TNA ACTIVITY REPORT

Universal Dependencies for Old Serbian and Serbian Church Slavonic: Creating a training data set for lemmatization and morphosyntactic annotation using UDPipe

Author: Vladimir Polomac

Current position: PhD, Full Professor

Affiliation: University of Kragujevac (Serbia), Faculty of Philology and Arts, Department of Serbian Language

Host institution: Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics

Mentor(s): Silvie Cinkova

Period of stay: 01.06.2024 – 30.06-2024.

Research Question

The recent development of the UDPipe tool enables users to do automatic lemmatization and morphosyntactic annotation for contemporary languages, for many classical languages (e.g. Ancient Greek, Paleo-Hebrew, Coptic, Gothic, Sanskrit, Latin), as well as for older versions of modern Indo-European languages (e.g. Old French). When it comes to older versions of Slavic languages, UDPipe enables automatic text processing with the aid of models for Old Church Slavonic and Old East Slavic languages. These models are created from materials extracted from different types of texts. Initial steps in automatic lemmatization and morphosyntactic annotation via UDPipe have been recently conducted for Old Czech.

Given these circumstances, the general aim of our project is to create a starting data set that would be used to train models for automatic lemmatization and morphosyntactic annotation of the oldest preserved Serbian texts from the 12th and 13th century with the help of the UDPipe tool. Special goals of the project include: a) defining the principles for the lemmatization of Old Serbian and Serbian Church Slavonic texts, b) creating a set of tags for morphosyntactic annotation in accordance with the Universal Dependencies standards for Old Church Slavonic and modern Slavic languages, c) manually creating a starting data set for the training model with UDPipe.

CLS INFRA has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101004984

The general and special goals of the project are defined with the help of the following theoretical and methodological frameworks: a) the basic principle of lemmatization should revolve around the reconstruction of Old Serbian or Serbian Church Slavonic forms of lemmas supposedly used in the 12th century; b) if we follow the principles of homogeneous diglossia, in Serbian texts of the 12th and 13th century many elements of Old Serbian vernacular (which greatly differs from contemporary Serbian language, especially in the system of verbs) and of Serbian Church Slavonic (Serbian version of the Old Church Slavonic language) are interwoven. Thus, when one is defining the set of tags, it is necessary to start with Old Church Slavonic, making certain corrections in accordance with the set of tags for modern Serbian language and other contemporary Slavic languages.

Research Visit and Outcomes

The visit to the Institute of Formal and Applied Linguistics was used primarily for (1) establishing principles for the lemmatization of Old Serbian and Serbian Church Slavonic manuscripts, (2) adapting the tagset for the morphosyntactic annotation of Old Serbian and Serbian Church Slavonic manuscripts, and (3) creating and evaluating the model using the UDPipe tool.

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. 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 Serbian Church Slavonic vocalism at the end of the 12th and beginning of the 13th centuries. The principles of lemmatization were developed not only for the specific needs of this work but also for the lemmatization of the 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.

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. 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. 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 is characteristic of other Slavic languages as well, 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, in addition to Case, Gender, and Number, also 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 Geo (geographical name, toponym), Giv (given name), Sur (surname), Pat (patronym), and 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 Dem, Rel, Int, Ind, Neg, and 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 Word for numbers written in words and 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.

Creating and Evaluating the model using UDPipe tool: preliminary results

After establishing the principles for lemmatization and morphosyntactic annotation, we proceeded to create the initial dataset for model training. For these purposes, we selected the oldest preserved Serbian charters from the late 12th and early 13th centuries. This selection is justified by the fact that these are the oldest preserved Serbian manuscripts with significant elements of the Serbian vernacular. The initial dataset was obtained by first lemmatizing and annotating the charters using the UDPipe2 model for Old Church Slavonic and then manually correcting them using the Conll-U editor. This resulted in an initial training dataset containing about 3,500 tokens.

The first experiment we conducted involved evaluating previously trained UDPipe2 models for Old Church Slavonic and Old East Slavic on our dataset of 3,500 tokens. This experiment showed that the Old East Slavic UDPipe2 model, based on birchbark letters (old_east_slavic-birchbark-ud-2.12-27077), had the best performance, at least regarding part-of-speech tagging, with an accuracy of 81%. The second experiment involved training specific UDPipe1 models that combine our data with the data from existing models for Old Church Slavonic and Old East Slavic. The results showed that none of the combinations achieved the accuracy demonstrated by the UDPipe2 model for Old East Slavic birchbark letters.

Future Work

From our experiments with three pre-trained UDPipe2 models and a selection of customtrained 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 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.