Retrieval of Parallelizable Texts Across Church Slavic Variants

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

The goal of our study is to identify parallelizable texts for Church Slavic, across chronological and regional variants. Next to using a benchmark text, we utilize a recently digitized, large text collection and compile new resources for the retrieval of similar texts: a ground truth dataset holding a small amount of manually aligned sentences in Old Church Slavic and in Old East Slavic, and a large unaligned dataset that has a subset of ground truth (GT) quality texts but contains noise from handwritten text recognition (HTR) for the majority of the collection. We discuss preprocessing challenges in the data and the impact of sentence segmentation on retrieval performance. We evaluate sentence snippets mapped across these two diachronic variants of Church Slavic, expressed by mean reciprocal rank, using embedding representations from large language models (LLMs) as well as classical string similarity based approaches combined with k-nearest neighbor (kNN) search. Experimental results indicate that in the current setup (short text snippets, off-the-shelf multilingual embeddings), classical string similarity based retrieval can still outperform embedding based retrieval.

1 Introduction

Despite recent successes of large language modeling and transformer-based representation of texts, for historical languages and dialectal varieties these techniques suffer from the lack of training data, leaving their text representation capabilities and generative functionalities weak. Furthermore, this field suffers from human insight due to the scarcity of historical linguists, making it challenging to compile benchmark resources and evaluate experimental results. Our work seeks to automatize and scale the mapping of parallel texts for diachronic variants of Old and Premodern Church Slavic. Since systematic standardization or normalization of Church Slavic has never taken place, we are confronted with the typical challenges associated with non-standard text variation in historical natural language processing (NLP).

Old Church Slavic got established in the 9th century C.E. during the christianization of Slavic language territories in Europe, primarily to translate from Byzantine (Koine) Greek into a language of the local people, and functioned as a liturgical written language with strong resemblance to Greek constructions as well as theological and philosophical terminology, sounding artificial to the Slavic ear. Despite conservative efforts and archaizing endeavours that regarded the texts as sacrosanct and thus unalterable, Church Slavic underwent considerable modification throughout its history: both spontaneous and dedicated adaptations occurred in morphosyntax and lexicon, as Slavic dialectal vernaculars themselves have evolved into separate languages. In addition to the changes resulting from the gradual divergence of dialects, a number of unintentional modifications occurred during the copying process, as well as a number of intentional redactions. These factors contributed to the emergence of a significant number of textual variants and manuscript copies.

1.1 NLP for Church Slavic

Two variants of Church Slavic are increasingly present in the NLP landscape: Old Church Slavic (ISO 639-3 language code: chu) and Old East Slavic (language code: orv), a.o. via the Universal Dependency Treebank and its tooling[1] and Stanza

[1] https://github.com/ufal/udpipe

Proceedings of the Twelfth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 105–114
January 19, 2025.

resources[2]. More recent work, primarily on language identification, reported about their incorporation in text classification models and downstream tasks (Kargaran et al., 2023) and a recent shared task focusing on evaluation of embeddings learned from historical language data included Church Slavic as well (Dereza et al., 2024). It is indicative that the authors of one of the systems submitted for the shared task, Dorkin and Sirts (2024), note that custom tokenizers as well as custom embeddings need to be created for these languages as off the shelf tokenizers do not cover chu and orv and output a large amount of unrecognized symbols.

Resources related to large language models (LLMs) such as benchmark data, tasks, or trained models for both of these Church Slavic variants are scarce. Typical benchmark tasks, e.g. for evaluating embeddings – cf. e.g. Muennighoff et al. (2022) –, are not applicable to the historical languages of our focus, for example since our type of data feature specific genres of religious texts and thus do not enable creating or translating texts for typical benchmark tasks for contemporary languages, such as product reviews, social media messages, image captions, etc.

Neither has it been systematically explored which generative capacities of LLMs may be relevant for this field, but we note that the shared task of Dereza et al. (2024) includes masked word and masked character prediction. Retrieval augmented generation and embedding-based similarity are powerful for modern languages, but likely less so for diachronic linguistic research purposes, since the primary goal of diachronic studies is to reveal orthographic and grammatical variation patterns and mechanisms in the data, and not to enable access to document content via semantic question answering as for historical and cultural studies.

Moreover, temporal and geographical variation within chu and orv are under-explored; our previous work includes a study using BERT (Devlin et al., 2018) to classify temporal-spatial dimensions of Church Slavic data on the sentence level, utilizing document level annotation as ground truth labeling of manuscript copying time and language region (Lendvai et al., 2023).

1.2 Our Goals and Contributions

In the current study we use the retrieval paradigm in order to identify parallelizable Church Slavic texts and to collect insights across two temporal-dialectal varieties, chu and orv. We create new datasets that can serve in future work as training resources both for machines and for Slavicists who can view and examine variation. Effectively, this could be considered a cross-lingual retrieval setting, as the textual variants exhibit significant differences due to temporal and regional distance: chu represents the original text tradition from the 10th-11th centuries in South Slavic regions, while orv represents later copies from the 15th-17th centuries, influenced by vernacular elements characteristic of East Slavic regions.

We use a set of classical string representation (character n-grams, TF-IDF) and similarity computation approaches (sequence matching, local alignment, and kNN similarity search). We contrast these with neural methods of string representation (text embedding vectors, BERT pooling and SBERT), and retrieve and rank candidates based on cosine vector similarity with kNN. We discuss the potentials and implications of our findings in the NLP parallel text compilation context.

1.3 Related Work

Measuring semantic textual similarity (STS), and more recently conditional STS, has been the topic of vast amounts of previous work, cf. e.g. Deshpande et al. (2023) and their references. Likewise, the construction of aligner systems and comparable corpora, such as those used in machine translation, has been a focus of research since several decades, cf. e.g. Zweigenbaum et al. (2017). Recent advancements in this area, including applications under sparse data conditions, have been explored b cf. e.g. Lin et al. (2024) and others. Dense text retrieval, particularly leveraging pretrained large language models (LLMs), is an emerging field of research. For a comprehensive survey, see cf. Zhao et al. (2022).

General purpose sentence representation learning has been extensively studied and is supported by a large body of literature, e.g. Artetxe and Schwenk (2018); Reimers and Gurevych (2019). Adaptation of LLMs to historical languages has been tackled by several works, cf. e.g. Dereza et al. (2023) and their references. Note that orthographic normalization, as utilized by the latter study, is not a feasible approach for us, since certain patterns of non-normalized orthography encode important temporal or geolocational attributes across diachronic language variants that can help retrieving paral-

[2] https://github.com/stanfordnlp/stanza

lelizable texts. Our work is rooted in a narrow, applied use case, focusing on the exploration of approaches that can be utilized for data filtering in order to boost resource compilation for historical variants of Church Slavic.

2 Data Preparation and Characteristics

We identified a (relatively) sizeable text, versions of which are present both in a chu manuscript and in an orv manuscript: the *Vita of Paul and Juliana*[3]. The goal of our initial experiments was to use the sentences of the older text version as queries and the newer text version as answers to be found, which we scaled up afterwards. To create a benchmark dataset, manual alignment was done first on the word level and subsequently on the (sub)sentential level. Neither steps were straightforward.

Identifying a text that occurs in several manuscripts is so far a manual process – until a robust retriever has been developed for Church Slavic –, since manuscripts typically do not have associated metadata on the individual text level, and in the digitized collection are often segmented only on word and manuscript page level, so it is not visible where texts or sentences start and end.

First and foremost, one needs to be able to read and understand the historical languages to a certain extent, and such experts are rarely available, e.g. to decide if the corresponding words are in a one-to-one or one-to-many/many-to-one relationship to each other across the two texts. Typically, we have seen one-to-one correspondences, but the two focus text variants are not completely parallel, thus there are phrases or sentences or entire passages that have no equivalents.

The text sources are in different initial formats; some in-house texts are plain text with linebreaks using hyphenation (inserted by human editors or HTR tools earlier), some are scattered across several consecutive page-based files, yet others are in CONLL-U format. We converted the texts to FoLiA format using the tooling from that ecosystem, cf. Lendvai et al. (2024), and reconstructed words split across manuscript pages using scripts.

2.1 Codex Suprasliensis

The *Vita of Paul and Juliana* is the first text in the collection *Codex Suprasliensis*, which is one of the oldest attestations of Church Slavic from the 10th century. The *Codex Suprasliensis* itself is part of the Universal Dependencies (UD) Treebank[4], encompassing 9,854 sentences compiled from 48 texts by different authors, serving to be a liturgical reader for the month of March. The manuscript's geographical origin in the strict sense is still disputed, it is likely from the South Slavic area, its language of its texts is said to be closest to the Old East Bulgarian literary language. Since the Suprasliensis contains translations of various origins, linguistic properties exhibited by the texts are heterogeneous and additionally chronologically ambiguous[5]. We had access to the Suprasliensis in ground truth (GT) quality, although we note that its character base is slightly different from online versions (cf. Figure 1).

2.2 Great Menaion Reader

Importantly, some texts that are part of the Suprasliensis, a.o. the *Vita of Paul and Juliana*, can also be found in a compilation of Church Slavic texts from ca. 500 years later (16th c.), originating from a different geographic-cultural area (Muscovy, East Slavic area): the Great Menaion Reader[6] (GMR). While the Suprasliensis only contains texts designated for readings for the month of March, the GMR is a collection of volumes for each month of the year, each consisting of a patchwork of translated and copied versions of biblical, hagiographic, ecclesiastic texts of Church Slavic. Of the three surviving copies of the GMR, the *Uspensky* copy preserved the monthly volume of March and is available to us in digital form. Consequently, we use the *Uspensky* version of the *Vita of Paul and Juliana*, from Weiher et al. (1997-2001), *sub mar. 4, fols. 33c 1 – 41b 19*, to explore parallels with its counterpart in the Suprasliensis manuscript. Note that the GMR text is much longer, as it holds a part that was lost from Suprasliensis, which we excluded from alignment.

The GMR March volume was prepared by us both in ground truth (GT) quality, based on Weiher et al. (1997-2001), as well as in raw HTR (handwritten text recognition) quality; for details about the latter cf. Rabus (2019); Rabus et al. (2023); Lend-

[3] https://en.wikipedia.org/wiki/Paul_and_Juliana

[4] https://torottreebank.github.io

[5] cf. https://textualheritage.org/bl/el-manusctipt-2012/codex-suprasliensis-full-text-electronic-corpus.html

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

vai et al. (2024). In the raw HTR data, noise includes character misrecognitions as well as falsely split or joined words.

2.3 Word Level Alignment

We manually aligned the text variants of the *Vita of Paul and Juliana* on the word token level. In fact, we aligned two different versions of each of the two text variants of the text, which enables pointing out similarities and differences of resources. The four column alignment is illustrated by Figure 1.

1. Of the chu *Vita of Paul and Juliana* text variant from the Suprasliensis, the character set is slightly different across resources, thus we aligned
   (a) in-house version based on http://suprasliensis.obdurodon.org
   (b) UD Treebank version[7].
2. Besides, of the orv *Vita of Paul and Juliana* text variant from the GMR we aligned
   (a) GT of the GMR text (Weiher et al., 1997-2001)
   (b) in-house raw HTR output of the GMR text (Rabus et al., 2023).

Altogether, the length of the word level aligned dataframe is 2,538. Word overlap between the in-house chu texts (1,169 words) and orv texts (1,256 words) is low (85 words), indicating that these two language variants differ substantially, most typically orthographically but often also lexically.

2.4 Breathmark Based Subsentential Snippet Segmentation

Next, we needed to create the same sentence boundaries across each of the text versions chu, orv, and orv-htr. This was not a trivial exercise, given that Church Slavic manuscripts do not use interpunction in the modern sense, neither whitespace between the words. We made an empirically based decision for the current study regarding sentence segmentation, since sentence boundaries in existing treebank data resp. created by such tools are not clearly defined, as we had earlier found (Jouravel et al., 2024). We note that some available sentence splitters create very long segments; these would clearly be suboptimal as input to string based similarity approaches. We tested simple chunking, e.g. creating snippets of word 6-grams and 10-grams (without overlap windowing), but these semantically random units did not prove to be robustly matchable in pilot experiments, neither convenient for human evaluation, nor well motivated by our core benchmark creation goal.

Therefore, snippet level segmentation was done using the following heuristics: (1) end-of-sentence full stop characters were manually inserted in the word aligned file for the UD Treebank column (column A in Figure 1, whenever the token was the last one of a sentence in the treebank data. After some noise cleanup, this yielded a sentence boundary count of 243.[8] (2) Subsequently, we observed the location of (presumed) breathmarks in the in-house Suprasliensis text (column B). These marks were coded as bullet point characters or as full stop char-

	A	B	C	D
1	Suprasliensis: UD treebank	Suprasliensis: in house	GMR: in house	GMR: in house HTR
146	о́тъвѣштавъши	о҆тъвѣштавъши	ѿвѣщавши	ѿвѣщапши
147	же	же	же	же
148				е
149	и́оулияни	и҆оулиꙗни	иоулїанїи	оулїанїи
150	рече.	рече.	рече.	рече.
151	не	не	не	не
152	о́тъвръгѫ	о҆тъвръгѫ	ѿвер'гоуса	ѿвергоуса
153	сѩ	сѧ		
154	аурилияне	аурилиа́не	аврилїане	аврилїа
155				не
156	томителю	томителю	моучителю	моучителю
157	и́	и҆		
158	непрѣподобьне	непрѣподобьне	неподѡб'не	не
159	.	.	.	подѡбне.
160	не	не	не	не
161	прѣльстиши	прѣльстиши	прельстиши	прельстиши
162	рабы	рабꙑ	рабы	рабы
163	бога	бога	бг҃а	бг҃а
164	вышьняаго.	вꙑшьнꙗаго.	вышнаго.	вышнаго.
165	не	не	не	не
166	примышляи́	примꙑшлꙗи́	помышлѧи	помышлѧи
167	ми	ми	ми	ми
168	съмрьти	съмрьти	смр҃ти	смр҃ти
169	вѣчьныꙗ	вѣчьнꙑꙗ	вѣч'ныѧ.	вѣчныѧ.
170	лишити	лишити	лишити	лишити
171	мѩ	мѧ	мѧ	мѧ
172	хотѩ	хотѧ	хотѧ	хотѧ
173	славы	славꙑ	славы	славы
174	хс҃овы	хс҃овꙑ	хвы҃	х҃вы
175	и	и	и	и
176	цѣсарьства	цѣсарьства	цар҃тва	цартва
177	небесьнааго	небесьнааго	нб҃наго.	нб҃наго.

Figure 1: Word level alignment: for each of the chu and orv variants of the *Vita of Paul and Juliana* text, we aligned two different versions: the chu versions from the UD Treebank resp. obdurodon.org, which show character encoding discrepancies (e.g. of superscript characters), as well as the in-house orv versions in ground truth (GT) vs. text recognition (HTR) quality. Note that across the chu and orv variants, the absence and presence of (presumed) breathmarks (rendered as full stops) differs. Breathmarks were used for snippet segmentation when they occurred in either the Suprasliensis (column A) or the GMR GT (column C).

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

[8] We tried to obtain information about the segmentation guidelines, see https://github.com/UniversalDependencies/UD_Old_Church_Slavonic-PROIEL/issues/3.

acters, or more rarely, as commas or colons (398 periods, 17 commas). About half of these boundaries overlapped with the UD Treebank sentence boundaries. (3) We observed the location of (presumed) breathmarks in the in-house GMR version of the ground truth *Vita of Paul and Juliana* text (column C). (4) We sliced each of the three token aligned texts in columns B, C, and D at the same positions, whenever there was a full stop character seen either at step (1) or (3). Note that this boundary setting method often (or typically) does not yield syntactically or semantically complete sentences but rather subsentential text snippets, which are typically coherent but short and out-of-context, which might be suboptimal input to LLMs, especially to sentence-based LLMs. After segmentation all punctuation marks were removed from the texts.

2.5 Snippet Level Ground Truth Alignment

Small Benchmark Dataset This slicing procedure created 409 snippets. The mean snippet lengths were uniformly 5 tokens across each of the three text versions. The mean edit distance between chu and orv snippets was 40. From this set, we removed snippets that were shorter than 3 words, in order to focus on creating parallel data with sizeable sentence snippets. The mean snippet lengths changed uniformly to 6 tokens across each of the three text versions (see Table 1).

Our resulting ground truth dataset consisted of 359 snippets, where chu, orv, and orv-htr are parallelized (i.e., columns B, C, and D). For examples see Figure 2.5. We provided English translations[9] for each snippet to additionally illustrate their semantics and syntactical complexity.

Large Benchmark Dataset We created breathmark based snippets from the entire large orv resource (GMR for the month March), both for the hand corrected quality (GT) and the uncorrected HTR version. These feature some orders of magnitude more data but similar snippet lengths as the small dataset. Table 1 shows a basic description of the resulting data sizes.

Note that in a recent shared task dataset based a.o. on the UD Treebank Dereza et al. (2024), reported mean sentence lengths are 9 words for chu and 10 words for orv[10]; the authors note that "sentences from historical texts are often much shorter than in modern language due to their genre or purpose.

[9] based on http://suprasliensis.obdurodon.org/01_paragraphed.html

[10] https://github.com/sigtyp/ST2024

Dataset	Lang ISO	Quality	Snippets	Words	mean W/S
Small	chu	GT	359	2,120	5.9
	orv	GT	359	2,037	5.7
	orv	HTR	359	2,090	5.8
Large	orv	GT	57,803	340,925	5.9
	orv	HTR	55,041	350,910	6.4

Table 1: Breathmark based snippet segmentation statistics for our chu and orv datasets.

3 Experimental Setup

For the task of identifying parallelizable snippets, we took the list of snippets from the chu Church Slavic language variant of our benchmark text as search queries. For each query, its aligned orv Old East Slavic language variant was regarded as the ground truth (or benchmark) reference answer in the retrieval process. We submitted each snippet from chu as a query to several retrieval procedures (or systems) that processed one of the orv datasets at a time, and we evaluated the top k retrieved orv snippets the systems returned as most similar matches, setting $k = 1$ as well as $k = 3$.

3.1 Evaluation

Below we list the evaluation metrics that were used to score retrieved snippets, as well as the five systems we tested for retrieval. For the task at hand, it is not straightforward to establish a baseline, since retrieval combines both similarity scoring as well as candidate ranking, and our results show that simple approaches currently outperform sophisticated ones.

3.1.1 Mean Reciprocal Rank

Top k snippets were evaluated using Mean Reciprocal Rank[11] (MRR). MRR is used for expressing retrieval quality in scenarios where there is a single relevant result to a query. Over all queries for a task for a system, MRR counts if the GT answer was present or not in the set of k most similar snippets that a system returned. According to our matrix of experiments, we measured MRR @1 and MRR @3, so the closer the corresponding MRR score is to 1, the more often the correct parallel snippet was returned as the highest ranked (top 1) answer, resp. was returned in the set of the top 3 highest ranked answers, over all queries.

[11] https://en.wikipedia.org/wiki/Mean_reciprocal_rank

	Suprasliensis: in house (*chu*)	GMR: in house (*orv*)	GMR: in house HTR (*orv*)	Translation
21	ѻтъвѣштавъши же іоулиꙗни рече	ѿвѣщавши же иоулїанїи рече	ѿвѣщапши же є оулїанїи рече	answering [him] Juliana said
22	не ѻтъврьгѫ сѧ аурилиꙗне томителю и непрѣподобьне	не ѿверꙵгоуса авриліане моучителю неподѡбꙵне	не ѿвергоуса авриліа не моучителю не подѡбне	I won't renounce, Aurelian, tormentor and impious one
23	не прѣльстиши рабъі бога въішьнꙗаго	не прельстиши рабы бг҃а вышнаго	не прельстиши рабы бг҃а вышнаго	you will not trick the servant of the most high God
24	не примъішлꙗи ми съмрьти вѣчьнъіꙗ	не помышлаи ми смр҃ти вѣчꙵныꙗ	не помышлаи ми смр҃ти вѣчныꙗ	do not plan eternal death for me
25	лишити мѧ хотѧ славъі хс҃овъі и цѣсарьства небесънааго	лишити мꙗ хотꙗ славы хвы҃ и цар҃тва нб҃наго	лишити мꙗ хотꙗ славы х҃вы и цартва нб҃наго	trying to deprive me of the glory of God and of the kingdom of Heaven
26	ѥгоже тъі штоуждъ ѥси	єгоже ты҄ тоуж҃ь еси	єгоже ты҄ тоуж҃ь єси	which you are alien to

Figure 2: Alignment of sentence snippets for the languages chu and orv, the latter in ground truth (GT) and HTR (handwritten text recognition) quality: six consecutive snippet pairs from our new dataset, created from the text *Vita of Paul and Juliana* present in the manuscripts *Codex Suprasliensis* and *Great Menaion Reader (GMR)*. The English translation is for illustrative purposes and was not part of the experiments.

3.1.2 Evaluative Similarity Score: Local Alignment

As a cumulative metric on the character level, we also expressed the mean similarity of all pairs of *query string – candidate string retrieved at rank 1* in terms of local alignment (Localign). We defined the Localign similarity as the proportion of characters in the query text that has been matched with the retrieved text by the following method.

Local alignment was carried out based on an adaption of the Smith-Waterman algorithm (Smith and Waterman, 1981). The chosen score function rewards zero substitutions by $+2$, punishes non-zero substitutions by -1 and insertions and deletions by -2, respectively. The minimum required length for aligned subsequences is set to 1 character, and cross alignment is prohibited. For details see Lendvai and Reichel (2016). In order to account for orthographic variation, we established single character equivalence classes in a joint table for both chu and orv, e.g. the numerous spelling variants of the '*i*' character, of the '*ya*' character, and so forth. We relaxed the zero substitution criterion not only to cover exact character matches but any match of characters within the same orthographic equivalence class.

3.1.3 Evaluation Quality: Gold, Silver, Bronze

Small Benchmark Dataset Besides evaluating retrieval between the small GT aligned data of chu and orv (rows 1 and 2 in Table 1), which we regard as having gold evaluation quality, we also assessed retrieval from noisy HTR data (row 3). This evaluation is however suboptimal – therefore we regards its representativeness as silver quality –, a.o. since the degree of HTR noise and the actual noisy strings may not be reproducible. e.g. if they originate from a different HTR engine across query and reference set.

Large Benchmark Dataset Next, we scaled up the orv data (rows 4 and 5 in Table 1) and assessed how this impacts retrieval quality. These data hold texts for the entire month of March, in both GT and uncorrected HTR quality. MRR scores for these experiments likely express tentative trends, therefore we regard these as having silver resp. bronze evaluation quality.

There are duplicate snippets in the data (e.g. 'and he said'), both due to the repetitive way of storytelling in the specific text genres at hand and the way how the snippets were segmented. During evaluation, in case a duplicate snippet was retrieved (i.e., its positional index was not the expected GT index), this was counted as if the snippet with the correct index value would have been matched.

3.2 Systems for Similarity Scoring and Ranking

Below we list the systems used for parallel snippet retrieval. Each of them perform similarity scoring and ranking between the chu queries and one of the orv reference datasets. We used two Python packages that implement classical approaches for representing string similarity, and three systems that utilize embedding vectors from LLMs for text representation. They transformed each snippet in the query resp. reference data into a fixed-length vector. For vector dimensions see column *Text encoding (dim)* in Table 2 resp. Table 3.

3.2.1 TF-IDF on Character 3-grams and kNN Search

A Python package[12] was used for n-gram-based string matching: splitting the orv reference corpus into character 3-grams and transforming it into a sparse matrix of features computed based on importance, i.e. on term frequency - inverse document frequency[13] (TF-IDF). An unsupervised nearest neighbor search model was fitted on this matrix[14], using

[12] https://github.com/LouisTsiattalou/tfidf_matcher

[13] https://scikit-learn.org/1.5/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

[14] https://scikit-learn.org/stable/modules/neighbors.html

cosine as distance metric between the *k*-matches nearest neighbors for the chu queries; queries got vectorized in terms of the TF-IDF sparse matrix features constructed from the orv reference corpus.

3.2.2 Character 3-gram Based Approximate Matching

The second system also used a Python library[15] and implemented character 3-gram based approximate matching. This system divided each snippet in the reference collection into character 3-grams and computed similarity based on common 3-grams, combined with an inverted index, mapping character 3-grams to the strings that contain them. For each query snippet, it retrieved a subset of snippets in the corpus based on shared n-grams, and used *SequenceMatcher* to calculate string similarity *ratio* only for the selected candidates, avoiding costly pairwise comparisons for unlikely pairs.

3.2.3 GlotLID Embeddings with PCA and kNN Search

The third system used *GlotLID*[16], a *FastText* language identification model that supports a large amount of languages, including chu and orv (Kargaran et al., 2023). Importantly, *FastText* allows to build vectors for nonstandard spellings since word vectors are built from character substring vectors[17]. *GlotLID* is a character n-gram embedding based model; we used version 3 to generate embeddings from our data. Next, we applied principal component analysis[18] (PCA) to reduce the dimensionality of the embeddings and found it to improve performance in general, so only scores with PCA incorporated are reported. The number of kept principle components was chosen to explain 95% of the reference data embedding variance. Cosine similarity and kNN search was used to retrieve and rank candidates.

3.2.4 mBERT Embeddings with PCA and kNN Search

The fourth system also expresses text similarity in terms of vector similarities, but of pretrained multilingual BERT embeddings (Devlin et al., 2018); we used *bert-base-multilingual-uncased* that had been pretrained on the largest 100+ Wikipedia languages. The vector representations of reference and query texts were derived by mean pooling of the final hidden layer output of the encoder over all tokens in a snippet, selected by the attention mask. We expected mean pooling (opposed to e.g. CLS pooling) to be more robust against the type of the processed text units – in our case snippets rather than sentences. Subsequently, we calculated the cosine similarity between the query and reference text embeddings. We fitted a PCA model on the reference data the same way as for the *GlotLID* based system explained in Section 3.2.3, and used kNN search.

3.2.5 SBERT with T5-based Dual Retriever Model

For the fifth system we evaluated several models from the SBERT framework (Reimers and Gurevych, 2019) applied in a zero-shot way, using the default cosine similarity. SBERT provides a large amount of sentence transformers models. For our task, the XL version of the pretrained community model – *gtr-t5-xl*[19] – outperformed others, thus we only report the scores for this specific model. It is a large-scale dual encoder retrieval model introduced by (Ni et al., 2021), initialized from the pretrained T5 model family that uses mean pooling gained from the encoder part of the T5 architecture.

4 Results and Discussion

For the tasks of identifying parallelizable candidates for our small set of queries, results are listed in Table 2 for the tasks using the *small* benchmark data and in Table 3 for the tasks using the large GMR orv data. The best MRR score was achieved by character 3-gram based approximate matching (2nd row, *difflib* system). The results indicate that systems using character n-gram based methods worked well for the tasks at hand. This is not very surprising, since the chu and orv text variants have strong character-level correspondences, being snapshots of a language taken at different times and locations.

The tested LLM-based systems and embedding representations seem not to be able to supersede classical string similarity based methods. This is likely due to chu and orv not being languages covered by the out of the box models we used, except for *GlotLID*. Similar to the finding of (Dorkin and

[15] https://docs.python.org/3/library/difflib.html#difflib.SequenceMatcher.ratio

[16] https://huggingface.co/cis-lmu/glotlid

[17] https://fasttext.cc/docs/en/faqs.html

[18] https://scikit-learn.org/dev/modules/generated/sklearn.decomposition.PCA.html

[19] https://huggingface.co/sentence-transformers/gtr-t5-xl

Similarity scoring & ranking	Text encoding (dim)	Eval quality	MRR @1	MRR @3	Localign @1 mean
kNN, Cosine	char3grams, tf-idf (3.4k) (3.3k)	GT (gold) HTR (silver)	.70 .66	.76 .74	.83 .80
Approx. seq. match (difflib)	char3grams, (inverted index)	GT (gold) HTR (silver)	**.87** .86	**.90** .89	**.93** .92
kNN, Cosine + PCA	GlotLID (256)	GT (gold) HTR (silver)	.10 .11	.14 .15	.40 .43
kNN, Cosine + PCA	mBERT (768)	GT (gold) HTR (silver)	.18 .18	.22 .21	.47 .45
SBERT, Cosine	gtr-t5-xl (768)	GT (gold) HTR (silver)	.58 .48	.62 .55	.73 .67

Table 2: Results from five systems for parallel snippet retrieval using the **small** datasets. Evaluation both in gold quality (on aligned GT pairs) and in silver quality (on gold-aligned pairs of noisy orv HTR data): each featuring 359 chu-orv query-answer snippet pairs.

Similarity scoring & ranking	Text encoding (dim)	Eval quality	MRR @1	MRR @3	Localign @1 mean
kNN, Cosine	char3grams, tf-idf (25k) (22.6k)	GT (silver) HTR (bronze)	.21 .19	.26 .23	.55 .53
Approx. seq. match (difflib)	char3grams (inverted index)	GT (silver) HTR (bronze)	**.58** .51	**.61** .54	**.83** .80
kNN, Cosine + PCA	GlotLID (256)	GT (silver) HTR (bronze)	.03 .02	.03 .02	.36 .36
kNN, Cosine + PCA	mBERT (768)	GT (silver) HTR (bronze)	.05 .03	.07 .04	.42 .40
SBERT, Cosine	gtr-t5-xl (768)	GT (silver) HTR (bronze)	.23 .14	.27 .18	.57 .51

Table 3: Results from five systems for parallel snippet retrieval using the **large** reference datasets. Evaluation both in silver quality, using gold-aligned pairs from the small dataset as reference, i.e. 359 chu query snippets used to retrieve answers from ca. 58k orv snippets, as well as in bronze quality: 359 chu query snippets used to retrieve answers from ca. 55k HTR orv snippets, for which we have HTR alignment in the small dataset as reference.

Sirts, 2024), the tokenizers typically yielded a vast amount of unknown tokens as well as character unigram or bigram tokens on our data, which could be detrimental for LLM based representation.

The snippets aligned in our benchmark datasets typically exhibit full semantic overlap by definition; however, due to historical semantic change as well as text modifications, they also regularly differ on the level of the lexicon or morphosyntax (e.g. when a prepositional phrase got modified into a construction involving a verbal prefix). It is left for future research to find ways to adapt LLMs to these specific languages and tasks. In qualitative evaluation, we noticed nevertheless that the LLM based systems tended to retrieve semantically closer matches than string based methods, yielding a more interesting pool of examples for humanist research on language change. We also note that filtering out short snippets (as described in Section 2.5) helped the systems improve their performance. HTR data quality had an expected lowering on the scores, which was slight for the small data and more impactful on the large data.

5 Conclusion

Our work is strongly anchored in the benchmark data compilation scenario: the goal was to devise ways to identify parallelizable text snippets from one historical variant to another across temporal and regional-cultural variants of Church Slavic, a low resource historical language. We recast this goal in a document retrieval setup and organized the data to allow for a two-step procedure: (1) snippet representation by classical as well as neural text representation techniques: n-gram vectors vs. embedding vectors, and (2) the retrieval and ranking of most similar snippets, as expressed by string distance metrics or by nearest neighbor vector distances.

We created and utilized a new data source for Church Slavic historical language variants: a large subset of the GMR corpus; we explored retrieval of similar snippets both from GT tokens and HTR versions of this subset, based on a new, manually aligned benchmark set of chu and orv subsentential snippets.

Our investigation provided insights into textual similarity and its representation for two diachronic, thus closely related, variants of the Church Slavic language. Experimental results indicate that on our Church Slavic data, the performance of tested LLMs is superseded by classical approaches, presumably since only customized tokenizers and embedding models would be able to create meaningful representations for these language variants; and perhaps partly because salient information for this particular language pair that are diachronic variants of each other is tied to the surface level and is less effectively expressed by composite sentence representation. This line of research should be given a focused effort in future work.

In the current setup, string-based classical methods combined with kNN search worked best, however, this method might not generalize to other data, or to other languages. Presumably, the current low LLM performance will in the future benefit from the emergence of large parallel resources involving historical Slavic languages, which is the goal we are working towards.

6 Limitations

Our evaluation scenario for the low resource language of Church Slavic was realistic, i.e. we had a large dataset from which to mine parallel sentences, and little ground truth to evaluate on, thus results, especially on the small aligned benchmark, might not be robust. The queries were created from a single text, and aligned resources were created by versions of this text by a single person manually. The resources are currently under revision, including the preparation of alignment guidelines, they can be released to the community with a delay.

Sentence segmentation was done on the basis of (presumed) breathmarks, which might be suboptimal for embeddings. Neither the LLMs nor their tokenizers were finetuned on the focus languages, which entails that character-level and UNK tokens were abunded and semantic information could not be utilized to full potential. Application of existing tools and previous approaches from the literature, including overlap-enabled text chunking or aligner systems, were beyond the scope of the current study.

7 Ethics Statement

The authors fully acknowledge the ACL Ethics Policy and strongly commit to circumventing bias and supporting respectful scientific debate, and 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 Federal Ministry of Education and Research in accordance with the guidelines for funding projects to strengthen the data skills of young scientists (Grant number: 16DKWN123B).

References

Mikel Artetxe and Holger Schwenk. 2018. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *CoRR*, abs/1812.10464.

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.

Oksana Dereza, Theodorus Fransen, and John P. Mccrae. 2023. Temporal domain adaptation for historical Irish. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 55–66, Dubrovnik, Croatia. Association for Computational Linguistics.

Ameet Deshpande, Carlos E. Jimenez, Howard Chen, Vishvak Murahari, Victoria Graf, Tanmay Rajpurohit, Ashwin Kalyan, Danqi Chen, and Karthik Narasimhan. 2023. C-STS: Conditional semantic textual similarity. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, page 5669–5690.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Aleksei Dorkin and Kairit Sirts. 2024. TartuNLP @ SIGTYP 2024 shared task: Adapting XLM-RoBERTa for ancient and historical languages. In *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 120–130, St. Julian's, Malta. Association for Computational Linguistics.

Anna Jouravel, Elena Renje, Piroska Lendvai, and Achim Rabus. 2024. Assessing Automatic Sentence Segmentation in Medieval Slavic Texts. In *Proc. of the Digital Humanities 2024 Conference, 6-9 August, 2024, Washington, DC, USA*.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schütze. 2023. GlotLID: Language identification for low-resource languages. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Piroska Lendvai and Uwe Reichel. 2016. Contradiction detection for rumorous claims. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics (ExProM)*, pages 31–40, Osaka, Japan. The COLING 2016 Organizing Committee.

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, Maarten van Gompel, Anna Jouravel, Elena Renje, Uwe Reichel, Achim Rabus, and Eckhart Arnold. 2024. A workflow for HTR-postprocessing, labeling and classifying diachronic and regional variation in pre-Modern Slavic texts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2039–2048, Torino, Italia. ELRA and ICCL.

Peiqin Lin, André Martins, and Hinrich Schütze. 2024. A recipe of parallel corpora exploitation for multilingual large language models. *Preprint*, arXiv:10.48550/arXiv.2407.00436.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. MTEB: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y. Zhao, Yi Luan, Keith B. Hall, Ming-Wei Chang, and Yinfei Yang. 2021. Large dual encoders are generalizable retrievers. *Preprint*, arXiv:2112.07899.

Achim Rabus. 2019. Recognizing Handwritten Text in Slavic Manuscripts: a Neural-Network Approach Using Transkribus. *Scripta & e-Scripta, The Journal of Interdisciplinary Mediaeval Studies*, 19:9–32.

Achim Rabus, Walker Riggs Thompson, and Daniel Stökl Ben Ezra. 2023. Generic HTR model for Old Cyrillic uncial and semi-uncial script styles (11th-16th c.). DOI 10.5281/zenodo.7755483.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Temple F. Smith and Michael S. Waterman. 1981. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147:195–197.

E. Weiher, S.O. Shmidt, and A.I. Shkurko. 1997-2001. *Die grossen Lesemenäen des Metropoliten Makarij: Uspenskij spisok. Bd. 1, 2, 3. [The Great Menaion Reader of Metropolitan Makary. Uspensky Version. Volumes 1, 2, 3.]*. Weiher.

Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji rong Wen. 2022. Dense text retrieval based on pretrained language models: A survey. *ACM Transactions on Information Systems*, 42:1 – 60.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2017. Overview of the second bucc shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 60–67.