary@fi.muni.cz

Abstract. Church Slavic is a low-resource historical language with limited resources and few experts. We explore the capabilities of off-the-shelf Large Language Models (LLMs) as Church Slavic translators by prompting multiple models in zero-shot and few-shot scenarios. We evaluate four LLMs of varying sizes on 262 sentence pairs translating Church Slavic into English and German, and conduct a second experiment examining the impact of model size using five Qwen2.5-Instruct variants. Our results show that on average EuroLLM-9B-Instruct achieves the best performance, outperforming much larger models. We find minimal benefit from few-shot prompting and performance gaps between English and German as target languages. The automated evaluation metrics suggest that LLMs can produce useful draft translations for Church Slavic, potentially assisting scholars in accessing historical texts.

Keywords: Large Language Models; Machine Translation; Church Slavic; Low-Resource Languages; Historical Languages

1 Introduction

Natural Language Processing (NLP) has made great progress in recent years with the advent of Large Language Models (LLMs). However, this advancement has mostly benefitted established, high-resource languages, such as English and Chinese. These represent only a tiny fraction of the 7000+ languages spoken now on the planet. The rest are still under-resourced and struggle to fill this ever-increasing digital divide.

Aside from living languages in need of NLP resources and tools, there is another class of under-resourced languages: historical ones. One such language is Church Slavic, a historical liturgical language used across several Slavic regions from the 9th century. NLP for this language is hindered by several challenges, first and foremost the scarcity of data and human experts [17].

The motivation of this paper is to explore the capabilities of off-the-shelf LLMs as Church Slavic translators. We conduct two experiments: prompting four LLMs in zero-shot and few-shot scenarios to assess the impact of in-context examples; and evaluating models of different sizes to understand the relationship between parameter count and translation quality. Our results show that

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2025, pp. 57–68, 2025.

contemporary LLMs can produce useful draft translations, however, significant performance disparities emerge between target languages.

The remainder of this paper is organized as follows. Section 2 provides background on Church Slavic and its historical development. Section 3 reviews related work on NLP for Church Slavic and historical languages. Section 4 describes our methodology, including the dataset and experimental setup. Section 5 presents our results for both the few-shot prompting and model size experiments. Finally, Section 6 concludes with a discussion of limitations and future work.

2 Church Slavic

Old Church Slavic (or Slavonic) is a Slavic language created by the 9th century missionaries Cyril and Methodius to christianize the Moravian Slavs and to translate the Bible and other Byzantine Greek religious texts in the local vernacular. They based it on the South Slavic Macedonian dialects around their native city of Thessalonica (modern Thessaloníki, Greece). The language retains strong influences from the original Greek text in syntax and lexicon [17,7].

This was the first Slavic literary language, written first in the Glagolitic and the in the Cyrillic alphabet.[1] Church Slavic, was adopted in other Slavic regions where it evolved and diverged, but remained in use as a religious and literary language during the Middle Ages, and as a liturgical language into modern times, being written until the 19th century by Serbs and Bulgarians [17,7,6,5].

Lendvai et al. (2025) [17] periodize Church Slavic in Early (10-11th cc.), Middle (12-16th cc.), and Late (17-18th cc.). Geographically, we have two main areas: South, covering Bulgaria (Early), Macedonia (Early and Middle), and Serbia (Middle and Late); and East, with the Kievan Rus' (Early and Middle), Novgorod, Suzdal, South Rus' (all in the Middle period), and Muscovy (Middle and Late).

3 NLP for Church Slavic

Church Slavic is not particularly prominent in NLP, however two major historical variants are gaining visibility, Old Church Slavic (`chu`) and Old East Slavic `orv`, as demonstrated by their inclusion in recent workshops and shared tasks, such as SIGTYP 2024 [2,4]. These variants are present in resources and tools such as Universal Dependencies [22], Stanza [27], UDPipe [31], and GlotLID [14]. Nevertheless, LLMs and embedding models struggle to process Church Slavic, because their tokenizers, trained on modern high-resource languages, are inadequate [17].

[1] The invention of Glagolitic is usually attributed to Cyril. The Cyrillic alphabet, despite its name, was devised by his disciples based on capital Greek letters, with some additions, when they were preaching in the South Slavic regions of the First Bulgarian Empire (modern-day Bulgaria and North Macedonia).

Recent work on Church Slavic includes datasets, and tools, but also wider studies.

Pedrazzini and Eckhoff (2021)[25] describe OldSlavNet, a variety-agnostic dependency parser for Church Slavic attained via manipulation of modern Slavic data to resemble the orthography and morphosyntax of pre-modern varieties.

Lendvai et al. (2023)[16] present a dataset of Pre-Modern Slavic text for cross-linguistic and diachronic analyses. They also finetune BERT [3] for these low-resource, historical Slavic variants to perform provenance attribution in terms of manuscript, century, and region of copying.

Lendvai et al. (2024)[15] outline their work in developing a workflow for the postprocessing, annotation, and classification of diachronic and regional language variation in Pre-Modern Slavic texts. They focus on the practical application of data obtained with handwritten text recognition and manual transcription. They adapt existing tools to make advanced NLP methods accessible for Humanists with limited computational experience.

Lendvai et al. (2025a)[18] evaluate sentence snippets in two diachronic variants, Old Church Slavic and Old East Slavic, using embedding representations from LLMs and string similarity approaches combined with k-nearest neighbour search to identify parallelizable text.

Cassese et al. (2025)[1] introduce DIACU (DIAchronic Analysis of Church Slavonic), a collection of existing corpora in Church Slavic, to be used as a unified corpus to investigate diachronic language phenomena and as a training set for machine learning attribution methods.

Janssen et al. (2025)[11] compile a sentence-level gold standard for alignments in five primary and derived texts, including ones in Church Slavic, related to *De Lepra ad Sistelium* by Methodius Olympius using the TEITOK corpus platform [12]. They then evaluate automated alignment methods, to show that for their historical text, Hunalign [10] performs better than deep learning methods. Notably, they prompt Llama 3.3 to produce translations needed to run MT-based methods such as Bleualign [30]. However, they do not specifically report about their experiments or translation quality.

Lendvai et al. (2025b)[17] address domain-specific text provenance classification for Church Slavic. They propose a new hierarchical and unambiguous labeling scheme and then finetune Vikhr [21], an LLM with knowledge of modern Russian, with instruction to classify sentence-level text, both of gold quality and with artificial noise, to simulate handwritten text recognition material. They show that the finetuned model is able to achieve good performances, with F-scores above .8 for all related tasks.

4 Methodology

4.1 Dataset

For these experiments, we rely on sections of the gold data compiled by Janssen at al. (2025)[11], to which we direct for a deeper analysis of the matter. The data are available at their repository[2].

As you may recall from Section 3, their dataset is focused on five different variants of a single historical work: *De Lepra ad Sistelium* by Methodius Olympius, a 4th century bishop active in Lycia (Asia Minor). The original text was written in Old Greek, and allegorically interprets the norms on leprosy found in the Old Testament (Leviticus 13) as a metaphor for moral and spiritual afflictions of one's soul and of the church as a community. The earliest manuscripts surviving in full are preserved in a 10th century Old Church Slavic translation.

The core of the dataset is the recent German edition of Jouravel et al. (2024)[13], which contains a critical edition of the Greek fragments, a reconstruction of the Slavic translation, the German translation of the Slavic text, and a diplomatic edition of the Slavic text. The gold standard was created starting from the Slavic reconstruction, to which the other text were gradually aligned first using an automated alignment method, and then manually checked and corrected.

From these, we use only the reconstructed Slavic and the German translation. The English translation was manually aligned from a public domain version.[3] We are left with 262 lines for two translation directions, Church Slavic-English (chu-eng) and Church Slavic-German (chu-deu).

4.2 Experiments

Zero and Few-Shot Prompting In our first experiment, we prompt three similar sized LLMs in a zero-shot, and two (1-shot and 3-shot) few-shot scenarios. The models we test are Qwen2.5-7B-Instruct [28], EuroLLM-9B-Instruct [19], and Vikhr-7B-Instruct_0.2 [21]. The latter model is adapted and developed specifically for Russian, and was previously used for Church Slavic by Lendvai et al. (2025b)[17]. To prompt these models, we use the HuggingFace framework. As a matter of comparison, we also prompt a much bigger model: a local instance of gpt-oss-120b [23].

We use the following prompt structure:

```
"You will be asked to translate a sentence from Old Church Slavonic
to {target_language}. You will output only the translation in
{target_language}, without any additional text. Here are some
examples:
Old Church Slavonic: {example}
```

[2] `github.com/ufal/histalign`

[3] https://www.roger-pearse.com/weblog/wp-content/uploads/2015/09/Methodius-De-Lepra-20151.pdf

```
{target_language}: {example}
...
Translate the following text from Old Church Slavonic to
{target_language}:
{example}
Write only the translation without any additional text."
```

For zero-shot prompting, we use only the section following the examples.

LLM Size In the second experiment, we prompt different sized LLMs in a zero-shot setup, to see to which extent the size of the model affects the quality of the output. For consistency and to keep variables at a minimum, in this experiment we prompt exclusively Qwen2.5-Instruct models [28], from 0.5 to 14 billion parameters in size. The prompts were the same as the first experiment.

Evaluation We evaluate the models with three automated metrics: BLEU [24], chrF [26], and COMET [29]. While COMET represents the state-of-the-art for high-resource language pairs [9], its reliability for low-resource historical languages like Church Slavic remains uncertain[8]. Mathur et al. (2020) [20] argue for the retirement of BLEU. We therefore use chrF as our primary reference metric, as it has shown better correlation with human judgments for morphologically rich and under-resourced languages.

5 Results

5.1 Zero and Few-Shot Prompting

We evaluated four LLMs across three prompting configurations (0-shot, 1-shot, and 3-shot) for translating Church Slavic into English and German. Table 1 reports the results for each target language.

Overall Performance EuroLLM-9B-Instruct achieved the best overall performance across both languages with a mean chrF score of 34.4, followed by gpt-oss-120b (32.7) and Qwen2.5-7B-Instruct (32.5). Vikhr-7B-instruct_0.2, despite being specifically adapted for modern Russian, performed significantly worse with a mean chrF score of 20.6 across both languages.

Model size alone does not guarantee better performance: the 9B parameter EuroLLM model outperformed the 120B parameter gpt-oss model by approximately 1.7 chrF points overall (34.4 vs. 32.7), suggesting that architecture and training data composition play important roles in translation quality.

The performance of the relatively small EuroLLM may lay in its training data, which covers a diverse set of European languages, including several Slavic ones. On top of Russian, EuroLLM has seen Polish, Czech, Slovak, and Bulgarian. This diverse mixture may have helped the model to better understand the complexities of Church Slavic.

Impact of Few-Shot Prompting Contrary to expectations, we observed minimal differences between 0-shot, 1-shot, and 3-shot configurations. The 0-shot approach achieved a marginally higher chrF score (30.5) compared to 1-shot (30.0) and 3-shot (30.4). This may suggest that the model's representations are already sufficient to obtain a decent performance, or that the few-shot prompt structure is suboptimal.

Translation Direction The translation direction had an impact on performance. For Church Slavic-to-English translation, models achieved better scores (mean chrF of 32.2 vs. 28.4 for German, -3.8 points). The best overall result was achieved by gpt-oss-120B in 3-shot mode for Church Slavic-to-English translation (chrF 36.3). For German translation, EuroLLM-9B-Instruct 0-shot achieved the best result (chrF 33.5, -2.8 points).

This disparity can be attributed to two main factors. First, LLMs have significantly more exposure to English in their training data, providing better target language generation capabilities. Second, the increased complexity of German morphology and word order compared to English may pose additional challenges when translating from a highly inflected historical language. These results expose that the average best model or setup, may not be the best for each language.

Table 1: Performance of all model-strategy combinations for Church Slavic translation. The best score is in **bold**.

(a) **Church Slavic to English**

Model	Shots	chrF	COMET	BLEU
EuroLLM-9B	0-shot	35.6	0.614	0.065
EuroLLM-9B	1-shot	35.8	0.598	0.077
EuroLLM-9B	3-shot	35.1	0.610	0.069
gpt-oss-120b	0-shot	33.7	0.608	0.061
gpt-oss-120b	1-shot	**36.3**	0.620	0.066
gpt-oss-120b	3-shot	**36.3**	0.617	0.064
Qwen2.5-7B	0-shot	34.1	0.596	0.050
Qwen2.5-7B	1-shot	34.1	0.597	0.046
Qwen2.5-7B	3-shot	35.6	0.617	0.043
Vikhr-7B	0-shot	25.3	0.496	0.015
Vikhr-7B	1-shot	21.4	0.435	0.008
Vikhr-7B	3-shot	22.8	0.449	0.010

(b) **Church Slavic to German**

Model	Shots	chrF	COMET	BLEU
EuroLLM-9B	0-shot	**33.5**	0.578	0.055
EuroLLM-9B	1-shot	33.0	0.564	0.051
EuroLLM-9B	3-shot	32.2	0.555	0.051
gpt-oss-120b	0-shot	31.7	0.560	0.038
gpt-oss-120b	1-shot	32.0	0.568	0.038
gpt-oss-120b	3-shot	32.3	0.569	0.044
Qwen2.5-7B	0-shot	30.8	0.535	0.025
Qwen2.5-7B	1-shot	29.6	0.537	0.025
Qwen2.5-7B	3-shot	30.6	0.523	0.032
Vikhr-7B	0-shot	19.4	0.406	0.004
Vikhr-7B	1-shot	17.9	0.338	0.003
Vikhr-7B	3-shot	18.1	0.351	0.001

5.2 LLM Size

To investigate the relationship between model size and translation quality, we evaluated five Qwen2.5-Instruct models ranging from 0.5B to 14B parameters in a zero-shot setup. Table 2 presents the results for both target languages.

Overall Performance There is positive correlation between model size and translation quality. The largest model, Qwen2.5-14B-Instruct, achieved the best performance with a mean chrF score of 34.5, followed by the 7B (32.5), 3B (28.9), 1.5B (24.5), and 0.5B (19.0) variants. A 15.5 chrF difference between the largest and smallest models demonstrates that parameter count significantly impacts translation performance.

However, the improvement pattern shows diminishing returns: upgrading from 0.5B to 1.5B yields a 5.5 chrF gain, while doubling from 7B to 14B only provides a 3.6 points improvement. This suggests that while larger models consistently perform better, the marginal benefit decreases at higher parameter counts. With the first experiment, we have already seen that size is not the only factor in play, as EuroLLM-9B-Instruct outperforms the bigger Qwen model.

Inference time increases with model size, from 0.70s for the 0.5B model to 1.90s for the 14B model. Diminishing returns become evident when comparing the results of the Qwen models to the previous experiment gpt-oss-120b, which fares comparably only to the 7B Qwen.

Performance by Target Language When looking at the results by target language, we find the same pattern observed in the few-shot experiments: all models performed better when translating to English compared to German. The mean chrF scores were 29.5 for English versus 26.3 for German, a gap of 3.2 points. This gap is consistent across all model sizes, ranging from 1.1 points for the 0.5B model to 3.8 points for the 14B model. The Qwen2.5-14B-Instruct model achieved the best results for both translation directions: chrF of 36.4 for English and 32.7 for German.

Table 2: Performance of Qwen2.5-Instruct models by size for Church Slavic translation (0-shot). The best score is in **bold**.

(a) Church Slavic to English

Model	chrF	COMET	BLEU
Qwen2.5-14B	**36.4**	0.629	0.075
Qwen2.5-7B	34.1	0.596	0.050
Qwen2.5-3B	31.2	0.578	0.037
Qwen2.5-1.5B	26.1	0.550	0.016
Qwen2.5-0.5B	19.5	0.479	0.004

(b) Church Slavic to German

Model	chrF	COMET	BLEU
Qwen2.5-14B	**32.7**	0.566	0.044
Qwen2.5-7B	30.8	0.535	0.025
Qwen2.5-3B	26.7	0.487	0.014
Qwen2.5-1.5B	23.0	0.455	0.008
Qwen2.5-0.5B	18.4	0.387	0.001

6 Conclusions

We investigated the capability of off-the-shelf Large Language Models to translate Church Slavic, a low-resource historical language, into English and German. Through two experiments examining few-shot prompting strategies and model

size effects, we evaluated multiple LLMs on 262 sentence pairs from a 10th century Church Slavic text.

This paper presents several insights: first, contemporary LLMs can produce useful draft translations of Church Slavic without specialized fine-tuning. Second, few-shot prompting provides minimal benefit over zero-shot approaches. Third, model size correlates with translation quality, though with diminishing returns. Finally, performance gaps persist between target languages.

These results suggest that LLMs can serve as assistants for scholars working with Church Slavic texts, potentially accelerating the translation process and improving accessibility to historical manuscripts.

Limitations Our study has some limitations. The evaluation dataset is small and drawn from a single text genre, which may not generalize to other Church Slavic texts and genres. We rely entirely on automated metrics, which may not fully capture translation quality for historical languages. Our model selection is limited to four families.

Future Work Several directions warrant further investigation. Human evaluation by experts would validate our automated metrics and provide qualitative insights into error types and translation usability. Fine-tuning models specifically for Church Slavic translation could substantially improve performance. Exploring inter-variant translation and orthographic normalization tasks would broaden the applicability of LLM-based approaches, especially to address data scarcity, variation across datasets, and noise.

Ethical Considerations The deployment of LLMs for translation tasks raises environmental concerns due to their substantial energy consumption during inference. Researchers should consider the carbon footprint of their computational experiments and seek energy-efficient alternatives when possible. More importantly, LLMs can produce errors, hallucinations, and culturally inappropriate content. They should be viewed as assistants rather than substitutes for human expertise. Scholars must carefully review and validate all LLM-generated translations before use in academic or practical contexts, particularly for historical texts where cultural and linguistic nuances are critical.

Acknowledgments. The authors would like to thank Piroska Lendvai and Maarten Janssen for helpful discussions and for providing the dataset used in this work. The work described herein has been supported by the Ministry of Education, Youth and Sports of the Czech Republic, Project No. LM2023062 LINDAT/CLARIAH-CZ.

References

1. Cassese, M., Puccetti, G., Napolitano, M., Esuli, A.: DIACU: A dataset for the DIAchronic analysis of church Slavonic. In: Piskorski, J., Přibáň, P., Nakov, P., Yangar-

ber, R., Marcinczuk, M. (eds.) Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025). pp. 101–107. Association for Computational Linguistics, Vienna, Austria (Jul 2025). https://doi.org/10.18653/v1/2025.bsnlp-1.12, https://aclanthology.org/2025.bsnlp-1.12/

2. Dereza, O., Doyle, A., Rani, P., Ojha, A.K., Moran, P., McCrae, J.: Findings of the SIGTYP 2024 Shared Task on Word Embedding Evaluation for Ancient and Historical Languages. In: Hahn, M., Sorokin, A., Kumar, R., Shcherbakov, A., Otmakhova, Y., Yang, J., Serikov, O., Rani, P., Ponti, E.M., Muradoğlu, S., Gao, R., Cotterell, R., Vylomova, E. (eds.) Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP. pp. 160–172. Association for Computational Linguistics, St. Julian's, Malta (Mar 2024), https://aclanthology.org/2024.sigtyp-1.19
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423, https://aclanthology.org/N19-1423/
4. Dorkin, A., Sirts, K.: TartuNLP @ SIGTYP 2024 shared task: Adapting XLM-RoBERTa for ancient and historical languages. In: Hahn, M., Sorokin, A., Kumar, R., Shcherbakov, A., Otmakhova, Y., Yang, J., Serikov, O., Rani, P., Ponti, E.M., Muradoğlu, S., Gao, R., Cotterell, R., Vylomova, E. (eds.) Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP. pp. 120–130. Association for Computational Linguistics, St. Julian's, Malta (Mar 2024), https://aclanthology.org/2024.sigtyp-1.15/
5. of Encyclopaedia Britannica, T.E.: Cyrillic alphabet. Encyclopedia Britannica (2018), https://www.britannica.com/topic/Cyrillic-alphabet
6. of Encyclopaedia Britannica, T.E.: Glagolitic alphabet. Encyclopedia Britannica (2018), https://www.britannica.com/topic/Glagolitic-alphabet
7. of Encyclopaedia Britannica, T.E.: Old church slavonic language. Encyclopedia Britannica (2018), https://www.britannica.com/topic/Old-Church-Slavonic-language
8. Falcão, J., Borg, C., Aranberri, N., Abela, K.: COMET for low-resource machine translation evaluation: A case study of English-Maltese and Spanish-Basque. In: Calzolari, N., Kan, M.Y., Hoste, V., Lenci, A., Sakti, S., Xue, N. (eds.) Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp. 3553–3565. ELRA and ICCL, Torino, Italia (May 2024), https://aclanthology.org/2024.lrec-main.315/
9. Freitag, M., Rei, R., Mathur, N., Lo, C.k., Stewart, C., Avramidis, E., Kocmi, T., Foster, G., Lavie, A., Martins, A.F.T.: Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In: Koehn, P., Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M.R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Jimeno Yepes, A., Kocmi, T., Martins, A., Morishita, M., Monz, C., Nagata, M., Nakazawa, T., Negri, M., Névéol, A., Neves, M., Popel, M., Turchi, M., Zampieri, M. (eds.) Proceedings of the Seventh Conference on Machine Translation (WMT). pp. 46–68. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid) (Dec 2022), https://aclanthology.org/2022.wmt-1.2/

10. Halácsy, P., Kornai, A., Nagy, V., Németh, L., Trón, V.: Parallel corpora for medium density languages, pp. 247–258 (01 2005). https://doi.org/10.1075/cilt.292.32var
11. Janssen, M., Lendvai, P., Jouravel, A.: Alignment of historical manuscript transcriptions and translations. In: Proc. of RANLP 2025, September 2025, Varna, Bulgaria (2025)
12. Janssen, M.: TEITOK: Text-faithful annotated corpora. In: Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp. 4037–4043. European Language Resources Association (ELRA), Portorož, Slovenia (May 2016), https://aclanthology.org/L16-1637/
13. Jouravel, A., Sieber, J.: The Greek and Slavonic Transmission of Methodius' De lepra, pp. 11–30. De Gruyter, Berlin, Boston (2024). https://doi.org/doi:10.1515/9783111350790-004, https://doi.org/10.1515/9783111350790-004
14. Kargaran, A.H., Imani, A., Yvon, F., Schuetze, H.: GlotLID: Language identification for low-resource languages. In: The 2023 Conference on Empirical Methods in Natural Language Processing (2023), https://openreview.net/forum?id=dl4e3EBz5j
15. Lendvai, P., van Gompel, M., Jouravel, A., Renje, E., Reichel, U., Rabus, A., Arnold, E.: A workflow for HTR-postprocessing, labeling and classifying diachronic and regional variation in pre-Modern Slavic texts. In: Calzolari, N., Kan, M.Y., Hoste, V., Lenci, A., Sakti, S., Xue, N. (eds.) Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp. 2039–2048. ELRA and ICCL, Torino, Italia (May 2024), https://aclanthology.org/2024.lrec-main.184/
16. Lendvai, P., Reichel, U., Jouravel, A., Rabus, A., Renje, E.: Domain-Adapting BERT for Attributing Manuscript, Century and Region in Pre-Modern Slavic Texts. In: Proceedings of the 4th International Workshop on Computational Approaches to Historical Language Change 2023 (LChange'23) co-located with EMNLP2023, Singapore (2023)
17. Lendvai, P., Reichel, U., Jouravel, A., Rabus, A., Renje, E.: Instruction finetuning to attribute language stage, dialect, and provenance region to historical church slavic texts. In: Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI era. pp. 654–662. INCOMA Ltd., Shoumen, Bulgaria, Varna, Bulgaria (September 2025), https://aclanthology.org/2025.ranlp-1.76
18. Lendvai, P., Reichel, U., Jouravel, A., Rabus, A., Renje, E.: Retrieval of Parallelizable Texts Across Church Slavic Variants. In: Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects. pp. 105–114 (2025)
19. Martins, P.H., Fernandes, P., Alves, J., Guerreiro, N.M., Rei, R., Alves, D.M., Pombal, J., Farajian, A., Faysse, M., Klimaszewski, M., Colombo, P., Haddow, B., de Souza, J.G.C., Birch, A., Martins, A.F.T.: Eurollm: Multilingual language models for europe (2024), https://arxiv.org/abs/2409.16235
20. Mathur, N., Baldwin, T., Cohn, T.: Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 4984–4997. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.448, https://aclanthology.org/2020.acl-main.448/

21. Nikolich, A., Korolev, K., Bratchikov, S., Kiselev, I., Shelmanov, A.: Vikhr: Constructing a State-of-the-art Bilingual Open-Source Instruction-Following Large Language Model for Russian. In: Proceedings of the 4th Workshop on Multilingual Representation Learning (MRL) at EMNLP-2024. Association for Computational Linguistics (2024)
22. Nivre, J., de Marneffe, M.C., Ginter, F., Hajič, J., Manning, C.D., Pyysalo, S., Schuster, S., Tyers, F., Zeman, D.: Universal Dependencies v2: An evergrowing multilingual treebank collection. In: Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Twelfth Language Resources and Evaluation Conference. pp. 4034–4043. European Language Resources Association, Marseille, France (May 2020), https://aclanthology.org/2020.lrec-1.497/
23. OpenAI, :, Agarwal, S., Ahmad, L., Ai, J., Altman, S., Applebaum, A., Arbus, E., Arora, R.K., Bai, Y., Baker, B., Bao, H., Barak, B., Bennett, A., Bertao, T., Brett, N., Brevdo, E., Brockman, G., Bubeck, S., Chang, C., Chen, K., Chen, M., Cheung, E., Clark, A., Cook, D., Dukhan, M., Dvorak, C., Fives, K., Fomenko, V., Garipov, T., Georgiev, K., Glaese, M., Gogineni, T., Goucher, A., Gross, L., Guzman, K.G., Hallman, J., Hehir, J., Heidecke, J., Helyar, A., Hu, H., Huet, R., Huh, J., Jain, S., Johnson, Z., Koch, C., Kofman, I., Kundel, D., Kwon, J., Kyrylov, V., Le, E.Y., Leclerc, G., Lennon, J.P., Lessans, S., Lezcano-Casado, M., Li, Y., Li, Z., Lin, J., Liss, J., Lily, Liu, Liu, J., Lu, K., Lu, C., Martinovic, Z., McCallum, L., McGrath, J., McKinney, S., McLaughlin, A., Mei, S., Mostovoy, S., Mu, T., Myles, G., Neitz, A., Nichol, A., Pachocki, J., Paino, A., Palmie, D., Pantuliano, A., Parascandolo, G., Park, J., Pathak, L., Paz, C., Peran, L., Pimenov, D., Pokrass, M., Proehl, E., Qiu, H., Raila, G., Raso, F., Ren, H., Richardson, K., Robinson, D., Rotsted, B., Salman, H., Sanjeev, S., Schwarzer, M., Sculley, D., Sikchi, H., Simon, K., Singhal, K., Song, Y., Stuckey, D., Sun, Z., Tillet, P., Toizer, S., Tsimpourlas, F., Vyas, N., Wallace, E., Wang, X., Wang, M., Watkins, O., Weil, K., Wendling, A., Whinnery, K., Whitney, C., Wong, H., Yang, L., Yang, Y., Yasunaga, M., Ying, K., Zaremba, W., Zhan, W., Zhang, C., Zhang, B., Zhang, E., Zhao, S.: gpt-oss-120b & gpt-oss-20b Model Card (2025), https://arxiv.org/abs/2508.10925
24. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Isabelle, P., Charniak, E., Lin, D. (eds.) Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA (Jul 2002). https://doi.org/10.3115/1073083.1073135, https://aclanthology.org/P02-1040/
25. Pedrazzini, N., Eckhoff, H.M.: Oldslavnet: A scalable early slavic dependency parser trained on modern language data. Software Impacts **8**, 100063 (2021). https://doi.org/https://doi.org/10.1016/j.simpa.2021.100063, https://www.sciencedirect.com/science/article/pii/S2665963821000117
26. Popović, M.: chrF: character n-gram F-score for automatic MT evaluation. In: Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Hokamp, C., Huck, M., Logacheva, V., Pecina, P. (eds.) Proceedings of the Tenth Workshop on Statistical Machine Translation. pp. 392–395. Association for Computational Linguistics, Lisbon, Portugal (Sep 2015). https://doi.org/10.18653/v1/W15-3049, https://aclanthology.org/W15-3049/
27. Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C.D.: 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 (2020), https://nlp.stanford.edu/pubs/qi2020stanza.pdf
28. Qwen, :, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., Qiu, Z.: Qwen2.5 technical report (2025), https://arxiv.org/abs/2412.15115
29. Rei, R., Stewart, C., Farinha, A.C., Lavie, A.: COMET: A neural framework for MT evaluation. In: Webber, B., Cohn, T., He, Y., Liu, Y. (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 2685–2702. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.emnlp-main.213, https://aclanthology.org/2020.emnlp-main.213/
30. Sennrich, R., Volk, M.: Iterative, MT-based sentence alignment of parallel texts. In: Pedersen, B.S., Nešpore, G., Skadiņa, I. (eds.) Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011). pp. 175–182. Northern European Association for Language Technology (NEALT), Riga, Latvia (May 2011), https://aclanthology.org/W11-4624/
31. Straka, M.: 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. pp. 197–207. Association for Computational Linguistics, Brussels, Belgium (Oct 2018). https://doi.org/10.18653/v1/K18-2020, https://www.aclweb.org/anthology/K18-2020