Electronic edition and linguistic annotation of Slavic fragments

Tsvetana Dimitrova
Institute for Bulgarian Language
Bulgarian Academy of Sciences
cvetana@dcl.bas.bg

Andrej Bojadžiev
Faculty of Slavic Studies
Sofia University
aboy@slav.uni-sofia.bg

0. Introduction

The paper presents some issues on a possible approach to combining electronic description, electronic edition and annotated corpus of medieval Slavonic texts. In the first section of the paper, the emphasis is on practical issues regarding the annotation of different types of information such as information about the text, attribution, dates, saints, images. The second part discusses an approach to annotation of anaphorically related elements that are decisive for organising the structure of the text and connections to specific elements that will be annotated at the level of the text (the linguistic annotation follows an already implemented approach). The examples given cover not only parchment fragments but also fragments of manuscripts from different periods of the South Slavonic Medieval and Early Modern Bulgarian tradition. The project will be implemented in an eXist database and will use approaches of the Reperorium (http://repertorium.obdurodon.org), PROIEL (http://www.hf.uio.no/ifikk/english/research/projects/proiel/) and TOROT (http://site.uit.no/ slavhistcorp / files / 2015/04 / Eckhoff.pdf) projects.

1. Electronic edition

The task of preparing an electronic edition of Medieval sources provides an unique opportunity to make a virtual re-collection of scattered pieces of cultural heritage kept in various repositories throughout the world. For fragments, especially, this opportunity is of specific value, as it will provide access to witnesses of unknown graphical or textual tradition. For the earliest South Slavic palaeography, this practice could help bridge the known separate facts and to reconstruct the history of writing, especially for the 12th and 13th centuries.

The most important part of the electronic form, along with the preservation of the physical original of the manuscript, is to propose different views for the presentation of its content and layout. At least three types of annotation could be included:

1. Diplomatic or palaeographic annotation. A combination of original page layout possibly linked to images of the original manuscript. This is the traditional way of representing cultural heritage in libraries or archives and it is closely related to the ideas of digital preservation and the approach used in digital libraries in general.

2. Textual history annotation. This covers annotation of textual variants concerning the history of the texts. It presupposes a special layer for textual annotation and metadata on text witnesses. Text-critical editing in an electronic form means constructing a text corpus of variants of the text. This annotation differs from the linguistic annotation *per se* in the ways it handles variant readings even if they reflect the history of the language and language changes.

3. Linguistic annotation covers marking information on various language levels including phonology, grammar or lexicon needed for establishing the properties of a language as attested in texts.

An essential part of each of the aforementioned annotation levels is to provide links to metadata information about the annotated source which covers description of the manuscript, links to terminological databanks, various types of references (Bojadžiev 2003: 81).

The approach to the preparation of an electronic edition of medieval Slavic fragments essentially involves two types of description. If the manuscript is already described in electronic format, the edition should include a link to its description, though the edition should involve description, although a brief one. An example is given in (1) from the description of the Manuil Apostolos from the 13th century:

(1)

```
<summary>The manuscript is a full (long) Aprakos (lectionary) Apostolos, beginning with the mobile cycle
after Easter. Parts of the Menologion are also preserved. The distribution of readings (lections) coincides with
what is published by C. R. Gregory for the Greek tradition …
<ref target=”http://repertorium.obdurodon.org/readFile.php filename=MDManoil.xml”>Description</ref>
</summary>
```

The example in (1) contains a brief overview along with the pointer to the extensive description of the fragment which was once part of the valuable Apostolos copy.

It should be noted here that in the lack of previous research on the text (and its filiation), the text critical apparatus cannot be provided, neither the variants from other text witnesses. However, the attribution of various parts of the source is essential, thus the description should provide attribution to each part of the content as in (2).

(2)

```
<msItemStruct xml:id="ACD11">
<locus n="11">NBKM 499, 5r</locus>
<title xml:lang="en">Acts 18: 22–28</title>
<note><date type="churchCal">Wednesday of the 6th Week in Easter</date></note>
<note type="parallel"><ref target=" http://prototypes.openscriptures.org/manuscript-comparator/?
passage=Acts.18.22-28">(ΠΡΑΞΙΣ)</ref></note>\
</msItemStruct>
```

In (2), identification of the text (*ACD11*) in the manuscript is given, along with its location (*n=”11”*), repository and folio (*NBKM 499, 5r*), the title of the text in English (*Acts 18:22–28*), the time of reading (*Wednesday of the 6th Week in Easter*), and a link to the corresponding Greek text. This information is valuable for various types of editions (diplomatic and text-critical) as well as for the linguistic annotation.

The text identification is closely related to the problem of its segmentation in both text edition and the linguistic corpus (Bojadžiev, Dimitrova 2008).

The conclusions from the text comparision could result in a short overview about the text recension in the copy that is described, as in (3).

(3)

```
<filiation type="litRedaction">The manuscript uses archaic lexis and is considered to belong to the archaic
group of Apostoloi representing the Cyrillo-Methodian translation; It uses several Preslavisms and shares
common reading with the Slepče Apostolos as a full Aprakos Apostolos and more rarely with the Matica Srpska
Apostolos; examples for common readings with the Slepče Apostolos as opposed to other archaic Apostoloi:
```

<foreign xml:lang="cu">**заповѣданое**</foreign> *Acts. 10: 32 vs* <foreign xml:lang="cu">**повелѣное**</foreign> *in Archaic Apostoloi;* <foreign xml:lang="cu">**оударившоу**</foreign>*Acts 12: 12 vs* <foreign xml:lang="cu">**тлькнѫвъшоу**</foreign>, <foreign xml:lang="cu">**алчѧщимь**</foreign> *Acts 13: 2 vs* <foreign xml:lang="cu">**постѧщимъ**</foreign>*, but* <foreign xml:lang="cu">**пощъше**</foreign>*Act 13: 3 as in archaic Apostoloi vs* <foreign xml:lang="cu">**алъкавъше**</foreign>
in the Slepče Apostolos. In several cases it has the readings of the Šišatovac Apostolos and Vranešnica Apostolos. Judging from the presence of some verses typical of the continuous Apostolos in the lections of the manuscript, one could suspect that its antigraph / protograph was based on a continuous Apostolos.
</filiation>

The information from the description such as the attribution of various text parts and palaeographic and codicological features will bring valuable information for the text editor and linguists. An example is given in (4) with a description of a particular headpiece that could be important for making a decision on the approach to text segmentation (an important prerequisite also for the linguistic annotation):

(4)
<decoNote type="headpieces" corresp="#NBKM499">
<locus>*f. 1r, NBKM 499*</locus>
<figure xml:id="NBKM499_1r_head">
<head>Narrow headband with interlace, drawn in red ink with yellow filling and ornamented with black dots</head>

An important benefit of the electronic edition is the opportunity to include links from various parts of the information to sources outside the edition or the corpus. Consider, for example the information about the days and saints for 27.07 (St. Panteleimon) in (5):

(5) link to DBPedia:
<ref target="http://dbpedia.org/page/Saint_Pantaleon">...</ref>

The pointer to popular and widely used resources could be useful for placing information about the date and the saint into a more general (or even specific) context. Here, we can also link this information to multilingual ontological resources and specialized resources where one may consult information about other saints days, as in (6)

(6) a. To Encyclopedia Slavica Sanctorum
<ref target="http://www.eslavsanct.net/mod_viewdate.php?day=27&month=7 ">27 July</ref>

b. To Menology project
<ref target="”http://menology.obdurodon.org/runSearch.php?mss[]=501&mss[]=882&mss[]=Arx&mss[]=B&mss[]=Baron&mss[]=Carp&mss[]=As&mss[]=Suprl&mss[]=H&mss[]=P&mss[]=C&mss[]=DC2&mss[]=En&mss[]=JT&mss[]=Bas&mss[]=Oh&mss[]=Os&mss[]=F72&mss[]=S&mss[]=Slep&mss[]=Strum&mss[]=ST&mss[]=ZT&type[]=saint&nametype=name&nameinput=Panteleemon&fqnameinput=&startdate=09%2F01&enddate=08%2F31&nationality=*&sex=*&archaic”>27 July</ref>

In such resources, one may choose among many copies, dates and saints, and even to compose a specific calendar of the information.

The most important obstacle for linking images in the edition or description is the way libraries provide internet addresses for their visualisation. The URLs are often dynamic so it is not possible to point to or embed the image in the edition or description. The images from the Slavic manuscripts are of poor quality with very few exceptions (see the Troicko-Sergieva Lavra web collection: http://old.stsl.ru/manuscripts/index.php).

2. Linguistic annotation

The linguistic annotation covers the process of description and labelling of any (linguistic) information to raw language data including morphological and syntactic (and morphosyntactic), semantic, pragmatic, referential, etc. information, with the aim of describing different text segments – a word or a larger segment such as phrase, sentence, paragraph, etc. The morphological annotation (including part-of-speech annotation) is a prerequisite for further annotation such as morphosyntactic annotation, syntactic segmentation (parsing) and annotation, information structure annotation, etc. (cf. Leech 1993; Garside et al. 1997; Müller, Strube 2006). The annotation of diachronic language data is a challenging task and various problems have been already discussed extensively through the years (Rissanen 1998; Rissanen 1992; Krause et al. 2012); corpora with Old Church Slavonic data have been also constructed and annotated (for a not so recent overview, cf. Dimitrova 2008), however the discussion on strengths and challenges of the approaches goes beyond the goal of this paper.

We propose to follow the (approach to) linguistic annotation of the PROIEL[1] project for construction a treebank of ancient Indo-European languages, including Latin and Ancient Greek, and the subsequent TOROT[2] project, which is an expansion of the Slavic part of the PROIEL corpus, as it proves to be viable and due to the volume of already annotated texts which are accessible and used for training of the available tools and platform. The approach to annotation has been documented (Haug 2010, Haug, Jøhndal 2008).

In (7), we give an example of an annotated segment (a sentence) from the Zograph Fragments (Lavrov Fragments) with the attributes (values are given in the Appendix) of the <token> element: 'id', 'form' (wordform), 'citation-part' (the title of the text), 'lemma', 'part-of-speech' (class and subclass of the token), 'morphology' (morphological annotation), 'presentation-after' (punctuation marks following the token). Here, we additionally propose two more attributes 'ref' (information about the referent; with values: person (PERS); higher being (HIGH); animal (ANIM); relative (REL); body part (BODY); possession / ownership (POSS) – marking alienable and inalienable possessee referents (домъ, имѧ); location (LOC): тамо, градъ; time expression (TIME): тъгда, кога; legal entity (LEG).), and 'ana-id' (id of the token which is (anaphorically) linked to the token at hand).

(7)

```
<sentence>
<token id="1" form="показати" citation-part="ZogrFolia" lemma="показати" part-of-speech="V-"
morphology="--pna----i" presentation-after=" "/>
<token id="2" form="братиѧ" citation-part="ZogrFolia" lemma="братьꙗ" part-of-speech="Nb"
morphology="-s---fn--i" head-id="1" relation="sub" ref="REL" presentation-after=" " ana-id="0"/>
```

[1] https://proiel.github.io/

[2] http://torottreebank.github.io/

```
<token id="3" form="гн҃ѧ" citation-part="ZogrFolia" lemma="господьнь" part-of-speech="A-" morphology="-s---fnpsi" head-id="2" relation="atr" ref="HIGH" presentation-after=" " ana-id="0"/>
<token id="4" form="възъванꙗ" citation-part="ZogrFolia" lemma="възъваньѥ" part-of-speech="Nb" morphology="-p---na--i" head-id="1" relation="obj" presentation-after=" " ana-id="0"/>
<token id="5" form="по" citation-part="ZogrFolia" lemma="по" part-of-speech="R-" morphology="---------n" head-id="1" relation="adv" presentation-after=" "/>
<token id="6" form="оучению" citation-part="ZogrFolia" lemma="оучениѥ" part-of-speech="Nb" morphology="-s---nd--i" head-id="5" relation="obl" presentation-after=" " ana-id="0"/>
<token id="7" form="гн҃ю" citation-part="ZogrFolia" lemma="господьнь" part-of-speech="A-" morphology="-s---ndpsi" head-id="6" relation="atr" ref="HIGH" presentation-after=" " ana-id="0"/>
<token id="8" form="гл҃авъшоу" citation-part="ZogrFolia" lemma="глаголати" part-of-speech="V-" morphology="-supamd-si" head-id="7" relation="atr" presentation-after="·"/>
</sentence>
```

The focus of this section will be on the additional information needed to encode the anaphoric relations in the text. Linguistic anaphora is the coreference of an expression with another expression found in the preceding text (the term cataphora refers to coreference with an element in the following text) and is used in maintaining the structure of the text. It can be also used in text segmentation to define the scope of propositional relations and referentiality. Anaphorically related elements are subject to morphosyntactic constraints including constraints on morphological and semantic characteristic of the anaphorically linked elements.

Identification of anaphoric relations covers marking of referents of specific textual elements (words, phrases, clauses), incl. referential noun phrases, adjectival phrases (headed by adjectives of referential content – denominal: член҃ьскы̏, агг҃льскы̏, possessive: антихрістова, pronominal: таквaꙃи, сѣкава), pronouns, adverbs (including pronominal adverbs: тамо, тако, тъгда, etc.) including those that function as subjunctions, etc. Types of anaphora involve nominal anaphora (including pronominal anaphora (Roberts 1989; Huang 2000), which is the most heavily researched type), as well as clausal anaphora (where the antecedent is a whole sentence or a proposition). The principles of automatic anaphora resolution (which covers identification and annotation) were also thoroughly studied and applied for building different anaphora resolution systems (cf. Mitkov 2014 (2002).

However, the encoding of the referential information in the text fragments we deal with requires some preliminary work which involves manual annotation to identify dependencies and trends and the extent to which approaches for antecedent and anaphora recognition applicable to contemporary text could be applied to historical data. Below, we propose an approach to annotation of referential elements with the aim of building a corpus (a collection of texts) with annotated referential elements that can be used to monitor the anaphora to find ways to automatically identify (anaphorically) related elements.

Syntactic segmentation is an important part in anaphora annotation. The text can be divided into larger segments; a segment should include at least one (or more) explicit antecedent binding at least one anaphoric element. More than one clause can be found within a given segment (at least one verb form must be present), with (one or more) subordinate clauses. This principle may test the anaphora scope and the distance between the antecedent and the anaphoric element, and barriers between the two. Cataphora is similarly marked at this stage.

In the text below we will give examples of annotation of the referents of pronouns. If a pronoun constitutes a noun phrase with its own referent, it shall have its own reference value (in the noun phrase *до(м) его* the noun *до(м)* is *'ref=POSS'*, while the pronoun *его* is *ref='HIGH'*). If the pronoun is used as modifier or determiner only, it may not have independent reference value (in the noun phrase *градоу томоу* only the head noun *градоу* is marked by *ref='LOC'*) - these are deictic or article-like demonstrative pronouns, as well as distributive pronouns. Below, we give examples of annotation of the referents and anaphoric relations in fragments of different texts (only the relevant attributes are marked).

a. *Personal (and anaphoric) pronouns.* First and second person pronouns are often found in direct speech, with deictic antecedent or an antecedent in the preceding clause - as in (8); referents are often PERS or HIGH. Third person (anaphoric) pronouns have antecedent that is often in the same segment or in an immediate clause but pronouns linked the same referent can be spread within a series of segments.

(8)

Рече же

<token id='3' form='**Моси**' ref='PERS'/>:

Покажи

<token id='5' form='**ми**' ref='PERS' ana-id='3'/>

<token id='6' form='**г҃и**' ref='HIGH'/>

славѫ

<token id='8' form='**твоѭ**' ref='HIGH' ana-id='6'/>•

да виждѫ

аще обрѣтъ благодѣть прѣдъ

<token id='15' form='**тобоѭ**' ref='HIGH' ana-id='6'/> (Exodus 33, 17-18)•

German Codex, Encomium to the Archangels Michael and Gabriel by Bishop Kliment

(9)

<token id='1' form='**Аврамь**' ref='PERS'/>

имаше вь

<token id='4' form='**ср(д)ци**' ref='BODY' ana-id='1'/>

<token id='5' form='**свое(м)**' ref='PERS' ana-id='1'/>

гостолюбство ·

не хотеше ꙗсти вь

<token id='11' form='**домꙋ'** ref='POSS' ana-id='1'/>

<token id='12' form='**свое(м)**' ref='PERS' ana-id='1'/>

дондеже не прїиде(т)

<token id='16' form='**гость**' ref='PERS'/>

вь

<token id='18' form='**домь**' ref='POSS' ana-id='1'/>

<token id='19' form='**его**' ref='PERS' ana-id='1'/> ·

<token id='20' form='**тог(д)а**' ref='TIME'/>

<token id='21' form='**дїаволь**' ref='HIGH'/>

затворы всe поутїе

да не прїиде

<token id='28' form='**гость**' ref='PERS'/>

вь
<token id='30' form='**до(м)**' ref='POSS' ana-id='1'/>
<token id='31' form='**его**' ref='PERS' ana-id='1'/>·

Adjar Codex (NBKM 326 (509), Sermon on the Holy Trinity

(10)

рече
<token id='2' form='**емоу**' ref='PERS'/>
<token id='3' form='**гъ̄**' ref='HIGH'/>:
Не можеши видѣт
<token id='7' form='**лица**' ref='BODY' ana-id='3'/>
<token id='8' form='**моего**' ref='HIGH' ana-id='3'/>•
Не можетъ бо
<token id='12' form='**члв̄къ**' ref='PERS'/>
видѣвъ
<token id='14' form='**лица**' ref='BODY' ana-id='3'/>
<token id='15' form='**моего**' ref='HIGH' ana-id='3'/>
[живъ] бъıг• нъ
<token id='19' form='**се**' ref='LOC' ana-id='20'/>
<token id='20' form='**мѣсто**' ref='LOC'/>
оу
<token id='22' form='**мене**' ref='HIGH' ana-id='3'/>
станеши при
<token id='25' form='**камени**' ref='LOC'/>
покръıѭ
<token id='27' form='**тѧ**' ref='PERS'/>
<token id='28' form='**рѫкоѭ**' ref='BODY' ana-id='3'/>
<token id='29' form='**моеѭ**' ref='HIGH' ana-id='3'/>•
доньдеже
<token id='31' form='**мимо**' ref='LOC'/>
идѫ
отъимѫ
<token id='34' form='**рѫкѫ**' ref='BODY' ana-id='3'/>
<token id='35' form='**моѭ**' ref='HIGH' ana-id='3'/>•
<token id='36' form='**тъгда**' ref='TIME'/>
видиши
<token id='38' form='**ʒадьнѣѣ**' ref='LOC'/>
<token id='39' form='**моѣ**' ref='HIGH' ana-id='3'/>•
<token id='40' form='**лице**' ref='BODY' ana-id='3'/>
же
<token id='42' form='**мо**e' ref='HIGH' ana-id='3'/>
не ѣвитъ
<token id='45' form='**ти**' ref='PERS'/>
сѧ
(Exodus 33, 20-23)•

German Codex, Encomium to the Archangels Michael and Gabriel by Bishop Kliment

b. Demonstrative pronouns: antecedents are found in a nearby clause (preceding or succeeding); here, the antecedent can be a whole sentence (in the discourse anaphora).

(11)

И видѣ
<token id='3' form='**то**’ ref='DISC'/>
<token id='4' form='**При[деш]ъ**' ref='PERS'/>
<token id='5' form='**кралъ**' ref='PERS' ana-id='4'/>,
ꙗко добро есть
и нача ꙁиздати
<token id='12' form='**град**' ref='LOC'/>,
и съꙁизда
<token id='15' form='градъ' ref='LOC' ana-id='12'/>
до старости
<token id='18' form='**своеѧ**' ref='PERS' ana-id='4'/>,
и нарече
<token id='21' form='**имѧ**' ref='POSS' ana-id='12'/>
<token id='22' form='**градоу**' ref='LOC' ana-id='12'/>
<token id='23' form='**томоу**' ref='PERS' ana-id='22'/>
<token id='24' form='**своимъ**' ref='PERS' ana-id='4'/>
<token id='25' form='**именемъ**' ref='POSS' ana-id-'4'/>,
да
<token id='27' form='**мꙋ**' ref='LOC' ana-id='12'/>
е
<token id='29' form='**имѧ**' ref='POSS' ana-id='12'/>
<token id='30' form='**Прижїа**' ref='LOC' ana-id='12'/>
<token id='31' form='**градъ**' ref='LOC' ana-id='12'/>.

Troya Legend

c. Reflexive pronouns: the antecedent is found in a nearby clause within the same segment.

(12)
А́
<token id='2' form='**Ко́йто**' ref='PERS'/>
<token id='3' form='**гы**' ref='PERS'/>
ꙋзме. и на
<token id='7' form='**ѡнía**' ref='PERS' ana-id='3'/>
погꙋбѣва добро́то, и
<token id='11' form='**себѣ**' ref='PERS' ana-id='2'/>
<token id='12' form='**си**' ref='PERS' ana-id='2'/>
нано́си
вѣчна мѫ́ка на
<token id='1' form='дша́та' ref='BODY' ana-id='2'/>.

Damascenus Troianensis

d. Possessive pronouns: antecedent is often found in a preceding clause but it can be in a preceding sentence or a segment, esp. with deixis.

(13)
рече
<token id='2' form='**емоу**' ref='PERS'/>
<token id='3' form='**г҃ь**' ref='HIGH'/>:
Не можеши видѣт
<token id='7' form='**лица**' ref='BODY' ana-id='3'/>

<token id='8' form='**моего**' ref='HIGH' ana-id='3'/>•
Не можетъ бо
<token id='12' form='**члв̅къ**' ref='PERS'/>
видѣвъ
<token id='14' form='**лица**' ref='BODY' ana-id='3'/>
<token id='15' form='**моего**' ref='HIGH' ana-id='3'/>
[живъ] бъıг•

German Codex, Encomium to the Archangels Michael and Gabriel by Bishop Kliment

d. Interrogative pronouns: the antecedent may be missing or found in a succeeding sentence or segment.

(14)

и
<token id='2' form='**тоги́ва**' ref='TIME'/>
ре́коше
<token id='4' form='**члци**' ref='PERS'/>
<token id='5' form='**ономȣва** ' ref='PERS' ana-id='6'/>
<token id='6' form='**ста́рцȣ** ' ref='PERS' ana-id='6'/>:
<token id='7' form='**Ста́рʿче**' ref='PERS' ana-id='6'/>,
с
<token id='9' form='**кого́**' ref='PERS'/>
дȣмашь, или́
<token id='12' form='**те**' ref='PERS' ana-id='6'/>
нѣ́що бла́зни.

Damascenus Troianensis

e. Relative pronouns: these originate from anaphoric or interrogative pronouns; the antecedent is often close, in a neighbouring clause, with a short path and no more than one barrier in between (these can be a previous word, the head of the previous (referential) phrase (noun or adjective phrase), or the head of the first phrase in the succeeding clause – in correlative constructions).

(15)

Ами́
<token id='2' form='**блже́нь**' ref='PERS'/>
<token id='3' form='**кой**' ref='PERS' ana-id='2'/>
се оупооⷣ́би съсʿ
мла́дыте
<token id='8' form='**дѣца́**' ref='PERS'/>.

Damascenus Troianensis

(16)

разȣмѣ́ й и за
<token id='4' form='**това́**'/>:
<token id='5' form='**Кого́**' ref='PERS'/>
почете́
<token id='7' form='**бь**' ref='HIGH'/>,
и
<token id='9' form='**члци**' ref='PERS'/>
<token id='10' form='**го**' ref='PERS' ana-id='5'/>
почи́тать.

Damascenus Troianensis

(17)

Ами́ w
<token id='3' form='ѻ́че' ref='PERS'/>
чтⷭ҇□´ны
сты́
<token id='6' form='**нико́лае**' ref='PERS' ana-id='3'/>:
не забꙋра́ вей
<token id='9' form='**ра́бы**' ref='PERS'/>
<token id='10' form='**свое́**' ref='PERS' ana-id='3'/>,
<token id='11' form='**дето́**' ref='PERS' ana-id='9'/>
<token id='12' form='**ти**' ref='PERS' ana-id='3'/>
по́честно чи́нать
<token id='15' form='**па́меть**' ref='POSS' ana-id='3'/>:

Damascenus Troianensis

f. Indefinite and negative pronouns: they are most often independent phrases.

(18)

И като́
<token id='3' form='**нѣкой**' ref='PERS'/>
<token id='4' form='**Кога́**' ref='TIME'/>
е ꙋмо́ рень ѿ длы́гъ пъ́ть и изгорѣ́ль ѿ пе́къ и ѻжъднѣ́ль:

Damascenus Troianensis

(19)

и
<token id='2' form='**тїа**' ref='PERS'/>
<token id='3' form='**мꙋ**' ref='PERS'/>
реⷱ҇:
Защо́ нѣ сьⷨ крⷭ҇тена,
и желаѧ да сьⷨ
<token id='13' form='**хртⷭ҇анʹка** ' ref='PERS' ana-id='2'/>:
ато́ не ще́
<token id='17' form='**никой**' ref='PERS'/>
да
<token id='19' form='**ми**' ref='PERS' ana-id='2'/>
стане
<token id='21' form='**крⷭ҇никь**' ref='PERS' ana-id='17'/>,
Защо́ сьⷨ грѣшна.

Damascenus Troianensis

Characteristics of the referent such as animacy / inanimacy, person / non-person, etc. are encoded in the type of the referent (person, animal, possession or ownership, etc.). The focus here is on the pronominal anaphora, though the anaphoric relations with other elements - nouns, adjective - are marked if they bind an anaphorically linked pronoun. With referents of possessee type such as body part, ownership and relatives, the relation points to the possessor.

0. Toward Implementation

Bringing together electronic description, various views of editing text and linguistic annotation is a challenging task. There is no productive way to make annotation in one file for all different approaches to the manuscript text. Therefore, diplomatic, text-critical editions and linguistic corpora should be stored in separate files but in one database with specific searches and queries. If we rely on XML technologies, one possible solution would be to use an XML Native database such as eXist (http://exist-db.org/). This will allow us to take advantage of the whole family of markup languages, such as XLST, XQuery, and XML itself.

An example of annotation of the Zograph Fragments (Les feuillets du Zograph) is given in the Appendix.

References

Bojadžiev, A. 2003. Electronic Student Editions of Medieval Slavic Texts. *Scripta & e-Scripta* 1. 75–88.

Bojadžiev, A., T. Dimitrova 2008. The Linguistic Information in the Electronic Corpus of Old Slavonic Texts. Scripta & e-Scripta 6. 105–149.

Dimitrova, T. The Old Bulgarian Noun Phrase: Towards an Annotation Specification. Doktoravhandlinger ved NTNU: 2008:99. Trondheim: Norwegian University of Science and Technology, 270.

Garside, R., G. N. Leech, T. McEnery, eds. 1997. Corpus annotation: linguistic information from computer text corpora. Taylor & Francis.

Haug, D. 2010. PROIEL guidelines for annotation. <15-07-2018>

Haug, D., M. L. Johndal. 2008. Creating a parallel treebank of the old Indo-European Bible translations. In: Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data. C. Sporleder and K. Ribarov (eds.), 27– 34.

Huang, Y. 2000. Discourse anaphora: Four theoretical models. Journal of pragmatics 32.2: 151-176.

Krause, T., A. Lüdeling, C. Odebrecht, A. Zeldes. 2012. Multiple tokenizations in a diachronic corpus. In Exploring Ancient Languages through Corpora Conference (EALC). 14.-16.06.

Leech, G. 1993. Corpus annotation schemes. Literary and linguistic computing 8.4: 275-281.

Mitkov, R. 2014 (2002). Anaphora resolution. Routledge.

Müller, C., M. Strube. 2006. Multi-level annotation of linguistic data with MMAX2. In: Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods 3: 197-214.

Rissanen, M. 1989. Three problems connected with the use of diachronic corpora. ICAME Journal 13: 16-19.

Rissanen, M. 1992. The diachronic corpus as a window to the history of English. In: Directions in Corpus Linguistics. Proceedings of Nobel Symposium, 82, 185-205.

Roberts, C. 1989. Modal subordination and pronominal anaphora in discourse. Linguistics and philosophy 12.6: 683-721.

Tsvetana Dimitrova is an Assistant Professor at the Department of Computational Linguistics of the Institute for Bulgarian Language, Bulgarian Academy of Sciences. Her research is focused on corpus linguistics, historical corpora, corpus annotation, diachronic syntax, lexical semantic networks.

Andrej Bojadžiev is Professor in Old Bulgarian (Old Church Slavonic) language at the University of Sofia. His publications are in the field of Slavic Cyrillic and Glagolitic paleography and orthography, Slavic historical linguistics and computational approaches to the study of Medieval Slavic culture.

Резюме

В статията се разглеждат възможните подходи за съчетаване на електронното описание, издание и оформянето на езиковите корпуси от средновековни славянски текстове. Особен акцент в публикацията са проблемите на анотирането на различна по тип информация – атрибуцията на текстове, данните за дати и светци, вмъкването или връзката с изображения. Данните за лингвистическата анотация обхващат местоименните форми и анафорични конструкции на морфологично и морфо-синтактично равнище. Примерите са от различни периоди на южнославянската средновековна и ранната новобългарска традиция. Проектът се опира на базата данни eXist, и инициативите Reperorium (http://repertorium.obdurodon.org), PROIEL (http://www.hf.uio.no/ifikk/english/research/projects/proiel/) и TOROT (http://site.uit.no/slavhistcorp/files/2015/04/Eckhoff.pdf).