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

Abstract

The paper reports on an ongoing work on a proposal of guidelines for unified annotation of the stages in the development of the Bulgarian language from the Middle Ages to the early modern period. It discusses the criteria for the selection of texts and their representation, along with some results of the trial tagging with an existing tagger which was already trained on other texts.

Keywords: tagging, historical data, Bulgarian language

1 Introduction

So far, research on Middle Bulgarian and early Modern Bulgarian texts has not followed systematically applied methods of corpus linguistics, and no attempts have been made to integrate the corpora with electronic descriptions (and editions) of texts and manuscripts. Separate corpora of texts reflecting different stages of the history of the Bulgarian language are part of other large corpora which also contain Old Russian and Old Serbian sources[1], as well as historical material of other Indo-European languages (Greek, Latin, Armenian, among others[2]). Numerous texts have been gathered in the Historical corpus of the Bulgarian language (from 10th to 17th century)[3].

In addition, the annotation schemes followed by the research efforts (if available at all) differ quite a lot as they reflect specific research purposes and cannot be uniformly applied. This means that the results of indexing and data analysis cannot be based on uniform criteria and cannot be used for comparison purposes. In fact, some textual collections contain only searchable texts (concordances can be made using other external tools) without any further linguistic annotation, active links to dictionaries or other useful tools.

The present paper reports on the development of an ongoing project which aims at offering a proposal for a unified annotation for the Bulgarian texts of all stages in the history of the Bulgarian language. We have started from the annotation principles of the PROIEL project[4] to extend the linguistic annotation to be further used for Middle Bulgarian and early Modern Bulgarian texts while considering other recent efforts in this direction (Šimko, 2021; Šimko et al., 2021). A dictionary of words and their wordforms is generated, and an applicable model for description of texts and manuscripts[5] is to be integrated to allow for electronic publication of the texts along with applicable metadata. A similar project[6] for Middle Bulgarian dealing with the translation of Philip Monotrop's work Dioptra in the 14th century, is

[1] http://www.manuscripts.ru/
[2] http://dev.syntacticus.org/#annotation-principles
[3] https://histdict.uni-sofia.bg/textcorpus/list
[4] https://github.com/proiel
[5] http://repertorium.obdurodon.org/
[6] https://m.pf.fwf.ac.at/en/research-in-practice/project-finder/57346

being developed at the University of Innsbruck. In this project, the Slavonic translation and the Greek original are lemmatized, morphologically annotated and aligned on the word level, thus yielding a searchable bilingual corpus.

2 The approach

The goal of the project *A Unified Annotation of the Stages of the Bulgarian Language (AUSBUL)* is to create a model for infrastructure which will made the texts and the annotated data online accessible and user-friendly for researchers and other potential users. 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 being enriched with linguistic annotation.

2. Linguistic annotation (morphological and syntactic annotation; 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 in the corpus with their electronic descriptions, along with a catalogue of their sources.

4. Metadata to the texts 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.

The project draws upon the idea that only such an integrated approach, devoted to sources from one era and resting on standardized practices and solutions, can provide new insights into the history of language and literature, to be further compared with similar phenomena from other cultures. The emphasis of our effort is to test for the possibility of applying the principles of corpus and computational linguistics to a selection of different types and genres of medieval texts. The result will be a unified model for describing the language and the texts of the Middle Bulgarian (13th – 14th centuries) and the Early Modern Bulgarian (17th – 18th centuries) periods with respect to both archaic and vernacular samples. By following uniform principles of electronic linguistic annotation, we may compare and analyze different phenomena in the development of the grammatical system of the Bulgarian language.

At this first stage of the project, work is being done in several directions:

1. Selection of the texts to be included in the corpus.

This part of the project requires a considerate analysis of the history of the texts. For example, some texts must be selected to make possible the comparison of the earlier (or archaic) variants with the Early Modern Bulgarian variants that reflect a more vernacular version of the language.

2. Brief description of the sources (manuscripts) in which they are found.

3. Compilation of a bibliography for the texts and their sources (manuscripts).

4. Trial annotation with texts of different genres and periods.

3 The texts

We will illustrate the approach with one of the texts currently being worked on, the Acts of the Apostle Thomas in India (BHG 1800 – 1801, CANT 245.II, Bonnet, 1903). The preliminary selection of the witnesses (sources) was made after first comparing the earlier witnesses from the 14th and 15th centuries with the copies from the 17th – 18th centuries. One representative of the text was selected – a manuscript from the Dragomirna Monastery No. 700 from the 15th century (Yufu, 1970; Nencheva, 2023). The manuscript contains a different, hitherto unknown redaction of the text (Velcheva, Bojadžiev, 2006), which would complement the observations made up to this point (Mitani, 2020). Transcripts from the 17th – 18th centuries are divided into two groups. The first group involving the archaic version of the text, dating back to the 14th – 15th century witnesses, which, is represented by the Damaskin of Kostenets (CHAI 503, second half of the 17th century, Petkanova-Toteva, 1965: 54; Hristova et al., 1982: 212). The second group contains the variants of the text according to the new Bulgarian damaskini and miscellanies from the 17th – 18th centuries. This second group is heterogeneous, with the following representatives:

1. *Koprivštitsa Damaskin*, second half of the 17th century (Miletič, 1908). The codex is now of unknown location; the text is used according to the Miletič's edition.

2. *Damaskin of Protopopintsi* from the 17th century (NBKM 708; Tsonev, 1923: 339–347).

3. *Miscellany of* Joseph Bradati from 1740 (NBKM 1058; Stoyanov, Kodov, 1963: 327–333; Dimitrova, Bojadžiev, 2009);

4. *Damaskin of Pope Todor Vračanski* from 1789 (NBKM 1062; Stoyanov, Kodov, 1963: 349–355).

All these witnesses are representatives not only of different Early Modern Bulgarian versions of the text, but also reveal different directions in the development of the Bulgarian language in the 17th – 18th centuries.

4 Resources and tests

We have experimented with applying linguistic annotation on two different versions of the texts – on the (archaic) *Damaskin of Kostenets* (Kosten) and the (vernacular) *Koprivštitsa Damaskin* (Kopriv). Before starting with the automatic linguistic annotation, we had to apply some preprocessing to our texts. The first step consisted of segmentation on the sentence level where we relied on the segmentation in the respective editions, i.e., we split the sentences wherever we found a punctuation mark indicating a sentence boundary (such as full-stop, colon, Georgian paragraph separator).

The problem of word segmentation (and tokenization) was more intricate as word segmentation in the manuscripts does not regularly correspond to modern practices. We adopt 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 taggers with linguistic input coherent with modern practices. Special signs were added to the text indicating word boundaries:

- A vertical line means that a token is written together with the following token in the manuscript, but the tokens are analyzed as two units for the morphological annotation (e.g., цѧр | же lit. ‘king thus’, where же is a discourse particle, stands for цѧрже in the manuscript but it is given as цѧр же in the annotated version).
- An underscore is added where a token which is analyzed as one unit in the morphological annotation, is divided into two tokens in the manuscript (e.g., the verb w̄_иⷣдéть ‘to go’ with the prefix w̄ stands for w̄ иⷣдéть in the manuscript but w̄иⷣдéть in the annotated version).

As we are currently in the starting phase of the linguistic annotation and aim at finding out the best way to tag our data, we tested two different models for the tagging process. At first, we used the damaskini texts annotated by I. Šimko (2021) to train a model using them as training data. The tagging was performed using the Stanza tagger version, which was modified to use bidirectional character-level LSTM by default and specifically adjusted for parts-of-speech (for low-resource languages) by Y. Scherrer (2021). As this tagger does only perform part-of-speech tagging (POS-tagging) and morphological annotation but no lemmatisation, we had to use another tool for this purpose – 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).

Before we could use the model based on the data of (Šimko, 2021) to tag our texts, we had to apply one further step of preprocessing. As all the training data was in the Latin alphabet, we created a script that transcribes the Cyrillic letters to their Latin counterparts and strips the texts of all the superscripts and diacritics. We, thus, performed the linguistic annotation on a graphically simplified version of the texts that matched the training data by (Šimko, 2021). The obtained results will henceforth be referred to as Tag1.

As the number of tokens in the training data was rather small (around 60.000 tokens), we performed a second round of tagging using other data from another source. In this second round, we used 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). When we trained our model, we did not use the original data from the PROIEL and the TOROT corpora but adapted it to some linguistic peculiarities of the Bulgarian language. Prior research has shown that data in such an adapted format provides better results for Middle Bulgarian (Maion, 2022). The result of this second round of annotation will be referred to as Tag2.

5 Results

The results from the tagging using the two annotated datasets differ in elements that may have different linguistic interpretation depending on the purpose of the intended corpus (considering further annotation). The Pop-Punčov dataset (Tag1) follows the MULTEXT-East annotation guidelines with a stricter focus on morphology while the Dioptra dataset (Tag2) closely (although not entirely) follows the PROIEL/TOROT annotation principles (which were directed toward the next step in the syntactic annotation for the purpose of building the PROIEL treebank). Results with Tag1 and Tag2 differ in those elements (and parts-of-speech) which may have different syntactic functions – pronouns, adverbs, particles, auxiliaries. Table 1 and Table 2 below give the differences in marking with each dataset for Kosten and Kopriv.

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

Table 1: Kostenets

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

Table 2: Kopriv

The accuracy of the tagger trained with the respective datasets on the two texts is given in Table 3.

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%

Table 3: Accuracy

Applied POS-tags were not considered erroneous when calculating the accuracy in the following cases:

a. When tags are not part of the dataset as in: PROPN that are marked as NOUN if there is no PROPN in the training dataset; AUX that are marked as VERB if there is no AUX in the training dataset.

b. Possessive pronouns that are marked as PRON or ADJ if there is such marking in the training dataset.

c. да ‘to’ when tagged as CCONJ, SCONJ or ADV.

d. The negative particle не when marked as PART or ADV.

e. Demonstratives marked as DET in noun phrases (mainly with the Dioptra dataset (Tag2).

f. Various conjunctions that are tagged as CCONJ, SCONJ, ADV depending on the tagset and the dataset.

The tagger achieves the greatest accuracy with tagging 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) – in the example below 4 of 9 wordforms are wrong.

И ‘and’	CCONJ
непрѣстàнно ‘ceaselessly’	ADV
бѹ ‘to-God’ ~~VERB~~	NOUN
сѐ ‘self’ (reflexive) ~~DET~~	PRON
млїаха ‘prayed’ ~~NOUN~~	VERB
и ‘and’	CCONJ
слàвѣха ‘praised’ ~~ADJ~~	VERB
сѓаа ‘saint’	ADJ
трo̊ица ‘Trinity’	NOUN

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.

Our trial tagging has shown results that 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.

6 Ongoing work

The next step in our effort is the development of a tagset and annotation principles. At this point, we have decided to keep morphology oriented marking, with some additions that can be beneficial for the further mark-up levels. We have decided to keep PROPN for proper names, and to mark all verbal forms as VERB (as in the Dioptra corpus) and all pronouns as PRON. Adverbials including pronominal adverbials will be marked as ADV while the pronominal adjectives formed with an adjectival suffix (as in въсѣкъ ‘each’, оногози ‘that’, etc.) are marked as ADJ. The forms with the article-like suffixes (such as жената ‘woman.DEF’ and цр҃ѹтомѹ ‘king.DEF’) will be marked as definite forms.

After correcting the results for POS, we expect to achieve better results with the morphological annotation but also with other texts that will be included in the database.

7 Acknowledgements

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 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).

Abbreviations

BHG: Bibliotheca Hagiographica Greaca (Halkin, 1957; Halkin, 1984).

CANT: Clavis apocryphorum Novi Testamenti (Geerard, 1992: 144–150).
CHAI: Church-Historical and Archival Institute, Sofia
NBKM: National Library "St. Cyril and Methodius", Sofia

References

Juliane Besters-Dilger. 2021. Neural morphological tagging for Slavic: Strengths and weaknesses. *Scripta & e-Scripta*, 21: 79–92.

Maximilianus Bonnet. 1903. *Acta Apostolorum Apocrypha.* Post C. Tischendorf denuo ediderunt R. A. Lipsius et M. Bonnet. Partius alterus. Volumen alterium. Acta Philippi et Acta Thomae accendunt Acta Barnabae. Ed. M. Bonnet. Lipsiae: Apud Hermannum Mendelssohn [Phototypische Ausgabe. Darmstadt: Wissenschaftliche Buchgeselschaft, 1959].

Margaret Dimitrova and Andrej Bojadžiev. 2009. Apokrifat za apostol Toma v kasnata damaskinarska traditsia. *Godishnik na Asotsiatsia "Ongal",* 8: 238–260.

Hanne Eckhoff, Kristin Bech, Gerlof Bouma, Kristine Eide, Dag Haug, Odd Einar Haugen, and Marius Jøhndal. 2018. The PROIEL treebank family: a standard for early attestations of Indo-European languages. *Language Resources and Evaluation* 52 (1): 29–65.

Hanne Eckhoff and Aleksandrs Berdičevskis. 2015. Linguistics vs. digital editions: The Tromsø Old Russian and OCS Treebank. *Scripta & e-Scripta*, 14–15: 9–25.

Maruritii Geerard. 1992. *Clavis apocryphorum Novi Testamenti.* Cura et studio Mauritii Geerard. Turnhout: Brepols

François Halkin. 1957. *Bibliotheca Hagiographica Graeca.* T. 2 (*Ioannes Calybita – Zoticus*). 3ème éd. (Subsidia Hagiographica 8a). Bruxelles: Société des Bollandistes.

François Halkin. 1984. *Novum Auctarium Bibliothecae Hagiographicae Graecae* (Subsidia hagiographica 65). Bruxelles: Société des Bollandistes.

Boryana Hristova, Darinka Karadzhova, and Siyka Ikonomova. 1982. *Balgarski rakopisi ot XI do XVIII vek zapazeni v Bulgaria. Svoden katalog*. Sofia: Narodna biblioteka "Sv. sv. Kiril i Metodiy".

Fabio Maion. 2022. Wege zur verbesserten automatischen Annotation des mittelbulgarischen Kirchenslawischen. *Scripta & e-Scripta,* 22: 365–390.

Keiko Mitani. 2020. Slavonic tradition of the Apocryphal Acts of Thomas in India and the MS 1789/700 of the Dragomirna Monastery (Moldavia, Romania). *Scripta & e-Scripta*, 20: 199–225.

Thomas Müller, Ryan Cotterell, Alexander Fraser, and Hinrich Schütze 2015. Joint lemmatization and morphological tagging with lemming. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, pages 2268–2274.

Elisaveta Nencheva. 2023. Panigirikat na dyak Andrey ot 1425 g. spryamo nay-blizkite po sastav yuzhnoslavyanski kodeksi. S*tudia Literaria Serdicensia*, 3(5): 120–176.
https://studialiteraria.eu/sites/default/files/ISSUES/2023_issue_5/pdf/2023_sls_5_120_176.pdf (Last access 2024-04-10)

Donka Petkanova-Toteva. 1965. *Damaskinite v balgarskata literatura*. Sofia: Izdatelstvo na BAN.

Slav Petrov, Dipanjanm, Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of LREC'2012*. Istanbul, pages 2089–2096.

Yves Scherrer, Achim Rabus, and Susanne Mocken. 2018. New developments in tagging pre-modern Orthodox Slavic texts. *Scripta & e-Scripta,* 18: 9–33.

Yves Scherrer. 2021. Adaptation of morphosyntactic taggers. In *Similar Languages, Varieties, and Dialects: A Computational Perspective, Studies in Natural Language Processing*, pages 138–166. Cambridge University Press.

Ivan Šimko. 2021. *Annotated Corpus of Pre-standardized Balkan Slavic Literature 1.1* [Online]. Slovenian language resource repository CLARIN.SI.
http://hdl.handle.net/11356/1441 (Last access 2024-04-08)

Ivan Šimko, Polina Mihova, Olivier Winistörfer, and Anastasia Escher. 2021. *Pop Punčov Sbornik – Digital Edition* [Online]. Zürich: UZH Institute of Slavic Studies.
https://www.punco.uzh.ch/ (Last access 2024-04-10)

Manyo Stoyanov and Hristo Kodov. 1963. *Opis na slavyanskite rakopisi v Sofiyskata narodna biblioteka*. Tom III. Sofia: Nauka i izkustvo.

Benyu Tsonev. 1923. *Opis na slavyanskite rakopisi v Sofiyskata narodna biblioteka*. Tom II. Sofia: Izdanie na bibliotekata.

Boryana Velcheva and Andrej Bojadžiev. 2006. The Slavonic text of Acta Thomae in India. *Scripta & e-Scripta*, 3–4: 95–119.

Zlatka Yufu. 1970. Za desettomnata kolektsia Studion. (Iz arhiva na rumanskia izsledvach Yon Yufu). In *Prouchvania po sluchay II Kongres po balkanistika (Studia Balcanica 2)*. Sofia: BAN, pages 299–343.