Studia Ceranea 8, 2018, p. 111–119
DOI: 10.18778/2084-140X.08.06

ISSN: 2084-140X
e-ISSN: 2449-8378

C O P E
Member since 2018
JM13714

Gergana Ganeva (Sofia)

Electronic Diachronic Corpus and Dictionaries of Old Bulgarian

The *histdict* electronic resource (available at *histdict.uni-sofia.bg*)[1], 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[2].

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[3]. At present, the

[1] The aim of this paper is to present the *histdict* system. The paper was written with the support from the project BG05M2OP001-2-009-0005 "Modern Palaeoslavonic and Medieval Studies", financed under the Operational Program "Science and Education for Smart Growth", co-financed by the European Union through the European Structural and Investment Funds.

[2] The development of the *histdict* system has been financed by several consecutive grants of the Fund for Scientific Studies at the Ministry of Education and Science, the Human Resources Development Operational Program, as well as the Center for Excellence in the Humanities "Alma Mater".

[3] А. Тотоманова, *Диахронный корпус болгарского языка. Состояние и перспективы* (in press: "Filologia", Zagreb).

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[4]. 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[5]. 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[6]. To this end, specialized software has been developed, namely two separate programs for creating and editing dictionary entries.

[4] *Старобългарски речник*, vol. I–II, София 1999–2009.

[5] A. Totomanova, *Digital Presentation of Bulgarian Lexical Heritage. Towards an Electronic Historical Dictionary*, SCer, 2, 2012, p. 221–234.

[6] 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 us 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, interpreted as changes of the semantic content. 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[7].

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 (fig. 1). 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

[7] А. Тотоманова, *Проектът "Информатика, граматика, лексикография" и дигиталната обработка на средновековни славянски текстове*, *Информатика, граматика, лексикография BG051-3.3-06-0024/2012*, [in:] *Информатика, граматика, лексикография BG051-3.3-06-0024/2012. Сборник доклади и материали от заключителната конференция, София, 29–30.06.2015 г.*, ed. А. Тотоманова, Т. Славова, София 2015, p. 5–16.

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, чоудимъ, богоꙁъванъ, 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[8].

This software for composing and editing dictionary entries proved to be far less convenient for implementing major changes, however (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. 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 (fig. 2). 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. In principle, this new convertor could be used to input an unlimited number of words, but in practice it starts slowing down when more than 100 entries are submitted. 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.

The electronic grammatical dictionary of the Old Bulgarian language (9th–15th century) is also part of the *histdict* system (fig. 3). Without a grammatical dictionary, it would not have been possible to create an adequate electronic system for

[8] 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.

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 *histdict* 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. For example, the verb дьрати only attests imperfect forms of the type дерѣхъ, and we could not be certain if there had ever existed any forms of the type дьраахъ, as traditionally stated in the grammars. A similar situation obtains for deficient noun paradigms (for example, *singularia* or *pluralia tantum*), aorist types in the verbs of the I and II conjugation, etc. 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?[9]

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 *histdict* 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.

[9] In our grammatical dictionary, defective paradigms are marked with a dash, while an asterisk is placed in front of reconstructed forms (which are themselves colored in red).

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 (fig. 4). 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 verb **дьрати** is not the root **дер-/дьр-**, but is **д-**. Next, the 'endings' (likewise interpreted superficially, and not grammatically) are pasted after the main part: for example, if **-ереши** is pasted after **д-**, one gets the form for the 2nd person singular present indicative. Similarly, **-еретъ**, **-ереть** and **-ерет** are the possible elements that can be pasted to obtain the 3rd person singular present indicative. To provide one more example: verbs such as **ковати**, **ꙁъвати** or **дьрати**, 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**, **ковеши** could not be generated using the model of **дьрати**, **дереши**, because the common part of the word forms of **ковати** is **ков-**, and therefore the following element could not be pasted as **-ереши**, **-еретъ** etc., but only **-еши**, **-етъ** etc. The verbs **ꙁъвати** and **дьрати**, for their part, also do not follow the same rule for generating their forms[10]. As regards **ꙁъвати**, manuscripts provide evidence for imperfect forms of the type **ꙁъваахъ**, **ꙁъвааше** and **ꙁовѣхъ**, **ꙁовѣше**, while for **дьрати** only the type **дерѣхъ**, **дерѣше** is attested.

Altogether, a total of 163 formal types have been specified for nouns, 22 for adjectives, and 230 for verbs[11]. 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, a fusional language.

In conclusion, we 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

[10] Compare the forms **ꙁ-ъвати**, **ꙁ-овеши** and **д-ьрати**, **д-ереши**.

[11] А. Тотоманова, Т. Славова, Г. Ганева, *Морфосинтактичен тагсет на старобългарския книжовен език*, [in:] *Информатика*…, p. 17–117.

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.

Bibliography

Primary Sources

Starobălgarski rečnik, vol. I–II, Sofija 1999–2009.

Secondary Literature

Totomanova A., *Diachronnyj korpus bolgarskogo jazyka. Sostojanie i perspektivy* (in press: "Filologia", Zagreb).

Totomanova A., *Digital Presentation of Bulgarian Lexical Heritage. Towards an Electronic Historical Dictionary*, "Studia Ceranea" 2, 2012, p. 221–234.

Totomanova A., *Proektăt "Informatika, gramatika, leksikografija" i digitalnata obrabotka na srednovekovni slavjanski tekstove, Informatika, gramatika, leksikografija BG051-3.3-06-0024/2012*, [in:] *Informatika, gramatika, leksikografija BG051-3.3-06-0024/2012. Sbornik dokladi i materiali ot zaključitelnata konferencija, Sofija, 29–30.06.2015 g.*, ed. A. Totomanova, T. Slavova, Sofija 2015, p. 5–16.

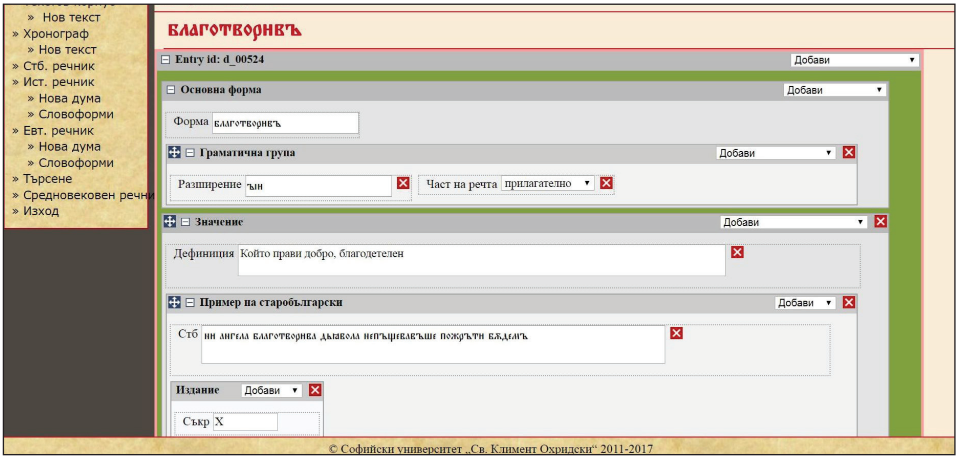
Totomanova A., Slavova T., Ganeva G., *Morfosintaktičen tagset na starobălgarskija knižoven ezik*, [in:] *Informatika, gramatika, leksikografija BG051-3.3-06-0024/2012. Sbornik dokladi i materiali ot zaključitelnata konferencija, Sofija, 29–30.06.2015 g.*, ed. A. Totomanova, T. Slavova, Sofija 2015, p. 17–117.

Abstract. The electronic system *histdict* is designed as a tool for research, adequate presentation and popularization of a part of Bulgaria's cultural and historical heritage: the Bulgarian language and its medieval literature. The article describes the various steps in the development of *histdict*. Attention is paid to each component of the resource: specialized Unicode fonts, electronic diachronic corpus, dictionary of Old Bulgarian, historical dictionary equipped with tools for writing and editing dictionary entries, grammatical dictionary, prototypical search engine, and virtual keyboard. The article also lays out the principles followed in the development of the diachronic grammatical dictionary of the Bulgarian language.

Keywords: *histdict*, historical dictionary, grammatical dictionary, electronic diachronic corpus.

Gergana Ganeva
St. Clement of Ohrid University of Sofia
15 Tsar Osvoboditel blvd. 1000 Sofia, Bulgaria
geriganeva@yahoo.com

ILLUSTRATIONS



благотворивъ

⊟ Entry id: d_00524

⊟ Основна форма

Форма благотворивъ

⊟ Граматична група

Разширение ꙑи ☒ Част на речта прилагателно ▾ ☒

⊟ Значение

Дефиниция Който прави добро, благодетелен ☒

⊟ Пример на старобългарски

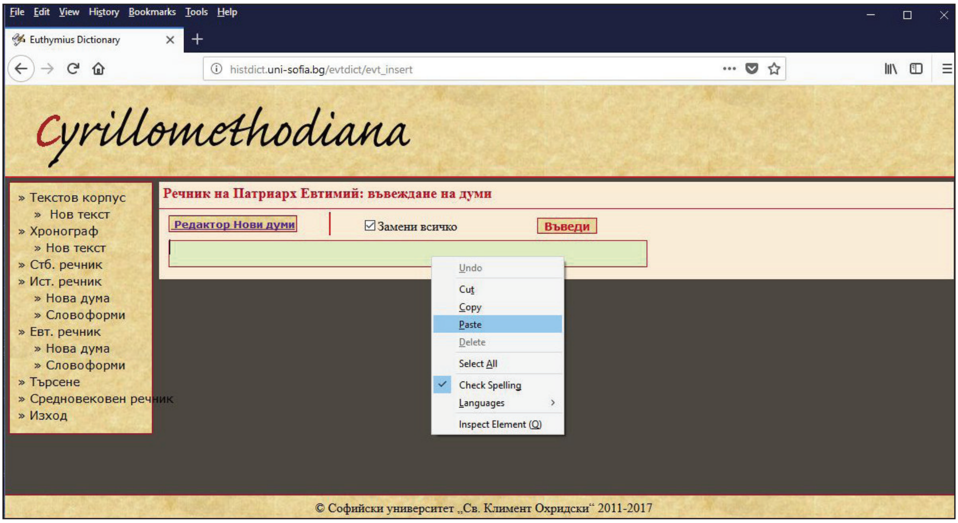
Стб ни ангєла благотворива дьꙗвола нєпъщєвавъшє пожрътн бѫдємъ ☒

Издание Добави ▾ ☒

Сткр X

© Софийски университет „Св. Климент Охридски“ 2011-2017

Fig. 1. Form-based software for dictionary entries

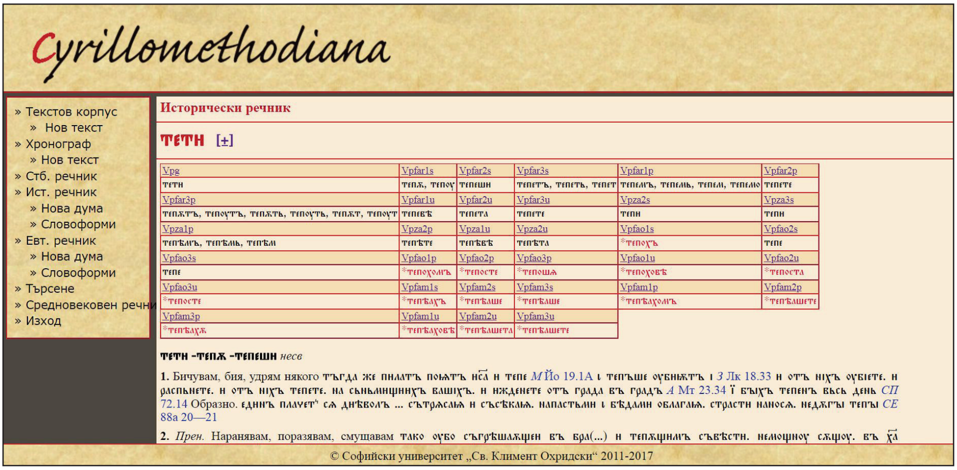


Речник на Патриарх Евтимий: въвеждане на думи

Редактор Нови думи ☑ Замени всичко Въведи

© Софийски университет „Св. Климент Охридски“ 2011-2017

Fig. 2. Convertor-type software for dictionary entries



Исторически речник

тетн [±]

Vpg	Vpfar1s	Vpfar2s	Vpfar3s	Vpfar1p	Vpfar2p
тетн	тепѫ, тепоу	тепешн	тепетъ, тепеть, тепет	тепемъ, тепемь, тепем, тепемо	тепете
Vpfar3p	Vpfar1u	Vpfar2u	Vpfar3u	Vpza2s	Vpza3s
тепѫтъ, тепоутъ, тепѫть, тепоуть, тепѫт, тепоут	тепевѣ	тепета	тепете	тепн	тепн
Vpza1p	Vpza2p	Vpza1u	Vpza2u	Vpfao1s	Vpfao2s
тепѣмъ, тепѣмь, тепѣм	тепѣте	тепѣвѣ	тепѣта	*тепохъ	тепе
Vpfao3s	Vpfao1p	Vpfao2p	Vpfao3p	Vpfao1u	Vpfao2u
тепе	*тепохомъ	*тепосте	*тепошѧ	*тепоховѣ	*тепоста
Vpfao3u	Vpfam1s	Vpfam2s	Vpfam3s	Vpfam1p	Vpfam2p
*тепосте	*тепѣахъ	*тепѣаше	*тепѣаше	*тепѣахомъ	*тепѣашете
Vpfam3p	Vpfam1u	Vpfam2u	Vpfam3u		
*тепѣахѫ	*тепѣаховѣ	*тепѣашета	*тепѣашете		

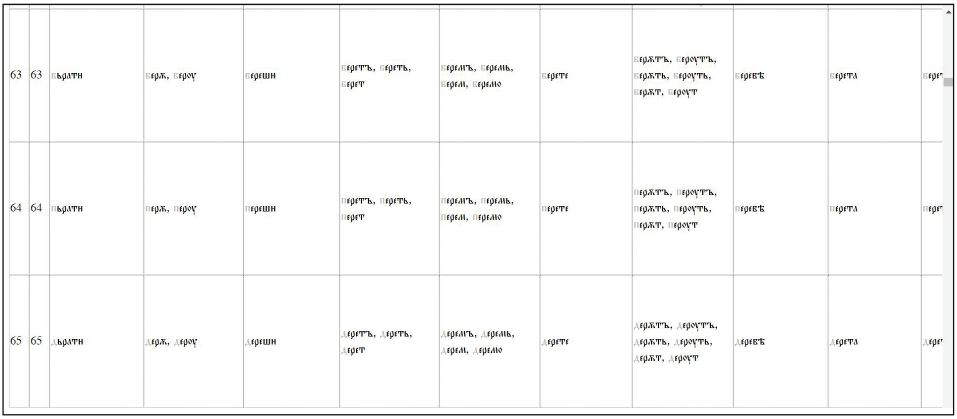
тетн -тепѫ -тепешн *несв*

1. Бичувам, бия, удрям някого тъгда же пнлатъ поѩтъ нс҃а н тепе *М* Йо 19.1А ι тепъше оуbнѩтъ ι *З* Лк 18.33 н отъ нıхъ оуbıете. н распьнете. н отъ нıхъ тепете. на сьньмнщнхъ вашıхъ. н нжденете отъ града въ градъ *А* Мт 23.34 ї бъıхъ тепенъ вьсь день *СП* 72.14 Образно. еднnъ плачет҇ сѧ днѣволъ ... сътрѧсаѩ н съсѣкаѩ. напастьмн ι бѣдамн облагаѩ. страстн наносѧ. недѫгъı тепъı *СЕ* 88а 20—21

2. *Прен.* Наранявам, поразявам, смущавам тако оубо съгрѣшаѫщен въ бра(...) н тепѫщнмъ съвѣстн. немощноу сѫщоу. въ х҃а

© Софийски университет „Св. Климент Охридски“ 2011-2017

Fig. 3. Electronic grammatical dictionary of the Old Bulgarian language



63	63	бьратн	берѫ, бероу	берешн	беретъ, береть, берет	beремъ, беремь, берем, беремо	берете	берѫтъ, бероутъ, берѫть, бероуть, берѫт, бероут	беревѣ	берета	берет
64	64	пьратн	перѫ, пероу	перешн	перетъ, переть, перет	перемъ, перемь, перем, перемо	перете	перѫтъ, пероутъ, перѫть, пероуть, перѫт, пероут	перевѣ	перета	перет
65	65	дьратн	дерѫ, дероу	дерешн	деретъ, дереть, дерет	деремъ, деремь, дерем, деремо	дерете	дерѫтъ, дероутъ, дерѫть, дероуть, дерѫт, дероут	деревѣ	дерета	дерет

Fig. 4. Rules for generating forms