Studia Ceranea 2, 2012, p. 221–232

Anna-Maria Totomanova (Sofia)

Digital Presentation of Bulgarian Lexical Heritage. Towards an Electronic Historical Dictionary

The project *ICT Tools for Historical Linguistic Studies*, funded by the European Social Fund, OP Human Resources, was designed and carried out with the idea to introduce ICT in such a conservative field as diachronic linguistics. The objective we pursued was twofold:

- to speed up the data collecting from the books created between 10th and 18th cent. and accelerate further data processing;
- to make diachronic linguistics more attractive for young people born in the Computer Age for whom computers are part of their natural habitat.

The Round Table *Interactive Methods in Historical Lexicology and Lexicography* held on 28.05.2010 played a crucial role for the project development. The participants reviewed and summarized the experience in the area of historical lexicography and made the following important decisions:

1. The project should focus on creating software tools for developing a web based Historical Dictionary of Bulgarian, which is the first literary and sacred language of the Slavs with a long written history.

2. *Старобългарски речник* (Old Bulgarian Dictionary), created by the Department of History of Bulgarian Language at the Institute for Bulgarian Language, will constitute the foundation for building a Historical Dictionary of Bulgarian. For this purpose the information it includes will not only be preserved but also enriched and upgraded with materials taken from the Electronic Corpus of Medieval and Early Modern Bulgarian texts.

The project target group participants (PhD and Post-Doc students, young researchers and interns) were assigned individual research tasks in compliance with the decisions made. The Round Table produced a preliminary list of electronic tools for digital processing of the texts. The ***Standard of the Dictionary*** took shape during the project course based on the decision that we are aiming at designing a ***Historical Dictionary of Diachronic Type***[1] ***that should present the history of the Bulgarian words***

[1] The terms *Diachronic* and *Synchronic Historical Dictionaries* were introduced and explained by: Г.А. Богатова, *Историческая лексикография как жанр*, ВЯ, 1981, p. 83–84.

from their first written occurrence until today. Such a Historical Dictionary has the following features:

- **Large chronological span,** starting from the beginning of the Slavonic writing in the 9th cent. up to the modern times;
- **Thematically unlimited text corpus** that includes: literary texts; non-literary texts (geographic and personal names, dialects, vernacular language, inscriptions, graffiti);
- **Open vocabulary** that will be enriched while the corpus building;
- **Diachronic presentation of the lexical material,** which implies the registration of the different meanings of the word and their genetic connection.

The Text Corpus of the Dictionary should include:

- **Bulgarian medieval texts:** works of the Old-Bulgarian writers; translations from Greek with proven Bulgarian origins (works of the Holy Fathers, Chronicles, monastic literature, Historical and Apocalyptic texts, juridical texts, miscellanies with stable and mixed content etc.);
- **Non-Literary texts:** notes of the copyists; inscriptions and graffiti; charts;
- **Early Modern Bulgarian texts** (mostly *Damaskins* and *Damaskin miscellanies*);
- **Dialectal texts.**

To create the electronic base of the Historical Dictionary the following electronic tools are needed:

- Digitalized *Старобългарски речник*;
- Specialized ***Diachronic Corpus of Medieval Bulgarian and Early Modern Bulgarian texts***;
- Other specialized corpora, such as the *Bulgarian National Corpus (Български национален корпус)*[2], ***dialectal corpora, BgSpeech Corpus (Корпус на българската разговорна реч)***[3] and so on.

Since the work on the other specialized corpora had already begun, the project team efforts concentrated on creating the Corpus of Medieval and Early Modern Bulgarian texts and on digitalizing the two volumes of *Старобългарски речник*. The creation of a new Old Bulgarian font was the first step towards the electronic processing of the medieval texts.

In the beginning of 2010 we already had at our disposal a new Old Bulgarian font based on Unicode, containing more signs than the previously existing Old Bulgarian Unicode fonts. The font has already successfully been used for the digital typing and publishing of some medieval texts. The medieval texts in the last three books of the series "History and Literature" were con-

[2] See the description and opportunities of using the BG National corpus on http://www.ibl.bas.bg/BGNC_bg.htm.

[3] The corpus was developed as a part of BgSpeech initiative and it is maintained by the Faculty of Slavic Studies at Sofia University at http://bgspeech.net/.

verted into the new font. The same font is being used for publishing the text of the Bulgarian, Russian and Serbian Synodika for the planned *Brepols* edition *COGD IV*[4] as well as for the electronic edition of the so called Архивский хронограф we are preparing under another project. The project team contributed a lot to the improvement of the font functionalities by providing valuable feedback to the software specialists.

The collaboration between the ICT specialists and project participants produced the synergy for the successful use of the font ***Cyrillica Bulgarian 10 U*** under different types of editing and publishing software and facilitated the Pre-print processing of medieval Slavonic texts. The font was initially elaborated under the project "The Concepts of History across the Orthodox Slavic World" but it was used for the first time and substantially improved under this project. The same font is used by the editorial project for publishing Slavic Synodica as well as by the project *Pragmatic Function Words: A Corpus-Based Description of Variation* run by O. Mladenova at University of Calgary, Canada. The technological development and the mass introduction of the so called *web fonts* in browsers allow the users to read the font without installing it in their own operating systems (fig. 1).

Together with the font a convertor was produced that converts the texts typed with the *Synthesis Soft* fonts into Unicode-based documents. All project participants contributed to the testing and improvement of the convertor and learned how to apply it, converting already typed texts for the diachronic corpus of Bulgarian. By the end of the project the convertor functionalities were expanded to all Synthesis Soft fonts plus the Italian Pop-Retkov font, which is of great importance since our Italian colleagues provided us with the digitally typed Alphabetical[5] and Roman[6] pateriks (fig. 2). Two additional Unicode fonts were included as well: Cyrillica Ochrid 10 U and Cyrillica Old Style 10 U, designed for typing Early Modern Bulgarian texts.

The font ***Cyrillica Bulgarian 10 U*** was used for digitalizing the two volumes of *Старобългарски речник*, produced by IBL. We express our gratitude to the ICT consultant Mr. Todor Todorov, who developed the font and the convertor and created a second specialized convertor/generator that successfully converted the dictionary containing 11000 entries into a structured XML document without losing a bit of existing information. This second convertor facilitates the process of converting other medieval texts already published on paper, such as *Германов сборник* for example. The software specialists from *Openintegra* elaborated software for editing, expanding and visualizing the

[4] COGD. I–VII. A Special Series of *Corpus Christianorum* by Brepols, 2006 – An International Research Program launched in Bologna and directed by †Giuseppe Alberigo and Alberto Melloni of FSCIRE, Fondazione per le Scienze Religiose Giovanni XXIII, Bologna.

[5] R. Caldarelli, *Il Paterik Alfabetico-Anonimo nella traduzione antico-slava*, Roma 1996.

[6] К. Диди, *Патерик Римский. Диалоги Григория Великого в древнеславянском переводе*, Москва 2001.

dictionary in web environment. It allows an easy and quick access to the media and contributes to popularizing the work of the team all over the world. It also enables data exchange between our institution and other universities since the dictionary is based on the globally recognized standard TEI in XML area. The digitalized Old Bulgarian Dictionary is located on the project web page and is accessible for all customers at ***histdict.uni-sofia.bg***. We are proud to say that it is the first digitally presented Palaeoslavonic lexicographic manual (fig. 3 and 4).

At the same address ***histdict.uni-sofia.bg*** one can find also the Diachronic Text Corpus, which already contains more than 75 texts of different length and the text collection is constantly growing. The corpus includes medieval Slavonic texts with proven Bulgarian origins and different orthography (Old Bulgarian – OCS, Middle Bulgarian, Resavian and Russian), Early Modern Bulgarian texts and notes of the medieval copyists. Translations and original works of the Old Bulgarian writers are equally represented in their genre variety – liturgical, exegetical, hagiographic, juridical, chronographic, historical and apocalyptical texts and so on. Some of them have not been published before.

Most project participants actively committed themselves to the workshop held on 20.11.2011, which was dedicated to the digital presentation of the medieval texts in the corpus. To our great satisfaction, in two weeks all interested parties – the project team, target group representatives, tutors and ICT specialists – all together managed to add the corpus a bigger number of texts than it was initially planned. The ICT specialists from *Openintegra company* supported our team, helping to alleviate errors that occurred during the testing while entering texts, and added new functionalities to the corpus software as suggested by the team. We consider that to be an enormous success, given the fact that this is the first diachronic corpus based on Slavonic material connected to the elaboration of a historical dictionary and provided with a program for linguistic annotation.

The software we developed is *user friendly* and very easy to use. The electronic tools for text commentaries (both paleographic and codicological) as well as for visualizing variant readings create new opportunities for the adequate presentation of the medieval Slavonic texts that will be included in the digital edition of the Chronograph of Archive, planned under the project "The Concepts of History across the Orthodox Slavic World", and other electronic publications (fig. 6–11 show the Corpus functionalities).

The software is fully transferable and may be used for digital processing of texts or for creating corpora and dictionaries of different languages. That is why the software developers and the team have the intention to publish it as an Open source material, so that our colleagues from abroad might access it. In return we hope to receive from them some ideas about its further improvement and application.

The corpus itself turned out to be a wonderful tool for the digital presentation of the Bulgarian lexical heritage in a diachronic perspective. The openness and accessibility of the data it contains provide opportunities for its expansion through adding new meanings and lexemes. Uploading texts is very simple and the copyright of the authors is preserved through the introduction of different access levels.

The corpus is also a study tool and could be easily Utilized in the teaching-learning process in the area of Palaeoslavonic and Medieval studies as well as in diachronic linguistics.

The corpus is supplied with a *Search engine* that allows searching the texts by metadata (author, genre, orthography etc.) as well as directly in the text content.

A programme for editing the articles of the digitalized *Старобългарски речник* was developed to make the dictionary the basis for creating the Historical Dictionary of Bulgarian. We have already started adding new lexemes that are not registered in the Old Bulgarian manuscripts and developed a number of new dictionary units using the experience and methodology of the authors of *Старобългарски речник* (fig. 5).

Yet the real work on the dictionary is only about to start. For this purpose we have to focus our efforts on the following directions: Developing new dictionary entries.

Expanding the chronological coverage of the existing dictionary entries.

Editing the units/articles of the Historical Dictionary.

In order to solve these problems we have to establish a connection between the Corpus and the Historical Dictionary, which shall allow us to discover both the missing lexemes and the new previously unregistered meanings. Producing glossaries and lists of lexemes for lexicographically unexplored texts from the corpus will be one of the project spin-off results. I do not think, however, that we should overlook the materials that can be found in already published lexicographic manuals. Adding new dictionary entries and new meanings in the existing ones will require a careful editing of *Старобългарски речник* entries, since the Historical Dictionary will rather focus on tracking the development of the word meaning throughout the centuries than on the exhaustive presentation of the lexical material. But we are still at the beginning and expect to gain valuable experience in this regard.

The set of electronic tools for creating corpora and dictionaries on medieval Bulgarian text material seems to be the most impressive and important project result. I am deeply convinced that the free access to both the corpus and the digital version of the dictionary will attract to our work many followers from both the country and abroad who will contribute to this extremely important lexicographic project.

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 since it 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

Abstract. The article presents the results of the project *"ICT Tools for Historical Linguistic Studies"*, funded by the European Social Fund, OP Human Resources. The main project goal was to elaborate electronic tools for creating a Historical Dictionary of Diachronic Type that should present the history of the Bulgarian words from their first written occurrence until today. By the end of the project the team (Faculty of Slavic Studies at Sofia University, Institute for Bulgarian Language, BAS and

PAM Publishing Company, Sofia) had at their disposal a set of *Old Bulgarian Unicode fonts*, meant for publishing medieval texts and a *convertor* that converts non-Unicode documents into the new standard. The convertor allowed the participants to create in a relatively short time a *Diachronic text corpus of Bulgarian medieval texts*, containing already more than 90 texts dated from the 10th to the 18th century. The corpus software enables editing the texts and turned out to be an excellent tool for preparing electronic editions of the Old Bulgarian (OCS) manuscripts. In addition to the corpus an *electronic dictionary of Old Bulgarian* is available, which contains the digitized version of *Старобългарски речник*, produced by IBL. Both tools are accessible on the project website at the address *histdict.uni-sofia.bg*. The *Standard of the Historical Dictionary* took shape during the project course and respective software for elaborating new dictionary entries was designed and tested. The article also displays screenshots that demonstrate the functionalities of both the corpus and dictionary software.

Anna-Maria Totomanova
St. Clement of Ohrid University of Sofia
15 Tsar Osvoboditel blvd.
1000 Sofia, Bulgaria
atotomanova@abv.bg

Figures:

Fig. 1. Cyrillica Bulgarian 10 U.

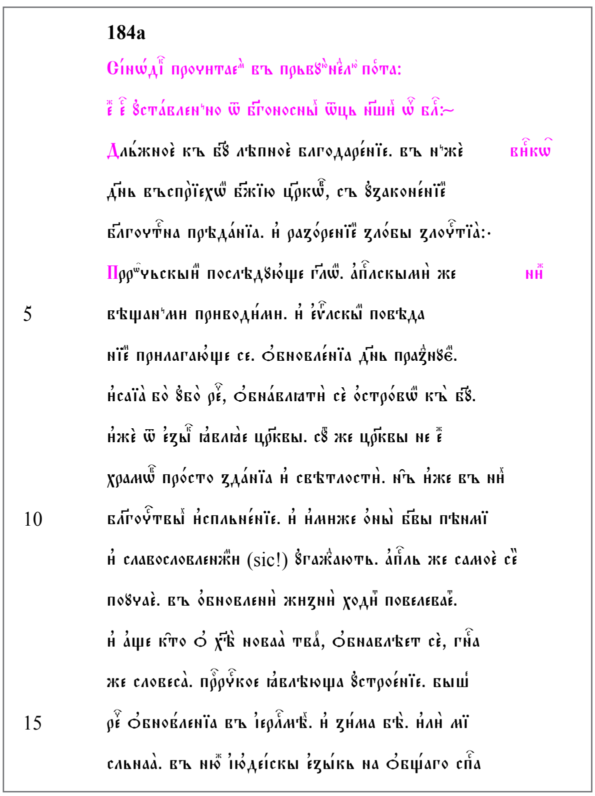

Fig. 2. Convertor interface.

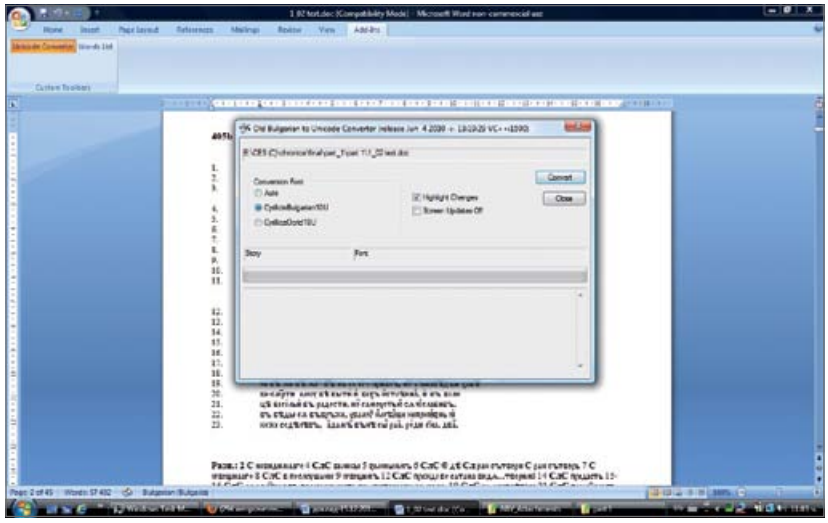

Fig. 3. Digitalized *Старобългарски речник* Interface (Lexeme search).

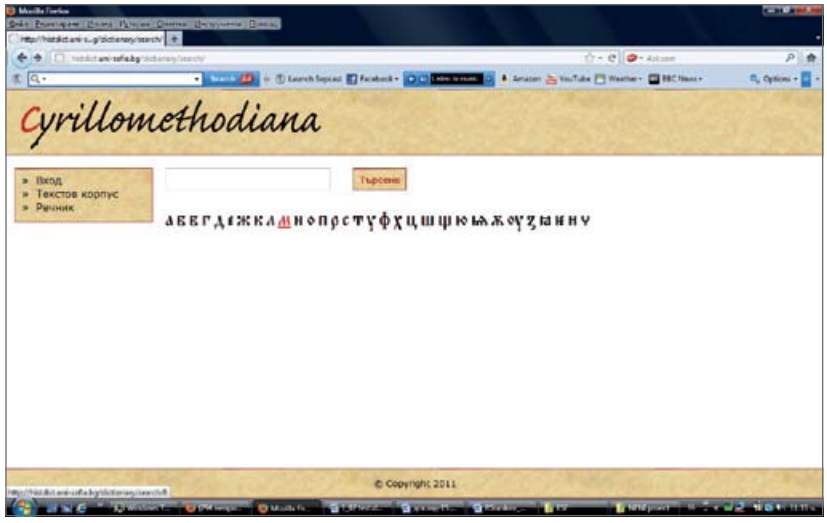

Fig. 4. Digitalized *Старобългарски речник* Interface (Dictionary entries).

алєксандровъ

алєксандровъ *прил притеж* от ЛИ

1. Александров, на Александър — синът на Симон Киринеец ı ӡадѣша мимо ходѧштоу єдиномоу симоноу кѵрѣниноу ... отцоу алєѯандровоу. и роуфовоу *М* Мк 15.21 *З А СК*

2. Александров, на Александър — презвитер в Сид [Памфилия], умр. мъченически при имп. Аврелиан [270—275 г.] съпричьникъ бѫди надълежѫштии съмрьти алеѯан'дровѣ *С* 161.2

Изч *М З А СК С* Гр [τοῦ] Ἀλεξάνδρου **алєѯандровъ алєксаньдровъ алєѯан'дровъ** Нвб александров ОА ВА Срв Александров *ФИ* СтИл,РЛФИ Александрово *ср МИ* ПК,Пр. в им

агньць

агньць *а м*

1. Агнец, агне ıдѣте се аӡъ посылаѭ вы. ѣко агньцѧ по срѣдѣ влъкъ *М* Лк 10.3 *Срв.* С 534.26 горы вьӡıграшѧ сѧ ѣко овьни. ı хлъми ѣко агньци овьчии *СП* 113.4 *Срв.* СП 113.6 *СЕ* 3b 10—11 мо на агньцемь на пасхѫ. приӡьри ги исхе. на си брашьна твоѣ. ı на агнець сь. ı сти и. ѣкоже стити иӡволи агнець. ıже приведе авелъ во в'сѣсъжагаемаа *СЕ* 16b 2, 4, 5 ıюдѣı же съвлаӡаѭште агнець ӡакалахѫ. а ıже отъ поганъ. въ плъть ба *К* 13b 14 *Срв.* С 450.21 акы овьча на ӡаколении ведень бысть. и акы агньць *С* 434.25—26 *Срв.* С 437.2 Образно. егда же обѣдоваша. гла симонови. петроу ис. симоне иониинъ любиши ли мѧ паче (сихъ). гла емоу еи ги. ты вѣси ѣко люблѭ тѧ. гла емоу паси агньцѧ моѩ *М* Йо 21.15 *З А*

2. В християнството — название на *Исус Христос*, който е принесен в жертва като изкупление за греха на човечеството ги бже нашъ. прѣдъложен сѧ самъ. агнець непороченъ. ӡа животъ в'сего мира. приӡьри на ны. ı на (х)лѣбъ сь. ı на чашѫ сиѭ. ı съ(т)вори ѭ прѣчистое тѣло твое. ха *СС* IIb 4 бако ги бже нашъ. в'седръжителю. ıстиннъны агньче. въӡемлаи грѣхы в'сего мира. не прѣӡьри дшъ молащь сѧ тебѣ *СЕ* 15a 4 стоıаше на крьстѣ агньць. и два влька *С* 437.15

агньць божьи
ὁ ἀμνὸς τοῦ θεοῦ Агнец божий — изкупителна жертва [за Исус Христос]

въ оутрѣи день видѣ иса грѧдѫшта къ себѣ. ı гла се агнецъ бжıи. въӡемлаи грѣхы мира вьсего *М* Йо 1.29 *З А СК Б* ı оуӡьрѣ иса грѧдѫшта. гла се агнецъ бжıи *М* Йо 1.36 *З А* си бо вьса быша. да отънемьли грѣхъ мироу. агньцъ и снъ божии. волѭ на сьпасьнѫѭ страсть съ вами придетъ. и на продании станетъ *С* 331.25

М З А СК Б СП СС СЕ К С Гр ἀρήν ἀρνίον ἀμνός **агнець ѣгньць агнецъ агньцъ** Нвб агнец *остар* ОА ВА НТ НГер ЕтБАН ЕтМл МлБТР АР РБЕ РРОДД

Fig. 5. Dictionary Entry Editing Tool Interface.

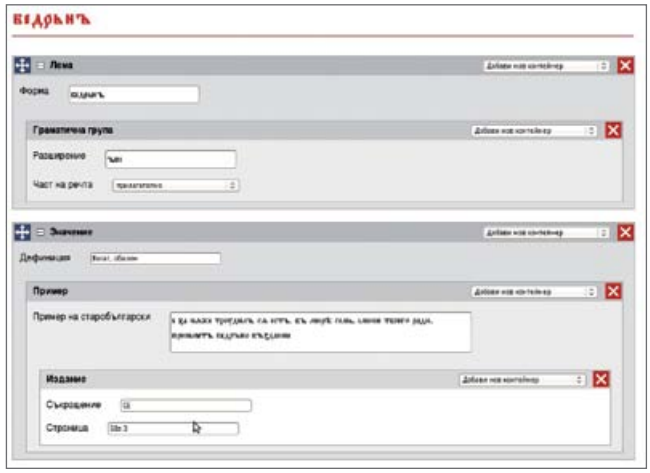

Fig. 6. Corpus Interface (Text search)

Fig. 7. Corpus functionalities (Metadata editing)

Fig. 8. Corpus Interface (Entering/editing texts)

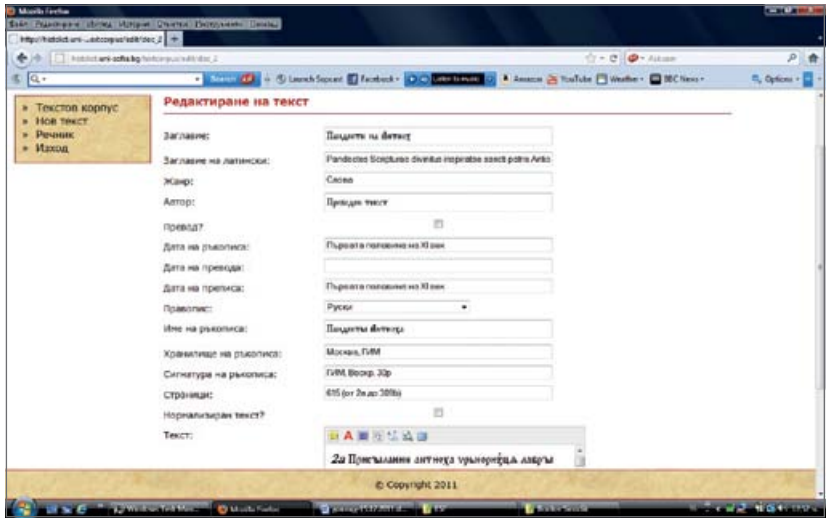

Fig. 9. Corpus functionalities (Footnote)

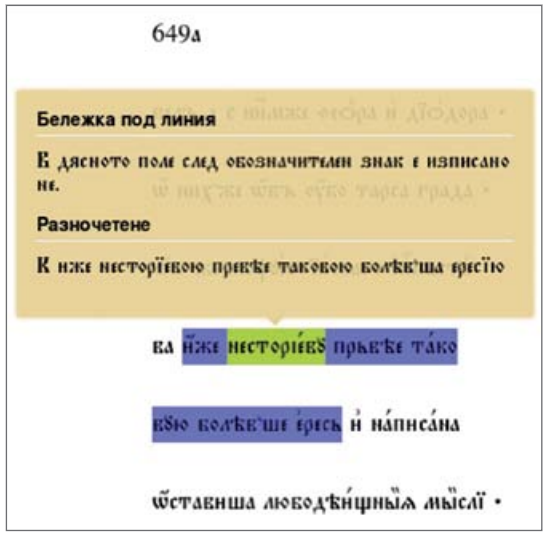

Fig. 10. Corpus functionalities (Variant readings)

Fig. 11. Corpus functionalities (Red letters)