

Dear guests, I am glad to welcome you here. In this first part I will give you a description of our *histdict* system. Afterwards together with Dr. Totmanova-Paneva we've prepared a short training, in which you will translate a phrase from Old Bulgarian. I'd like to apologize for reading the lecture. Please also excuse my less than ideal English.

The *histdict* electronic resource (available at [histdict.uni-sofia.bg](http://histdict.uni-sofia.bg)), launched over nine years ago, is the first attempt at creating a representative diachronic corpus of the Bulgarian language, as well as a historical dictionary and set of electronic tools for processing medieval texts. The work on the *histdict* system started with the development of specialized Old Bulgarian Unicode fonts with a diverse inventory of letters and diacritic marks. Three such fonts were developed at that time: Cyrillica Bulgarian 10U, Cyrillica Ohrid 10U and Cyrillica OldStyle 10U, which is designated for Early Modern Bulgarian texts. The fonts also feature a convertor that converts texts typed in non-Unicode fonts into Unicode. Naturally, the objective was for the electronic resources to be accessible to everyone, not requiring the user to have the respective font.

The objective of the *histdict* electronic diachronic corpus is to present 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 corpus comprises texts of certain Bulgarian provenance – both original works and translations by Bulgarian scribes (including ones preserved in later Russian and Serbian copies). 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 diplomas, 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. At present, the electronic corpus does not include any texts from the classic Old Bulgarian (Old Church Slavonic) corpus, 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. The dictionary has been digitalized and incorporated in the *histdict* system. Inasmuch as the electronic version of the dictionary is a type of annotated corpus, it may be used for automated searching and extracting of information.

The documents in the corpus normally reflect the original spelling of the manuscripts or editions they have been drawn from. The corpus is freely accessible and includes certain text annotation tools: comments, variant readings, paleographic and codicological notes, etc. Footnotes are marked in yellow, variant readings in blue, while words marked for both footnotes and variant readings are displayed in green. For our users' convenience, some of the text titles have also been translated into Latin.

As regards the contents of the electronic diachronic corpus, it may be noted that – in the perfect scenario – this should be a matter of policy rather than a given person's enthusiasm or subjective assessment. The matter of representation and sampling is a highly contentious issue in diachronic corpora. Only a diachronic corpus with a highly diverse and rich content may be claimed to have a truly representative character. Thus, it is mandatory for a diachronic corpus to have a clear underlying concept of its nature and of the identity of the texts that it is supposed to include – so that it does not omit any works of importance for the relevant language and its literary history, and so that a certain type of texts does not dominate over others. Simply put, the diachronic corpus of a particular language should never be a mere mechanical collection of works, and it should by no means be used as a tool for including and sharing unwarranted texts based solely on their availability.

Two dictionaries are currently included in the *histdict* system – the digitalized Old Bulgarian dictionary, already mentioned above, and a historical dictionary of the Bulgarian language. The latter traces the history of words and their meanings from the time of their first attestation in medieval manuscripts until the present day. The concept of the *histdict* system is to create a historical dictionary of the Bulgarian language through editing and supplementing the digitalized Old Bulgarian dictionary. To this end, specialized software has been developed, namely two separate programs for creating and editing dictionary entries. In order to accommodate our users, new words and words edited on the basis of their appearance in the Old Bulgarian dictionary have been marked in blue and green, respectively, in the historical dictionary.

Firstly, let me briefly describe the principles behind our electronic historical dictionary of the Bulgarian language. A historical dictionary should be regarded as a lexicographical manual which follows the changes in the meaning of words. The dictionary in question is not interested in all contextual variations, nor does it claim to be exhaustive in terms of manuscript attestations. The historical dictionary of the *histdict* system is based on four principles: 1) the history of words is reviewed in a wide chronological perspective; 2) meanings are retrieved from a language corpus unlimited from the thematic point of view; 3) it has an open glossary; 4) the meanings are ordered according to their occurrence, and according to the genetic connections among them.

The historical dictionary of the Bulgarian language is constructed according to a thematically-oriented principle. After the separate lexical fields have been processed, such a thematic approach to the lexicon of the Bulgarian language in a diachronic perspective allows to reach certain conclusions regarding the pathways of the intellectualization of the language during the Middle Ages as well as regarding the development of the literary vocabulary. At the present stage, the section on the Christian terminology has been developed, including approximately 800 new and revised dictionary entries. The selection of this particular lexical field was dictated by the fact that Old Bulgarian (Old Church Slavic) was the sacred and literary language of the Orthodox Slavs.

During the development of the program for creating and editing dictionary entries, we have encountered a number of problems. The first software we developed operates according to a form-based principle. It displays a sequence of boxes featuring a dropdown menu, or typing boxes. It is possible to complete and/or modify their content. The development of this program was the result of a prolonged and meticulous effort spanning many years. From the very beginning, we were aware of the fact that we needed a piece of software which would allow changing a single letter in the longest and most complex dictionary entry, and saving it without any other changes. This meant that forms had to be structured in such a manner that they could cover all entries from the digitalized Old Bulgarian dictionary, which we intended to edit and supplement in order to create our historical dictionary. This idea, which at first glance appeared simple and appropriate, proved to be challenging to implement, because the authors of the Old Bulgarian dictionary had allowed for certain ambiguities and inconsistencies in their entries. However, even in dictionaries created with the finest level of precision certain entries will most likely have an ambiguous structure, and this group is prone to be exceedingly difficult to digitalize using the model of the remaining entries. However, we still make use of this program anyway: our experience has shown that it is quite convenient for making small changes, such as correcting printing errors. The software is also highly useful for editing the information contained in the dropdown menus – in our case, this corresponds to grammatical information. For example, there was a group of words in the Old Bulgarian dictionary which were inadequately marked as participles in their ‘part of speech’ field (ОГЛАШЕНЪ, ПОВЕЛѢНЪ, ЧОУДНИНЪ, БОГОЗЪВАНЪ, БОГОНАОУЧЕНЪ, НЕВѢДОМЪ etc.). It became evident that, using the above-described form-based software, it was possible to change the grammatical category quickly

and easily, marking these words as adjectives or verbs, which they indeed are. As a quick aside: The changes we introduce are only saved in our historical dictionary. The digitalized Old Bulgarian dictionary does not use this software, and it remains a precise copy of the printed version of the Old Bulgarian dictionary.

This software for composing and editing dictionary entries proved to be far less convenient for implementing major changes (for example, for merging meanings, adding new ones, interchanging them, etc.), as well as for creating wholly new entries. Currently, we are working on a dictionary of the language of Patriarch Euthymius, which requires writing entirely new dictionary entries (which we will talk about in a minute). The form-based software makes it necessary to compose the text in Word files and subsequently to distribute the information across the relevant boxes by copying and pasting. This is overly labor-intensive, and therefore we created new software specifically designed for dictionary entries, which is of the convertor type. The text has to be composed in a Word file nevertheless, but in accordance with special conventions – there are requirements regarding the formatting of the headline, grammatical information, meanings, examples, etc. Afterwards, the authors copy and paste their entries into a special box in *histdict*. When converting, the software automatically arranges the words in alphabetical order. If we copy and paste an already existing headword into the dictionary of Patriarch Euthymius, the new content is automatically substituted in place of the pre-existing one. It is possible to copy a word from the electronic dictionary into a Word file, edit it there, and then replace it in the dictionary by copy-pasting. This new convertor can be used to input an unlimited number of words. Although this software has been created for the dictionary of Patriarch Euthymius, it is suitable for creating new dictionaries. *Histdict* now incorporates both the legacy software and the new software for dictionary entries; users can choose the one they consider easier and more comfortable to work with.

Now I'd like to explain why we chose to create a dictionary of the language of Patriarch Euthymius. Saint Euthymius of Tarnovo was Patriarch of Bulgaria between 1375 and 1393. Regarded as one of the most important figures of medieval Bulgaria, Euthymius was the last head of the Bulgarian Orthodox Church in the Second Bulgarian Empire. In the spring of 1393 the son of Ottoman Sultan Bayezid I, Suleyman Çelebi, laid siege to the Bulgarian capital Tarnovo with his sizable forces. With Tsar Ivan Shishman out of the city, Euthymius was the one entrusted with the defense of Tarnovo. After a three-month siege on 17 July 1393, the Ottomans captured the capital despite its ferocious resistance.

Arguably the best esteemed of all Bulgarian patriarchs, Euthymius was a supporter of hesychasm and an authoritative figure in the Eastern Orthodox world of the time. The term Hesychasm derives from the Greek Hesychia (ἡσυχία, Greek pronunciation: [isi'çia]) meaning "stillness, rest, quiet, silence". Hesychasm, in Eastern Christianity is a type of monastic life in which practitioners seek divine quietness through the contemplation of God in uninterrupted prayer.

Euthymius initiated a language reform in Old Bulgarian that was widely influential to the standardized literary languages of Serbia, Walachia, Moldova, and the Ukrainian principalities. He established orthographic rules and corrected inaccurately translated religious books by comparing them to the Greek originals. These corrected texts became models for the Orthodox churches using Church Slavonic language: Bulgaria, Serbia, Romania, and Ukraine.

There are 15 known works by Euthymius: Four of these are designated as vitae: of St John of Rila, of St Parasceve (Petka) of Tarnovo, of St Hilarion of Moglena and of St Philothea of Temnić; the rest occur as eulogies/encomia: of St John of Polybotum, of St Michael the Warrior of Bulgaria, of SS Constantine and Helena and of Great Martyr Kyriaki (Nedelia).

Patriarch Euthymius has been canonized and his memory is honored on 20 January.

The electronic grammatical dictionary of the Old Bulgarian language (from 9<sup>th</sup> to 15<sup>th</sup> century) is also part of the *histsdict* system. Without a grammatical dictionary, it would not have been possible to create an adequate electronic system for presenting the Bulgarian literary heritage. Considerations of a technical nature make it essential to include such an element. Search engines may only operate efficiently and provide reliable data when working on an annotated corpus, and the morphological annotation of parts of speech is necessary at the very least. An automated analysis tool (tagger) is indispensable for such morphological annotation, and in turn the tagger could not have been created without a grammatical dictionary. To compile such a dictionary, it was necessary to generate all possible word forms from our medieval manuscripts. Initially conceived as a mere ancillary tool for the future tagger, the grammatical dictionary of *histsdict* in fact took on an existence of its own. Currently, it presents complete paradigms of words, taking into account any phonetic and morphological changes in the endings; it also displays the shape of the relevant word forms according to the Russian and Serbian recensions of the Old Bulgarian (Old Church Slavic) language. The grammatical dictionary is incorporated in the historical dictionary. The information may be retrieved either by a special search, using the ‘word forms’ button, or by clicking on the respective word in the historical dictionary. It is natural that the grammatical dictionary needs to provide actually reliable grammatical information. Let’s take for example the verb **дѣрати** (meaning scratch, tear): According to grammars its imperfect forms should be of the type **дѣраахъ**. However our corpus only attests imperfect forms of the type **дѣрѣхъ**, and we could not be certain if there had ever existed any forms of the type **дѣраахъ**. A similar situation obtains for nouns that have only singular or plural forms (*an example in English would be milk for singular and jeans for plural only words*). A question arises how to proceed in such cases: is it justified, for the sake of creating a comprehensive resource, to mechanically generate word forms whose existence cannot be confirmed by sufficient evidence? Furthermore, if such forms are to be reconstructed, is it necessary to mark them in any special way? In the end we decided defective paradigms to be marked with a dash, while an asterisk is placed in front of reconstructed forms (which are themselves colored in red).

Thus, in order to create a reliable grammatical dictionary, it is necessary to verify (using the search engine) which forms are attested in the electronic diachronic corpus and which ones are not. But the reason for creating the grammatical dictionary is precisely because no reliable search engine could be created in the corpus. The salvation in this case was the ingenious solution of adopting the functionalities of browsers in order to allow searches in the electronic corpus and in the historical dictionary. Currently, the search engine of the *histsdict* system only displays the texts in which a given search string is attested and the frequency of the string in each of those texts. Subsequently, by clicking on the title of a given text, a search inside the text itself can be performed, which allows locating all of the occurrences. The search engine has a virtual keyboard, and it works flawlessly with the historical dictionary, which is a type of annotated corpus.

In order to create an electronic grammatical dictionary, it was necessary to specify all possible formal types (rules for generating forms) of the Old Bulgarian language. What we have in mind here is not a traditional grammatical description; instead, we operate on the principle of ‘cutting and pasting’. The common part of the word forms – interpreted literally, not in a strictly linguistic sense – is separated; e.g., the ‘basic part’ of the word forms of the aforementioned verb **дѣрати**: present tense **дѣрѣши**, **дѣра** (aorist), **дѣрѣше** (imperfect) is not the root **дѣр-/дѣр-**, but is **д-**. Next, the ‘endings’ (likewise interpreted superficially, and not grammatically) are pasted after the main part: for example, if **-рѣши** is pasted after **д-**, one gets the form for the 2nd person singular present

indicative. Similarly, -ѡра is pasted to obtain aorist and -ѡрѣше to obtain imperfect forms. To provide one more example: verbs such as ковати, върати or дърати, which belong to the III subtype of the I conjugation, could not be described in the grammatical dictionary using a single rule. The verb ковати: коवेशи (present), кова (aorist), ковааше (imperfect) could not be generated using the model of дърати, дѣреши, because the common part of the word forms of ковати is ков-, and therefore the following element could not be pasted as -ѣреши, etc., but only -ѣши, -а, -ааше etc. The verbs върати and дърати, (д-ѡръше — в-ѡраше; д-ѡра — в-ѡра) for their part, also do not follow the same rule for generating their forms. As regards върати, manuscripts provide evidence for imperfect forms of the type в-рахъ, в-раше, while for дърати only the type д-ѡрѣхъ, д-ѡрѣше is attested.

Altogether, a total of 163 formal types have been specified for nouns, 22 for adjectives, and 230 for verbs. Once created, a formal rule can be applied to an unlimited number of new words. This means that it is possible to automatically generate paradigms of new words later to be included in the historical dictionary. Another advantage of the electronic grammatical dictionary is the possibility to edit the rules at any time, in case yet different types of forms were to occur in a newly added manuscript.

A total of 16 cells of the system are allocated for nouns, 129 for adjectives and 33 for verbs. Each cell can be filled with several word forms, because both language change and spelling variations are considered. Furthermore, this number is not final; rather, it constantly increases with the inclusion of new texts in our corpus. In fact, the large number of patterns in the grammatical dictionary reflects the different formal types in declension and conjugation, the changes that have occurred in the history of the language, as well as the natural anomalies and exceptions in the inflection of Bulgarian.

In conclusion, I would like to state that we have always been guided by the aspiration to make the *histdict* system an open, rich and well-structured platform, which would guarantee its longevity. We are trying to make this electronic resource provide its users with as much information as possible. It is clear to us that *histdict* has a representative function, and it can potentially be utilized by a wide range of users.