Morphological Annotation of Old Serbian in Universal Dependencies

Vladimir Polomac[1], Silvie Cinková[2]

University of Kragujevac, Faculty of Philology and Arts[1], Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics[2]

v.polomac@filum.kg.ac.rs[1], cinkova@ufal.mff.cuni.cz[2]

Abstract

We report on the morphological tagging of Old Serbian in the Universal Dependencies framework. To facilitate the manual annotation, we pre-processed the data with the Old Church Slavonic 2.12 UDPipe model. The decision was based on the known similarity of these two languages as well as on the declared performance of this model compared to other models for historical varieties of Slavic languages. With over 3,000 manually annotated tokens, we evaluated the performance of the relevant pre-trained UDPipe2 models of historical Slavic languages. Besides, we also trained and evaluated custom models with UDPipe1 containing the annotated Old Serbian data. We have found that: (1) for this particular domain and amount of training data, the most suitable model is UD Old East Slavic – Birchbark 2.12, although its declared performance is much lower than that of Old Church Slavonic; (2) even 3,000 tokens of Old Serbian increase the performance of UDPipe1 models almost to the level of the Birchbark 2.12 model. The dataset is publicly available at https://doi.org/10.5281/zenodo.19317842.

Keywords: Old Serbian, Old Church Slavonic, Universal Dependencies, PoS Tagging, UDPipe, Serbian Medieval Charters, tagger evaluation.

1. Introduction

Corpus searches, as well as diverse text-mining methods, benefit from linguistic markup (lemmatization, part-of-speech tagging, and syntactic relations). Universal Dependencies (McDonald et al., 2013; Zeman, 2018; de Marneffe et al., 2021) is a largely theory-agnostic cross-lingual scheme providing principles of tokenization and sentence splitting, as well as a set of morphological tags and syntactic labels with definitions, all well-grounded in the language typology. A cross-lingual annotation standard paved the way for powerful language-agnostic tools, e.g., UDPipe (Straka et. al., 2016; Straka 2018), SpaCy (Honnibal et al., 2020), or Stanza (Qi et al., 2020), which can process diverse languages (including their specific varieties and domains) by plugging in individual language models trained on dedicated data sets.

The goal of our project is to contribute to the Universal Dependencies GitHub repository a valid UD-annotated data set of the oldest preserved Serbian texts from the 12th and 13th century. This paper reports on its first steps – the annotation pre-processing and lemmatization decisions.

The paper is organized as follows: Section 2 provides information about the Old Serbian language, its relationship to Proto-Slavic and Old Church Slavonic, as well as the criteria for selecting the manuscripts used in the study. Section 3 elaborates on the principles of text lemmatization in more detail, while Section 4 covers the principles of morphosyntactic annotation. Section 5 presents the initial results of tagging, automatic lemmatization, and morphosyntactic annotation using UDPipe. The final chapters briefly summarize the results (section 6) and point to the prospects for further work (section 7).

2. Old Serbian: Data Selection and Preprocessing

Serbian belongs to the group of South Slavic languages that developed as distinct languages from the common Proto-Slavic language in the second half of the first millennium AD, following the migration to the Balkan Peninsula. The early history of the Serbian language can be reconstructed based on comparisons with other Slavic languages, especially Old Church Slavonic, which emerged in the 9th century for liturgical purposes among the Western Slavs and later spread throughout the entire area of Pax Slavia Orthodoxa. The earliest preserved texts in the Serbian language date from the late 12th century: the Miroslav's Gospel was written in Serbian Church Slavonic, a variety of Old Church Slavonic that incorporated features of the 12th-century Serbian vernacular; the Charter of Ban Kulin (1189), was written in the Serbian vernacular with minimal elements of Serbian Church Slavonic. This means that in the earliest Serbian texts from the late 12th century and the first half of 13th century, features of the Old Serbian vernacular and Serbian Church Slavonic are mixed. Since these are different varieties of the same language, rather than two different languages (e.g., Czech and Latin in the medieval Bohemia), it is necessary to develop a unified

Proceedings of LT4HALA 2026 @ LREC 2026, pages 1–6.

methodological framework for the lemmatization and annotation of both varieties.

For the purposes of this research, we selected a corpus of Serbian medieval charters from the late 12th and early 13th centuries. This selection is justified by the following criteria: a) they are the oldest preserved Serbian texts, b) they are texts with the most elements of Old Serbian vernacular. The list of texts manually annotated[1] in the CoNLL-U Editor (Heinecke, 2019) is given in Table 1.

ID	Title	Tokens
№1	The Signature of Grand Prince Stefan Nemanja and Prince Miroslav on the Latin-written charter to Dubrovnik (1186)	16
№2	The Charter of Ban Kulin to Dubrovnik (1189)	208
№3	The Signature of Prince Miroslav on the Latin written charter to Dubrovnik (1190)	4
№4	The Charter of Grand Prince Stefan Nemanja (Monk Simeon) to the Hilandar Monastery (1198–1199)	898
№5	The Charter of Grand Prince Stefan Nemanjić (Saint Sava) to the Hilandar Monastery (1207–1208)	1519
№6	The Charter of Grand Prince Stefan Nemanjić to Dubrovnik (1214–1217)	220
№7	The Charter of King Radoslav to Dubrovnik (1234)	489
№8	The Charter of Prince Andrej to Dubrovnik (1214–1235)	177
	TOTAL	3531

Table 1: Initial document collection (manually annotated)

The table shows that these texts are of varying lengths. Documents №1 and №3 are only fragments with a small number of tokens. On the other hand, there are more extensive documents (e.g., №4 and №5) with several dozen sentences and several hundred tokens. The documents addressed to Dubrovnik (№1, №2, №3, №6, №7, №8) are related to trade and therefore are written in Old Serbian with minimal presence of Serbian Church Slavonic features. Conversely, the documents addressed to the Hilandar Monastery on Mount Athos (№4, №5) predominantly feature Serbian Church Slavonic, with Old Serbian vernacular present to a lesser extent.

The potentially relevant UDPipe models were old_church_slavonic-ud-2.12-2307017 (PROIEL), old_east_slavic-birchbark-ud-2.12-230717 (Birchbark), and old_east_slavic-torot-ud-2.12-230717 (TOROT), considering the similarity of language and text domains to our data, and their declared performance. The domains and vernaculars on which these models were trained were very similar to our data, but with corpus-specific transcription conventions, none of which quite matched our data. Hence, we opted for the Old Church Slavonic model due to its best declared performance (see Table 2).

UDPipe2 Model	UPOS	Lemma
old_church_slavonic-proiel-ud-2.12-230717 (PROIEL)	**96.39**	**90.18**
old_east_slavic-birchbark-ud-2.12-230717 (Birchbark)	88.50	65.58
old_east_slavic-torot-ud-2.12-230717 (TOROT)	95.40	88.09

Table 2: Declared performance of selected UDPipe2 models on UPOS and lemma (F1 score)

3. Lemmatization Principles

Lemmatization of historical language varieties is complex due to lack of knowledge, semantic homonymy, and diachronic/orthographic lemma variation as well as diglossia, with variants often co-occurring within the same text (Pichhadze 2017). Semantic homonymy affects even very frequent words, e.g., *съ* represents both the preposition *with* and the determiner *this*.

When lemmatizing the texts from the corpus, we adhered to three main principles: a) the principle of orthographic normalization, b) reconstruction of the presumed linguistic state from the late 12th and early 13th centuries, and c) the principle of preserving variant lemma forms that reflect specific phonetic features of Old Serbian and Serbian Church Slavonic.

The principle of orthographic normalization implies that lemmas are given in the modern Cyrillic script, with the addition of several specific letters for phonemes present in Old Serbian and

[1] The annotation of the texts was carried out solely by the first author of the paper, and therefore the evaluation of interannotator agreement was not applicable.

Serbian Church Slavonic vocalism at the end of the 12th and beginning of the 13th centuries: the letters <ь>, <ъ>, and <ы> for the semivowel [ə], the vowel *yat* [ě], and the vowel *yeri* [y].

The principles of lemmatization were developed not only for the specific needs of this work but also for the lemmatization of the future electronic corpus of Old Serbian and Serbian Church Slavonic as a whole. Therefore, the second principle involved reconstructing the lemma always according to the presumed state of the phonological system common to Old Serbian and Serbian Church Slavonic at the end of the 12th and beginning of the 13th centuries, regardless of later linguistic development. Since both linguistic varieties are present in the charter texts, the third principle involved retaining variant lemma forms that reflect their specific phonetic features: for example, Old Serbian *noć* and Serbian Church Slavonic *nošt* for *night* or Old Serbian *ja* and *jaz* and Serbian Church Slavonic *az* for the first-person singular pronoun. These lemmas will be linked through a ModernLemma attribute in the MISC column.

4. Universal PoS and Morphosyntactic Features

Old Serbian and Serbian Church Slavonic are grammatically much closer to Old Church Slavonic than to modern Serbian. Therefore, when adapting the tag set for the annotation of parts of speech and morphosyntactic features, we started from the tag set for Old Church Slavonic developed within the PROIEL project (Eckhoff 2018). Unlike the Old Church Slavonic tag set, which does not include PART, PUNCT, and SYM, our tag set encompasses all universal PoS tags. The category of determiner (DET), which is not common in traditional grammatical descriptions of the Serbian language and its historical varieties, has been introduced as a tag for pronominal adjectives to ensure consistency with Old Church Slavonic and modern Slavic languages (Zeman, 2015). The SYM tag is used to mark the cross that appears in the texts either in place of a ruler's signature or as a symbol of divine invocation at the beginning of charters. The issue of tagging participles, which are characteristic of other Slavic languages as well (Zeman 2016: 145), has been resolved by consistently tagging them as VERB, even though they also have nominal features such as Case, Gender, Number, or Variant. We deviated from this principle only in a few instances where the verb lemma could not be reconstructed for adjectives derived from participles, but even in such cases, the feature VerbForm=Part was included with the adjective.

Partial adaptations of the Old Church Slavonic set were also implemented when marking morphosyntactic features and values. In the category of nouns, one can find Polarity=Neg for nouns with lexical negation. A small number of indeclinable nouns have the feature InflClass=Ind. Proper nouns, in addition to the features Case, Gender, and Number, also have the feature NameType, which can take the values NameType=Geo (geographical name, toponym), NameType=Giv (given name), NameType=Sur (surname), NameType=Pat (patronym), and NameType=Nat (ethnonym). In the category of first and second-person personal pronouns, reflexive pronouns for all persons (*sebe*, *se*), and nominal pronouns (*kto* and *što*), the Gender feature is omitted. Adjectival pronouns are always tagged as DET, with the feature PronType, which can have the values PronType=Dem, PronType=Rel, PronType=Int, PronType=Ind, PronType=Neg, and PronType=Tot. To the morphosyntactic features of adjectives that are marked in Old Church Slavonic (Case, Degree, Gender, Number, and Variant), we have added Poss=Yes for possessive adjectives, as well as Polarity=Neg for adjectives with lexical negation. In the category of numerals, we consistently marked the features NumForm (with values NumType=Word for numbers written in words and NumType=Cyrill for numbers written with letters in numerical value) and NumType (with values Card, Ord, and Sets). The most significant change in the category of verbs involved the method of marking aorist and imperfect. For aorist, we used VerbForm=Past as in Old Czech, and for imperfect, VerbForm=Imp (Zeman et al., 2023)[2].

5. Tagging Results

Among the three pre-trained UDPipe2 models in consideration (Table 2), Old Church Slavonic has best declared performance, and therefore we used it for the pre-processing.

Our initial manual annotation batch consisted of 3,531 tokens (eight complete documents). We evaluated the three pretrained models on this batch (henceforth *3k*), with results as listed in Table 3 and 4, using gold tokenization.

Although we expected some decrease of model performance on *3k*, it being a different vernacular, the margin was surprising: in UPOS, the Old Church Slavonic model dropped by 32.05. Similarly, TOROT dropped by 15.62 points, while Birchbark by mere 6.6 points, suddenly ranking first among the pre-trained UDPipe2 models.

In lemmatization, Old Church Slavonic ranked again worst, dropping by almost 60 points.

[2] For marking these forms, we cannot use the Aspect feature because it would conflict with the lexical aspect that has developed in Slavic languages (Zeman et al., 2023).

Birchbark and TOROT outperformed it by ten points and with almost identical results.

UDPipe2 Model	UPOS	UPOS diff
old_church_slavonic-proiel-ud-2.12-230717 (PROIEL)	64.34	**-32.05**
old_east_slavic-birchbark-ud-2.12-230717 (Birchbark)	**81.90**	-6.6
old_east_slavic-torot-ud-2.12-230717 (TOROT)	79.78	-15.62

Table 3: Performance of selected UDPipe2 models on UPOS (F1 score), on the *3k* Old Serbian dataset, with differences from the declared performance

UDPipe2 Model	Lemma	Lemma diff
old_church_slavonic-proiel-ud-2.12-230717 (PROIEL)	30.95	**-59.23**
old_east_slavic-birchbark-ud-2.12-230717 (Birchbark)	**40.83**	-24.75
old_east_slavic-torot-ud-2.12-230717 (TOROT)	40.81	-47.28

Table 4: Performance of selected UDPipe2 models on lemma (F1 score), on the *3k* Old Serbian dataset, with differences from the declared performance

Facing such a performance drop in the pretrained UDPipe2 models, we trained several UDPipe1 models to include *3k* in different combinations with Old Church Slavonic, Birchbark, and TOROT train data.[3] We used the default parameters of the udpipe_train function in the R library udpipe (Wijfels, 2023). We measured the performance of the UDPipe1 models by a ten-fold cross-evaluation, where the train data contained all train data of the UD selected treebanks and 9/10 of *3k* and the test data contained 1/10 of *3k* (Figure 1).

6. Discussion

The experimental results show that none of our UDPipe1 models reliably outperforms the pre-trained old_east_slavic-birchbark-ud-2.12-230717 UDPipe2 model in UPOS. The greatest advantage of Birchbark is possibly the domain similarity: practical written communication, administrative notes, business and church records written in a vernacular. When training UDPipe1 models, no combination with 3k outperforms a model trained on 3k alone. We have also included a cross-validated result for a UDPipe1 model trained just on the Birchbark 2.12 training data to assess the performance difference between the UDPipe1 and UDPipe2 models trained on the same data and applied to 3k. The performance of 46.29 was poor (note that the UDPipe2 Birchbark model reached 81.9). Adding *3k* raised the performance of the Birchbark based UDPipe1 model to 78.54, which is anyway less than 3k alone, yet with a somewhat smaller dispersion. The pretrained UDPipe1 models old_church_slavonic-proiel-ud-2.5-191206 and old_russian-torot-ud-2.5-191206 also performed on *3k* much worse than their UDPipe2 counterparts.

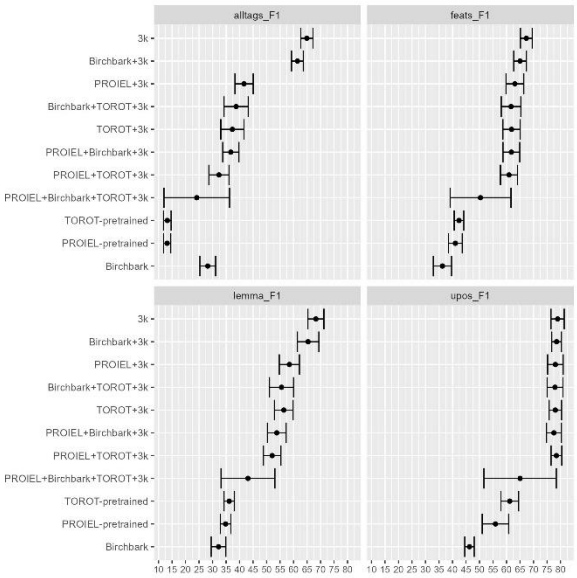

Figure 1: Performance of UDPipe1-trained models (average F1 in a ten-fold evaluation)

We analyzed the errors made by the pre-trained models for individual parts of speech. The most striking finding was that Old Church Slavonic 2.12 fails to identify punctuation, classifying colons as ADJ and periods as NOUN, due to a specific lemmatization, which interprets them in words and the parts of speech are then parts of speech of those words. However, it was not its only issue on the Old Serbian domain. When we compared the proportions of correct guesses except punctuation marks (once for the entire dataset), the models ranked the same, and the difference from TOROT 2.12 (second) remained apparent, although it had decreased.

All three models were above 90% in CCONJ and NOUN. Birchbark and TOROT also reached this mark in ADP (Old Church Slavonic below 80%) and PUNCT. Birchbark fails to detect INTJ (which are rare) and almost fails to detect ADV and AUX (but all models perform poorly here); it is also

[3] Unfortunately, only the older UDPipe1 tool has a user-friendly API to train custom models, whereas doing so with UDPipe2 requires a significant programming expertise and processing capacity.

noticeably worst at PRON. For more details see Figure 2.

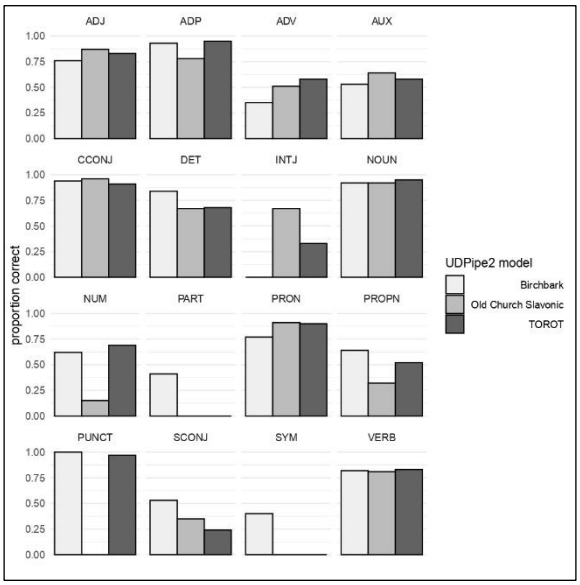

Figure 2: Proportion of correct guesses by UPOS and model

Regarding lemmatization, the performance of Old Church Slavonic and Old East Slavic models is sub-optimal (as partly expected), which can be explained by different principles of orthographic normalization of lemmas. More precisely, they assign the Old Serbian tokens to a correct lemma according to the conventions of their train data, which differ from those for Old Serbian. A good example is numerals written in letters (in Old Church Slavonic, numbers can be represented by letters with the corresponding index in the Cyrillic alphabet, e.g. *а* would stand for 1, *в* for 2, etc.). They pose no problem for UPOS tagging: only the lemmatization is completely different. On the other hand, the orthography of many pronouns is so different that they do not even get recognized as pronouns.

Another difference (among many) is that the traditional normalization of Old Serbian does not systematically use *ъ* and *ь* as endings. For illustration, we can mention the token *богъ* (*God*), which we lemmatize as *бог*, while the Old Church Slavonic and Old East Slavic models lemmatize it as *богъ*.

7. Conclusion and Future Work

From our experiments with three pre-trained UDPipe2 models and a selection of custom trained UDPipe1 models we conclude that our next annotation batch of Old Serbian texts will benefit from pre-processing with the old_east_slavic-birchbark-ud-2.12-230717 model – at least regarding UPOS. We intend to run the same experiment with the Stanza (Qi et al., 2020) models, whose declared performance on the three relevant languages is higher than UDPipe2's.

We will repeat this series of experiments incrementally with each new annotation batch of 5,000 tokens, including the evaluation of lemma and the universal features. As soon as the manually annotated corpus reaches 20,000 tokens, we will gradually add syntactic dependencies, using the most suitable model for their preprocessing, and finally contribute it to the UD github repository to make it eligible for the regular model training workflow of the UDPipe2 developers. The current dataset is available in the Zenodo repository (Polomac, 2026).

8. Acknowledgments

The work was carried out within the research grant *Universal Dependencies for Old Serbian and Serbian Church Slavonic: Creating a Training Dataset for Lemmatization and Morphosyntactic Annotation Using UDPipe*, supported by the Computational Literary Studies Infrastructure (CLS INFRA) project. This output was created with the support of the Ministry of Education, Youth and Sports and the Operational Program Johannes Amos Comenius within the project Open Science II (reg. No. CZ.02.01.01/00/24_030/0015041). The work described herein has also been using data, tools, and services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062).

9. Bibliographical References

de Marneffe, M.-C., Manning, C. D., Nivre, J., Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 47(2): 255-308.

Eckhoff, H., Bech, K., Bouma, G., Eide, K., Haug, D., Haugen, O. E. and Jøhndal M. (2018). The PROIEL treebank family: a standard for early attestations of Indo-European languages. *Language Resources and Evaluation*, 52: 29–65. URL: https://doi.org/10.1007/s10579-017-9388-5.

Heinecke, J. (2019). CoNLLU-Editor: a fully graphical editor for universal dependencies treebank files. In *Proceedings of the Third Workshop on Universal Dependencies*, pages 87–93, Paris, France. Association for Computational Linguistics. URL: https://aclanthology.org/W19-8010.

McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., Bedini, C., Bertomeu Castelló, N. and Lee, J. (2013) Universal Dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–

97, Sofia, Bulgaria. Association for Computational Linguistics. URL: https://aclanthology.org/P13-2017.

Pichhadze, A. (2017). Razmetka cerkovnoslavjanskih i drevnerusskih tekstov: problemy lemmatizacii. *Filologija*, 68: 143–155.

Polomac, V. (2026). UD_Old_Serbian-DiaC [Data set]. Zenodo. https://doi.org/10.5281/zenodo.19317842

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C.D. (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, Stroudsburg, USA. *Association for Computational Linguistics.* URL: https://aclanthology.org/2020.acl-demos.14/

Straka, M., Hajič, J. and Straková, J. (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, Poreč, Slovenia. European Language Resources Association. URL: https://aclanthology.org/L16-1680.

Zeman, D. (2015). Slavic languages in universal dependencies. In K. Gajdošová and Adriána Žáková (eds.), *Natural Language Processing, Corpus Linguistics, E-learning*, RAM-Verlag, Lüdenscheid: RAM-Verlag, pp. 151–163.

Zeman, D. (2016). Universal Annotation of Slavic Verb Forms. *The Prague Bulletin of Mathematical Linguistics*, 105 (2016): 143–193.

Zeman, D. (2018). *The World of Tokens, Tags and Trees*. Studies in Computational and Theoretical Linguistics, vol. 19, first ed. ÚFAL MFF UK, Praha.

Zeman, D., Kosek, P., Březina, M. and Pergler, J. (2023). Morphosyntactic Annotation in Universal Dependencies for Old Czech. *Jazykovedný časopis/Journal of Linguistics*, 74: 214–222.

10. Language Resource References

Honnibal, M., Montani, I., Van Landeghem, S. and Boyd A (2020). spaCy: Industrial-strength NaturalLanguage Processing in Python. doi:10.5281/zenodo.1212303.

Straka, M. (2018): UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. In *Proceedings of CoNLL 2018: The SIGNLL Conference on Computational Natural Language Learning*, pp. 197-207. Association for Computational Linguistics, Stroudsburg, PA, USA.

Wijffels, J. (2023). udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit. URL: https://CRAN.R-project.org/package=udpipe