Towards Benchmarking Old Church Slavonic Lemmatization

Usman Nawaz[1] **Marianna Napolitano**[2] **Iris Karafillidis**[3]
Liliana Lo Presti[1] **Marco La Cascia**[1]

[1]Department of Engineering, University of Palermo, Palermo, Italy
[2]University of Modena and Reggio Emilia / FSCIRE, Italy
[3]University of Padua, Padua, Italy
mariannanapolitano@unimore.it

Abstract

Lemmatization is an important preprocessing step in Natural Language Processing (NLP); however, annotated resources for medieval languages such as Old Church Slavonic (OCS) are limited in scope, size, and diversity. This paper presents the annotated resources for OCS lemmatization, including annotation process, design choices and non-standard Unicode related issues. The annotated corpus is used to evaluate existing lemmatization tools (Stanza and UDPipe-2 models trained on the UD 2.12 treebank, and a dictionary-based approach) both in cross-dataset and on a corpus obtained by merging the new annotations with existing UD V2.12 OCS data. Pretrained models perform poorly ($\approx$ 15–16%), below a dictionary baseline ($\approx$ 38%), while retraining on the new data improves performance (up to $\approx$ 51%) and shows different cross-dataset generalization. Experiments in cross-dataset and on the combined corpus demonstrate that lemmatization performance depends strongly on dataset similarity, annotation conventions, and orthographic mismatch. Overall, the findings show the value of the newly annotated resources and the importance of extending OCS lemmatization benchmarks for historical Slavic NLP.

1 Introduction

OCS holds a unique place in Slavic linguistic history, diachrony, and manuscript variation (Picchio, 1972; Garzaniti, 2019; Naumow, 1983; Zhivov, 2009) and plays an important role in the study of early Slavic written culture and the evolution of the Cyrillic and Glagolitic scripts (Picchio, 1972; Garzaniti, 2019; Hetényi and Ivanič, 2021).

In recent times, OCS texts have become more prominent in computational linguistics with the development of treebanks, corpora, and other applications of NLP (Eckhoff and Berdicevskis, 2015). Although several annotated OCS resources are available, benchmark-quality lemmatization datasets remain limited in size, fragmented across sources and editorial traditions, and difficult to transfer across annotation schemes. Moreover, because OCS is a language associated with the early Slavic literary tradition from the 9th century, expert annotation is substantially harder to scale than for modern languages, where larger annotator pools and standardized resources are more readily available.

Lemmatization transforms words into their base or canonical form, which helps reduce data sparsity and improve downstream tasks, including machine translation (Goldwater and McClosky, 2005; Koehn et al., 2007), information retrieval (Kanis and Skorkovská, 2010; Zeroual and Lakhouaja, 2017), and text analysis for historical texts (Piotrowski, 2012; Manjavacas et al., 2019) and sentiment analysis (Abdul-Mageed et al., 2014). Lemmatization is particularly critical for OCS because the language is characterized by regional and historical diversification and exhibits numerous and significant orthographic and textual variants. Lemmatization can support corpus search, comparison of textual witnesses, analysis of orthographic and morphological variation, and the study of lexical distributions across sources and textual variants. For example, the noun ѩзыкъ, "tongue, language, speech, nation, people", may appear in several orthographic forms, including ꙗзыкъ, ꙗзикь, езыкъ, езикь, езыкь, ѥзыкь, ꙗзыкъ, ꙗзыкъ, ѩз҆ыкъ, ѩзыикь, ѩзыкъ, ѫзыкъ, ѫзыкъ, and ѣзикь, as well as in abbreviated forms such as ѩӟкъ, ѫӟкъ, ꙗӟкъ, and ѣӟкъ. These forms must also be considered across plural and dual number and across the seven cases, as reflected in the Digital OCS Dictionary at GORAZD (SJS, 1966–1997).

Third Workshop on the Bridges and Gaps between Formal and Computational Linguistics, pages 82–93

While modern languages are equipped with standardized datasets (i.e., GLUE (Wang et al., 2018), SuperGLUE (Wang et al., 2019), XGLUE (Liang et al., 2020), etc), historical languages have significantly limited benchmark-quality resources (Dereza et al., 2024). This is particularly applicable to Church Slavonic, which has recently experienced an increase in digital support and tool development, while lemmatization benchmarks remain limited. Among the available resources, PROIEL (Haug and Jøhndal, 2008) introduced treebank-style annotation for early Indo-European texts, including OCS, and TOROT (Eckhoff and Berdicevskis, 2015) for Old Russian and OCS texts. Universal Dependencies (UD) also provides a shared annotation framework for such resources (de Marneffe et al., 2021). Toolkits such as Stanza and UDPipe are also available for UD-style processing and lemmatization (Qi et al., 2020; Straka et al., 2019). More recently, the heterogeneity and fragmentation of Church Slavonic resources have also been highlighted by DIACU (Cassese et al., 2025). However, the availability of digital or annotated resources does not by itself guarantee reliable evaluation on newly prepared OCS texts.

In this work, we describe an ongoing effort in annotating texts for OCS lemma-level evaluation. We present the normalization and encoding steps required for the computational stability of the texts, the annotation process used in the creation of the dataset, and evaluate off-the-shelf systems on the newly annotated datasets. The results show that pretrained systems struggle with the new data, but in-domain retraining leads to significant improvements. At the same time, the results across datasets are low, suggesting that, once again, orthographic, editorial, and annotational differences are a major challenge in OCS lemmatization.

The contributions of this paper are as follows:

- We introduce a newly annotated OCS dataset for lemma-level evaluation comprising 29,678 tokens total, 25,517 unique forms, and 23,803 unique lemmas.
- We document the preprocessing and annotation decisions required to convert heterogeneous OCS materials into a stable evaluation resource, including a normalization workflow for handling source-specific PUA characters and related encoding inconsistencies.
- We benchmark pretrained and retrained models (Stanza, UDPipe-2, and a dictionary baseline) on the new dataset. We compare in-domain and combined-corpus training across test sets. We show that lemmatization performance depends on dataset similarity, orthographic variation, and annotation mismatch, highlighting the difficulty of cross-resource generalization in historical Slavic NLP. Code, model weights, dataset access details, and license information are publicly available on GitHub[1].

The remainder of the paper is structured as follows. Section 2 reviews related work on OCS resources and historical lemmatization. Section 3 describes the source material, annotation scheme, preprocessing decisions, and workflow used to construct the dataset. Section 4 presents the experimental setup, including data splits, systems, and evaluation metrics. Section 5 reports the main benchmarking results and cross-dataset findings. Section 6 discusses the implications of the results, and Section 7 concludes the paper.

2 Related Work

The computational study of OCS has been supported by a limited set of resources including PROIEL for historical Indo-European languages (Haug and Jøhndal, 2008), TOROT for Old Russian (Eckhoff and Berdicevskis, 2015), and the UD annotation framework (de Marneffe et al., 2021). These resources collectively form the foundation for computational analysis of medieval Slavic languages, but it remains unexplored how well they support the generalization of models to unseen datasets.

A recent work aggregates multiple resources in DIACU, a diachronically and geographically annotated Church Slavonic dataset (Cassese et al., 2025). It directly addresses the fragmentation in Church Slavonic resources, which are scattered across corpora, and editorial practices. A similar concern appears in retrieval-based research on Church Slavonic variants, which points to the difficulty of working across chronological and regional variation, as well as heterogeneous source conditions (Lendvai et al., 2025). In other words, the existence of digital text does not automatically imply access to benchmark-quality resources for lemmatization and similar tasks.

[1] https://github.com/usmannawaz01/ocs-lemmatization-benchmark

Several methods have been proposed for lemmatization. Rule-based and finite-state approaches are commonly used in settings with substantial expert knowledge and manually constructed resources (Schmid et al., 2004). Dictionary-based methods remain attractive because of their simplicity and usefulness in settings with high lexical overlap (Nawaz et al., 2024). Edit-based approaches have shown strong performance across languages (Straka and Straková, 2017; Straka et al., 2019), while recent neural methods model lemmatization as a character-level transduction problem (Schnober et al., 2016; Kanerva et al., 2021; Bergmanis and Goldwater, 2018). The Stanford lemmatization system (Qi et al., 2018) combines lexicon lookup with neural sequence-to-sequence generation. Lemmas are first retrieved from lexicons built from the training data, and neural generation is used when lookup fails. This approach later became part of the Stanza toolkit (Qi et al., 2020). Stanza and UDPipe are of particular interest with regard to the OCS language, but their performance is correlated with the resources used for training and evaluation, and is sensitive to orthographic variation.

Recent results from the SIGTYP 2024 shared task on historical languages (Dereza et al., 2024) highlight ongoing challenges in evaluation. Even robust approaches such as TartuNLP (Dorkin and Sirts, 2024), Heidelberg-Boston (Riemenschneider and Krahn, 2024), and UDParse (Heinecke, 2024) are affected by scarcity and tokenization differences in historical languages (Dereza et al., 2024). This is particularly relevant for OCS lemmatization, where recently available resources may be used for both training and evaluation.

In light of these challenges, our work focuses not on proposing a new lemmatization architecture, but on resource construction and evaluation. We present a newly annotated OCS dataset for lemma-level benchmarking, describe the normalization and annotation decisions behind it, and use it to assess the portability of existing OCS lemmatization systems.

3 Data Collection and Annotation

The annotated material used in this work includes texts from diverse OCS sources to reflect differences in editorial practices, orthographic systems, and textual genres. The current corpus includes texts from the Cyrillomethodiana[2] portal, creed texts from Pravmir[3], and dictionary texts from Azbuka[4]. Table 1 describes the composition of the annotated sources.

Table 1: OCS source of the annotated texts in the dataset

Source group	Origin	Tokens
Cyrillomethodiana	Cyrillomethodiana portal	5,789
Creed	Pravmir: *Simvol very*	77
Dictionary texts	Azbuka	23,812
Total		29,678

The dataset was created by including texts from sources with varying origins, formats, and textual conventions. This variety supports the evaluation of lemmatization systems under more realistic generalization conditions.

3.1 Normalization and Annotation Policy

The study focuses on lemma annotation at the token level, where each surface form is mapped to its canonical lexical form. This is particularly relevant for OCS due to its extensive inflection and orthographic variations.

Prior to annotation, the source texts were normalized to a consistent Unicode representation. Source-specific Private Use Area (PUA) characters were addressed using manually validated mapping tables (Napolitano and Nawaz, 2025). The purpose of the normalization process was not to eliminate the natural variations that exist in the source texts, but rather to avoid technical characters that become corrupted after scraping and may affect the annotation and evaluation processes. Additional details and representative examples are provided in Appendix A.

Table 2 presents representative examples of PUA characters and their standardized Unicode equivalents: this required systematic character-level remapping rather than simple formatting cleanup.

Table 3 illustrates how raw source text is converted into a stable Unicode representation prior to annotation. PUA characters are resolved into standard Unicode Cyrillic forms, easing downstream tokenization and lemma annotation.

As the next step, lemma annotation was applied to the normalized texts. Where possible,

[2] https://histdict.uni-sofia.bg/textcorpus/list
[3] https://www.pravmir.ru/simvol-very/
[4] https://azbyka.ru/otechnik/Spravochniki/polnyj-pravoslavnyj-tserkovnoslavjanskij-slovar/1

Table 2: Representative examples of source-specific PUA characters and their standardized Unicode equivalents used during normalization.

PUA character	PUA code point	Normalized character	Unicode code point
	U+E205	и	U+0438
~	U+E012	Ѣ	U+0462
	U+E20D	Ꙋ	U+0427
н	U+E201	Ѥ	U+0464
	U+E342	◌͡◌	U+0487
	U+E343	◌҃	U+0488
	U+E20C	Ꙝ	U+041D
⌐	U+E018	ъ	U+044A
	U+E20E	Ѳ	U+042E
	U+E213	Ѥ	U+0464

lemma forms follow existing scholarly conventions and treebank best practices, with contextual and grammatical considerations applied in ambiguous cases, while punctuation and other non-lexical tokens were retained. The present data is restricted to token-level lemmatization and does not introduce a separate multi-word expression annotation layer. This follows UD OCS tokenization, where there are no multi-word tokens[5].

Table 3: Example of text before and after Unicode harmonization.

Stage	Text snippet
Before normalization (Source Text)	Странꙗнолюбⷭемь цѣлꙋ ѣкоже древле Авраамъ· вꙑшьнѣго доⷣе свѣтла селенⷭѣ· въ немьже блаженѥ прѣдъстоа· сⷭѣшⷭ ѣко ⷭ слъньце· Тѣмьже ⷭ насъ просвѣтⷭ нꙑнѣ ⷭьтжщⷭхъ прѣсвѣтлоѥ твоѥ трьжьство
After normalization (PUA characters remapping)	Странꙗнолюбиемь цѣлꙋ ѣкоже древле Авраамъ· вꙑшьнѣго доиде свѣтла селениѣ· въ немьже блаженѥ прѣдъстоа· сиѣеши ѣко и слъньце· Тѣмьже и насъ просвѣти нꙑнѣ ѣьтжщиихъ прѣсвѣтлоѥ твоѥ трьжьство:

3.2 Annotation Workflow

The annotation workflow includes preprocessing, candidate organization, expert annotation, and multiple-stage verification. The data collection was performed by scraping the relevant websites. During preprocessing, texts are cleaned and normalized to handle PUA characters and encoding issues. This process results in a consistent text representation suitable for tokenization and annotation.

Annotation was conducted semi-automatically by taking advantage of a lookup lexicon built by merging lemma entries from UD 2.12 OCS with the lemmas manually annotated and iteratively obtained through the dataset creation process. To improve efficiency, lemma candidates are grouped based on difficulty by distinguishing between single-match, ambiguous, and unseen forms. In the initial candidate-generation stage for the manually annotated Cyrillomethodiana running-text subset, 1,838 tokens received a single lexicon match, 1,432 tokens received multiple possible lemma candidates, and 2,519 tokens were not covered by the UD OCS lookup lexicon. These correspond to 31.75%, 24.74%, and 43.51% of the 5,789 tokens in this subset, respectively.

This allows for easier prioritization of difficult cases and minimizes redundant work on simpler cases. When a reliable lexicon match was available, the lexicon-based lemma was used as the initial annotation. For tokens not covered by the lexicon, the lemma prediction produced by the pretrained Stanza model (Version 1.5.1) for OCS was used. The initial lemma annotations were manually verified and, where necessary, corrected and reviewed by an expert linguist, resulting in a dataset verified through expert verifications.

The overall annotation workflow in Figure 1 produces human-verified lemma annotations that can be extended over time as more texts are included. Although the present work targets lemma annotation only, the corpus is already organized in CoNLL-U format (Straka et al., 2016) to ensure compatibility with the UD framework and to support future extensions. Whenever possible, sentence boundaries are also preserved from the source material, which facilitates future expansion toward fuller UD-style annotation (Straka et al., 2016).

3.3 Dataset Statistics

The current annotated dataset consists of 29,678 tokens drawn from multiple OCS source groups. In addition to its overall size, the dataset exhibits substantial lexical diversity; it contains 25,517 unique word forms and 23,803 unique lemmas. Beyond the annotated subset used in the present experiments, the broader source collection is substantially larger. It includes more than 200 texts comprising approximately more than 5 million tokens, and may support future annotation efforts.

As shown in Table 4, although the present dataset is smaller than UD 2.12 OCS in total number of tokens, it exhibits substantially greater lemma diversity, as reflected in its much larger number of unique lemmas. Indeed, beyond to-

[5] https://universaldependencies.org/cu/index.html

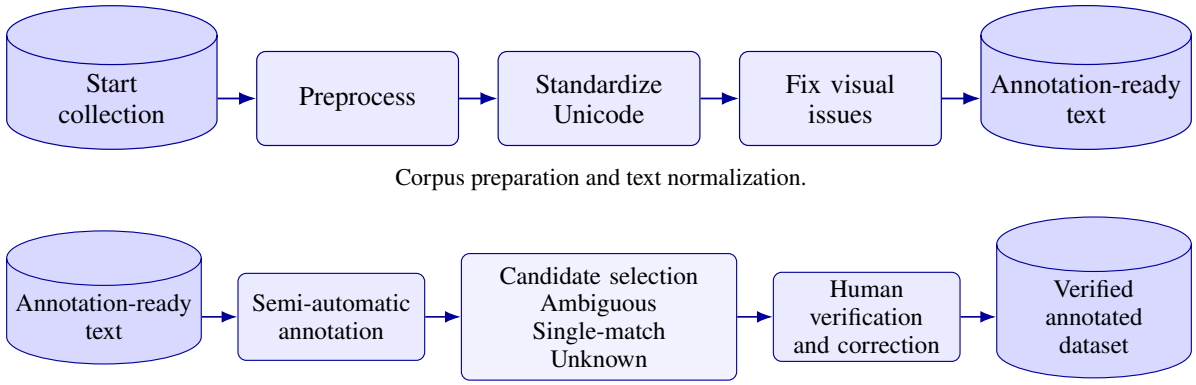

Figure 1: Workflow for corpus preparation and lemma annotation. The upper panel shows data collection and preprocessing, where source-specific Unicode and rendering issues are resolved to produce annotation-ready text. The lower panel shows the annotation stage, moving from semi-automatic annotation through candidate selection and human verification to the final verified annotated dataset.

Table 4: Lexical statistics for our dataset and UD 2.12 OCS. The UD 2.12 OCS values are taken from the official dataset statistics, while the values for our dataset were computed by us.

Dataset	Tokens	Forms	Lemmas
Our	29678	25517	23803
UD 2.12	198843	43815	8247

ken counts, the dataset reflects variation across sources and highlights more challenging generalization conditions. The high ratio of unique forms and lemmas is mainly due to the composition of the current data. In particular, the Azbuka portion consists largely of lexicographical entries rather than continuous running text, which increases lexical diversity and reduces token repetition. Therefore, the dataset should not be interpreted as a frequency-balanced sample of natural OCS. Instead, it provides a challenging lemma-level evaluation set with broad lexical coverage and many low-frequency forms.

4 Experimental Protocol and Baseline Methods

The annotated OCS material is partitioned into training, development, and test sets using an approximate 80/10/10 split at the sentence level, preserving sentence boundaries within each source file. This results in 23,768 training tokens, 2,979 development tokens, and 2,931 test tokens (Table 5).

Table 5: Train/dev/test split of the annotated OCS material.

Split	Tokens
Train	23768
Dev	2979
Test	2931
Total	29678

Following common practice in UD-based and historical NLP (Dereza et al., 2024), the training and development splits are used for model training and selection, while the test split is reserved for evaluation. Additionally, we propose cross-dataset evaluation on existing OCS resources to assess generalization under variation in orthography and editorial conventions.

In the experiments presented in this paper, four setups for lemmatization of OCS texts are compared. First, a pre-trained Stanza lemmatizer for OCS based on the UD 2.12 treebank (Qi et al., 2020). The Stanza lemmatizer uses lexicon-based memorization in conjunction with a neural character-level seq2seq model in its backoff stage. Second, a pre-trained UDPipe 2 model trained on the same treebank release (Straka and Straková, 2017; Straka et al., 2019), uses an edit-tree classification approach for lemmatization, predicting a word-to-lemma transformation rule. Third, a dictionary-based baseline constructed from the training split. For any surface word form in the test set, it predicts the most frequently associated lemma seen in this training data; for any unseen word form, it predicts the unchanged surface form itself. Finally, two retrained Stanza settings are evaluated. In the first setting, the Stanza pipeline is trained on the annotated OCS training split only. In the second setting, the Stanza pipeline is trained

on a combined corpus formed by merging the existing UD 2.12 OCS data with the new annotations. In both settings, the tokenizer and lemmatizer are trained from scratch, the development split is used for model selection, and evaluation is conducted under both gold tokenization and raw-text settings.

Evaluation is performed using the official CoNLL-U 2018 UD script. In the gold-tokenized setting, we report lemma accuracy, and in the raw-text setting, we report lemma F1-score. Beyond overall score, we also evaluate performance on unseen tokens to assess generalization, which is especially important in the context of historical lemmatization. Evaluation is conducted both on the annotated test set and across datasets where appropriate.

5 Results

We evaluate all systems on the annotated test set. Results are reported in Table 6 under both gold tokenization and raw-text settings.

Under gold tokenization, the pre-trained Stanza model achieves 15.56% lemma accuracy, while the pre-trained UDPipe 2 model achieves 16.27%, indicating that both pre-trained systems generalize poorly on the annotated test set.

The best-performing model is the Stanza model trained on the annotated data, reaching 51.42%, while training on the combined corpus yields to 45.92%. In both retraining settings, the Stanza models are trained from scratch rather than initialized from the publicly available pretrained model. The annotated training and development data provide task-specific supervision that improves performance on the held-out annotated test set. However, the scores remain far below perfect accuracy, which suggests that the test set remains challenging for current lemmatization systems.

The dictionary baseline achieves 37.67%, outperforming both pre-trained models. Because the dictionary baseline is built only from the annotated training split and evaluated on held-out test material, this score does not reflect evaluation on the training data itself, although the method can directly memorize form–lemma pairs observed during training.

5.1 Cross-Dataset Generalization

To better illustrate out-of-domain evaluation beyond our annotated material, we also evaluate the systems across additional test sets, as shown in Table 6. This setting complements the in-dataset evaluation reported in the previous section and helps assess how far adaptation to our annotation scheme generalizes to an existing benchmark resource. Figure 2 visualizes the same cross-dataset pattern: the model retrained on the new data performs best on the annotated test set, whereas the combined-data model transfers much better to the UD 2.12 and combined test sets.

On the UD 2.12 OCS test set, the Stanza model trained on the new annotated data reaches 46.54% under gold tokenization. The results show that training on the new data alone does not transfer strongly to the UD benchmark, which is much larger. Overall, the pretrained Stanza model generalizes poorly on our dataset, while the combined-data retrained model shows improvements on our dataset and, albeit slightly, on the UD 2.12 test set.

These results are consistent with broader observations in historical NLP (Dereza et al., 2024) and, in the case of OCS, this degradation is plausible because of differences in orthography, corpus composition, and lemma conventions. The cross-dataset results therefore suggest that OCS lemmatization quality should not be assessed only on a single familiar benchmark.

5.2 OOV Analysis

To further understand system behavior on this annotated test set, we also examine system performance on out-of-vocabulary (OOV) words with respect to each system's own training vocabulary. This is particularly relevant in historical lemmatization, where problems often arise due to rare words, spelling variations, and vocabulary differences between sources.

The annotated test set has one important property, the proportion of OOV tokens remains high across all systems when OOV is defined with respect to each model's own training vocabulary.

OOV performance is shown in Table 7. The table reports, for each system, the size of the training and test sets, the number and proportion of OOV tokens in the annotated test set, the number of OOV tokens lemmatized correctly and incorrectly, and the resulting OOV accuracy. Lower values for OOV tokens and OOV% are better, whereas higher values for Correct OOV and OOV Accuracy are preferable. Figure 3 provides a visual summary of the same trend, showing that retraining on the new annotated data improves OOV accuracy substantially, although unseen forms remain difficult overall.

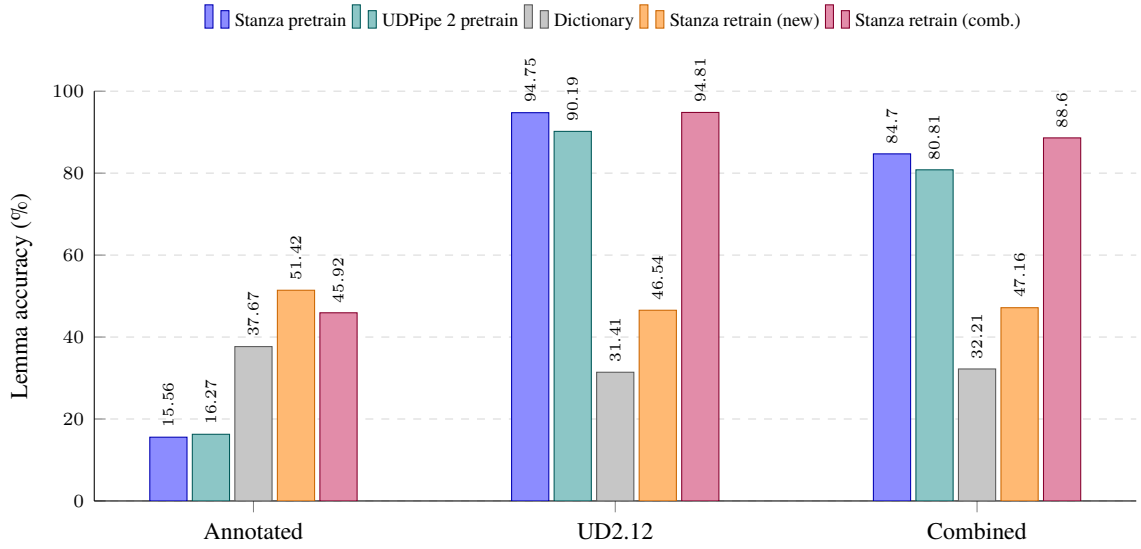

Figure 2: Cross-dataset lemma accuracy under gold tokenization. The x-axis indicates the evaluation test set. The figure highlights strong dataset dependence: the new-data retrained model performs best on the annotated test set, while the combined-data retrained model generalizes much better to the UD 2.12 and combined test sets.

The pre-trained Stanza and UDPipe 2 systems each have 2,617 OOV tokens out of 2,931 test tokens, corresponding to an OOV rate of 89.29%. Their OOV accuracy is correspondingly low, at 7.60% for both Stanza and UDPipe 2. After retraining on the annotated data, Stanza reaches 24.16% OOV accuracy, with 1,854 OOV tokens (63.25%). The combined-data retrained model has 1,778 OOV tokens (60.66%) and reaches 14.17% OOV accuracy. This indicates that retraining on annotated in-domain data is helpful in adapting to lexical and spelling properties in a new dataset. Nonetheless, OOV words are a major source of error in this case. The main challenge in this dataset is not simply lemmatization accuracy in general. Instead, it is more closely related to how well out-of-vocabulary words are handled. Further annotation and lexical coverage are likely important in improving lemmatization accuracy on this OCS data.

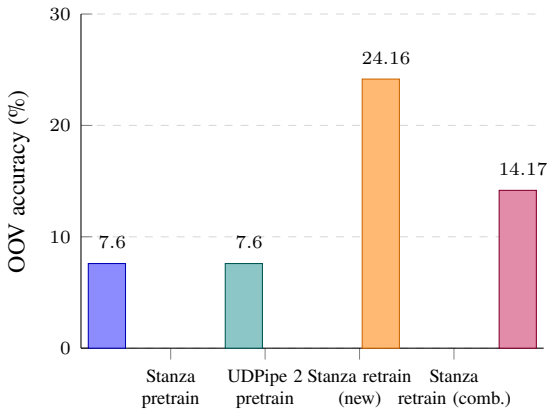

Figure 3: OOV lemma accuracy on the annotated test set under gold tokenization. Retraining on the new annotated data substantially improves performance on unseen forms, although OOV items remain difficult.

6 Discussion

The results suggest that evaluation on a limited, well-known OCS resource does not suffice to understand the behavior of modern lemmatization systems. In particular, the two models trained on the UD 2.12 OCS treebank perform poorly on our annotated test set (Stanza reaches 15.56%, UDPipe 2 reaches 16.27%). The results suggest that these models struggle to generalize to the newly annotated test set, despite all systems being trained on OCS. This result is not unique to our specific case and is in line with other observations in historical NLP, where dataset shift, orthographic variation, and limited supervision remain open challenges in the field (Dereza et al., 2024; Dorkin and Sirts, 2024; Riemenschneider and Krahn, 2024; Heinecke, 2024).

The analysis of OOV highlights the test set difficulty. The pretrained Stanza and UDPipe 2 models each treat 89.29% of the annotated test set as OOV and achieve only 7.60% OOV accuracy. In contrast, the retrained Stanza model on the new annotated data reaches 24.16% OOV accuracy, while the combined-data retrained model reaches 14.17%. This shows that in-domain train-

Table 6: Evaluation of all systems across available test sets. For each test set, we computed the gold tokenization lemma accuracy, while Raw→CoNLL-U denotes the lemma F1 score from end-to-end UD-style evaluation on raw running text. The UD 2.12 OCS and combined test sets contain 20,149 and 23,080 tokens, respectively.

System	Annotated GOLD Tokenization	Annotated Raw→CoNLL-U	UD 2.12 GOLD Tokenization	UD 2.12 Raw→CoNLL-U	Combined GOLD Tokenization	Combined Raw→CoNLL-U
Pretrained Stanza	15.56	15.91	94.75	94.65	84.70	84.55
Pretrained UDPipe 2	16.27	16.38	90.19	90.18	80.81	80.80
Stanza (train on new)	**51.42**	**51.42**	46.54	46.52	47.16	47.14
Stanza (train on comb.)	45.92	46.57	**94.81**	**94.74**	**88.60**	**88.62**
Dictionary	37.67	—	31.41	—	32.21	—

Table 7: Model-centered OOV analysis on the annotated test set under gold tokenization. OOV is defined with respect to each system's own training vocabulary.

System	Train tokens	Test tokens	OOV tokens ↓	OOV % ↓	Correct OOV ↑	Incorrect OOV ↓	OOV Acc. ↑
Pretrained Stanza	159710	2931	2617	89.29	199	2418	7.60
Pretrained UDPipe 2	159710	2931	2617	89.29	199	2418	7.60
Stanza (train on new)	23768	2931	1854	63.25	**448**	**1406**	**24.16**
Stanza (train on comb.)	183478	2931	**1778**	**60.66**	252	1526	14.17

ing improves generalization to unseen forms, although OOV items remain a major source of error.

Overall, this implies that the main difficulty of this dataset lies in generalization to unseen and orthographically variable forms.

The cross-dataset experiments are also significant in this context. When the model is trained on our annotated corpus and evaluated on the UD 2.12 OCS test set, the Stanza model reaches 46.54%, while the dictionary baseline reaches 31.41%. This occurs despite the annotated dataset being considerably smaller than the UD 2.12 corpus, suggesting that the new annotations provide useful supervision, while also confirming that cross-dataset performance is strongly affected by distributional, orthographic, and annotation differences. The experiments show that our corpus improves the model beyond simple lexical lookup, but also highlight the limited degree to which cross-dataset generalization is possible.

The experiments also demonstrate the close relationship between the construction of resources and the evaluation of models in medieval Slavic NLP. In the high-resource scenario, annotation is often treated as part of the background, while evaluation is treated as a distinct step in the process. In the case of OCS, these processes are much more closely integrated. The decisions regarding normalization, tokenization, lemma choice, and source materials all directly influence the behavior of models and the interpretation of benchmarking experiments.

7 Conclusion and Future Work

This work presented a newly annotated OCS corpus for lemma-level evaluation and used it to benchmark off-the-shelf systems, retrained Stanza settings, and a dictionary baseline. The results show that in-domain retraining improves performance, while cross-dataset evaluation highlights the impact of domain, orthographic variation, and annotation mismatch on OCS lemmatization. More broadly, the study shows that carefully constructed annotation can provide a useful benchmark and a valuable first step toward a more comprehensive OCS lemmatization resource.

The current work has a number of limitations. First, although this work provides a preliminary contribution towards a more comprehensive OCS lemmatization resource, it does not yet cover the full range of textual, orthographic, and textual variations found in OCS sources. Further limitations concern the annotation process itself. Due to project time constraints, the annotations presented in this study were reviewed by a single expert linguistic annotator. As a consequence, it was not possible to calculate an inter-annotator agreement score. Addressing this limitation and implementing a multi-annotator validation procedure constitute important objectives for future stages of the project. Nevertheless, it should be noted that the philological methodology underlying the annotation process was developed and discussed collaboratively by the expert linguistic annotator and the WP(4) Product Owner, who is also responsible for the research unit dedicated to OCS studies.

Second, while a number of benchmark systems are included in this experimental design, Stanza is

tested in both off-the-shelf and retrained configurations, while UDPipe is tested in its available configuration only (Qi et al., 2020; Straka et al., 2019). Moreover, the retrained lemmatizers rely on silver POS tags from the official Stanza model. Accordingly, this study's comparison is informative rather than exhaustive. Future work could include a broader range of systems and additional historical language resources.

Third, although this study's annotation policy is designed with specific choices regarding normalization, tokenization, and canonical lemma selection, these choices may not align fully with other available resources. This may have contributed to degradation across datasets in addition to any underlying linguistic difficulty. A related limitation concerns the orthographic and character-level discrepancy between the provided corpus and the data used to train the pretrained models. This mismatch is also visible at the character level in the training splits, UD 2.12 and the new data share 111 unique characters, while 61 characters occur only in UD 2.12 and 42 only in the new data. In particular, differences among the redactions of OCS and other textual variations may have introduced representation mismatch in addition to genuine lemmatization difficulty. Further research is therefore needed to examine whether improved normalization and orthographic standardization would improve results on OCS resources.

Lastly, this study's evaluation is centered on token-level lemma accuracy and includes both gold-tokenized and end-to-end raw-text settings. A substantial part of the current resource is derived from Azbuka lexicographical material, which consists mainly of lemma-based entries rather than continuous sentence-level text, future work should evaluate tokenization more explicitly on sources with richer sentence-level structure. Although this is appropriate in a benchmark setting and is in accordance with UD-style evaluations (Straka et al., 2019; Qi et al., 2020), it does not provide a complete picture of how historical NLP resources are ultimately used. Tokenization, normalization, and lemmatization are closely coupled in a number of practical scenarios. Future work could therefore explore how this annotated material can be used in a more comprehensive pipeline evaluation. We will also work toward expanding lemma annotation for OCS by covering a wider range of orthographic and textual variation, and will benchmark additional lemmatization systems using more comprehensive evaluation metrics, with the goal of establishing a stronger public benchmark for OCS.

Acknowledgments

This work was supported by the PNRR (National Recovery and Resilience Plan) project Italian Strengthening of ESFRI RI Resilience (ITSERR), funded by the European Union—NextGenerationEU (CUP:B53C22001770006).

References

Muhammad Abdul-Mageed, Mona Diab, and Sandra Kübler. 2014. Samar: Subjectivity and sentiment analysis for arabic social media. *Computer Speech & Language*, 28(1):20–37.

Toms Bergmanis and Sharon Goldwater. 2018. Context sensitive neural lemmatization with lematus. In *16th annual conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1391–1400. Association for Computational Linguistics (ACL).

David J Birnbaum, Ralph Cleminson, Sebastian Kempgen, and Kiril Ribarov. 2008. White paper on character set standardization for early cyrillic writing after unicode 5.1.

Maria Cassese, Giovanni Puccetti, Marianna Napolitano, and Andrea Esuli. 2025. Diacu: A dataset for the diachronic analysis of church slavonic. In *Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025)*, pages 101–107.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Oksana Dereza, Adrian Doyle, Priya Rani, Atul Kr Ojha, Pádraic Moran, and John Philip 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.

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.

Hanne Martine Eckhoff and Aleksandrs Berdicevskis. 2015. Linguistics vs. digital editions: The tromsø old russian and ocs treebank.

Marcello Garzaniti. 2019. *Gli slavi. Storia, culture e lingue dalle origini ai nostri giorni. Nuova edizione*. Carocci, Roma.

Sharon Goldwater and David McClosky. 2005. Improving statistical mt through morphological analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 676–683.

Dag TT Haug and Marius Jøhndal. 2008. Creating a parallel treebank of the old indo-european bible translations. In *Proceedings of the second workshop on language technology for cultural heritage data (LaTeCH 2008)*, pages 27–34. Prague.

Johannes Heinecke. 2024. Udparse@ sigtyp 2024 shared task: Modern language models for historical languages. In *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 142–150.

Martin Hetényi and Peter Ivanič. 2021. The contribution of ss. cyril and methodius to culture and religion. *Religions*, 12(6):417.

Jenna Kanerva, Filip Ginter, and Tapio Salakoski. 2021. Universal lemmatizer: A sequence-to-sequence model for lemmatizing universal dependencies treebanks. *Natural Language Engineering*, 27(5):545–574.

Jakub Kanis and Lucie Skorkovská. 2010. Comparison of different lemmatization approaches through the means of information retrieval performance. In *International Conference on Text, Speech and Dialogue*, pages 93–100. Springer.

Sebastian Kempgen. 2008. Unicode 5.1, old church slavonic, remaining problems–and solutions, including opentype features. In *Slovo: Towards a Digital Library of South Slavic Manuscripts; Proceedings of the International Conference, 21–26 February 2008, Sofia, Bulgaria*. Otto-Friedrich-Universität.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

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.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, and 5 others. 2020. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

Enrique Manjavacas, Ákos Kádár, and Mike Kestemont. 2019. Improving lemmatization of non-standard languages with joint learning. In *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)*, pages 1493–1503.

Marianna Napolitano and Usman Nawaz. 2025. From characters to meaning: Challenges and limitations in the semantic analysis of church slavonic texts. In Alberto Melloni and Francesca Cadeddu, editors, *The Digital Turn in Religious Studies Research, Services, Infrastructures*, volume 6 of *FSCIRE Research and Papers*, pages 41–65. Vandenhoeck & Ruprecht Verlage.

Aleksander E Naumow. 1983. *Biblia w strukturze artystycznej utworów cerkiewnos· lowiańskich*. Uniw. Jagielloński.

Usman Nawaz, Liliana Lo Presti, Marianna Napolitano, and Marco La Cascia. 2024. Automatic lemmatization of old church slavonic language using a novel dictionary-based approach. In *International Workshop on Document Analysis Systems*, pages 408–421. Springer.

Riccardo Picchio. 1972. *Questione della lingua e Slavia cirillometodiana*. Edizioni dell'Ateneo,[sd].

Michael Piotrowski. 2012. *Natural language processing for historical texts*. Morgan & Claypool Publishers.

Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D Manning. 2018. Universal dependency parsing from scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual parsing from raw text to universal dependencies*, pages 160–170.

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*, pages 101–108.

Frederick Riemenschneider and Kevin Krahn. 2024. Heidelberg-boston@ sigtyp 2024 shared task: Enhancing low-resource language analysis with character-aware hierarchical transformers. In *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 131–141.

Helmut Schmid, Arne Fitschen, and Ulrich Heid. 2004. Smor: A german computational morphology covering derivation, composition and inflection. In *LREC*, pages 1–263. Lisbon.

Carsten Schnober, Steffen Eger, Erik-Lân Do Dinh, and Iryna Gurevych. 2016. Still not there? comparing traditional sequence-to-sequence models to encoder-decoder neural networks on monotone string translation tasks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1703–1714.

SJS. 1966–1997. Slovník jazyka staroslověnského / Old Church Slavonic Dictionary. 4 vols. Prague. Online version available through GORAZD: An Old Church Slavonic Digital Hub, http://www.gorazd.org/?q=en/node/21; accessed 10 June 2026.

Milan Straka, Jan Hajic, and Jana Straková. 2016. Udpipe: trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297.

Milan Straka and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies*, pages 88–99.

Milan Straka, Jana Straková, and Jan Hajic. 2019. Udpipe at sigmorphon 2019: Contextualized embeddings, regularization with morphological categories, corpora merging. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 95–103.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP*, pages 353–355.

Imad Zeroual and Abdelhak Lakhouaja. 2017. Arabic information retrieval: Stemming or lemmatization? In *2017 Intelligent Systems and Computer Vision (ISCV)*, pages 1–6. IEEE.

Victor Zhivov. 2009. *Language and culture in eighteenth century Russia*. JSTOR.

Appendix

A Text Collection and Encoding Harmonization

A substantial part of the resource construction effort concerned not lemma annotation itself, but the preparation of source texts into a representation suitable for reliable computational processing. The collected OCS materials originate from heterogeneous digital sources and therefore exhibit differences in formatting, orthographic conventions, and character encoding. In particular, some sources contained Private Use Area (PUA) characters and other non-standard Unicode representations that rendered correctly only in source-specific environments, but became corrupted or unreadable after extraction, editor import, or format conversion.

The collection workflow was grounded in the acquisition of machine-readable text from digital repositories, especially the Cyrillomethodiana portal, followed by standardization steps. During preprocessing, unwanted symbols, spacing noise, and source-specific encoding inconsistencies were removed or corrected before annotation. A central problem was that text that appeared visually acceptable on the source website did not remain stable across environments such as text editors, XML workflows, and downstream NLP tools. In our earlier processing experiments, this problem was observed both after web scraping and after TEI/XML handling, where some characters remained missing or corrupted because the source used PUA or otherwise non-standard encodings (Kempgen, 2008; Birnbaum et al., 2008).

To address this, we constructed manually verified mapping tables that replaced source-specific PUA symbols and other problematic forms with standardized Unicode equivalents before tokenization and manual annotation. This harmonization process included: (i) identifying characters in the Unicode PUA, (ii) assigning standardized Unicode correspondences based on the intended graphemic value, (iii) resolving confusions between visually similar characters and combining marks, and (iv) rechecking rendering across multiple environments after conversion (Napolitano and Nawaz, 2025). This step was necessary because character-level incompatibilities can create artificial token differences unrelated to the linguistic problem of lemmatization.

The normalization effort was particularly important for OCS because orthographic distinctions may be obscured by encoding errors as in preprocessing work. For example, we observed confusion between *titlo* (U+0483) and *pokrytie* (U+0487). More broadly, the collection process revealed over more than 43 PUA characters in the analysed material, requiring explicit remapping be-

fore the text could be processed consistently across environments.

Although standardization reduces artificial representational variation, it does not eliminate genuine historical and textual variation. Rather, it provides a stable representational layer on which such variation can be studied more reliably in annotation and evaluation.