

Annotated Corpus of Pre-Standardized Balkan Slavic Literature

Technical Description

corpus link - <http://hdl.handle.net/11356/1441>

version date - 02.07.2021

Ivan Šimko, University of Zurich

1. Introduction.....	2
2. Background.....	2
3. Data Sources.....	4
4. Data Structure	5
5. Morphological Annotation	6
5.1. Nouns.....	7
5.2. Adjectives	8
5.3. Pronouns	10
5.4. Numerals	12
5.5 Verbs.....	12
5.6. Adverbs.....	14
5.7. Prepositions.....	14
5.8 Particles	15
5.9. Other Parts of Speech.....	16
6. Syntactic Annotation	16
6.1. Standard Tags	16
6.2. Tag Extensions.....	17
7. Annotation Examples	19
8. Character Sets.....	22
Literature	23

1. Introduction

The following document describes the properties of our corpus of pre-standardized Balkan Slavic literature. The sources for this corpus are *pre-standardized*: they can be dated in the period before the debate on the standardization of modern Bulgarian began to level the existing dialectal differences between the literary traditions. In the present stage, the earliest sources can be dated to 16th century, the latest to the first half of the 19th century.

The linguistic category of *Balkan Slavic* encompasses roughly the Bulgarian, Macedonian and adjacent dialects in Serbia (Torlak). While this category can be understood in geographic terms, it is primarily a linguistic one: it represents varieties of Slavic, which have undergone convergent developments together with neighboring non-Slavic varieties in the area, including Albanian, Romani, Romance, modern Greek and Gagauz Turkish. Not all of the sources have been produced in the area with a majority of populace speaking a Balkan Slavic dialect. Philological metadata for each source specify approximate affiliation of authors or scribes within broader dialectal areas (e.g. West Macedonia, Northwestern Bulgaria...) or literary standards (Church Slavonic and its redactions).

Finally, the corpus includes works of *literature*. Each source can be seen as an integral product of a single scribe, but it also can be situated within a broader scope of a literary tradition, which in many cases reaches beyond the borders of a national literature or the respective period. Each source reflects not only phonetic and grammatical features of the scribe's individual language variety, but also the relation between the scribe and existing literary traditions and the local dialect: the scribe's desire to sanctify the text by using archaic grammatical features, or to democratize it by stylistics of secular storytelling and the use of popular loanwords. Thus, each source includes implicit information about its purpose: enlightenment of a non-educated auditory, preservation of a collective memory and symbolics, and so on.

The purpose of the corpus is to provide a resource basis for philologically informed quantitative studies concerning the development of Balkan Slavic languages, focusing on the features characteristic of this group within the Slavic family.

2. Background

Balkan Slavic dialects have undergone complex morphosyntactic changes in the last (at least) five centuries. For example, case inflection, which can be observed in the oldest (Old Church Slavonic, OCS) texts from the Balkan area and other Slavic languages, has been limited to the pronominal paradigms, which are also simplified to the point of a simple distinction between a nominative and a single oblique case, with some dative forms. Morphosyntactic functions formerly expressed by inflectional cases received new analytic means of expression. Synthetic possessive constructions with genitive (e.g. *duxъ otъca vašego* 'the spirit of your father' in OCS; Lunt 2001:146), dative (*syni světu* 'the sons of the light'; Lunt 2001:149) or adjectives (*učenici ioanovi* 'the disciples of John'; Lunt 2001:146) have been replaced by constructions with clitic dative pronouns (e.g. *nai mladīa+ mu sīnъ* 'his youngest son', PPS 80v), prepositions (e.g. *sīnъ na+ stārago crĕ* 'son of the old king', *Berl.d.* 182r) or both (e.g. *tāmo pripādna na+ ikona+ ta+ i na+ presfetāe bogorōdīza* 'there she knelt in front of the icon of the Holy Mother of God', NBKM 1064 31r).

Another example: Balkan Slavic dialects developed a system of obligatory definiteness marking. The syntactic position of the definiteness marker has been fixed to the first element of the noun phrase (the postponed article), or as the first element itself (extended demonstrative pronouns like *tozi*, indefinite marker *edin*). Similar markers are attested already in the oldest OCS texts, but rather for

deictic definiteness than for anaphoric relations and the definiteness of generally known objects. Instead, OCS marked anaphoric definiteness by the distinction of long and short form of adjectives (cf. §5.2.3.). Translations of Greek texts also prefer the use of long forms for the translation of Greek articles, the demonstratives translating other determiners (e.g. *ho kakos doulos ekeinos* -> *zъi rabъ tъ* 'that bad servant'; Mt 24:48 in *Cod.Zogr.*). Sources from the 17th century reflect a change in this practice, freely using clitic demonstratives along long forms and distal extended demonstratives (e.g. *isplъzě on'zi smókъ wt+ město+ to+ si* 'the dragon crawled out from his lair', *Tixon.d.*, Demina 1972:57).

The function of the mentioned archaic and innovative forms is not the same. The modern replacements for old possessive constructions can be used for marking indirect objects (as in the mentioned sentence *támo pripádna na+ ikona+ ta+ i*) as well. Different forms of articles may be used (at least in literature) to mark the syntactic position of the noun phrase (e.g. FEM.SG article in *koito gi turixa vъ patriarcheskoтo cъrkvо* 'who put them into the Patriarch's church', *Nedělnik* 1856:258). Thus, it would not be fully correct to describe them as merely "possessive" resp. "definiteness" markers.

Thanks to the amount of available pre-standardized literature, it is possible to observe the individual developments of these features over time. However, literary conservatism of the scribes complicates the precise dating of the changes. A possible solution is to statistically measure the frequency of particular features in various periods of time. Especially popular works with many editions throughout the centuries are very fruitful for this kind of research. Such works can be used to construct parallel corpora, which give results less distorted by the differences in genre or writer's style. Any attempt at the construction of such a corpus, however, requires the ability to unambiguously distinguish the aforementioned changes. The corpus must be able to reflect the full diachronic (as well as diatopic) variation of morphs (or, rather, their written forms) employed throughout the observed period in its annotation. As each morph carries lexical, morphological and syntactic information, the struggle is to make this annotation economic.

The preparation of such a corpus requires the following steps:

- i. **Selection of a sample.** The sources are chosen from various dialectal areas and periods. The sample should represent major trends in both literature and dialects of the Balkan Slavic area.
- ii. **Digital edition.** The variety of fonts and orthographic freedom of the scribes requires a character set capable of reflecting most of the characters used in the period. Some of these (e.g. spirits) can be omitted due to their mostly ornamental function. A common UTF-8 Latin-based character set is used to reflect works written in the alphabets used in our sources - Cyrillic, Greek and Latin. Most of our texts have clear word boundaries, but unlike today, they write clitics, articles and other monosyllabic words/particles together with longer words. Such complex orthographic words are transcribed as separate tokens (e.g. *и ѱпѱстиніѱто* -> *i+ u+ pustinіѱ+ to* 'and in the desert', *Berl.d.* 180r). The transcription is supplemented by additional columns with original (if Cyrillic) and diplomatic forms of the tokens. The diplomatic layer uses a simplified character set, removing sets of graphemes used for a single phonem (*и ѱ ѱ ѱ ѱ ѱ* -> *i ѱ ѱ ѱ ѱ* -> *i*), as well as accentuation, interpunction, and auxiliary markers (e.g. for orthographic words, numbers). This layer makes the lemmatization easier, und also helps us to train automatic recognition of morphological markers.
- iii. **Lemmatization.** Individual literary sources have different vocabularies. The words uncommon in modern literary Bulgarian were lemmatized mostly according to the specific dictionary based on the *Tixonravov damaskin* (Demina et al. 2012), supported by Church Slavonic (Miklosich 1865, Cejtin 1996) and dialectal Bulgarian (*Bulgarian Etymological Dictionary* 1972-

2006) dictionaries. In rare cases of loanwords not reflected in the dictionaries mentioned above, new lemmas were introduced using roots from Turkish or other donor languages. The lemmatization helps us also to cope with orthographic differences, especially in works radically adhering to phonetic principles.

- iv. **Annotation.** Tags are assigned to individual words (tokens) to describe their morphological and syntactic features. The tagset has been customized to reflect not only productive features of modern Bulgarian and Macedonian, but also those of Church Slavonic and of other Balkan Slavic dialects. Morphological annotation contains a tag based on the MultextEast (ME) standard. Syntactic annotation is based on the Universal Dependencies (UD) system. As these are the main tools for capturing the morphosyntactic variety of Balkan Slavic, we will explain this step in more detail in the following paragraphs.

3. Data Sources

The sources of the corpus are digital transcripts of various manuscripts and printed books originating (or written by authors or scribes from) the Balkan Slavic area. The sources were chosen in order to cover as much dialectal areas and periods as possible. However, an exact localization in space and time is not always possible, especially in case of older manuscripts. So far the sources are labelled by the conventional name (abbreviated) of the edition. Details about the respective sources are published in a separate file.

source	language variety	text
<i>Vel.s.</i>	Church Slavonic	<i>Tale of Alexander the Elder</i>
Vuković 1536	Church Slavonic	<i>Life of St. Petka</i>
<i>Kiev.d.</i>	Church Slavonic	<i>Life of St. Mary of Egypt</i>
<i>Tixon.d.</i>	Lukovit-Teteven	<i>Life of St. Petka</i>
<i>Lov.d.</i>	Trojan	<i>Homily against Drinking</i>
<i>Ljub.d.</i>	Kotel-Elena	<i>Life of St. Petka</i>
NBKM 328	Samokov	<i>Legend of St. Thais</i>
<i>Sv.d.</i>	Moesian	<i>Life of St. Mary of Egypt</i>
<i>Jan.s</i>	Samokov	<i>Tale of Fathers slain at Sinai and Raithu</i>
<i>Temski r.</i>	Torlak	<i>Homily on Punishment of Children</i>
NBKM 1069	Panagjurište	<i>Homily on Divination</i>
NBKM 370	Slavenobulgarian	introduction
<i>Ioann.d.</i>	Vratsa	<i>Life of St. Petka</i>
<i>Berl.d.</i>	Moesian	<i>Life of St. Petka</i>
NBKM 1423	Paulician	various
PPS	Vidin-Lom	various
<i>Nedělnik</i> 1806	Slavenobulgarian	<i>Life of St. Petka</i>
Krčovski 1814	Slavenobulgarian	<i>Second Homily of Daniel</i>
NBKM 1081	Pirdop-Koprivštica	<i>Daniel's Prophecy</i>
NBKM 1064	Sliven	<i>Life of St. Petka</i>
NBKM 728	Tetovo	<i>Life of St. Petka</i>
<i>Rai.d.</i>	Smoljan	<i>Second Homily of Daniel</i>
<i>Nedělnik</i> 1856	standard Bulgarian	<i>Life of St. Petka</i>

The corpus contains multiple versions of the *Life of St Petka* (or *Parascheva*), a Church Slavonic hagiography composed by Patriarch Euthymius of Tarnovo in the late 14th century. The oldest source (Vuković 1536) contains a shortened version, which was later added to manuscripts containing a

translation of homilies by Damaskēnos Stouditēs (†1577), hence called *damaskini*. Among these, the damaskini of Tixonravov (*Tixon.d.*), Ljubljana (*Ljub.d.*), Lovech (*Lov.d.*), and Berlin (*Berl.d.*) were included to represent the four major editions of damaskini manuscripts in the popular language ("simple Bulgarian"), as described by Demina (1968:53-64). The late Greek-script damaskin *NBKM 1064* also belongs to this text tradition. Manuscript *NBKM 1423*, written with Latin letters, represents another, Catholic tradition of hagiographic texts. Most of the other manuscripts among our sources are also eclectic compositions of hagiographies and homilies from various authors (*NBKM 328, 728, 1069, 1081, PPS, Temski r.*). Among these, the *sbornik* of Pop Punčo (*PPS*) will be released with full annotation as a whole later. The printed *Nedělnik 1806* by Sophronius, Bishop of Vratsa (†1813) shows a similar eclectic structure as the damaskini too.

Although the language of the damaskini is also sometimes considered a "literary standard", as it combines features of various Balkan and Western dialects of Bulgaria (cf. Velčeva 2001, Mladenova 2007). However, later examples (*Berl.d.*, *NBKM 1064*) deviate much from the 17th century norms. For this reason, we classify them according to dialectal areas of the place of their approximate origin (based on Stojkov 2002). The language of the collections *NBKM 1069* and *1081* shows considerable influence of the language of the early damaskini, too.

The language varieties include two quasi-standards. The first is the traditional Church Slavonic, used in Vuković 1536, *Kiev d.* and *Vel.s.*, the last of which also shows strong influence of the Macedonian dialects. The second is *Slavenobulgarian*, a variety developed on the basis of Western Bulgarian dialects with many archaic features in the 18th century. Its major example is the *Slavenobulgarian Chronicle* by hieromonk Paisius of Hilandar. The corpus includes the introduction of the *Chronicle* from the transcript produced in Elena (*NBKM 370*). The texts of Josif Bradati (*NBKM 328*) and his students (*Jan.s.*, *Ioann.d.*) represent a variant of this norm. The choice is aimed at covering an as wide as possible area (also see the map 9.1. at the end of the article):

dialectal area	source
Balkan	<i>Ljub.d.</i> , <i>Lov.d.</i> , <i>NBKM 1064, 1069, 1081, Tixon.d.</i>
Macedonia	<i>NBKM 728, Vel.s.</i>
Moesian	<i>Berl.d.</i> , <i>Sv.d.</i>
Northwestern	<i>PPS</i>
Rup	<i>NBKM 1423, Rai.d.</i>
Torlak	<i>Temski r.</i>
Western	<i>Ioann.d.</i> , <i>Jan.s.</i> , <i>NBKM 328</i>
Slavenobulgarian	Krčovski 1814, <i>NBKM 370, Nedělnik 1806</i>
Church Slavonic	<i>Kiev d.</i> , Vuković 1536

The majority of the sources were provided as digital copies by the respective libraries (*Ljub.d.*, Krčovski 1814, *NBKM 370, 728, 1064, 1069, 1081, 1423, Nedělnik 1806, PPS, Tixon.d.*, Vuković 1536) or other researchers in the field (*Berl.d.*, *Kiev d.*). The rest (*Ioann.d.*, *Jan.s.*, *Lov.d.*, *Sv.d.*, *Temski r.*, *Vel.s.*) was transcribed on the basis of critical editions.

4. Data Structure

Each source is provided in a separate file. The files use the text encoding UTF-8, which is likely to stay as a standard for a while. They are provided in two formats: the first format is a simple tab-separated text file, which can be open both in table (e.g. MS Excel) and text editors (e.g. Notepad), and is easy to convert to other formats. The files are also provided in the CoNLL-U format, which is compatible with the interface of the repository.

Text tables are per default organized in 12 columns. The first column (TOKEN) contains the main transcript, including interpunctuation, suprasegmental markings, and distinguishing graphematic variants. The second column contains a transcript either to Cyrillic (CYRILLIC) or a simplified character set (DIPLOMATIC), which is more easily searchable. In Cyrillic texts, the diplomatic transcript is in the next column. The column LEMMA provides lemmas in basic forms - nominative singulars for nouns, pronouns and adjectives, and 1st person present for verbs. The basic form was chosen according to widely available Bulgarian and Macedonian dictionaries. If the word is not mentioned in them, other (Church Slavonic, Turkish; cf. above §2.iii) dictionaries were used to determine the lemma. Sources for less common lemmas were included in the column REF.

The following columns include codes (tags) for morphological and syntactic information about the token (POS_TAG, POS_EXT, SENT_ID, UD_ID, UD_NCY, UD_TYPE, UD_EXT). The following column (FOLIO) shows the page number at the first token of the page. The pages are usually numbered per folio on the right side, but multiple numberings (e.g. a Cyrillic one by the author, Arabic by the library) may be given as well. The following column (TRANSLATION) contains the English translation of the whole sentence. It is usually placed at the first token of the sentence. If the sentence is complex or its syntax is broken, the translation may be provided for each logical part (clause), each starting at the first token of these parts. Some of the sources adapted for critical philological editions also have an additional column with information marking end of lines in the original edition (EOL). If multiple chapters are included from a source (like in *PPS* and *NBKM 1423*), beginnings of particular chapters are marked in another column (CHUNK), which also can be used to mark narrative components of a story (e.g. for parallel comparisons between various editions). The last column (REF) contains references to dictionaries, Biblical citations and other relevant sources.

The CoNLL-U files use the standardized structure ([link](#)). Primary morphological tag has been reflected as the "language specific" tag (XPOS). However, format-specific limitations prevent the use of all information, which is available in text tables.

5. Morphological Annotation

The corpus uses a tagset based on the Multext-East system of morphosyntactic annotation ([link](#)). It is roughly based on the specifications for modern Bulgarian, Macedonian and Croatian. The tag is composed of a string of (max. 6) characters, in which the first capital letter defines the basic part-of-speech category, followed by lower-case characters for morphological attributes. For example, the word *jasna* FEM.SG 'clear' receives the tag AFSNN, containing the following information based on the ending *-a*:

A	adjective (part-of-speech)
F	gender: FEM (i.e. not <i>*jasen</i> or <i>*jasno</i>)
S	number: SG (not <i>*jasni</i>)
N	case: NOM (not <i>*jasnu</i> , <i>*jasnyę</i> etc.)
N	(old) definiteness: short-form (not <i>*jasnaa</i>)

Due to transitional character and diversity of the language of our sources, the set was customized to reflect both the features, which became unproductive later, as well as the ones, which are recent. The morphological tags do not reflect the function of the morphem (e.g. definiteness, possession), which is thus left for the interpretation of measurements. They are strictly form-based, reflecting the realization of the word in the annotated text. Some desinences are ambiguous: for example the noun ending *-a* can denote not only FEM.SG.NOM, but also MASC.SG.GEN/ACC in older texts, MASC.SG shortened

article, MASC.DL.NOM, NEUT.PL.NOM; they can be distinguished only by lemmas or context. Thus all the tag positions after the category marker interpret the desinence together.

5.1. Nouns

The tag includes four attributes. We distinguish grammatical gender (1), number (2), case (3), and animacy (4). Gender, number, and (at least in CS) case are marked at all elements of the noun phrase. Animacy is a lexical attribute, but it may have an effect on the paradigm. The case attribute serves to classify the graphic shape of the ending; it does not necessarily denote the morphosyntactic role of the noun. Nouns are classified either according to gender and historical stem classes (*osnova*), based on pre-Slavic SG.NOM endings (cf. Lunt 2001:54), or according to the generalized paradigms for respective genders. This is usually a choice specific for a text, but it may vary between various texts within a source (e.g. in *Tixon.d.*).

5.1.1. Tagset for nouns

Bold script marks options commonly met in more innovative (i.e. non-CS) texts.

- N-gender-number-case-animacy

- Gender:

M - masculine

F - feminine

N - neutral

- Number:

S - singular

P - plural

D - dual

- Case:

- singular:

N - MASC -**ъ**, FEM -**a/b**, NEUT -**o** (hard stem)/**e** (soft stem)

G - MASC/NEUT -**(j)a**, FEM -**e/i**

D - MASC/NEUT -**(j)u**, FEM -**e/i**

A - FEM -**u**

L - MASC -**e**, NEUT -**i**

I - MASC/NEUT -**om/em**, FEM -**oju/eju**

V - MASC -**e/-(j)u**, FEM -**o**

O - MASC -**o/ija**, FEM: -**ъ**

- plural:

N - MASC/FEM -**i/e/ove**, MASC -**ove**, NEUT -**a/i**

G - MASC -**ov**, MASC/FEM -**ъ**, NEUT -**ii**

D - MASC/NEUT -**om/em**, FEM -**am**

L - MASC/FEM/NEUT -**ex/ax**

I - MASC/FEM/NEUT -**mi**

- dual:

N - MASC -**a**, FEM -**e**, NEUT -**i**

G - MASC/FEM/NEUT -**u**

D - MASC/FEM/NEUT -**ama**

- Animacy: **Y/N**

5.1.2. Differences from legacy tagsets

The Bulgarian tagset (BG) marks five attributes. The first attribute distinguishes common and proper nouns. This distinction was deemed redundant, being merely semantic one. Our corpus also contains a number of words ambiguous from this aspect (*Gospod* 'Lord [God]', *Bogorodica* 'Mother of God'). The last position, instead of

marking animacy, denotes definiteness. In our set, the postponed article is usually not treated as a suffix. As long as they include the demonstrative root (see below §5.3.4.), articles are handled as separate tokens. This enables us to compare their use with short demonstrative pronouns used in Church Slavonic (CS). Only shortened articles (-a, dialectal -ъ, -o) are marked on the noun or adjective. When they are graphically identical to a case marker, they are annotated as such (e.g. in *diavola se prestruvaše* 'the Devil clothed himself': NMSGY, *Nedělnik* 1856: 257).

The BG tagset distinguishes only two options for the case attribute: nominative and vocative. Additionally, the Macedonian (MK) set distinguishes an oblique case, too. Our marking of cases is based on the Croatian (HR) tagset, which is used also in sets for other Slavic languages, like Russian and Slovak. The situation in HR is comparable to an earlier stage reflected in CS literature. The option "oblique" has been adopted for situations, where the shape of the ending is different from the CS cases, as defined above.

1. shortened MASC.SG article	-o	<i>stolpo</i> 'the pillar' (PPS)
2. anlaut of a MASC.SG article	-o(t)	<i>posniko+ t</i> 'the hermit' (NBKM 728)
3. borrowed agent nouns	-ija	<i>gemuĭia</i> 'sailor' (Ljub.d.)
4. Bogorov's FEM.SG.OBL marker	-ŏ	<i>čistotŏ</i> '[due to her] purity' (Nedělnik 1856)
5. <i>damaskini</i> FEM.SG.OBL marker	-b	<i>dšb+ta</i> '[she gave] her soul' (Tixon.d.)

5.1.3. Phonetic and graphic ambiguities (e.g. differences between *i/y* or *e/ě*) are resolved by a convention based on the development of the sounds in the approximate area of origin of the text, the writer or his orthographic school (e.g. *y* → *i*, *ě* → *e* for the majority of today's Bulgaria). Ambiguous case markers are marked conventionally according to their final shape, even if the ending would be syntactically interpreted otherwise. The convention is based on the traditional order of cases in paradigms (NOM-GEN-DAT-ACC-LOC-INST). Locative is marked only if its realization would be different from the dative (e.g. *v'+ životě svoémъ*: NMSLN; but *na zemli* 'on Earth': NFDN; *Nedělnik* 1806). The short forms for FEM.SG.GEN and DAT are distinguished according to the stem. Old *a*-stems (ending in a hard consonant) have *-i* (< OCS *-y*) as their SG.GEN ending (e.g. *pet'ki sláva* 'glory of Petka': NFGY; *Tixon.d.*). Old *ja*-stems (with a palatal consonant) have *-e* (< OCS *-e*) here (e.g. *do zemlje* 'to the ground': NFGN; *Tixon.d.*). If an OCS fem.sg.acc *-ŏ* would occur (if we include an older CS text like *Dobr.ev.* or *NBKM 667*), it would be marked as an *o*-case. Old FEM.SG *i*-stems ending in an *-i* are marked always as genitives. Various FEM.ACC.SG forms (*-u*, *-b*) are handled with separate markers; both A and O forms may occur in the same text. Vocatives (which may be graphically indistinguishable from locatives) are marked as such, if their function is clear from the context.

5.1.4. Old *a*-stem MASC nouns (e.g. *bašta* 'father') are handled as FEM nouns.

5.1.5. Marking of dual forms follows gender-specific rules. Some originally dual forms were generalized as plurals (*kraka* 'legs'; *ručě* 'hands'; *oči* 'eyes'), appearing alongside plural forms of adjectives and pronouns in noun phrases. The MASC ending *-a* is handled as a dual in count forms (cf. Mirčev 1978:195; Maslov 1981:149), i.e. when numerals are included in the noun phrase (e.g. *četèri děla* 'four parts': NMDNN; *NBKM 1081*).

5.1.6. Adverbs based on old oblique case forms are also tagged according to the shape of the ending. Adverbs with the full ending are tagged as nouns, if the ending is equal to one present in the paradigm of the respective gender and stem class (e.g. *rédom po+ síčky svěťb* 'all over the world': NMSIN; *Berl.d.*; *zime ot studb tręperaše* 'in the winter she trembled from cold': NFDN; *Ioann.d.*). Participles with such endings are analyzed as adjectives or verbs (e.g. *tičiš'kom* 'running': AMSIN, VMPP-SI; *Berl.d.*; cf. §5.5.6. too). If the adverbial ending differs from those of the known nominal or adjectival paradigms, the token is tagged as an adverb (e.g. *noštjá* 'in the night': R; *NBKM 1064*). Syntactically, these forms are handled as adverbial modifiers (cf. §6.1.3.4.)

5.1.7. The option N for the case attribute represents the nominative or a default form with an unmarked case. The option is used also for abbreviated or incompletely written forms without the case ending (e.g. *iv xà* 'of Jesus Christ': NMSNY; *Kiev d.*). Default option for the number attribute is S. If grammatical gender of the word cannot be specified, for example in Turkish loanwords (e.g. *štè da+e berekètb* 'there will be abundance': N-SNN; *NBKM 1081*), it is left empty.

5.2. Adjectives

Similarly as nouns, adjectives distinguish four attributes. The gender (1) and number (2) are in the same positions. Case (3) is distinguished in Church Slavonic texts in all genders, but in the more innovative texts only in MASC.SG adjectives systematically. Adjectives also vary according to complexity, the old expression of definiteness (4). Degree (5) is used to mark the synthetic comparative used in CS.

5.2.1. Tagset for adjectives

Bold script marks options commonly met in more innovative (i.e. non-CS) texts.

- A-gender-number-case-complexity

- Gender: F/M/N

- Number: S/P/D

- Case:

- shortform singular:

N - MASC -*ъ*, FEM -*а*, NEUT -*е*

G - MASC/NEUT -*а*, FEM -*е/и*

D - MASC/NEUT -*у*, FEM -*е/и*

A - FEM -*у*

V - vocative -*е*

L - MASC/NEUT -*е*

I - MASC/NEUT -*ом*, FEM -*оју/еју*

- shortform plural:

N - MASC/FEM -*и*, NEUT -*а*

G - MASC/FEM/NEUT -*ъ*

D - MASC -*ом/ем*, FEM -*ам*

A - MASC -*е*

L - MASC/NEUT -*ех*, FEM -*ах*

I - NEUT -*и*, FEM -*ами*

- longform singular:

N - MASC -*и*, FEM -*аа*, NEUT -*ое*

G - MASC -*аго*, FEM -*ije*

D - MASC -*ому*, FEM -*еи*

A - FEM -*оју/-ују*

L - MASC -*ом*

I - MASC -*им*

O - MASC -*а/ъ*

- longform plural:

N - MASC -*ии*, FEM -*ие/ее*, NEUT -*аа*

G - MASC/FEM/NEUT -*их*

D - MASC/FEM/NEUT -*им*

A - MASC -*ее*

I - MASC/FEM -*ими*

- (old) Definiteness: Y/N

- (old) Degree:

C - comparative (MASC -*ѣи*, FEM -*ѣиши*, NEUT -*ѣе*)

5.2.2. Differences from legacy tagsets

Adjectives are tagged differently between the respective tagsets: e.g. HR and MK sets mark attributes "type" and "degree" at first two positions, unused in the BG set. The case system is comparable to the HR set, with an additional *o*-case for specific situations (cf. §5.2.5.). Animacy, marked in the HR set, is not included in our system, as it does not have an influence upon the morphology of adjectives.

Degree is marked in HR and MK sets at position (2). We have moved it to the last position due to its rare occurrence in the corpus. Unlike in MK set, it is limited to old comparative suffixes, used in CS texts (e.g. *что" + бо í+ xóтѣše*

tó mou inò ljúbьzněše býti 'for what else would make him more glad?': ANSNNC; Vuković 1536). Positive degree is not explicitly marked by the tag.

Old definiteness or complexity is marked as "Definiteness" in the code, basing on the HR set, although it is not equal: it denotes the difference between short and long forms only. The term is used to ensure the compatibility of the tagset with the infrastructure of the repository. Articles known from innovative varieties are tagged as separate tokens.

5.2.3. In standard Bulgarian, the short form is prevalent throughout the paradigm; besides few lexicalized exceptions (e.g. MASC.SG *vtori* 'second'), only MASC.SG adjectives with an article follow a long form (e.g. MASC.SG.DEF *visokijat* 'the high'). The definiteness marking by the variation of short and long forms seems to be limited to CS and Serbian texts (Lunt 2001:59, 66), but it was also reported in some Moesian dialects up to the modern times as well (Mladenov 1963). However, doubling of letters representing /i/, likely inspired by the MASC.SG.LF, can also be found on unexpected places in the texts (e.g. adverbs like *junaški* 'heroically'). Non-nominative short-forms are rare in the damaskini; the long forms were likely generalized before the loss of inflection. As with nouns, archaic, non- or pre-standardized variants (e.g. *visokъt*, *visokii*, *visoka*) should be all covered by combinations of the boolean longform marker and case.

5.2.4. Ambiguous shortforms are resolved in the same way as nouns, i.e. according to stem and etymology. The MASC.SG.LOC ending *-om* appearing in Serbian sources and *Nedělnik* is handled as I (e.g. *na presvětlomъ prestólě*: AMSIN; Vuković 1536). Abbreviated endings (e.g. *въ+ stmb evglīi* 'in the Holy Gospel': AMSIN; *Kiev d.*) and the irregular adjective *vsem* 'all' (< OCS MASC/NEUT.SG.LOC.SF/LF *vbsemъ* or MASC/NEUT.SG.INST.LF *vbšěmъ*) are resolved in the same way (as AMSIN).

5.2.4.1. The one vocative form is marked only if other elements of the noun phrase have a vocative ending, considered a short form (e.g. *fariseju slěpe* 'o blind Pharisee!': AMSVN; Lunt 2001:142).

5.2.5. The so-called *o*-cases are marked according to the rules defined for nouns. It can be used only with SG long forms: for example, *zliio+ tь* 'the bad' (*Dobr.ev.*) is analyzed as a long form adjective with an *o*-case ending (AMSOY) and as a separate article (PD-MSN)

5.3. Pronouns

All pronouns are tagged according to their type (1), distinguishing e.g. demonstrative and indefinite ones. Person (2) is tagged only in personal pronouns. Gender (3), number (4) and case (5) are tagged in nominal pronouns of relevant types. For economic reasons, adjectival pronouns (e.g. *svoi*, *njakakъv*) are tagged as adjectives (e.g. *někoja carica* 'a queen': AFSNY; but *koi može iskaza* 'who could tell': PQ--N; both *Tixon.d.*). The determiner function of such elements is expressed in the syntactic annotation.

5.3.1. Tagset for pronouns

- P-type-person-gender-number-case

- Type:

D - demonstrative (e.g. *tazi*, *onova*, *-tъ*, *onъ*)

I - indefinite (e.g. *njakoi*)

P - personal (e.g. *az*)

Q - interrogative (e.g. *koi*, *što*, *kolko*)

R - relative (e.g. *deto*, *štoto*, *toko*, *takъvzi*)

X - reflexive (*sebe*, *si*, *se*)

Z - negative (e.g. *nikoi*)

- Person: 1/2/3

- Gender: F/M/N

- Number: S/P/D

- Case:

N - unmarked/general case (e.g. *az*, *ti*, *nie*, *koi*)

G - long G/A forms: *mene*, *tebe*, *nego*, *neja/nju*, *nas*, *vas*, *těx/nix*, *kogo*

D - *mi, ti, nemu/mu, nei/i, nam, vam, im/gim/xim, komu*
 A - short G/A forms: *me, te, go/ego, ja/eę/gu, ni, vi, gi/ix/xi*
 L - e.g. *mně, něm, nix, kom*
 I - e.g. *mnoju, nimi, kym/cěm*
 O - e.g. *tbzi*

5.3.2. Differences from legacy tagsets

The BG and HR sets use longer strings (up to 13 characters) to reflect various nominal and adjectival attributes at once. Our system is closer to the MK set, but it removes the clitic attribute at position (6), which is obsolete due to use of syntactic marking, as well as the definiteness attribute (7), which is only relevant for adjectival pronouns (which we handle as adjectives). The number of options for the case attribute is extended according to the HR set. The *o*-case attribute is used only for FEM.SG.OBL demonstratives (cf. §5.1.2.).

5.3.3. Conventions are set to eliminate paradigmatic and phonetic ambiguities: e.g. *ti* is always tagged PP2-SN as a regular reflex of CS 2SG.NOM *ty*, even if the token is the homonymous dative form. Similarly, 2PL.ACC *vi* is tagged as an accusative (PP2-PA) also if it is used in a possessive or indirect object construction.

5.3.3.1. Such ambiguities can be observed in our CS texts too. *Vel.s.* does not distinguish *i* and *y* consistently, e.g. *ti priličen esi tomu koupcu* 'you look like that merchant'. *Kiev d.* (as well as in Vuković 1536 and partly also in *Tixon.d.*) does distinguish them, so an annotation of both nominative and dative is possible. On the other hand, it does not distinguish 1SG.GEN *meně* (*ne+ sьkrivai ničtōže+ wt+ méně* 'do not hide anything from me') and 1sg.acc *mene* (*edà+ xotet+ li i+ méně* 'would they want me too?'). In other persons, also dative may be ambiguous from this point of view (*těbě+ podbaet+ blsvi_ti+ me* 'it is appropriate for you to bless me'). Thus these forms are annotated in the same way, as in the later sources (PP1-SG).

5.3.3.2. 3rd person forms are lemmatized according to the attested forms. If the text is in CS or simply shows multiple locative and instrumental forms, the preferred lemma for MASC.3SG is *jb*; in texts with the simplified declension, the lemma is *toi*.

5.3.4. Articles, i.e. postponed short demonstrative pronouns, are marked as separate tokens, so that the tagset is compatible with Church Slavonic sources, where their clitic character is unclear. Unlike adnominal demonstratives they occur after the word, for which they serve as a determiner. Marking of articles as separate tokens allows us to analyze their declension, as well as the (dialectally relevant) phonetic character of the hiatus vowel, emerging at the end of MASC.SG.NOM nouns carrying the articles. The position after the noun or adjective is tagged by an extension in the syntactic annotation (§6.2.2.).

5.3.5. Gender and number are conventionally not marked on indefinite, interrogative, and negative pronouns. Case can be marked, if the referent is a person (e.g. *kogò*: PQ---G; *Berl.d.*). Individual sources (*NBKM 370*, *Nedělnik 1806* and *Temski r.*) make a difference between SG and PL.NOM (e.g. *ktò*: PQ---SN, while *koí*: PQ---PN; *Nedělnik 1806*). If the form is ambiguous (e.g. because of using unaccented *koi* for both numbers), number is excluded from the tag. Tags for demonstrative and relative pronouns, as far as analyzable as nominal pronouns, include gender and number (e.g. *koěto*: PR-NSN; *Berl.d.*) too.

5.3.6. Pronouns used in possessive constructions are handled as adjectives (e.g. *negov* 'his') or as personal pronouns (e.g. M.3SG.DAT *mu*). Possessive function is marked in the syntactic annotation as a tag extension: the pronoun is tagged as a dependent on the noun expressing the possessed object (NMOD:POSS, cf. §6.2.1.). The attributes "owner number" and "gender" (used in the HR tagset) were not included in the tagset.

5.3.7. The type of the pronoun is marked according to the attached morpheme to the main root. A pronominal root without a suffix is always marked as interrogative pronoun, even if the author uses it e.g. as a relative pronoun (e.g. *kogí po_tražixu 'i+ nařdoxu čášu* 'as they searched, they found the cup'; PPS):

interrogative	(none)	<i>koga</i> 'when'	PQ
negative	<i>ni-</i>	<i>nikoga</i> 'never'	PZ
indefinite	<i>ně-, (v)sě-</i>	<i>někoga</i> 'sometimes'	PI
CS relative	<i>e-, -ž(d)e</i>	<i>egda</i> 'as'	PR
BG relative	<i>-to</i>	<i>kogato</i> 'as'	PR

The following roots are tagged as pronouns: *gde, egda, eliko, (do)ideže, kak(o), kakъv* (with suffix *-to*), *koga, kogda, kolko, kъde, otneliže, poslěžde, taka, takъvzi, togda, tolko, toliko*.

5.3.7.1. *kako* is tagged Pq also in those sources from the West (e.g. in *NBKM 728, Temski r.*), where it is used instead of the conjunction *kato* (e.g. *se+ eñilb kako srašni zveri* 'he appeared in the form of terrible beasts', *NBKM 728*; *se" přetváraše katô žvěrb strážňb* 'he masked himself as a terrible beast', *Tixon.d.*).

5.3.7.2. As in case of possessives, adjectivally used pronouns (e.g. *tak'vzbi síla* 'such power', *Tixon.d.*; *nekoj wtcb* 'one father/priest', *Jan.s.*) are analyzed as adjectives.

5.4. Numerals

Only one attribute (type) is included in the tag. Due to their lexical and morphological traits, ordinal numerals are marked as adjectives in our system (e.g. *vtóri* 'the second one': AMSNY). If the category is ambiguous, a secondary tag can be added (e.g. PL.GEN *oboix* 'both': ML and AMPGY; count form *dvama* 'two': ML and NFSNN). As with nouns and adjectives, collective numerals and articles are marked as separate tokens (e.g. *dvama+ ta* 'the two'). Compound numerals (e.g. category 11-19: *četir+ na+ iset* '14') are split into multiple tokens (see §6.1.2.). Cyrillic numerals are marked by two asterisks at the beginning and the end of the string in the corpus; the marking in the original text may vary, usually containing a tilde over the letters.

5.4.1. Tagset for numerals

- M-Form

- Form:

- c - alphabetic (e.g. *dl*)
- D - digital/arabic (e.g. 14)
- L - letter (e.g. *četir+na+iset*)
- R - roman (e.g. XIV)

5.4.2. Differences from legacy tagsets

The form attribute is used in the MK set. The "alphabetic" option has been added to mark Cyrillic numerals. In future, it might be used to mark other alphabetic numbering systems (e.g. Greek, Hebrew...) as well. Due to our handling of ordinal numerals as adjectives, other attributes were considered redundant for our system.

5.5 Verbs

We distinguish six attributes for verbs: their syntactic type (1), mood (2), tense (3), person (4), number (5) and aspect (6). The syntactic type (main/auxiliary) is a legacy from standard ME tagsets; it does not denote any morphological feature. Attribute "mood" distinguishes finite forms (indicative, conditional) from non-finite ones (proper infinitive, participles). Tense is distinguished only in synthetic forms (aorist, imperfect, present). Negative prefix *ne* is separated even if it fuses with the stem (e.g. *nja+mam* 'I do not have'). As verbal voice is usually expressed by analytical means, it is not a part of the tag.

5.5.1. Tagset for verbs

- V-type-mood-tense-person-number-aspect

- Type:

- M - main
- A - auxiliary

- Mood:

- I - indicative
- M - imperative
- N - infinitive
- O - conditional

- P - participle
- Tense:
 - A - aorist
 - I - imperfect
 - P - present
- Person: 1/2/3
- Number: s/P
- Aspect:
 - E - perfective
 - I - imperfective

5.5.2. Differences from legacy tagsets

Our set resembles the BG and the HR, but the tag ends at position (5). Among other attributes, gender is not considered a verbal attribute (participles may be tagged as adjectives to reflect it), while voice and negation (i.e. presence of a *ne*-marker) are marked by syntactic annotation. Position (6) is newly filled with an aspect attribute (cf. below §5.5.7).

5.5.3. As with other categories, analytic constructions (future tense, infinitive) are not marked on the morphological level. Thus the tense of the main verb of the future tense is marked with P ("present"). Analytic constructions are marked at auxiliary verbs by syntactic annotation. As with the attribute "type" used for nouns and adjectives, the attribute is preserved in our system to ensure compatibility with other corpora annotated by MultextEast-based tagsets. For example, in *bgъ šte da proslavi stica+ta svoę* 'God will glorify His saint' (*Berl.d.*), the main verb *proslavi* would be tagged as VMIP3SE and the future marker *šte* as VAIP3SI.

5.5.4. In comparison to standardized Bulgarian, pre-standardized varieties also include synthetic infinitives (e.g. *ni+mózi postigna* 'cannot reach': VMN---E; *NBKM 1064*). The same tag is used also for older infinitives with the full desinence (e.g. *móžeš+li+mi dáti* 'could you give me', *PPS*).

5.5.5. Modern BG/MK use (1.) resultative (or *I-*) participles as main verbs of some compound tenses, (2.) passive past participles as main verbs with a passive meaning, and (3.) gerunds (bulg. *deepričastie*), based on old present active participles. In later texts, (4.) narrative constructions with *I-*participles built from imperfect stems occur as well. The resultative and passive suffix is extended by an adjectival ending expressing gender and number; this may be reflected by a secondary tag (only number is reflected by the verbal tag). Separate marking of the voice is unnecessary in newer varieties: it is fully expressed by the tense of the participle, for present passive/past active participles were abandoned:

1. resultative	-I-	<i>što+è tija stóрила</i> 'what she did' (<i>Tixon.d.</i>)	VMPP--E
2. gerund	-ki/-št-	<i>kopaiki grobo</i> 'digging the grave' (<i>NBKM 728</i>)	VMPP-I
3. passive	-n/-t-	<i>što+sà smiréni</i> 'who are meek' (<i>Berl.d.</i>)	VMPA-E
4. narrative	-I-	<i>tečálъ na stáę</i> 'he ran to the saint' (<i>Nedělnik 1806</i>)	VMPI-I

5.5.6. In archaic (CS) texts, which still use both present and past forms and both voices, voice can be explicitly distinguished in the syntactic annotation as a secondary tag of the auxiliary verb (cf. below §6.2.3.2.). If the participle is used as the main predicate of the sentence, inflection can be reflected in the secondary tag (all examples from Vuković 1536):

1. PRS active	-ę/-št-	<i>skázoujušte</i> 'saying'	VMPP-PE	AMPNN
2. PRS passive	-m-	<i>istávajema</i> 'being exposed'	VMPP-SE	AFSNN
3. PST active	-v(š)-	<i>stéžav'sii</i> 'having conquered'	VMPA-SI	AMSNY
4. PST passive	-n/-t-	<i>blsvéнь</i> 'blessed'	VMPA-SI	AMSNN

5.5.6.1. Participles and gerunds directly dependent on nouns (e.g. *crstvujuštomu gradu* 'to the ruling city') are tagged as adjectives. A secondary tag may be mentioned to reflect the verbal attributes. If the participle serves as the main predicate of the sentence, the verbal tag may be used as the primary one. For example, *crstvujuštomu* would receive AMSDY (adjectival) as the primary tag and VMPP-SI (verbal) as the secondary one. In *što+sà smiréni*, the verbal tag is primary.

5.5.6.2. Rare dialectal gerunds with extensions like *-mi* or *-mica* are marked as present active participles as well. The extension was likely based on M.SG dative absolute (cf. Mirčev 1978:243) and generalized for other genders and numbers. Generally, it

is not considered in the analysis of the adjectival case of the participle (e.g. *pláčjuštumi xoděx pōtem* 'I walked the way weeping': AFSAN, VMPP-SI; *Kiev d.*). For similar adverbial gerunds based on instrumentals, see §5.1.6.

5.5.7. Aspect is handled as an inherent feature of the lemma. Previous ME tagsets used different methods: e.g. MK/UA/PL show it on the 2nd position, RU on the 9th, BG/SC do not reflect it at all. These tagsets usually list E and P as the options, with P denoting the imperfect ("progressive") aspect; this option was replaced with an I to avoid confusion. If the aspect cannot be unambiguously determined (e.g. in loanwords), the position remains empty.

5.5.8. Imperfect and aorist plural forms have likely undergone analogical levellings already in the Middle Bulgarian period (Mirčev 1978:215). Imperfect endings and aorist stems seem to have been preferred in the process. For this reason some of the modern forms, already attested in the earliest damaskini, are ambiguous (e.g. 3PL.AOR/IMPF *poslušaxa* 'they obeyed'; *Tixon.d.*). As the generalized plural endings (-*xme*, -*xte*, -*xa*) reflect the old imperfect, tokens with them are tagged as imperfects too. Only if the token shows a clear reflex of an old aorist ending (-*xom*, -*ste*, -*šę*), it is marked as aorist (e.g. *tako satvoriša* 'thus they did': VMIA3PE; NBKM 328). Ambiguous 1SG.AOR/IMPF forms are resolved according to etymology: verbs with uncontracted IMPF stems (e.g. *ímějaxъ* 'I had': VMII1SI; Vuković 1536) are tagged as imperfects, the rest as aorists (e.g. *imax*: VMIA1SI; *Tixon.d.*).

5.6. Adverbs

As an indeclinable category, adverbs are tagged without any attributes in most cases. The single optional attribute is degree (1), denoting comparative suffixes like -*ěe* or -*ěišeje* (cf. Lunt 2001:77). This marking is limited to Church Slavonic texts (e.g. *istíněišeje réšti* 'to say it more precisely': RC; Vuković 1536). Preposed degree markers used in more innovative texts are handled as separate tokens (e.g. *pó+ skóro* 'more quickly': QC; *Berl.d.*).

5.6.1. Tagset for adverbs.

- R-degree
- Degree:
 - c - comparative (-*ěe*)

5.6.2. Differences from legacy tagsets

The attribute "degree" is used in HR and MK sets at position (2). BG, HR and MK sets have an attribute "type" at position (1). We did not include it in our tagset, because the difference between respective types is arbitrary. For example, MK set uses *dobro* 'well' as an example of a "general" and *sigurno* 'surely' as an "adjectival" adverb, although both are constructed from adjectival roots in the same way. Some of the types may be interpreted as other parts of speech, e.g. verbal adverbs (gerunds) are marked as participles (e.g. *kopaiķi* 'while digging': VMPP-I; NBKM 728; also cf. §5.5.5.).

5.6.3. For adverbs based on old instrumentals, cf. §5.1.6. If the adverbial root can be extended with all of the pronominal affixes (*ně-*, *ni-*, *-to*, CS *e-*), it is handled as a pronoun (cf. §5.3.7.). Inflected CS forms of *mnogo* (e.g. F.SG.INST *mnógojù* in Vuković 1536) are classed as adjectives under the lemma *mnog*. For distinguishing conjunctions and adverbs cf. §5.9.2.

The following lemmas are always marked as adverbs: *abie*, *dnes*, *dokolě*, *doselě*, *edva*, *zautra*, *zde*, *žělo*, *izdrevlě*, *inogda*, *legomi*, *mal(k)o*, *mnogo*, *množiceju*, *naedno*, *naprasno*, *nyně*, *ošte*, *pak*, *paki*, *pače*, *poslě(di)*, *potom*, *proče*, *prězde*, *rano*, *sega*, *sele*, *skoro*, *tamo*, *toko*, *tokmo*, *tuk*, *tъi*, *tъkmo*, *tъčiju*, *velmi*, *včera*, *въh*, *въsma*, *въtre*, *въkupě*, *vynu*, *štom*.

5.6.3.1. *tъi* (or *tui*) is marked as an adverb also when it occurs as a phonetic variant of *toja*, i.e. as a determiner of a noun phrase (e.g. *trъva+tъ na túi pušténie* 'the grass of this desert'; *Sv.d.*), which may be etymologically related. Forms based on *toi* (e.g. in *Nedělnik* 1806) are marked as a personal pronoun. The difference is reflected on the syntactic annotation.

5.7. Prepositions

Marked as "adpositions" in the tagset. The single attribute marks the nominal case or row, called by the adposition in Church Slavonic and other Slavic languages with preserved nominal inflection. By this we can distinguish e.g. *na* denoting location (tagged SL, e.g. *na zemli* 'on Earth') from those denoting direction or possession (both tagged SA, e.g. *na+ kúšta* 'of the house'). The choice is based on the classification by Lunt (2001:151 f.). This attribute helps us distinguish some homonymes. More recent prepositions (e.g. *spored* 'according to') have no case marking.

5.7.1. Tagset for prepositions

- S-case

- Case:

G - e.g. *bez, vьz, iz, do, ot, radi, za* (causal)

D - e.g. *k, po* (direction), *prěmo, protivno*

A - e.g. *na* (direction or possession), *nad, za* (direction), *po* (causal)

L - e.g. *na, po, pri, v* (location)

I - e.g. *meždu, nad, za*, (location), *s*

5.7.2. Differences from legacy tagsets

The case marking has a precedent in the HR set, where it is used at position (3). Separate attributes for distinguishing of pre- and postpositions (usually marked at the first position), as well as between simple and compound ones, were considered redundant for our corpus. The first attribute is relevant for only two words used in CS (*radi* and *děľma* both 'because'), the second one is inferred in the syntactic annotation.

5.8 Particles

The single attribute reflects the type, which is a lexical category. This category is generally avoided due to its ambiguous definition. For example, *da* can be classified as both a conjunction ('so that') and a particle used in analytic verbal constructions. We prefer to mark it as a conjunction (C) to avoid confusion. This is a conventional choice: its syntactic function is reflected in another layer of annotation. The verbal particle *šte* (which can still be inflected by tense in the standard Bulgarian and by person in pre-standardized varieties) is marked as an auxiliary verb (i.e. VAIP3SI, for *šteše*: VAIIP3SI etc.). Ambiguous forms can also be explained by a secondary tag (e.g. *e* 'this' in *NBKM 1423*: both as QD and PP3NSN).

5.8.1. Tagset for particles

- Q-type

- Type:

C - comparative (e.g. *po, nai*)

D - definitive (e.g. *eto, že, si[cev]*)

G - general (e.g. *bare, daže, dori, elma, makar, na, pone, sireč*)

O - modal (e.g. *mai*)

Q - interrogative (e.g. *dali, li, nali, eda*)

Z - negative (e.g. *ne*)

5.8.2. Differences from legacy tagsets

The particles are marked similarly as in the HR set. The number of options was extended with general and comparative ones, present in the BG set. Separate attribute for distinguishing between simple and compound particles (present in the BG and MK sets) was considered redundant: it can be inferred from the syntactic annotation. Particles denoting anaphoric and deictic relations are marked as "definitive", an option adopted from the Albanian tagset.

5.9. Other Parts of Speech

Other parts of speech have no further specifications: conjunctions (C; e.g. *i* 'and'), interjections (I; e.g. *o* 'oh') and non-verbal elements (X; e.g. †, …). With such tokens, only the part-of-speech category is marked in the tag.

5.9.1. Differences from legacy tagsets

BG, HR and MK sets distinguish two additional attributes for conjunctions: type (coordinating or subordinating) and formation (simple and compound). As these attributes reflect rather syntactic properties, the both were removed from our tagset.

Although preserved for special cases (e.g. markers of omission), the representation of non-verbal elements as separate tokens was by rule avoided during the process of annotation. In many our sources, punctuation is not used systematically by the scribe (especially in *PPS*). For this reason, punctuation and other non-verbal symbols were attached to the preceding token. Unless analyzed as separate tokens, they are removed from the diplomatic transcripts and the CoNLL version of our files.

5.9.2. Some words are functionally ambiguous. The following lemmas are always marked as conjunctions: *a, ako, ala, ali, ami, andžak, ašte, bo, da, dano, zaedno, zam, zatova, zašto* (unless context would rather denote *za+što*), *zaštoto, i, ibo, ili, jako* (only for CS 'if'; newer *jako* 'much, very, strongly' tagged as R), *jakože, kato, nelo, neti, ni, niže, nito, no, нь, obače, oti, pa, počto, poneže, ta, ubo, vьskuju, če, čunki*.

6. Syntactic Annotation

6.1. Standard Tags

Syntax is marked on every token in three columns. The first column (UD_ID) contains a number, which denotes the position of the word in the sentence. The second column (UD_NCY) contains a number denoting the dependency of the element. Roots (sentence predicates) are tagged as 0. The third column (UD_TYPE) defines the type of the relation. The corpus uses the standard tagset defined at the Universal Dependencies website ([link](#)).

- ACL - adjectival clause (root of a subordinate clause dependent on a nominal)
- ADVCL - adverbial clause (root of a subordinate clause dependent on a verb)
- ADVMOD - adverbial modifier (e.g. *síč'ko+to naednò oustávì* 'she left everything at once')
- AMOD - adjectival modifier (e.g. *stáę pétka* 'St Petka')
- APPOS - appositional modifier, modifying the immediately preceding noun (e.g. *sъs bráta+i evéimie* 'with her brother Euthymius')
- AUX - auxilla (usually 'be', 'have' or 'want' auxiliary verbs)
- CASE - analytic marking of an oblique nominal (usually an adposition)
- CC - coordinating conjunction (e.g. *dobrini i+zakrílinie* 'virtues and protection')
- CONJ - conjunct (dependent on first conjunct, e.g. in *dobrini i+zakrílinie*, *dobrini* will be tagged as OBJ and *zakrílinie* as CONJ)
- COP - copula
- CSUBJ - clausal subject
- DET - determiner (demonstratives, articles)
- DISCOURSE - discourse marker (no meaning within the syntax of the sentence)
- EXPL - expletives: reflexive pronouns and doubled clitics
- FIXED - fixed multiple word expression (e.g. *ne+ ště da trъpi* 'she wouldn't wait')
- MARK - marker of subordinated clauses (usually a subordinating conjunction or relative pronoun, e.g. *sъzi katò wt kládenъcb istíčaxa* 'tears poured as from a well', *da*-particle in *ne+ bě' neí támb da+ sa zatrížì* 'there was no one there to care')
- NSUBJ - main sentence subject
- NMOD - nominal modifier (e.g. *xódinie to ã* 'her journeys')

NUMMOD - numeric modifier (e.g. *čtíry pšířšta* '[distance of] four shots')

OBJ - direct sentence object

OBL - nominal used as an oblique argument (e.g. in *Krstíteľ ĩwán'na oupríličáva sšs póstenie* 'she followed John the Baptist by fasting', *Krstíteľ* would be OBJ, *póstenie* OBL, and *ĭwán'na* APPOS)

ORPHAN - used in case of a head ellipsis, where simple promotion (to a head) would result in unnatural/misleading dependency relation (e.g. due to missing object *da+ iz'rečē něĭ, déto+ gĭ na epivatš struvaše* 'to tell of her [deeds?], which she did in Epibates')

PUNCT - non-verbal elements

REPARANDUM - reparandum: stricken tokens (depend on any adjacent non-stricken element in the sentence)

ROOT - head/predicate of the sentence

VOCATIVE - vocative

6.1.1. Placeholder items are marked with the DISCOURSE tag, if they mark an interjection or non-grammatical repeating of a word, which is meaningful within the context of the discourse (e.g. *amí já idí na lordánska+ ta rěká* 'so, hey, go to the river Jordan'; *Sv.d.*), marking the clause or sentence root as a head. If a word is consciously stricken as a whole, it can be ignored or annotated as REPARANDUM, with any adjacent (NON-REPARANDUM) token as a head.

If the text is damaged by external influences, the word may be reconstructed (using brackets) if there are any reliable basis to do so (another transcript of the source, preserved parts of a graphem, traces visible by UV light etc.).

6.1.2. Compound numerals are analyzed into separate tokens and handled as quasi-adjectival clauses. The last element of a numeral phrase not following a preposition or conjunction is considered the head. Thus *pe+desětš i trí godíni* '53 years' (lit. 'fifty and three') we annotate in the following way:

<i>pe+</i>	1	2	NUMMOD
<i>desětš+</i>	2	5	NUMMOD
<i>i</i>	3	4	CC
<i>trí</i>	4	2	CONJ
<i>godíni</i>	5	0	ROOT

In numerals for 11-19, the 'ten' is considered dependent on the other number. For example, a clause *sedem+ná+desětš godíni* '17 years' (lit. 'seven on ten') would be tagged in the following way:

<i>sedem+</i>	1	4	NUMMOD
<i>ná+</i>	2	3	CASE
<i>desětš</i>	3	1	NUMMOD
<i>godíni</i>	4	0	ROOT

6.1.3. Tags do not reflect the word category:

6.1.3.1. The particle *ne* 'not' can be used as a DISCOURSE marker, adverbial modifier (ADVMOD) when dependent on a verb or adverb, or adjectival (AMOD) modifier when dependent on an adjective.

6.1.3.2. The conjunction *i* 'and' can be used as CC, but also as an AMOD with the meaning 'also, too' (e.g. *dobří sa legomí i tvóí+te dobríný* 'because your virtues are good too'; *Sv.d.*).

6.1.3.3. An adjective can be used as an AMOD, but also as a ROOT, if it is a qualitative predicate (cf. §6.2.1.2.). Adjectives sometimes appear as heads of noun phrases, receiving the appropriate tag (NSUBJ, OBJ, OBL, NMOD etc.) according to the sentence hierarchy (e.g. NMOD in *da sa pokrĭĭ tš'lo+to na svetáę* 'to cover the body of the saint'; *Sv.d.*).

6.1.3.4. An adverb based on archaic instrumental or locative forms of a noun or an active verbal participle (cf. §5.1.6) is syntactically marked as ADVMOD. If the participle has any dependent elements, it is handled as a head of a subordinate clause (ADVCL or ACL).

6.2. Tag Extensions

Some tags can optionally be extended for more specific relations, using the fourth column (UD_EXT). Multiple extensions can be added, separated by colon <:>. When an elision causes the tagged word to be promoted to a clause head (i.e. ACL or ADVCL) or when it is marked as a conjunct (CONJ), it may receive the extension (e.g. ADVCL in *sľzi katò wt kládenš ističaxa* 'tears flew as from a well'; *Tixon.d.*) too.

The customization of our tagsets aims to unambiguously capture features, which are expressed in one variety synthetically (e.g. by a case ending) and analytically in another one (e.g. by a clitic). In this way, the corpus can include archaic, pre-standardized and contemporary dialectal material, allowing us its comparison across space and time.

6.2.1. One set of extensions is used with nominal modifiers (NMOD) and oblique arguments (OBL), describing their relation in a more specified way (e.g. as a spatial or possessive relation). This helps us distinguish meanings of constructions using polysemic prepositions like *na*, or to analyze the marking of syntactic functions formerly represented by different nominal case endings. The extension is chosen according to the meaning of the sentence. The tag extensions are marked in a separate column (examples from *Berl.d.*):

- ABL - ablative relation (e.g. *běše wt+ město epiváti* 'she was from the town of Epibates')
- IOBJ - indirect object (e.g. *bědí+ mī drugár'* 'be my comrade', lit. 'be to me a comrade')
- LAT - allative relation (e.g. *wtídoxa v' nbsny" ográdī* 'they departed to the heavenly gardens')
- LOC - locative relation (e.g. *po+ síč'ky svě'tb* 'all over the world')
- POSS - possessor (e.g. *сѣс'+ bráta+ sí* 'with her brother')
- PRED - predicative modifier (e.g. *bědí+ mī drugár'* 'be my comrade')

6.2.1.1. The difference between IOBJ (recipient) and POSS (possessor) is roughly analogous to that of LAT/ABL (movement in a direction to/from) and LOC (location or position). The choice depends on the meaning of the clause within the sentence.

6.2.1.2. The extension PRED is used, when the attribute of the nominal subject is another noun. It denotes a category instead of quality. In CS, such instances are sometimes marked with an instrumental or another oblique case (e.g. *ne bōdi niktože ijudojō* 'let no one be a Judas'; Lunt 2001:151). The verb is considered the root in such sentences. Only adjectives, as qualitative predicates, are promoted to roots (*Tózi stárecb běši tvrb 'dě dobrodě'telēb* 'this old man was very virtuous'; *Sv.d.*) according to the UD convention.

6.2.1.3. The verb *pítam* 'ask' can have two object dependents (*rábota ma si pítáľb* 'you have asked me a thing'; *Sv.d.*). In such a case, the 'topic' of the question (i.e. *rábota*) is marked as OBL with LAT extension, the 'asked' (i.e. 1SG.GEN/ACC *me*) as a direct object (OBJ). The reason is that the 'topic' can be marked with an additional preposition (*ako+ lí píta ně'koi za+ xubos+ta* 'if someone asks about the beauty'; *Berl.d.*), the 'asked' usually not.

6.2.2. One of the first uses of the corpus was the research of postposed articles and other definiteness markers. The variation between specific determiners (extended and short clitic demonstratives) is marked by additional extensions of the DET relation. As the articles are in most cases analyzed as separate tokens, the P_NOM or P_ADJ extension reflects the morphological category of the preceding token. The extension EXT denotes morphology of the determiner itself (examples from *Tixon.d.*):

- EXT - an extended demonstrative (e.g. *tázi miriz'mà* 'the stench')
- P_ADJ - a postponed article following an adjective (e.g. *móí+ te móšti* 'my relics')
- P_NOM - a postponed article following a noun (e.g. *kov'čeg'+ t'+ sí* 'their coffin')

6.2.2.1. The extension can be used after tags for nouns and adjectives, if the determiner is expressed by a vowel only (e.g. *a vinaro crev' reče* 'but the king's cupbearer said'; *PPS*). Such an extension is likely to be combined with other ones: for example, in *života čoveko* 'the life of the human' (*NBKM 1069*) both endings may reflect definiteness, but in *čoveko* the relation is also that of a possessor. Thus its UD tag would be extended with P_NOM:POSS (extensions are ordered alphabetically).

6.2.2.2. When the determiner is written separately from the noun (e.g. *slóvo w'nd prrw'čbskoè* 'the word of prophecy'; Vuković 1536). Preposed short demonstratives (*vč+ tómb monastíri* 'in the monastery'; *Kiev.d.*) have no extension.

6.2.2.3. The article is sometimes attached to words of other categories than nouns and adjectives. When following an adjectival pronoun or the numeral '1' (which are tagged as adjectives anyway), the extension is P_ADJ. If they follow any numeral other than '1' (e.g. *dva+ ta* 'the two'; *Tixon.d.*) or an adverb (*mnógo+ tu moé grěxové* 'my many sins'; *Sv.d.*), P_NOM is used.

6.2.3. Periphrastic verbal constructions like perfect tenses or conditionals can be marked by adding the function tag to the auxiliary verb (AUX relation), markers of subjunctive relations (MARK) or their complements of fixed phrases (FIXED):

CON - conditional (e.g. *dší+te+sí bix'me dàlǐ* 'we would give our souls away'; *Berl.d.*)
 FUT - periphrastic future (e.g. *šte+ da+ reče+ tája+ réčb* 'he will say the word'; *NBKM 1069*)
 INF - "infinitives" after auxiliary verbs (e.g. *šte+ da+ reče+ tája+ réčb* 'he will say the word'; *NBKM 1069*)
 OPT - optative (e.g. *da ne viditb+ nas neko* 'nobody should see us'; *NBKM 238*)
 PASS - passive voice (e.g. *ispálnena mpési* 'she was filled'; *NBKM 1064*)
 PRF - perfect/narrative mood (e.g. *togaī dīavoľb e+zavīdīľb* 'then the devil was jealous'; *NBKM 728*)
 PPRF - plusquamperfect (e.g. *tǐ+i+se běše pod_kánila* 'thus she had decided to live'; *Tixon.d.*)

6.2.3.1. The tag extension INF is applied to *da*-markers auxiliary verbs, which complement synthetic infinitives in some dialects (cf. Mirčev 1978:235). Such constructions are attested in older sources within our corpus too, e.g. *polovina ot crstvoto si da štem* 'we would give a half of our kingdom (*Tixon.d.*). The auxiliary verbs thus marked are the following:

moga 'can',
šta 'want' (including variants like *xoču, cú* etc.),
ima 'have' (esp. negative future marker *njama*).

We have also included a number of verbs with the meaning 'begin' or 'stop', which function in a similar way:

načena or *počna* 'begin' (cf. Lunt 2001:154),
podbra 'begin' (*puđmpra da struva tzjudesa* 'she began to do miracles'; *NBKM 1064*),
(v)zema 'begin', lit. 'take' (e.g. *ze da vuni* 'it began to stink'; *NBKM 1064*),
xvana 'begin', lit. 'take' (*xvana po-mnogo da běga* 'began to run faster', *Sv.d.*; incl. variants like *fana*),
stiga 'stop',
prěstana 'stop',
nedei 'do not' (e.g. *ne+dei se gnusi* 'do not be disgusted; *Tixon.d.*).

The extension serves to distinguish instances, where *da*-constructions have likely replaced synthetic infinitives, from other types of subordinated clauses. It is also able to recognize phrases with multiple main verbs after a single auxiliary verb, e.g. *Poněže ničto drúgoe táko dšu člčskoe ne+ móže [1] da očísti i+ na pérvo_obražie [2] da+ privedè kakvotb pustýnnoe i+ bezmólvnoe živénie*. 'Because nothing else is able to purify the human soul and to make it perfect so well as the silent life in a desert.' (*Nedělnik 1806*).

6.2.3.2. The current design of the morphological tagset does not allow us to distinguish the voice of verbal participles (cf. §5.5.6.). Although the problem is of less relevance to non-CS texts, the extension PASS can be used with present and past participles used as roots of subordinate clauses (ACL, ADVCL).

7. Annotation Examples

7.1. Vuković 1536 (Venice, Church Slavonic, 1536), 194v

Врѣмениже немáлѣ мѣмошьдшоу. своѣ ѡхóжденіе ѣже ѡсоудоу разоумѣ. 'and after some time passed, she understood that she will depart soon from this world'

text	diplomatic	lemma	PoS tag	UD tag		
<i>Vrě'meni+</i>	<i>vrěmeni</i>	vrěme	NNSDN	1	5	CSUBJ
<i>že</i>	<i>že</i>	že	QD	2	12	CC
<i>ne+</i>	<i>ne</i>	ne	QZ	3	4	AMOD
<i>málu</i>	<i>malu</i>	malo	ANSDN	4	1	AMOD
<i>mímošb'd'shoù.</i>	<i>mimošb'dšou</i>	mimoiti	VMPPA-SE ANSDN	5	12	ADVCL
<i>svoje</i>	<i>svoje</i>	svoi	ANSNY	6	7	AMOD POSS
<i>wtxóždenie</i>	<i>otxoždenie</i>	otxoždenie	NNSNN	7	12	OBJ
<i>je+</i>	<i>je</i>	je	PP3NSN	8	9	NMOD
<i>že</i>	<i>že</i>	že	QD	9	11	MARK
<i>wt+</i>	<i>ot</i>	ot	SG	10	11	CASE
<i>soúdou</i>	<i>soudou</i>	sъдѣ	R	11	7	ACL OBL
<i>razoúmĕ.</i>	<i>razoumĕ</i>	razumĕti	VMIA3SE	12	0	ROOT

7.2. *Tixon.d.* (Central Balkan area, 1650s-1660s), 55r

áďšátatiščè dabŭde nakrásnyĭ ĭnevещestŭvnyĭ ráĭ. 'and your soul shall be in the beautiful and immaterial Paradise'

text	diplomatic	lemma	PoS tag	UD tag		
<i>a+</i>	<i>a</i>	a	C	1	7	CC
<i>dšá+</i>	<i>dša</i>	duša	NFSNY	2	7	NSUBJ
<i>ta+</i>	<i>ta</i>	тъ	PD-FSN	3	2	DET P_NOM
<i>ti+</i>	<i>ti</i>	ti	PP2-SD	4	2	NMOD POSS
<i>šte"</i>	<i>šte</i>	šta	VAIP3SI	5	7	AUX FUT
<i>(da)+</i>	<i>da</i>	da	C	6	5	FIXED INF
<i>búde</i>	<i>bude</i>	bŭda	VMIP3SE	7	0	ROOT
<i>na+</i>	<i>na</i>	na	SL	8	13	CASE
<i>krásnyĭ</i>	<i>krasnĭi</i>	krasen	AMSNY	9	13	AMOD
<i>i+</i>	<i>i</i>	i	C	10	12	CC
<i>ne+</i>	<i>ne</i>	ne	QZ	11	12	AMOD
<i>veštestŭvnyĭ</i>	<i>veštestŭvnyĭ</i>	veštestven	AMSNY	12	9	CONJ
<i>ráĭ.</i>	<i>rai</i>	rai	NMSNN	13	7	OBL LOC

7.3. *NBKM 1069* (Beljovo, Panagjurište area, 1776), 137r

такамѣ билó писáпw на wногŭзи члwвѣка 'thus it was ordained for that man'

text	diplomatic	lemma	ME tag	UD tag		
<i>taka+</i>	<i>taka</i>	taka	PR	1	5	ADVMOD
<i>mu+</i>	<i>mu</i>	toi	PP3MSD	2	8	EXPL
<i>e</i>	<i>e</i>	e	VAIP3SI	3	4	AUX PRF
<i>biló+</i>	<i>biló</i>	съм	VAP--SI	4	5	AUX PASS
<i>pisáпw</i>	<i>pisano</i>	pisan	VMPPA-SI ANSNN	5	0	ROOT
<i>na+</i>	<i>na</i>	na	SA	6	8	CASE
<i>wno_gw'zi</i>	<i>onogozĭ</i>	onzi	PD-MSG	7	8	DET EXT
<i>člw'véka</i>	<i>človeka</i>	človek	NMSAY	8	5	OBL IOBJ

7.4. *Berl.d.* (Moesian area, 1791), 179v

ámĭkóyščè damóže da ĭskáže néĭnyte póčestĭ, rábotĭ ĭ ĭčюдесата 'but who could retell her virtues, works and wonders?'

text	diplomatic	lemma	PoS_tag	UD tag		
<i>ami+</i>	<i>ami</i>	ami	C	1	7	CC
<i>kóĭ+</i>	<i>koi</i>	koi	PQ-MSN	2	7	NSUBJ
<i>štè</i>	<i>šte</i>	šta	VAIP3SI	3	5	AUX FUT
<i>da+</i>	<i>da</i>	da	C	4	3	FIXED
<i>móže</i>	<i>može</i>	moga	VAIP3SI	5	7	AUX
<i>da</i>	<i>da</i>	da	C	6	5	FIXED INF
<i>iskáže</i>	<i>iskaže</i>	iskaža	VMIP3SE	7	0	ROOT
<i>néĭni+</i>	<i>neini</i>	nein	AFPNN	8	10	AMOD POSS
<i>te</i>	<i>te</i>	тъ	PD-FPN	9	8	DET P_ADJ
<i>póčestĭ,</i>	<i>počesti</i>	počest	NFPNN	10	7	OBJ

<i>ráboti</i>	<i>raboti</i>	rabota	NFPNN	11	10	CONJ	
<i>ì</i>	<i>i</i>	tja	PP3FSD	12	11	NMOD	POSS
<i>i+</i>	<i>i</i>	i	C	13	14	CC	
<i>čjudesa+</i>	<i>čjudesa</i>	čjudo	NPNPN	14	11	CONJ	
<i>ta</i>	<i>ta</i>	тъ	PD-NSN	15	14	DET	P_NOM

7.5. PPS (Mokreš, Vidin-Lom area, 1796), 75v

ѡвстави бащѣ имайкѣ ѡдрѣги срѡдници све цо бехѣ ѡмѣрь 'and she left her father and mother and other relatives, everybody in the world'

text	diplomatic	lemma	PoS tag	UD tag			
<i>'i+</i>	<i>i</i>	i	C	1	2	CC	
<i>wstavi</i>	<i>ostavi</i>	ostavja	VMIA3SE	2	0	ROOT	
<i>baštu</i>	<i>baštu</i>	bašta	NFSAY	3	2	OBJ	
<i>i+</i>	<i>i</i>	i	C	4	5	CC	
<i>maïku</i>	<i>maïku</i>	maïka	NFSAY	5	3	CONJ	
<i>'i+</i>	<i>i</i>	i	C	6	8	CC	
<i>drugi</i>	<i>drugi</i>	drug	AMPNN	7	8	AMOD	
<i>srѡdnici</i>	<i>srodnici</i>	srodnik	NMPNY	8	5	CONJ	
<i>sve</i>	<i>sve</i>	vse	NNSNN	9	8	APPOS	
<i>što</i>	<i>što</i>	što	PQ	10	11	MARK	
<i>bexu</i>	<i>bexu</i>	сѣм	VMII3PI	11	9	ACL	
<i>u+</i>	<i>u</i>	u	SG	12	13	CASE	
<i>mѣрь</i>	<i>mѣрь</i>	mir	NMSNN	13	11	OBL	LOC

7.6. NBKM 1064 (Sliven, Subbalkan area, 1820s), 40r

ἀμὴ δασα ὑστὰβιμ ἕτ τῆη μπράτῖα μὴ χριστένι 'but let us abandon this, my Christian brothers'

text	diplomatic	lemma	PoS tag	UD tag			
<i>amì</i>	<i>ami</i>	ami	C	1	4	CC	
<i>da+</i>	<i>da</i>	da	C	2	4	AUX	OPT
<i>sà</i>	<i>sa</i>	se	PX---A	3	4	EXPL	
<i>ustávìm</i>	<i>ustavim</i>	ostavja	VMIP1PE	4	0	ROOT	
<i>ut</i>	<i>ut</i>	ot	SG	5	6	CASE	
<i>túi</i>	<i>tui</i>	тѣи	R	6	4	ADVMOD	
<i>mprátia</i>	<i>bratia</i>	brat	NMPNY	7	4	VOCATIVE	
<i>mói</i>	<i>moi</i>	moi	AMPNN	8	7	AMOD	
<i>xristéni.</i>	<i>xristeni</i>	xristianin	NMPNY	9	7	APPOS	

7.7. NBKM 728 (Tetovo, West Macedonia, 1830s), 8r

κοπαϊκί γροβο ναϊδοα тѣло неїзгниєно 'as they dug, they found an undecayed body'

text	diplomatic	lemma	PoS tag	UD tag			
<i>kopaiiki</i>	<i>kopaiiki</i>	koraja	VMPP-PI	1	3	ADVCL	
<i>grobo</i>	<i>grobo</i>	grob	NMSON	2	1	OBJ	P_NOM
<i>naïdoa</i>	<i>naïdoa</i>	naida	VMII3PE	3	0	ROOT	
<i>tělo</i>	<i>tělo</i>	tělo	NNSNN	4	3	OBJ	

Literature

- Berl.d.*: [Berlin damaskin]. Cracow: Library of Jagellonian University, sign. Slav. fol. 36.
- Cod.Zogr.*: *Codex Zographensis*. On the basis of the edition by V. Jagić, *Quattuor evangeliorum codex glagoliticus olim Zographensis nunc Petropolitanus, characteribus cyrillicis transcriptum notis criticis prolegomenis appendicibus auctum*, Berolini/Sanktpeterburg, 1879. (available online - [link](#))
- Dobr.ev.*: Conev, Benjo (1904) *Dobrějšovo četveroevangele: srědnobălgarski pametnik ot XIII věk (= Bălgarski starini kn. I)*. Sofia: Ministerstvoto na narodnata prosvěta.
- Ioann.d.*: [Pop Ioann's damaskin]. Sofia: Archive of the Bulgarian Academy of Sciences, sign. № 3312.
- Jan.s.*: *Jankulov sbornik ot 1755 g.* Sofia: National Library of Bulgaria "Sv. Sv. Kiril i Metodii", sign. HБKM 689 (272).
- Kiev d.*: Ilievski, Petar Xr. & Ilievska, Krasimira (2015) *Kievski damaskin*. Skopje: MANU.
- Krčovski, Joakim (1814) *Pověst radi strašnago i vtorago prišestvia Xristova*. Budin-grad: Kraljev. vseučilište ungarskoe.
- Ljub.d.*: [Ljubljana damaskin]. *Zbornik "Damaskin", bolgarski*. Ljubljana: National and University Library of Slovenia, sign. NUK Cod. Kop. 21. (available online - [link](#))
- Lov.d.*: *Loveški damaskin*. Sofia: Sofia University "Sv. Kliment Oxridski". (available online - [link](#))
- NBKM 328: Damaskin ot 1750 g.* Sofia: National Library of Bulgaria "Sv. Sv. Kiril i Metodii", sign. HБKM 328.
- NBKM 370: Paisieva Istorija slaveno-bolgarskaja, prepis (Elenski) ot 1784 g.* Sofia: National Library of Bulgaria "Sv. Sv. Kiril i Metodii", sign. HБKM 370.
- NBKM 728: Otkāslek ot prostonaroden damaskin, vek XIX. Otkāslek ot prostonaroden damaskin, vek XIX.* Sofia: National Library of Bulgaria "Sv. Sv. Kiril i Metodii", sign. HБKM 728.
- NBKM 1064: Damaskin, ot kraja na XVIII ili načaloto na XIX v.* Sofia: National Library of Bulgaria "Sv. Sv. Kiril i Metodii", sign. HБKM 1064.
- NBKM 1069: Damaskin (Beljovski), ot 1776 g.* Sofia: National Library of Bulgaria "Sv. Sv. Kiril i Metodii", sign. HБKM 1069.
- NBKM 1081: Sbornik ot slova i poučenija, ot načaloto na XIX v. (1821 g.)*. Sofia: National Library of Bulgaria "Sv. Sv. Kiril i Metodii", sign. HБKM 1081.
- NBKM 1423: Pavlikjanski sbornik na Petăr Carski, ot 1793 g.* Sofia: National Library of Bulgaria, "Sv. Sv. Kiril i Metodii", sign. HБKM 1423.
- Nedělnik 1806: Sofronii ep. Vračanskii (1806) Kyriakodromion sireč Nedělnik – Poučenie*. Râmnicu: ep. Nektarii. (available online - [link](#))
- Nedělnik 1856: Sofronij ep. Vračanskij (avt.); Xrulev, Todor T. (ed. 1856) Evangelie poučitelno*. Novi Sad: Knigopečatnjata Dr. Dan. Medakov. (available online - [link](#))
- PPS: Pop-Punčov sbornik ot 1796 god.* Sofia: National Library of Bulgaria "Sv. Sv. Kiril i Metodii", sign. HБKM 693. (available online - [link](#))
- Rai.d.*: [Raikovski damaskin]. Plovdiv: National Library of Bulgaria "Ivan Vazov", sign. HБИВ 160 (600). (available online - [link](#))
- Temski r.*: Vasilev, Vasil P. (1986) Temskijat rākopis - bălgarski ezikov pametnik ot 1764 g. *Starobălgaristika X* (1986), 1. 49-72.
- Tixon.d.*: [Tixonravovskij damaskin]. Moscow: Russian State Library, sign. OP Ф.299 № 702.
- Sv.d.*: Miletič, Ljubomir (1923) *Svištovski damaskin. Novobălgarski pametnik ot XVIII věk (= Bălgarski starini. Kniga VII)*. Sofia: BAN.
- Vel.s.*: *Sborniče (Veleško) ot tropari, mesecoslov i pr., redakcija srăbska, vek XV-XVI*. Sofia: National Library of Bulgaria "Sv. Sv. Kiril i Metodii", sign. HБKM 667 (45).

- Vuković 1536: [Zbornik za putnike]. Venecia: Božidar Vuković. (available online - [link](#))
- BAN I-VII: Georgiev, Vladimir I. (red., et al. 1972–2010) *Balgarski etimologičen rečnik. Tom I–VII*. Sofia: BAN.
- Cejtlin, Rajla M.; Blagova, E.; Večerka, R. (et al. 1994) *Staroslavjanskij slovar' (po rukopisam X-XI vekov)*. Moskva: Russkij jazyk.
- Demina, Evgenija I. (1968) *Tixonravovskij damaskin. Bolgarskij pamjatnik XVII v. Issledvanie i tekst, 1*. Sofia: BAN.
- (1972) *Tixonravovskij damaskin. Bolgarskij pamjatnik XVII v. Issledvanie i tekst, 2*. Sofia: BAN.
- (1985) *Tixonravovskij damaskin. Bolgarskij pamjatnik XVII v. Issledvanie i tekst, 3*. Sofia: BAN.
- (et al. 2012) *Rečnik na knižovnija balgarski ezik na narodna osnova ot XVII vek*. Sofia: Valentin Trajanov.
- Lunt, Horace G. (1952) *Grammar of the Macedonian Literary Language*. Skopje: Državno Knigoizdatelstvo na NR Makedonija.
- Maslov, Jurij S. (1981) *Grammatika bolgarskogo jazyka*. Moskva: Vysšaja škola.
- Miklosich, Franz (1865) *Lexicon Palaeslovenico-graeco-latinum*. Vindobona: Guilelmus Braumueller. (available online - [link](#))
- Mladenov, Maksim Sl. (1963) Členuvani prilagatelni formi na -ij v severoiztočnite balgarski govori. *Balgarski ezik*, 1963, kn. 4-5. 404-410.
- Mladenova, Olga M. (2007) Otnovo za lokalizacijata na pŕvonačalnija novobalgarski damaskinov prevod. *Tova čudo: ezikät! Izsledvanija v čest na prof. d-r Živko Bojadžiev*. Sofia: Sv.Kliment Oxridski. 309-316.
- Mirčev, Kiril (1978) *Istoričeska gramatika na balgarskija ezik*. Sofia: Nauka i izkustvo.
- Stojkov, Stojko (1993, fototip. izd. 2002) *Balgarska dialektologija*. Sofia: Akad. izd. "Prof. Marin Drinov". (available online - [link](#))
- Velčeva, Borjana (2001) Damaskinite ot XVII vek i načaloto na novobalgarskija knižoven ezik. *Starobalgaristikata* XXV (2001), 4. 64-81.