E-edition and linguistic annotation of Slavic fragments

Tsvetana Dimitrova (Institute for Bulgarian Language)

Andrej Bojadžiev (Sofia University)

Work on experimental annotation of referentiality and anaphora of a collection of fragments has been done along the project "European Identity of Bulgarian Language: In Search of New Research and Methodological Approaches", funded by the National Science Fund of the Republic of Bulgaria.

Plan

- Task
- Layers
- Information from external sources
- Linguistic annotation
- Additional phenomena
- Issues at the preliminary stage

Work on experimental annotation of referentiality and anaphora of a collection of fragments has been done along the project "European Identity of Bulgarian Language: In Search of New Research and Methodological Approaches", funded by the National Science Fund of the Republic of Bulgaria.

The task

- A collection of mediaeval (South) Slavic **text fragments (!)** with access to available information (**even if not filliated yet**; useful also to non-specialists).
- The ultimate goal: to combine electronic description, electronic edition and linguistic annotation of Slavic text fragments **using available tools / sources**.
- Integrated implementation in eXist database: Repertorium project + PROIEL / TOROT approach to linguistic annotation.
- Relational database – for performance reasons!!!!!

Layers of description

1. Diplomatic or paleographic (e-description): information from 'external' sources; links to manuscript if available in digital library resources.

2. Text critical (e-edition): variants linked to one another, with some metadata – **if available**.

3. Linguistic (corpus): various language levels including grammar, lexicon needed for establishing the properties of a language as attested in the fragments;

- **other phenomena** expressed linguistically but can be related to 'external' phenomena.

E-description

If the manuscript is already described in electronic format: a link + a brief on-hand description (for non-specialists).

In lack of previous research – some attribution of various parts of the source, plus useful links:

<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>

Repertorium

A short overview about the text recension in the copy (extracted from the Repertorium database):

<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; (...). In several cases it has 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>

Additional information, links

Links to various sources of information such as:

DBPedia: <ref target="http://dbpedia.org/page/Saint_Pantaleon">...</ref>

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

Menology project: <ref target="http://menology.obdurodon.org/tableSearch.php?mss%5B%5D =H&mss%5B%5D=P&mss%5B%5D=En&mss%5B%5D=Bas&mss%5B%5D=Oh&mss%5B%5D=Os&startDate=07%2F27&endDate=07%2F27">27 July</ref>

Linguistic annotation

Stick to available and **tested** approaches! (PROIEL / TOROT… Why? **Open source,** available tools, a **considerable amount of data** has been annotated to train the tools.)

Levels covered: lemma (reference / semantic tags); class / subclass of part-of-speech; morphology; syntax (dependencies); information status; alignment; …

The annotation scheme makes it possible to include additional elements if necessary (as an additional xml-tag attribute; or as part of the value of the xml-tag attribute – in the positional tag)?

Work on experimental annotation of referentiality and anaphora of a collection of fragments has been done along the project "European Identity of Bulgarian Language: In Search of New Research and Methodological Approaches", funded by the National Science Fund of the Republic of Bulgaria.

Anaphora-related marking

Experimental marking of pronominal anaphora – information needed:

1) Referents – most are already part of the PROIEL / TOROT reference tags (person, higher being, animal, body part, kinship, general relatedness, location, time, legal entity).

Elements to consider: nouns (both common and esp.! proper – reference tags), adjectives (referential – denominal, etc. – reference tags), adverbs (time, location – reference tags), and pronouns (reference information can be inherited along the syntactic tree).

2) Anaphoric relation: anaphorically-related elements (linked via id).

Examples from Zograph Fragments only here; other examples – in the article.

Referents

Person (PERS; humrel, human):
<token id="9" form="иже" lemma="иже" part-of-speech="Pr" morphology="-s---mn--i" relation="sub" **ref="PERS"** ana-id="15"/>

Higher being (HIGH; humrel, human *again*):
<token id="97" form="бжїѧ" lemma="божии" part-of-speech="A-" morphology="-s---fgpsi" relation="atr" **ref="HIGH"**/>

Kinship (REL; kinship, human):
<token id="101" form="братиеѣ" lemma="братьꙗ" part-of-speech="Nb" morphology="-s---fi--i" relation="ag" **ref="REL"**/>

Referents

Body part (BODY; body_part):
<token id="31" form="оумѣ" lemma="оумъ" part-of-speech="Nb" morphology="-s---ml--i" relation="obl" **ref="BODY"** />

Location (LOC; location)
<token id="86" form="сьде" lemma="сьде" part-of-speech="Df" morphology="---------n" relation="adv" **ref="LOC"**/>

Time (TIME; time)
<token id="233" form="когда" lemma="когда" part-of-speech="Du" morphology="---------n" relation="adv" **ref="TIME"**/>

General relatedness (POSS; *varia!*)
<token id="29" form="житии" lemma="житиє" part-of-speech="Nb" morphology="-s---nl--i" relation="obl" **ref="POSS"** />

Example

Part of the same sentence (иже…, тъ…); reference information can be inherited syntactically

<sentence> <token id="9" **form="иже" ref="PERS" ana-id="15"**/>
<token id="10" form="створитъ"/>
<token id="11" form="волѭ" ana-id="0"/> <token id="12" form="оцѣа" ref="REL"/>
<token id="13" form="моего" ref="PERS" ana-id="18"/> <token id="14" form="небескаго"/>
<token id="15" **form="тъ" ref="PERS" ana-id="9"**/>
<token id="16" form="естъ"/>
<token id="17" form="мати" ref="REL"/> <token id="18" form="моѧ" ref="PERS" ana-id="13"/><token id="19" form="и" ana-id="0"/><token id="20" form="оць" ref="REL"/><token id="21" form="и" /><slash target-id="15" relation="xsub"/></token>
<token id="22" form="братъ" ref="REL"/></sentence>

Example

Across sentences (the unique id of the token keeps it):

<sentence><token id="71" form="вьторое"/><token id="72" form="же" />
<token id="73" form="ꙗко" />
<token id="74" form="недостоинѣе" ref="PERS"/><token id="75" form="приемати"/>
<token id="76" form="съмѣетъ" /><token id="77" form="достоиньіимь" ref="PERS"/>
<token id="78" form="съготованьіихъ"/><token id="79" empty-token-sort="V"/>
</sentence>

<sentence><token id="80" **form="тѣмь" ref="PERS" ana-id="77"**/> <token id="81"
form="же" /><token id="82" form="потрѣбъно" />
<token id="83" form="естъ" /><token id="84" form="въспомѣнѫти"/>
(...) </sentence>

Example

124 and 125 are syntactically linked. No antecedent – 123; Links to external sources can be included...

<quote type="Апостол" n="Ефес. 4:1">
<sentence><token id="122" form="молѭ"/>
<token id="123" form="вьі" ref="PERS"/>
<token id="124" form="азъ" ref="PERS"/>
<token id="125" form="съвѧзаньіи" ref="PERS"/>
<token id="126" form="о" /><token id="127" form="гꙗи" ref="HIGH"/>
<token id="128" form="подобнѣ"/> <token id="129" form="ходити"/>
<token id="130" form="по"/> <token id="131" form="зъванию"/>
<token id="132" form="имьже" **ana-id="131"**/> <token id="133" form="позъвасте"/>
<token id="134" form="сѧ"/>
</sentence></quote>

Issues

- Integration between the levels of description (between eXist and PROIEL annotation files).
- An obstacle for linking the images from the manuscripts: dynamic URLs.
- Linguistic annotation: only one interpretation, hence marking;
- Anaphora-related annotation:
  - Redundant if the elements are already syntactically linked – information can be inherited within the syntactic tree. Some redundancy or partial redundancy (PERS and HIGH; ANIM can be POSS; etc.).
  - More than one anaphorically-related element.

Хвала!