Diachronic Corpus of the Old Bulgarian Language – content and research opportunities

A. Totomanova, Sofia University

The ***Diachronic Corpus of the Bulgarian Language*** is the core of the Histdict system, which is developing as a unique electronic research infrastructure. It was meant initially to serve only for extracting lexical material for composing the historical dictionary of Bulgarian. By the end of our first project however we realized that the diachronic corpus is an excellent tool for presenting the Bulgarian literary heritage of the period from the 10th to the 18th century in all of its genres and across its thematic diversity. The software of the corpus developed by ICT specialists from *Openintegra company* is *user friendly* and very easy to use. It includes also electronic tools for text commentaries (both paleographic and codicological) as well as for visualizing variant readings and creates new opportunities for adequate presentation of the medieval Slavonic texts. It was applied to the digital edition of the Chronograph of Archive, that was planned under the project "The Concepts of History across the Orthodox Slavic World" and to other electronic publications. Just a few days ago we received a cooperation proposal from some German colleagues who would like to publish the electronic edition of the Slavonic texts related to Moses in our corpus. (Fig. 6-11 show the Corpus functionalities.)

It is fully transferable and could be used for digital processing of texts, for creating corpora and dictionaries of different languages.

The corpus contains over150 texts of proven Bulgarian origin from different genres of the 10^{th} to 18^{th} centuries. Given the fact that Bulgarian literature transmitted the Byzantine cultural and literary model to the other Orthodox nations in our part of Europe, the corpus contains both translated and original Medieval Bulgarian texts. The corpus also includes Early Modern Bulgarian texts – mostly Damascenes and other compilations as well as some non-literary texts such as scribal notes, inscriptions and juridical documents.Thus, the corpus contains works by Clement of Ohrid, John the Exarch, Constantine of Preslav, Patriarch Euthymius, and Constantine of Kostenets; also included are the texts of the Manasses chronicle and the Troyan parable, the Philippi Monotropi Dioptra, the Wallachian-Bulgarian charts, Paisii Hilendarski's Slavonic-Bulgarian History, the Lovech and Troyan Damascenes, etc. Furthermore, the corpus features chronicles, pieces of monastic literature, historical and apocalyptic texts, legal texts, miscellanies with stable and mixed content, and codicils. The texts are digitally typed, and all of them (excluding the works of St. Kliment Ohridski) are reproduced in the orthography, in which they survived (Bulgarian, Serbian or Russian). At present, the electronic corpus does not include any texts from the, such as Codex Marianus, Codex Zographensis, Codex Assemanius, Sava's book, Codex Suprasliensis, etc.; these manuscripts have been lexicographically processed and their material included (with contexts) in the two-volume Old Bulgarian dictionary of the Institute for Bulgarian Language at the Bulgarian Academy of Sciences.

For less than 10 years we managed to double the number of the texts in the corpus and we still continue uploading new texts. Some of them are provided from colleagues abroad who also use the corpus. The above-mentioned Chronograph of the Archive that includes the oldest text of the Octateuch and Kingdoms translated in Bulgaria during the reign of Simeon the Great (893-927) is one of the largest text in the system. Our ambition is to include into the corpus all medieval texts produced or translated into Bulgarian between 10^{th}-18^{th} cc.

The font Cyrillica Bulgarian10U, that was the first UTF font we produced was used for compiling the corpus. The first version of the corpus was created through converting already keyboarded texts and uploading them onto the system. The third version of the converter we are using now has more functionalities and allows us to change the outlook of the texts we prepare for editions if needed.

The corpus interface displays the original titles of the writings, the date of the manuscripts, their genre and orthography. For our users' convenience, the text titles of the translated worus have also been translated into Latin.

Each text included in the corpus is preceded by metadata set including title, provenance, author, dating, orthography, etc. As you can see the information in the interface rubrics has been extracted from these archaeographic descriptions.

Uploading the texts is very easy and fast with Copy and Paste options using the Menu *New text.*

The software of the corpus provides different functionalities for codicological, palaeographic, and textological commentaries.

- red letters
- make paleographic comments on the handwriting, abbreviations, peculiar shape of some characters and diacritics etc.
- make codicological comments on the paper/parchment quality, ink, binding etc.
- comments on scribal errors
- comments on errors in translation
- report missing parts of the text (that might be a result of physical damages like a loss of folia in the codex, or of a scribal distraction, or might reflect some editorial intervention)
- identify biblical quotations
- report variant readings

Variant readings appear in blue on the screen, while other comments are marked in yellow. In case of overlapping comments, the text becomes green. As you can see, the above listed functionalities of the corpus software allow for producing electronic editions supplied with proper apparatus criticus. Here it is worth mentioning that the tradition of producing critical editions of Slavonic manuscripts differs from the Western one. While the western scholars tend to reconstruct original version of the text preserved in a certain number of witnesses, the eastern editors choose one of the witnesses as a basis of the edition and after reproducing it adequately select the variant readings to the main text. To the great extent it predetermined the fact that in our corpus almost all texts reflect the original spelling.

A few years ago, we started realizing that we would have to expand the functionalities of the corpus towards other kinds of annotation. In the first place it applies to the morphological annotation if our efforts to produce an automatic morphological tagger prove to be successful. In the second place, we need a content annotation for the texts translated form Greek that will not only facilitate the work of the researchers involved in Palaeoslavic studies but will also attract the Byzantinists that would like to know whether a certain Greek text or author was translated into Old Bulgarian (OCS). The latter would boost up the Slavonic-Byzantine studies and strengthened the cooperation between the researchers coming from different traditions and countries.

The corpus is searchable through the Histdict system search engine. The engine allows for finding the letter strings in all electronic resources. The user can also choose the type of resource he wants to explore. However, the search engine needs to be upgraded and refined. In its present state it displays the documents in which the string we are searching for is found and the number of occurrences. In order to accede the lexical material, one should open the document and using the browser search options retrieve the needed information. The new search engine should be able to search by wordforms, beginnings and endings of the words, and at the end by lemmas. The latter will become possible once the automatic morphological annotation is set in place.

One of the great advantages of the corpus software is the opportunity to edit the texts and the metadata any time you find it necessary.

Yet not all users are given this right.

Common users can read the texts they are interested in without having installing the the font, which is readable from all web-browsers. In case someone wants to copy the text and place it in a Word document the font Cyrillica Bulgarian should be installed on his computer.

Common users can also use the search engine in order to collect rapidly and easily the information they need

Editors are allowed to:

Upload texts using functionalities for introducing different kinds of commentary

Edit texts and metadata

The highest level of editors can decide on publishing, hiding or deleting already published texts.

The Diachronic Corpus of Bulgarian we created is the first of this kind since it is connected to a dictionary and supplied with respective electronic tools for text processing. The electronic source might have many applications and could be used for:

1. **Producing e-based lexicographic manuals of different types :**

- Diachronic Historical Dictionaries
- Historical Dictionaries of synchronic type (Dictionaries of Literature or of different authors, different periods etc.)
- Glossaries
- Thematic dictionaries
- Etymological dictionaries

2. **Historical Linguistic Studies in the area of:**

- Morphology and Morphosyntax
- Morphonology
- Phonetics
- Lexicology
- Etymology

- Derivation
- Phraseology
- Textology
- Orthography

3. **University education on all levels (bachelor, master, doctor) in the field of:**

- Palaeoslavonic and Old Church Slavonic Studies
- History of Bulgarian Language
- History of Literary Bulgarian
- Old Bulgarian Literature
- Medieval History
- Computer and Corpus based linguistics

4. **Preparing the editions (both traditional and electronic) of :**

- Medieval texts
- Dictionaries, Glossaries etc.
- Textbooks, Handbooks, Manuals etc.

5. **Presenting Bulgarian Cultural Heritage**

In its present shape the corpus represents an excellent research tool for the scholars in the area of Slavic studies who can read Cyrillic. A tender for upgrading the histdict system was announced and very soon we will have the results of the bid, hoping that some serious and reliable ICT company will take the chance to be involved in producing such non-commercial software.

In order to open the corpus to users from outside the Slavic world and make the full use of it as a tool for popularizing Bulgarian cultural heritage we need first of all to translate the menu of the Corpus and other digital resources as well and to combine the Medieval texts with translations in modern languages. For this purpose we intent to use projects like ReIReS and RESILIENCE thus integrating our research and resources in large European infrastructures.