A Unified Annotation of the Stages of the Bulgarian Language. First Steps

Fabio Maion, Leopold-Franzens-Universität Innsbruck, Fabio.Maion@uibk.ac.at
Tsvetana Dimitrova, Institute for Bulgarian Language, Bulgarian Academy of Sciences, cvetana@dcl.bas.bg
Andrej Bojadžiev, St. Kliment Ohridski Sofia University, aboy@slav.uni-sofia.bg

THE PROJECT

The goal of the project *A Unified Annotation of the Stages of the Bulgarian Language (AUSBUL)* is to create a model for infrastructure for researcher-friendly online access to texts and annotated data.

The infrastructure integrates several components:
1. A corpus of texts in Cyrillic that are formatted according to uniform criteria, suitable both for electronic publication and for linguistic annotation.
2. Linguistic annotation (morphological and syntactic; plus lemmatization with reference to earlier (attested) and modern variants of the words) that follows standardized methods adopted in corpus linguistics and established by practice.
3. Linking the texts with their electronic descriptions, along with a catalogue of their sources.
4. Metadata such as: information about the authors (and editions of the manuscripts and/or texts), references to dates and places found in the texts or other information that is necessary for understanding their context.

TEXTS FOR EXPERIMENTING

Tagging two different versions of the Acts of the Apostle Thomas in India (BHG 1800 – 1801, CANT 245.II, Bonnet, 1903) available in: the (archaic) Damaskin of Kostenets (Kosten) and the (vernacular) Koprivštitsa Damaskin (Kopriv).

The texts were segmented in sentences following the respective editions with the end of a sentence being the respective punctuation mark (full-stop; colon; etc...).

We adopted a method developed by Šimko et al. (2021) for the edition of the Pop-Punčov Sbornik that allows us both to keep information on word boundaries in the manuscript and to provide the tagger with linguistic input coherent with modern practices. Special signs were added to indicate some word boundaries (as in цѧ̀р| же lit. 'king thus': a token is written together with the following token in the manuscript, but the tokens are analyzed as two units for the annotation; and ѿ_и́де́ть 'to go': a token which is analyzed as one unit, but it is divided in the manuscript.

THE TAGGING

We used the Stanza tagger version modified to use bidirectional character-level LSTM by default and specifically adjusted for POS (for low-resource languages) by Y. Scherrer (2021). For lemmatisation, we used Lemming (Müller et al., 2015). The annotated texts are stored in the CoNLL-U format and follow the conventions for Universal Dependencies (Petrov et al., 2012). We tested two different datasets to train a model using them as training data:
Tag1: the damaskini texts annotated by I. Šimko (2021). Here, the training data was in the Latin alphabet, so we created a script that transcribes the Cyrillic letters into their Latin counterparts and strips the texts of all the superscripts and diacritics to tag a graphically simplified version of the texts.
Tag 2: the annotated Old Church Slavonic texts from the PROIEL (Eckhoff et al., 2018) and the TOROT (Eckhoff and Berdičevskis, 2015) corpora. The data is linguistically less similar to our texts than the data by Šimko (2021) but contains much more tokens (around 357.000). We did not use the original data but adapted it to some linguistic peculiarities of the Bulgarian language following Maion (2022).

Acts of the Apostle Thomas in India, Kostenečki		
Element	**Tag1 (Pop-Punčov)**	**Tag2 (Dioptra)**
не 'not'	PART	ADV
же 'thus'	PART	ADV
бо 'because'	CCONJ	ADV
ли (interrogative particle)	PART	ADV
Demonstrative pronouns (сь 'this (over here) ', тъ 'this', онъ 'that')	PRON, ADJ, DET	PRON, ADJ
Possessive pronouns (мои 'my', твои 'your'...)	ADJ	PRON
да 'to'	CCONJ	ADV, SCONJ
Auxiliaries	AUX	VERB
Passive participles	ADJ	VERB; ADJ
Proper names	NOUN	PROPN

Acts of the Apostle Thomas in India, Koprivštenski		
Element	**Tag1 (Pop-Punčov)**	**Tag2 (Dioptra)**
не 'not'	PART	ADV
же 'thus'	PART	ADV
бо 'because'	CCONJ	ADV
ли (interrogative particle)	PART	ADV
Demonstrative pronouns (сь 'this (over here) ', тъ 'this', онъ 'that')	PRON, ADJ	PRON, ADJ, DET
Possessive pronouns (мои 'my', твои 'your'...)	ADJ	PRON, ADJ
да 'to'	CCONJ	ADV, SCONJ
Auxiliaries	AUX	VERB
Passive participles	ADJ	VERB; ADJ
Proper names	NOUN	PROPN

Text	POS	Morphology
Kosten – Tag1 (Pop-Punčov dataset)	91.44%	82.29%
Kosten – Tag2 (Dioptra tagset)	92.36%	89.56%
Kopriv – Tag1 (Pop-Punčov dataset)	95.03%	93.62%
Korpiv – Tag2 (Diotpra tagset)	73.97%	65.18%

SOME RESULTS

The tagger achieves the greatest accuracy with the vernacular Korpiv when trained with the Pop-Punčov dataset (Tag1) and the lowest accuracy with Kopriv and trained with the Dioptra dataset (Tag2). The POS-tagging of the archaic Kosten was better when the tagger was trained on the Dioptra dataset (Tag2) than with the (vernacular) Pop-Punčov dataset (Tag1). When the tagger was trained with the Pop-Punčov dataset (Tag1) comprising texts from the same period, its results on both texts were much closer than when it was trained with the Dioptra dataset (Tag2).

Most errors on POS-level are found when the vernacular Kopriv text was tagged with the tagger trained with the Dioptra dataset (Tag2). The results for morphological annotation are lower (and for lemmatisation are even lower) but they are also linked to the accuracy of the POS-tagging.

IN PERSPECTIVE

Results are similar to those from previous attempts at tagging early Slavic texts but are still lower due to the character of the texts (they are Bulgarian and from a later period). Except for the normalization method with statistical CRF-tagger MarMoT and a neural network tagger, Scherrer et al. (2018) experimented with applying Modern Russian resources to pre-modern data to show that transfer experiments did not improve tagging performance significantly, but state-of-the-art taggers still reached between 90% and more than 95% tagging accuracy even without normalization. J. Besters-Dilger (2021) applied neural network tagger CLStM to the Old Russian Žitie Evfimija Velikogo (GIM, Chud. 20), a copy of the second half of the 14th century. The tagger was successfully applied on non-normalised text with high accuracy – however, unknown words (which means those that had not been "seen" by the tagger before) still showed a higher error rate.

This research is carried out as part of the project "A Unified Annotation of the Stages of Bulgarian Language (AUSBU)" funded by the Bulgarian National Science Fund and the OeAD under the Programme Bulgaria: Competitions for Financial Support for Bilateral Projects, Science & Technological Cooperation (WTZ) Austria / Bulgaria No. КП-06-Австрия / 2, 18.07.2023 / OeAD-GsmbH (Österreichischer Austauschdienst) (BG 09/2023, WTZ Bulgarien S&T Bulgaria 2023-25).