Instruction Finetuning to Attribute Language Stage, Dialect, and Provenance Region to Historical Church Slavic Texts

Piroska Lendvai
Bavarian Academy of Sciences
Munich, Germany
piroska.lendvai@badw.de

Uwe Reichel
Hungarian Research Centre for Linguistics
Budapest, Hungary
uwe.reichel@nytud.hu

Anna Jouravel and **Achim Rabus** and **Elena Renje**
Department of Slavic Languages and Literatures
University of Freiburg, Germany
anna.jouravel,achim.rabus,elena.renje@slavistik.uni-freiburg.de

Abstract

Our study addresses domain-specific text provenance classification for the historical *Church Slavic* language. The downstream task is to attribute the language stage and its dialectal and regional varieties to texts compiled from newly curated sources, including digitally unpublished manuscripts, in addition to established Church Slavic resources from the Universal Dependencies Treebank. We aim to harmonize previously used tag sets pertaining to textual provenance, and construct a new, hierarchical, multi-layer provenance labeling scheme. For the classification task, we finetune Vikhr (Nikolich et al., 2024), a generative LLM with knowledge of modern Russian, with the instruction to generate labels to classify the provenance of sentence-level text units. Besides gold standard manuscript transcriptions, we test the finetuned model on character-corrupted data that emulate the quality of noisy, handwritten text recognition material. The experiments show that the Vikhr base model has low provenance attribution knowledge of Church Slavic, whereas our finetuned model achieves above .9 F-scores on Language stage labeling and Dialect labeling, and above .8 F-score on generating the label that jointly classifies all three provenance layers. The task of classifying the fine-grained geographical region from which a manuscript originates proves harder (but still performs above .8), and is negatively impacted by character level noise injection.

1 Introduction

In recent years, transformer-based large language models (LLMs) have achieved notable success in a variety of natural language processing (NLP) tasks. However, their application to historical and low-resource languages remains severely limited. For Church Slavic — a historical liturgical language used across Slavic territories from the 9th century onward — the development of reliable NLP tools is hindered by the scarcity of digitized, annotated training data expressing its complex diachronic and regional variation. Moreover, human expertise in historical Slavic linguistics is scarce, making the compilation of high-quality annotations and evaluation benchmarks especially challenging.

A recent study on the Dead Sea Scrolls (Popović et al., 2025) vividly demonstrates how curated data combined with AI methods can lead to paradigm-shifting results in the interpretation of historical texts. Earlier work in NLP on chronological attribution of texts with deep learning methods for historical languages include Assael et al. (2019); Liebeskind and Liebeskind (2020).

Two major historical Slavic variants are increasingly visible in the NLP landscape: Old Church Slavic (ISO 639-3 code: `chu`) and Old East Slavic (`orv`). These variants are represented in resources such as the Universal Dependencies (UD) Treebank (Nivre et al., 2020) and integrated into toolkits such as Stanza (Qi et al., 2020) or UDPipe (Straka, 2018). Despite these initial efforts, substantial barriers remain. LLM tokenizers and embedding models trained on contemporary languages fail to handle Church Slavic adequately, e.g. they produce large quantities of character-level tokens, or oddly split sentences. As shown in the recent SIGTyp shared task (Dereza et al., 2024), successful systems needed custom tokenizers and embeddings.

Church Slavic was artificially developed during the christianization of the Slavs, primarily for translating Byzantine Greek religious texts. For the history of the Slavic written heritage see Marti (1989); Birnbaum and Schaeken (1999); Trunte (2014). It retained strong influence from its Greek source texts in terms of both syntactic structures and vocabulary. Church Slavic underwent con-

Proceedings of Recent Advances in Natural Language Processing,pages 654–662
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-076

tinuous evolution in lexicon and morphosyntax. This occurred due to both the natural divergence of Slavic dialects into distinct modern languages as well as manuscript copying—introducing both intentional editorial revisions and unintentional scribal errors. The result is a large body of text material transmitted through manuscripts that exhibit extensive orthographic, lexical, and structural variation.

Challenges and our Contributions We address three core challenges in this domain and describe our respective contributions.

(1) First, reliable supervised classification requires labeled ground truth data for the *initial* geographical provenance and temporal context of texts, but such metadata is often missing or is not consistent. For example, in the Universal Dependencies repository, there is the PROIEL dataset[1] for `chu` and the TOROT dataset[2] for `orv`. However, the texts labeled by the language code `orv` in fact cover seven language variants: (i) the *vernacular* texts are written in *Old East Slavic* (traditionally called *Old Russian*) as well as in *Middle Russian*, *Old Novgorodian* and *Ruthenian*; (ii) the *religious* texts pertain to different transmitted language variants (so-called *recensions*) in *East Slavic* and in *Russian*, but also encompass canonical *Old Church Slavic* (cf. the text listed in the Syntacticus treebank[3]).

Furthermore, currently there is a metadata interoperability issue: the labels for language or dialect names that pertain to Church Slavic are ambiguous both within and across communities of Philology, Digital Humanities, and NLP. E.g., a specific Church Slavic language stage can be referenced by diverse terms which are often interchangeably used in the literature; cf. e.g. Keipert (2014). There is also granularity inconsistency in such metadata, where macro- and micro levels of temporal or regional designations or specifications are occasionally mixed; e.g. *Serbian Church Slavic* vs. *Church Slavic, Southern recension* vs. *Church Slavic, Serbian recension* may label one and the same text type. As a result, it is not transparent how the different (named or unnamed) constraints split these groups of related historical languages, and whether Church Slavic temporal pairs or variants or dialects have been or can be fully represented in a consolidated metadata scheme.

To overcome such ambiguity, we establish and uniformly apply three hierarchical levels of metadata for the provenance labeling task at hand: Language stage, Dialect, and Region (cf. Figure 1), and map the available manuscript provenance information for the collections at hand to these axes. Such metadata clarity is of additional importance in generative techniques, so that label attribution techniques can keep apart the category names. This labeling method serves as an operational classification framework for the development and evaluation of automated attribution techniques, without challenging (established) Slavicist research terminologies or existing scholarly positions on dialectal boundaries and periodizations.

(2) Second, reuse of textual material without explicit annotation across manuscripts, including quotations from older sources or from other recensions, is a common phenomenon in historical texts. This practice makes it difficult to track text reuse, and to determine whether it applies at the paragraph, sentence, or sub-sentential level, which remains an open problem. NLP-based provenance attribution of Church Slavic texts has so far been scarcely covered, a.o. addressed in our previous studies (Lendvai et al., 2023, 2025). In the current study, our approach to provenance classification is to recast it as a label generation task, via performing systematic instruction finetuning of *Vikhr* (Nikolich et al., 2024), an open-source, bilingual, instruction-tuned, dense, decoder-only Large Language Model (LLM). The model we used had been built upon the *LlamaForCausalLM* architecture that underwent continued pretraining on modern Russian and English instruction data, enhancing the model's ability to follow instructions effectively. Its *SentencePiece* tokenizer also underwent adaptations to better support the modern Russian language. In our experiments, after our instruction finetuning (IFT) process, we used the IFTed model for our downstream task of provenance attribution: the labeling of 11 variants of Church Slavic on the sentence level.

(3) Third, morphological complexity and orthographic variability of Church Slavic, e.g., due to multiple legitimate grapheme variants for a single phoneme or morpheme, introduces a high level of language data sparsity for our applied end task. To-

[1] https://github.com/UniversalDependencies/UD_Old_Church_Slavonic-PROIEL

[2] https://github.com/UniversalDependencies/UD_Old_East_Slavic-TOROT

[3] https://github.com/syntacticus/syntacticus-treebank-data/tree/main/torot

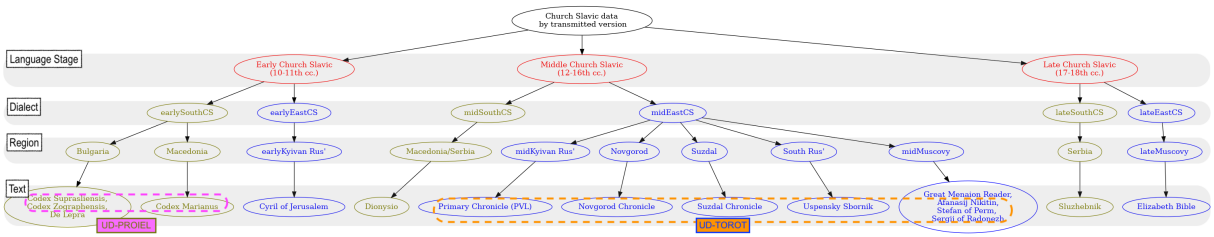

Figure 1: Our newly constructed, hierarchical text labeling scheme. Its three layers represent Language stage (Early, Middle, Late), Dialect (South, East), and Region of manuscript provenance (Bulgaria, Macedonia, Kyivan Rus', Macedonia/Serbia, Novgorod, Suzdal, South Rus', Muscovy, Serbia). 11 leaf nodes hold the names of manuscripts where our data originate from. Dashed areas delimit data from two benchmark collections of the Universal Dependencies (UD) Treebank: UD-PROIEL (pink) and UD-TOROT (orange).

kens often appear in many low-frequency forms that may be philologically correct but computationally problematic for data-driven methods. Attempts to normalize such variation are fraught with trade-offs: normalization may obscure important orthographic signals that are valuable for identifying linguistic trends or regional features. In addition, handwritten text recognition (HTR), the method used to digitize large volumes of source material, introduces further character- or token-level variation due to recognition errors. To account for noise in the data, we artificially emulated HTR errors to test our classification approach on data resembling HTR-induced noise, inspired by a character-level data corruption technique described in Aepli and Sennrich (2022).

2 Data and Labeling

Our data come from both online and so far unreleased historical Church Slavic texts. We mapped existing provenance metadata of our texts to three layers of a newly created, hierarchical provenance scheme, cf. Figure 1):

1. Language Stage (Early, Middle, Late)
2. Dialect, defined by the spreading of Church Slavic, resulting in local recensions (South, East)
3. Region, i.e. country or culturally distinct geographical space (Bulgaria, Macedonia, etc.).

We note that the availability of provenance information is such that these metadata typically designate values tied to the place where the text/collection had been *copied*, i.e., they may not disclose provenance beyond that, potentially obscuring cases when e.g. an older text was verbatim copied over from a source that was created at a different region and likely at another language stage.

The data are compiled from 16 different text collections or manuscripts (see the leaf nodes of the labeling scheme), which are as follows:

- Early Church Slavic language stage
  - The *earlySouthCS* texts come from the PROIEL UD Treebank: Codex Suprasliensis and Codex Zographensis (region of origin: Bulgaria), Codex Marianus (region: Macedonia), except for *De Lepra* (Jouravel et al., 2024) (region: Bulgaria).
  - The *earlyEastCS* Cyril of Jerusalem[4] comes from our own corpus (region of origin: Early Kyivan Rus').
- Middle Church Slavic language stage
  - The *midSouthCS* Dionysio corpus comes from the Digital Archive of the University of Kragujevac[5] (region of origin: Macedonia/Serbia)
  - The *midEastCS* texts come from the TOROT UD Treebank and originate from the Middle-stage Kyivan Rus' (Primary Chronicle), from Novgorod (Novgorod Chronicle) from Suzdal (Suzdal Chronicle), from the South of Rus' (Uspensky Sbornik), from Muscovy (Journey of Afanasij Nikitin, resp. Life of Stefan of Perm, resp. Life of Sergij of Radonezh), and from our in-house corpus of the *Great Menaion Reader (GMR)*[6], where we sampled data from its

[4] www.weiher-verlag.de/publikationen/tom-lxiv.html

[5] https://scidar.kg.ac.rs/handle/123456789/17759

[6] https://en.wikipedia.org/wiki/Great_Menaion_Reader

Language Stage	Collection (Dialect) / Text	Snippets	Mean snippet length			Word count		
			train	dev	test	train	dev	test
Early CS	UD-PROIEL (South)	17,639	9.6	9.3	10.8	135,587	15,752	17,118
	Lepra (South)	261	15.3	14.7	11.8	3,140	369	308
	Cyril (East)	4,284	14.7	13.9	13.6	50,368	5,940	5,847
Middle CS	Dionysio (South)	10,000	14.0	14.0	13.8	111,646	13,956	13,820
	UD-TOROT (East)	18,377	9.5	9.2	9.2	137,778	15,429	19,312
	GMR March (East)	10,000	6.8	6.9	6.9	54,099	5,971	6,889
Late CS	Sluzhebnik (South)	3,352	16.8	17.8	17.0	45,087	5,971	5,699
	Elizabeth Bible (East)	10,000	18.2	17.8	18.1	145,534	17,737	18,138
Total		73,913				683,239	82,021	87,131

Table 1: Input data statistics: sentence snippets by text, as well as mean snippet length (in tokens) and total tokens by partition for the instruction finetuning task (using the train and dev sets) as well as for the provenance attribution task (using the test set).

Uspensky version, March section texts (region: Muscovy).

- Late Church Slavic language stage
  - The *lateSouthCS* texts comprise the Sluzhebnik[7] (region of origin: Serbia)
  - The *lateEastCS* texts come from the Elizabeth Bible[8] (region: Muscovy).

Text segmentation into sentences was done the following way. For PROIEL and TOROT data, we took over the sentence boundaries that were present in CONLL-U format. On the remaining texts we uniformly used the Old Church Slavic sentence model of Stanza (Qi et al., 2020). We refer to the resulting units as *text snippets*, and use this term uniformly throughout the paper for all datasets. The segmentation of texts for downstream analysis is particularly challenging due to the use of *scriptura continua*, a writing style without whitespace between words that characterizes many of the manuscripts in our dataset. Additionally, the lack of systematic punctuation makes establishing ground truth sentence boundaries in Church Slavic disputable.

For data partitioning, we took over the given splits from the UD treebanks for PROIEL and TOROT data. For all other sources we randomly selected 10,000 sentences, and split them into train/development/test sets by the ratio 80/10/10.

For detailed data statistics see Table 1, where word counts and mean snippet lengths are given per partition and dataset. For snippet length distributions in terms of probability density curves, comparable per partition in each data subset, see the Appendix[9].

3 Method

Our labeling of texts is regarded as Ground Truth (GT) provenance classification data for the current study. Since the text transcripts had been manually created, these are also of GT quality.

3.1 Instruction Finetuning (IFT)

For the downstream tasks of joint language, dialect, and region labeling of text snippets, we adapted the LLM *VikhrmodelsVikhr-7B-instruct_0.2* (Nikolich et al., 2024) by means of instruction finetuning (IFT), cf. Ouyang et al. (2022). For IFT we defined the system prompt as follows:

> "You are a historical linguist who can differentiate three stages of Church Slavic: Early, Middle, Late, and their respective regional dialects. You can reproduce the type of orthographic, grammatical, and lexical variation that is characteristic for specific cultural-geographical areas for all these variations of Church Slavic.".

We maintained the instruction–output pair format during training, for which we constructed the user prompt for each snippet from the following text template:

> "You will see a text. Identify its Church Slavic language stage, regional dialect, and the historical geographic-cultural area where the text was written. Attribute one of the following labels to

[7] https://lib-fond.ru/lib-rgb/256/f-256-401

[8] https://en.wikipedia.org/wiki/Elizabeth_Bible

[9] https://github.com/pirolen/ranlp2025-churchslavic-ift/#sentence-length-distributions

identify the language and regional origin of the text: *LABELSET*. This is the sentence to be annotated: *TEXT*"

TEXT is to be replaced with the respective text snippet to be classified. *LABELSET* is to be replaced by the list of 11 labels comprising all available language-dialect-region combinations and serves to constrain the model to behave like a classifier with a closed answer vocabulary:

1. Language stage (3 classes: Early, Middle, Late)
2. Dialect (2 classes: South, East)
3. Region (9 classes: Bulgaria, Macedonia, Kyivan Rus', Macedonia/Serbia, Novgorod, Suzdal, South Rus', Muscovy, Serbia)
4. Language stage and Dialect jointly (6 classes: Early South, Early East, etc.)
5. Language stage, Dialect, and Region jointly (11 classes: Early South Bulgaria, Early South Macedonia, etc.).

The *Vikhr-7B-instruct_0.2* tokenizer was initialized with maximum length 1024 and left-padding, using the *bos* token as *pad* token.

IFT was realised by means of parameter efficient fine tuning (PEFT) based on low-rank adaptation (LoRA). LoRA was applied with task type *Causal LM* on the following modules: *q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj* and *lm_head*. The rank was set to 32, alpha to 64, and the dropout to .05. For IFT we used paged AdamW 8 bit optimization with an initial learning rate of $2.5e-5$. Cross entropy loss was used as the metric to be minimized. Batch size was set to 8, the number of finetuning steps to 10,000.

3.2 Evaluation

We evaluated model performance in terms of the F1-score metric on the gold label compared to the generated label. The labels were generated as follows. After IFT, we selected the checkpoint that performed best on the development set in terms of minimum loss. Using cross entropy loss for this purpose was motivated by training a generative model that can output a (theoretically) unconstrained stream of natural language from its multilingual token vocabulary, especially in initial training cycles, while the label set that should classify Church Slavic sentence snippets comprise a finite set of English words. This means that not all types of mismatches between gold and generated labels could be meaningfully expressed in terms of F1 during instruction finetuning.

On both the clean and the corrupted test set dataset, we applied both the base LLM *Vikhr-base*[10] as well as the finetuned LLM *Vikhr-IFTed*; meaning that – in separate experiments — the vanilla Vikhr resp. the selected checkpoint of *Vikhr-IFTed* was used for generating the labels for the held out test set, using the same system and user prompt templates as for IFT (cf. Section 3.1). Generation was configured with greedy decoding and a maximum number of 100 new tokens.

Testing on noise injected data Aepli and Sennrich (2022) introduced a character-level data corruption technique that helped LLM tokenizers to be more robust towards spelling variations, such as character substitutions, insertions, and deletions, inspired by previous work on surface-level noise injection that was found to improve e.g. machine translation. Note that whereas Aepli and Sennrich (2022) set corruption values between 10%-15%, Blaschke et al. (2023) systematically explored a range of noise injection rates for a large amount of language pairs, including dialects, and found that increasing the noise level from 0% to 15% improves embedding-based models' robustness on part-of-speech tagging.

Although for a different end task, we adopted this method to emulate HTR quality data. In the test phase, we injected character-level noise for the same snippets that were used as GT test data. It was not an option to use real HTR data, since reproducing the exact same snippet segments on actual HTR data would have required significant manual work. Corruption ratio was set to affect 15% of a snippet's characters. Contrary to the above studies where random characters were added as noise, we introduced specific types of errors that we previously observed HTR systems to produce.

Corruption operations were the following:

1. Delete a correct whitespace
2. Insert a false whitespace
3. Substitute a character with an random, but phonetically equivalent grapheme, based on an internally compiled grapheme correspondence table of Church Slavic variants.

[10] https://huggingface.co/Vikhrmodels/Vikhr-7B-instruct_0.2

4 Results

4.1 Quantitative Analysis

Table 2 shows the Unweighted Average F1 scores for the base model and the finetuned model on the clean and the noisy test set variants for the three individual classification tasks: *Language Stage*, *Dialect*, *Region*; and task combinations: *Language Stage* + *Dialect* jointly and *Language Stage* + *Dialect* + *Region* jointly. For language, dialect and region evaluation we compared the corresponding parts of the reference and answer labels, respectively.

We report Precision, Recall, and F-scores for each task separately in the Appendix[11]. Confusion matrices for the finetuned model on the ground truth ('clean') test set and on the noisy set are also shown in the Appendix[12].

Classification is highest (.98) for the tasks of Language Stage and Dialect, as well as their combination (.97). On the Region task, the lower F-score (.82) originates in confusing Early Church Slavic texts from Macedonia with Bulgaria, resp. Middle Church Slavic texts from Kyivan Rus' with Muscovy.[13] Note that that Language stage property within the Region task is only discernable from the confusion matrix of the full label generation[14]. This plot shows that the largest group of misclassifications (in total 287 data points) is observed for Middle Church Slavic, Eastern dialect, Suzdal region, from where only a small amount of ground truth data was available.

These confusion trends get more prominent on the noisy data, also impacting the fully combined label that incorporates all three annotation layers (.846 F on clean data vs. .659 on noisy data). The difference between these scores may indicate that character level noise injection is not beneficial for historical language variant identification for our specific languages at hand. In Church Slavic historical texts, character variation is by definition high; these variations typically encode characteristic writing styles or morphological features that can be important cues in classification. It remains to be manually verified to what extent texts from these confusion regions exhibit cross-fertilization, i.e. whether the causes of these confusions are due to the practice of text reuse and borrowing.

Task	Vikhr–base		Vikhr–IFTed	
	clean	noisy	clean	noisy
Language Stage (LS)	.210	.230	.985	.958
Dialect (D)	.380	.380	.972	.908
Region (R)	.065	.074	.828	.644
LS + D	.042	.051	.949	.862
LS + D + R	.026	.030	.846	.659
# Fail	0	0	1	6

Table 2: Unweighted average F1 scores by the **Vikhr-base** and **Vikhr-instruction-finetuned (IFTed)** models on the clean and noisy (character-corrupted) test sets for the classification tasks per labeling components: *Language Stage*, *Dialect*, *Region*, *Language Stage* + *Dialect* jointly, and *Language Stage* + *Dialect* + *Region* jointly. *# Fail* gives the absolute number of times the models fail to answer.

In the test set, the mean length of misclassified snippets is 10 words and in correctly classified snippets 12 words. Both averages are in line with the distribution of UD treebank data values, cf. Figure 1 as well as the Appendix[15]; yet, shorter snippets are likely more difficult to correctly classify.

4.2 Philological Interpretation

Even though the snippets often exhibit characteristic elements that indicate temporal-regional origin, *Vikhr–IFTed* was not always able to adequately classify these according to the gold standard labels. Remarkably, frequent errors involve contexts where transmission history blurs linguistic borders, pertaining to canonically shared texts and formulaic passages. Such borderline cases often expose scribal convergence zones or genuinely ambiguous segments.

Frequent confusions occur between Early Church Slavic, Southern dialect, Macedonia region and vs. its Bulgaria region, in both directions. Since all the Early Church Slavic, Southern dialect data in this study (except for the later treatise *De lepra*) belong to the Old (i.e., Early) Church Slavic canon, where canon membership is defined by linguistic features, the texts naturally exhibit similar characteristics. Furthermore, *Codex Zographensis* and *Codex Marianus* are essen-

[11] https://github.com/pirolen/ranlp2025-churchslavic-ift#precision-recall-and-f-scores

[12] https://github.com/pirolen/ranlp2025-churchslavic-ift#confusion-matrices

[13] https://github.com/pirolen/ranlp2025-churchslavic-ift/#region

[14] https://github.com/pirolen/ranlp2025-churchslavic-ift/#joint-language-stage-dialect-and-region

[15] https://github.com/pirolen/ranlp2025-churchslavic-ift/#sentence-length-distributions

tially versions of the same text, i.e. (incomplete) Tetraevangelia. Aside from easily confusable Bible-based sequences such as Пилатъ же посѫди бъıти прошениѥ ихъ, these manuscripts also contain many formulaic expressions or collocations that may appear in any text, including much later ones, such as азъ ѥсмъ, ı г҃лаахѫ or рече же и дроугъı.

In some cases, snippets are likely difficult to analyze, e.g. because they are too short to contain forms that can be reliably assigned to a particular region or period: e.g., characteristic forms are entirely absent, making the snippet generally compatible with any of the labels. For example, the snippet что бо имоуть ꙁаꙁь contains a word segmentation error, likely caused by a false line break or other segmentation issue in the data: the word ꙁаꙁь is incomplete and appears to be only the first part of the lexeme зазьрѣти or зазьдати. Snippets that are too short or lack diagnostic dialectal features include и рече имъ or писано бо ѥстъ, both gold labeled as Early Church Slavic, Southern dialect, Macedonia region and misclassified as Middle Church Slavic, Eastern dialect, Kyivan Rus' region.

In other cases, misclassifications could be comprehensible. For instance, in the snippet дъждь бо на роуно сьходѧ (gold label: Early Church Slavic, Eastern dialect, Kyivan Rus' region) was falsely attributed to Early Church Slavic, Southern dialect, Bulgaria region. Here, the lexeme дъждь (Uspenskij, 2002), could be confusing, tokenized as _д|ъ|ж|дь, where the reduced vowel -ъ- is not realized as -о- (see above). This notation is typically attested in manuscripts of South Slavic provenance including Early Church Slavic (cf. Kurz (1958) [SJS]), but also appears in Middle Church Slavic, Eastern dialect manuscripts such as the *Uspenskij Sbornik* or the *Novgorod chronicle*.

An example of an East Slavic text misclassified as South Slavic is the snippet етероу оухъ мнѣ ꙗко погыбе доброговѣниѥ оть ꙁемлѧ и оуправѧштааго въ чл҃вцѣхь нѣсть (gold label: Early Church Slavic, Eastern dialect, Kyivan Rus' region, i.e., attested in the *Cyril of Jerusalem* dataset), which was confused with Early Church Slavic, Southern dialect, Bulgaria region. Here, the participle оуправѧштааго has a typical South Slavic adjectival ending –аго (East Slavic: –ого, cf. Uspenskij (2002), pp. 207–208) and the *l-epentheticum* is omitted (cf. Diels (1932), p. 131; Chaburgaev (1974), p. 104; Krivko (2016), pp. 132-137), which points to the text's South Slavic and especially Bulgarian region provenance. Furthermore, the lexeme доброговѣниѥ is attested only in the Bulgarian-origin *Codex Suprasliensis*, besides its occurrence in the *Cyril of Jerusalem* dataset. We note that the model's tokenizer splits up the morphological components that could have served as cues, obscuring information that could be key to provenance assignment: _о|у|прав|ѧ|шта|а|го resp. _добро|гов|ѣ|ни|ѥ, whereby the iotized graphemes ѧ and ѥ are additionally split into two.

The snippet Сѐ а҆́зъ даю̀ пред҃ ва́ми дне́сь бл҃гве́нїе и҆кла́твꙋ [...] (gold label: Late Church Slavic, Eastern dialect, Muscovy region), which holds characteristic morphological features such as the first person singular personal ending –ю (as in даю; заповѣдаю; молю) that support East Slavic origin (vs. South Slavic: даѭ, заповѣдаѭ, молѭ, cf. Trunte (2005), p. 113), however, it was misclassified as Late Church Slavic, Southern dialect, Serbia region. Likewise, the snippet кро́вь бо є҆гѡ̀ вмѣ́стѡ дꙋшѝ оу҆мо́литъ (gold label: Late Church Slavic, Eastern dialect, Muscovy region), in which phonological features such as the East Slavic fully vocalized form of reduced vowels in combination with liquids are present (as in кровь, South Slavic: кръв, cf. Uspenskij (2002), pp. 150–155), was labeled as Early Church Slavic, Southern dialect, Bulgaria region.

Conversely, there are cases in which South Slavic texts were incorrectly classified as East Slavic. An example is вѣдѣаше бо ѣко зависти ради прѣдашѧ ı арх҆иереı, (gold label: Early Church Slavic, Southern dialect, Bulgaria region), confused with Middle Church Slavic, Eastern dialect, Muscovy region. The text features the notation ѣко, attested only in Early Church Slavic manuscripts. The snippet и идѫ въı инѫ весь (gold label: Early Church Slavic, Southern dialect, Macedonia region) was misclassified as Middle Church Slavic, Eastern dialect, Novgorod region, despite the typically South Slavic first person singular ending –ѫ in идѫ (East Slavic: идоу, see the converse case above). The cue words are tokenized as ѣ|ко and _и|дѫ.

5 Conclusions

Our goal was to implement a scalable, data-driven investigation of textual provenance in medieval and early modern Slavic texts, performing systematic instruction finetuning of Vikhr (Nikolich et al., 2024), an open-source, bilingual LLM. We cre-

ated annotated data for Church Slavic provenance classification, and showed a methodology for historical text processing in a non-Latin script that can be applied to other low-resource and under-researched languages. Our study demonstrates that instruction finetuning can become a valuable tool for historical linguistics and digital philology, since on the downstream task of provenance attribution the IFTed Vikhr model attained high scores. More research is required to explore the outcome that character level noise injection proved harmful on the task of classifying the fine-grained geographical region from where a specific text originates. The hierarchical annotation scheme we constructed aims to support domain-specific metadata interoperability, and demonstrates the need for upgrading existing Church Slavic benchmarks. In future work we aim to create new datasets, where besides the Church Slavic macrolanguage fine-grained language variety labels, covering dialect, are included.

6 Limitations

On all data, we uniformly used a sentence splitter that was optimized for Old Church Slavic, and did not adapt the Vihkr model's *SentencePiece* tokenizer, meaning that its representations are typically not on the lexical but on the character(-sequence)-level. We are not reporting scores by vanilla LLMs, even though we tested a few of them, since their results are very low, similar to the base Vikhr model. Since we complemented the available UD benchmark data with a (massively downsampled) in-house dataset, the biggest class for the present experiments covers Middle Church Slavic in Eastern dialects. Discrepancies in the size and distribution of the training data across geographical and diachronic labels may play a role in misclassifications, potentially biasing the classification performance of the present model despite distinctive provenance markers that might be present in a text.

7 Ethics Statement

The authors fully acknowledge the ACL Ethics Policy and strongly commit to using their skills for the benefit of society, its members and the environment surrounding them.

8 Acknowledgments

The QuantiSlav project is funded from the EU's Recovery and Resilience Facility and by the German Federal Ministry of Research, Technology and Space in accordance with the guidelines for funding projects to strengthen the data skills of young scientists (Grant number: 16DKWN123B).

References

Noëmi Aepli and Rico Sennrich. 2022. Improving Zero-shot Cross-lingual Transfer between Closely Related Languages by Injecting Character-level Noise. In *60th Annual Meeting of the Association for Computational Linguistics*, pages 4074–4083. Association for Computational Linguistics.

Yannis Assael, Thea Sommerschield, and Jonathan Prag. 2019. Restoring ancient text using deep learning: a case study on Greek epigraphy. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6368–6375, Hong Kong, China. Association for Computational Linguistics.

Henrik Birnbaum and Joseph Schaeken. 1999. *Die altkirchenslavische Schriftkultur: Geschichte - Laute und Schriftzeichen - Sprachdenkmäler: Altkirchenslavische Studien 2*, volume 382 of *Slavistische Beiträge*. Verlag Otto Sagner, München.

Verena Blaschke, Hinrich Schütze, and Barbara Plank. 2023. Does Manipulating Tokenization Aid Cross-Lingual Transfer? A Study on POS Tagging for Non-Standardized Languages. In *The Tenth Workshop on NLP for Similar Languages, Varieties and Dialects*.

Georgij Aleksandrovič Chaburgaev. 1974. *Старославянский язык*. Просвещение, Москва.

Oksana Dereza, Adrian Doyle, Priya Rani, Atul Kr. Ojha, Pádraic Moran, and John McCrae. 2024. Findings of the SIGTYP 2024 Shared Task on Word Embedding Evaluation for Ancient and Historical Languages. In *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 160–172, St. Julian's, Malta. Association for Computational Linguistics.

Paul Diels. 1932. *Altkirchenslavische Grammatik*, volume 6,1 of *Sammlung slavischer Lehr- und Handbücher, I. Reihe, Grammatiken*. Winter, Heidelberg.

Anna Jouravel, Janina Sieber, and Katharina Bracht, editors. 2024. *Methodius von Olympus: De lepra*. De Gruyter, Berlin, Boston.

Helmut Keipert. 2014. Kirchenslavisch-Begriffe / Conceptions of Church Slavonic. In Sebastian Kempgen, Peter Kosta, Tilman Berger, and Karl Gutschmidt,

editors, *Die slavischen Sprachen*, Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science, pages 1211–1252. Mouton de Gruyter, Berlin and New York.

Roman Nikolaevič Krivko. 2016. Орфография рукописи как свидетель текстологической преемственности. *Труды института русского языка им. В.В. Виноградова*, 9:124–148.

Joseph Kurz, editor. 1958. *Slovník jazyka staroslovenského: Lexikon linguae palaeoslovenicae*. Nakladatelství Českovlovenské Akademie věd, Praha.

Piroska Lendvai, Uwe Reichel, Anna Jouravel, Achim Rabus, and Elena Renje. 2023. 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*.

Piroska Lendvai, Uwe Reichel, Anna Jouravel, Achim Rabus, and Elena Renje. 2025. Retrieval of Parallelizable Texts Across Church Slavic Variants. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 105–114.

Chaya Liebeskind and Shmuel Liebeskind. 2020. Deep learning for period classification of historical Hebrew texts. *Journal of Data Mining & Digital Humanities*, 2020.

Roland Marti. 1989. *Handschrift, Text, Textgruppe, Literatur: Untersuchungen zur inneren Gliederung der frühen Literatur aus dem ostslavischen Sprachbereich in den Handschriften des 11. bis 14. Jahrhunderts*, volume 68 of *Veröffentlichungen der Abteilung für slavische Sprachen und Literaturen des Osteuropa-Instituts [Slavisches Seminar] an der Freien Universität Berlin*. Harrassowitz, Wiesbaden.

Aleksandr Nikolich, Konstantin Korolev, Sergei Bratchikov, Igor Kiselev, and Artem Shelmanov. 2024. 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.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Mladen Popović, Maruf A. Dhali, Lambert Schomaker, Johannes van der Plicht, Kaare Lund Rasmussen, Jacopo La Nasa, Ilaria Degano, Maria Perla Colombini, and Eibert Tigchelaar. 2025. Dating ancient manuscripts using radiocarbon and AI-based writing style analysis. *PLOS ONE*, 20(6):1–14.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Nicolina Trunte. 2005. *Slavenskij jazyk: Ein praktisches Lehrbuch des Kirchenslavischen in 30 Lektionen. Bd. 1, Altkirchenslavisch*. Sagner.

Nicolina Trunte. 2014. *Slavenskij jazyk: Ein praktisches Lehrbuch des Kirchenslavischen in 30 Lektionen; zugleich eine Einführung in die slavische Philologie, Bd. 2, Mittel- und Neukirchenslavisch*, 2 edition, volume 494 of *Slavistische Beiträge*. Sagner, München.

B. A. Uspenskij. 2002. *История русского литературного языка (XI –XVII вв.)*, 3 edition. Москва: Аспект Пресс.