Digital Edition of the *Life of St. Petka*

Ivan Šimko
Slavic Seminary
University of Zurich
ivan.simko@uzh.ch

Abstract

This paper presents the construction of a digital edition of multiple versions of the hagiography of St. Petka of Tarnovo. Two related versions are uploaded at first: a Church Slavonic print edition and its later damaskini redaction. Both texts are adapted for user-friendly reading with side-by-side facsimiles. Translations and additional data concerning separate tokens and sentences can be shown up by the cursor on fly. Further metadata will be available for search. Annotation has been adapted for the transitionary status of the language of the texts: it allows us to compare similar morphological forms with various functions. The edition has already been published online and can be used for both teaching and studying.

The texts have been digitalized as a part of a larger project concerning the development of the Balkan areal features.

Keywords: linguistic annotation, corpus linguistics, text analysis

1. Introduction

Resources for quantitative research of the phenomena distinguishing Middle and Modern Bulgarian, which have been considered Balkan areal features (like e.g. postpositional definiteness marking, analytic infinitive or renarrative mood), are limited. The existing digital resources for study of Bulgarian contain primarily modern standardized variety (e.g. BNC, link), while Church Slavonic literature is online represented mostly by older sources (TITUS, link; Obdurodon, link). Furthermore, the requirements of their genre prevent us to observe the contemporary language shifts: medieval scribes did their best to preserve both the contents and the form of the original. The earliest texts, which show the phenomena mentioned above, can be dated back to the 17th century. They have attracted considerable attention by slavists (e.g. Petkanova-Toteva 1965, Demina 1968) and balkanists (e.g. Sonnenhauser 2016) as well. Although some of these texts have been digitalized (e.g. the damaskin of Loveč: link; cf. Mladenova & Velčeva 2013), users have often struggled with basic problems such as character encoding. No edition so far could satisfyingly combine the searchability of a digital corpus, philological exactness of a critical edition, and user-friendliness of a webpage.

To address these needs, we are developing a method to create a unified Balkan Slavic diachronic corpus, which will contain texts from various areas and epochs. Such a corpus will enable us not only to compare these texts with older varieties, but also with present-day dialectal data[1]. To demonstrate the capabilities of this method, we have created a webpage with two model texts processed in this way. This paper will explain the selection of sources and their processing into a simply accessible website.

[1] The paper has been written within the framework of the project *'Ill-bred sons', family and friends: tracing the multiple affiliations of Balkan Slavic*, led by Prof. Barbara Sonnenhauser of the University of Zurich, funded by the Swiss National Science Foundation (SNSF project grant IZRPZ0_177557/1), to whom I would like to express my thanks for support.

2. Source Texts

Both texts are based on the hagiography of St. Petka of Tarnovo (Parascheva of Epibates), written in the second half of the 14th century by Patriarch Euthymius of Bulgaria. This text is not only a precious artefact of the Middle Bulgarian literature, but it also played an important political role as well, being a crucial part of the cult of this ascetic saint, whose relics resided in the royal capital of Tarnovo. For a 17th-century reader (or audience), it preserved the memory of the former glory, kindling early sparks of national consciousness. In the Christian topology, St. Petka presented a figure of a universal ascetic role model; she personalized the idea that the sainthood can be attained by anyone.

The choice of text offers multiple advantages from the philological point of view. First, the differences between Church Slavonic and "simple Bulgarian"[2] of the damaskini are smaller than between Slavic and Greek. The translator could concentrate more on the styllistics than on translation itself. Thus, unusual grammatical constructions were less likely to appear because of the influence of a foreign language: they would more likely appear due to innovations in the target language. Second, there is a relatively large number of editions (so far 19 versions from the 16th to 19th century are known to the author) from various dialectal areas and epochs due to the popularity of the text, allowing us to use it as a reference, without interferences of the differences in content and genre.

The first version processed for our website is the shortened print edition by the monk Moses. The text was adapted to the Church Slavonic of the Resava redaction and published by Božidar Vuković in his *Traveller's miscellany*, first in 1521 in Venice. An online copy of the 1536 edition provided by the Matica Srpska in Novi Sad has been used for our purposes. As few pages of the copy were damaged, the missing words were completed according to other sources, including the related manuscript NBKM 665 of the National Library in Sofia[3], the modern critical edition based on Vuković's prints by Novaković (1877) or the edition based on multiple manuscript versions of Euthymius' original hagiography by Kałużniacki (1901). Vuković's print editions of the text likely reached the early damaskini circles.

First translations of the hagiography into "simple Bulgarian" appear in the 17th century (Kenanov 2009:59). These were often transcribed along the Church Slavonic version of the story. Curiously, the damaskin NBKM 709 from Sliven begins with the "simple Bulgarian" version, switching to Church Slavonic at the end. Our edition is from the Berlin damaskin, which was likely composed in 1803 in Pleven or Svištov (Ciaramella 1996). It is based on a later, 18th-century edition of the hagiography, reflecting a Moesian dialect. Unlike the earlier editions, this version does not end at the translation of St. Petka's relics by King John Asen II, but continues with their fate after the Ottoman conquest of Tarnovo[4], and in the end adds an original exegesis. An earlier fragment of this edition can also be found in the damaskin of the Church Archive in Sofia (CIAI) 133 from Pleven. The source copy was provided by the Jagiellon University of Cracow (signature Slav. fol. 36).

3. Processing

The texts were manually transcribed by the method developed for the diachronic corpus. As they are quite short (Vuković's edition: 2222 tokens, Berlin edition: 4852), use of automatic tools (like

[2] The term is based on the headings of the texts authored by Damaskēnos Stouditēs: *metaphrastheis eis tēn koinēn glōssan* 'translated into the common language' (e.g. Stouditēs 1751:5). Church Slavonic editions translate the phrase literally *ob'štymь ęzykomь*, while their translations into early modern Bulgarian use adjectives *prostymь* 'simple' or *bolgarskymь* 'Bulgarian'.

[3] NBKM 665 from Serbia includes both a synaxar- and the shortened panegyric (few pages are missing) version of the hagiography, as well as the liturgical service for the saint. Vuković's edition reflects the shortened panegyric *Life* from NBKM 665 or a similar handwritten source. Conev (1923:176) and recently Mineva (2005:5) date the manuscript already to the second half of the 15th century, at least a century before the emergence of the damaskini tradition.

[4] The source for this information was likely the printed *Menaion* of Demetrius of Rostov (1689), which was known among damaskini circles (cf. Kenanov 2009:59). This source included a new shortened edition, based on the original panegyric hagiography by Patriarch Euthymius (according to personal communication with Jürgen Fuchsbauer).

Transkribus, link) was unnecessary. For the sake of cross-platform compatibility, a set of Latin-based characters has been used, which is compatible with the UTF-8 format and which can easily be converted to the Cyrillic alphabet. The digital transcripts do not reflect graphemes serving rather ornamental functions (e.g. spirits). Broad initials (e.g. <ѻ> for /o/) and space-saving variants (e.g. the <*T*>-like character for /t/, adopted from the Greek cursive) were not distinguished; only *paerčik* (reflected as apostrophe <'>) and double gravis (<">, often representing a word-final /i/) are reflected in the transcript. Accent markers are mostly written with vowel characters together (e.g. *á*). If the accents are not fully compatible with the vowel character, they are written as separate characters following the vowel (e.g. *ę́* for a <ѧ> with an acute).

Additionally, the transcripts include auxiliary markers. The character <+> at the end of a token marks orthographic words: clusters of monosyllabics written together due to orthography, including articles and the negative particle *ne*. For example: ᾽иѳпѳстинїе́то 'and in the desert' is rendered as four separate tokens (*i+ u+ pustinïé+ to*). The marker <_> reflects the separation of a token over two lines or pages (e.g. *mnó_go* 'much'), or the separation of verbal or adjectival prefixes, which is common in the damaskini (e.g *zlató_juzdnïĭ konè* 'golden-bridled horses'). If the line break is marked in the text with a hyphen, the marker <-> is used instead of <_> (e.g. *pro-lét'nĭ* '[of the] spring'). Cyrillic numbers are marked with two asterisks <*> (e.g **a** '1'), as their actual marking in the source varies, but they do not occur in the two selected texts.

The texts are split into a table of tokens, which can be converted into an .xml file. A token possesses the following structure in the source table:

token	diplomatic	lemma	PoS tag	sentence id	syntactic position	syntactic dependency	dependency type
pétky	*petki*	*Petka*	NFSGY	1	5	4	NMOD

A diplomatic transcript is created for each token to simplify search queries, using a smaller set of Latin characters. In this layer, sets of graphemes representing the same phoneme are removed. For example, letters <и й ı ї ы> are all reflected by specific characters in the first column, but diplomatized as *i* in the second. Accentuation, punctuation, and auxiliary markers are also removed (e.g. *i+u+pustinïé+to* > *i u pustinie to*). This layer makes the lemmatization easier, and helps us to train automatic recognition of morphological markers, which is relevant for further work on the diachronic corpus.

Lemmatization itself is based on various sources. The most specific dictionary, based on the Tixonravov damaskin (Demina et al. 2012), was supplemented by Church Slavonic (Miklosich 1865, Cejtlin 1994) and dialectal Bulgarian dictionaries (*Etymological Dictionary of BAN* 1972-2006). Lemmatization helps us to cope with orthographic differences, especially in texts radically adhering to phonetic principles and using non-Cyrillic scripts (e.g. *ζιδβέενιτδ* for *živenie+ to* 'the Life' in NBKM 1064), as well as in texts, which follow the orthographic norms only loosely (such as the writing of <ы> in the damaskini).

Individual tokens are annotated of morphological and syntactic features. The tag set has been customized not only to reflect the "simple Bulgarian" of the damaskini era, but also Church Slavonic. Morphological annotation (marked as "PoS tag" on the page) contains in most cases a single tag based on to the MultextEast standard. The tag set corresponds closely to the variant developed for Croatian (link). The tags include a marker for the part-of-speech category of the token (e.g. an N for a noun, A for an adjective etc.), which is followed by further information on number, tense, and other morphological categories.

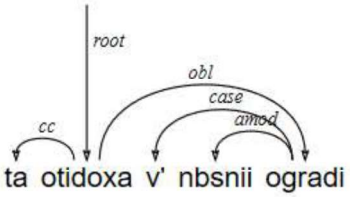
The new tag set was designed to limit the number of semantic and syntactic features. Thus a *da* would be tagged always as a "conjunction" (i.e. C), even if it serves as a marker of the infinitive or future tense, for such compounds are reflected by the syntactic annotation well enough. If the token bears features of two part-of-speech categories, it can receive an additional PoS tag ("alt.PoS tag" on the

page). For example, Church Slavonic verbal participles show features of both verbs (tense) and adjectives (gender, case). Thus the clause *kъ crstvoujuštomou prišьd'ši grádou* 'as [she] came to Tsarigrad' (Vuković) receives tags as follows:

diplomatic	lemma	PoS tag	alt. PoS tag
kъ	*k*	SD	
crstvujuštomou	*carstvuvati*	AMSDY	VMPP-S
prišъdъši	*priiti*	VMPA-S	AFSNN
gradou	*grad*	NMSDY	

Morphological ambiguities are resolved according to Church Slavonic inflection paradigms and phonological shifts. For example, an *-i* in *ja*-inflected words like *dši* 'soul'.DAT.SG is handled as a proper dative, and the token is tagged as NFSDY. In words of the *a*-inflection it is handled as a genitive marker, i.e. [*Žitie*] *Petki* 'Petka'.GEN.SG is tagged as NFSGY. If the same marker is used for multiple options within the same paradigm, the first option in the traditional order (e.g. N-G-D-A-V-L-I for nominal cases) is used. Thus, an *-i* in a phrase like [*ou*] *kašti* 'house'.LOC.SG (NBKM 328) is tagged as a "dative" (NFSDN). This procedure allows us to mark the shape of morphemes with ambiguous functions: e.g. a MASC.SG (*o-/jo*-inflection) word ending in an *-a* is always tagged as a "genitive" (e.g. *bga* 'God'.OBL.SG is tagged as NMSGY), even if the ending reflects a shortened article (e.g. *diavola se prestruvaše* 'the Devil was changing himself', Xrulev 1856) or a nominal count form (e.g. *četýri pъ́prišta* 'four shots', Tixon. d.). The PoS tag can be viewed on the website in a hover box, when moving the cursor over a word.

Syntactic annotation is based on the Universal Dependencies (link) scheme, which was designed in order to be applied to any human language. It works well enough for the transitionary varieties of early modern Bulgarian. The scheme is based on marking dependency relations (e.g. NMOD reflects a nominal modifier of the head: e.g. *pétky* in *pámétь pétky* 'remembrance of Petka'), using numbered syntactic positions within the range of a single sentence (denoted by the sentence id number). An extension layer was further added to provide closer information on the position of articles and demonstratives (e.g. P_NOM if they follow a noun), on the spatial relations denoted by an oblique modifier (e.g. LAT for the lative relation, LOC for locative etc.). For example, the dependency relations in the sentence *ta wtídoxa v' nbsny" ográdĭ* 'and they went into the heavenly gardens' (Berlin d.) can be visualized in the following way (using Arborator; link):



The scheme can be represented in an .xml table:

diplomatic	sentence	position	dep.	dep. type	dep. ext.
ta	19	1	2	CC	
otidoxa	19	2	0	ROOT	
vъ	19	3	5	CASE	
nbsnii	19	4	5	AMOD	
ogradi	19	5	2	OBL	LAT

While this layer of annotation is already used in the analyses of our corpus, it has not yet been integrated to the webpage. Only sentence numbers are shown in the text. The first token of the sentence also shows the translation of the whole sentence or of the following subordinate clause. Last

tokens of the line or page were marked with respective signs as well. First tokens of each page also contain the link to the original scan of the page.

An .xml table for both texts was generated by Excel. The .xml is transformed into an .html webpage by the Oxygen editor (link), using a customized stylesheet. Two versions were produced: one using the Cyrillic script, another one with the Latin transcription. After minor manual modifications, the webpages and scans of the originals were provisionally uploaded to the webspace provided by the University of Vienna. The website with both scans and the transcripts can be accessed via the following link - https://homepage.univie.ac.at/ivan.simko/. The description page (link) also contains source files (both in Excel and .xml format), as well as further information about the tag set and the stylesheet

4. Perspective

The existing webpage is very basic: it does not include any scripts, nor source data can be accessed by now. The next step will be the implementation of a visual representation of the syntactic annotation (e.g. using exported images from Arborator or a tabular diagram) and the development of the search widget, which would be able to process the annotation data, provided at the website. It is possible the website will be supplemented by additional pages reflecting other versions of the hagiography, given the facsimile are available to the author. These may include the print editions by Demetrius of Rostov (1689) and in Sophronius' *Nedělnik* (1806), as well as the well-preserved damaskini of Sliven (NBKM 709), Drjanovo (NBKM 711), Pop Punčo (NBKM 697) or Hadži Gendo (NBKM 1064) and other sources. In this way the website could find use both for teachers of Bulgarian literature to illustrate diachronic developments to students in a modern way, and for scholars studying these developments themselves.

References

Cejtlin, R.M., et al. (1994). *Staroslavjanskij slovar': po rukopisjam X-XI vekov*. Moskva: Russkij jazyk.

Ciaramella, R. (1996). Novi danni za Berlinskija damaskin. *Palaeobulgarica* XX (3), 120-129.

Conev, B. (1923). *Opis na slavjanskite răkopisi v sofijskata narodna biblioteka. Tom II*. Sofia: Narodna biblioteka.

Damaskēnos Stouditēs (1751). *Thēsauros Damaskinou tou Hypodiakonou kai Stouditou tou Thessalonikeōs*. Enetia: Nikolaos Glykeos.

Demetrius: Dmitrij ep. Rostovskij (1689) *Kniga žitij svjatyx*, 1. Kiev: Lavra Pečerskaja.

Demina, E.I. et al. (2012). *Rečnik na knižovnija bălgarski ezik na narodna osnova ot XVII vek*. Sofia: Valentin Trajanov.

Etymological Dictionary of BAN: Georgiev, V.I. ed., (1972-2006). *Bălgarski etimologičen rečnik, tom I-V*. Sofia: BAN.

Kałużniacki E. (hrsg., 1901). *Werke des Patriarchen von Bulgarien Euthymius (1375-1393)*. Wien: Carl Gerold's Sohn.

Kenanov, D.V. (2009). Žitija i služby na sv. Petka Tărnovska v staropečatnite slavjanski knigi. *Biblioteki. Četene. Komunikacii. Sedma nacionalna konferencija V. Tărnovo 21.-22.11.2008*. Veliko Tărnovo: UI, 57-65.

Miklosich, Fr. (ed., 1865). *Lexicon palaeoslovenico-graeco-latinum.* Vindobona: Guilelmus Braumueller.

Mineva E. (2005). *Pet ximnografski tvorbi za sv. Petka Tărnovska (XIII-XV v.)*. Academia.edu (link), 18.06.2020.

Novaković, St. (1877). *Život sv. Petke patrijarha bugarskoga Jeftimija*. Predano u sjednici filologičko-historičkoga razreda jug. ak. 30.5.1877. 48-59.

Petkanova-Toteva, D. (1965). *Damaskinite v bălgarskata literatura*. Sofia: BAN.

Velčeva, B. & O.M. Mladenova (2013). *Loveški damaskin: novobălgarski pametnik ot XVII vek*. Sofia: Nacionalna biblioteka "Sv.Sv. Kiril i Metodij".

Vuković, Božidar (1536). [*Traveller's miscellany*]. Venice. (link)

Sophronius: Sofronij ep. Vračanskij (1806). *Kyriakodromion sireč Nedělnik - Poučenie*. Râmnic: ep. Nektarij.

Sonnenhauser, B. (2016). "The Balkan Manner of Narration": Narrative Functions of the *l*-Periphrasis in Pre-Standardized Balkan Slavic. *Balkanistica* 29, 1-42.

Tixon.d.: Demina, E.I. (1968). *Tixonravovskij damaskin. Bolgarskij pamjatnik XVII v. Issledvanie i tekst, 1*. Sofia: BAN.

Used Manuscripts of the National Library in Sofia: NBKM 328, NBKM 665, NBKM 697 (Pop Punčo's miscellany), NBKM 709, NBKM 711 (Drjanovo B damaskin), NBKM 1064

Used Manuscripts of the Church Historical and Archive Institute in Sofia: CIAI 133.

Used Manuscripts of the Jagiellon University Library in Cracow: Slav. fol. 36 (Berlin damaskin).

Transcriptions from Cyrillic are based on Church Slavonic ISO 9 standard (link) with minor customizations (link).

All hyperlinks used in the text refer to the state of 18.06.2020.