GRAMMATICAL DICTIONARY

A. Totomanova, Sofia University

Launching the Histdict system 10 years ago, we were dreaming of producing an automatic morphological annotator (tagger) in order to facilitate the work on compiling the Historical dictionary of Bulgarian language, using the lexical material of the Diachronic corpus. At that time, we had not yet realized that a grammatical dictionary was also a prerequisite for an efficient electronic search engine, for reliable search engines work only on annotated corpora. That awareness came to us only during our second project, when we found out that the search engine we were trying to produce without considering the grammatical characteristics of the wordforms, might be mere a temporary solution, since it does not meet the needs of the researchers in the area of diachronic linguistic. And that was the moment we decided to dedicate ourselves to producing a grammatical dictionary of Old Bulgarian language. In principle such a dictionary is supposed to take into account all possible representations of a single form or to put it simply, it should describe all possible forms of inflectable words, taking into account the nomenclature of the uninflectable words as well. Given the fact that Old Bulgarian is a language with complicated inflectional morphology, especially in the nominal paradigm (six cases, three genders, three grammatical numbers, plus vocative, simple and compound forms of the adjectives and declinable participles), which further was reduced to 2 to 5 forms, depending on grammatical gender, the production of the grammatical dictionary represented one of the biggest challenges we faced.

Morphological annotation according to the parts of speech is not possible without defining their morpho-syntactic features and producing a set of morphological tags, that reflect the grammatical characteristics of the Old Bulgarian literary language. Being a mnemonic description of a certain part-of-speech and its grammatical features, the tag contains a string of characters (small, capitals) and/or numbers and each position corresponds to a determined grammatical feature.

The Old Bulgarian tagset was based on the already existing tagset of the Modern Bulgarian, that was defined and used on the syntactic corpus BulTreeBank (on 215 149 words of the Modern Bulgarian). A detailed description of the tags could be found in Simov, Osenova, Slavcheva. BTB-TR03: BulTreeBank Morphosyntactic Tagset, 2004 (http://www.bultreebank.org/TechRep/BTB-TR03.pdf).

The tag structure looks as follows: the first letter is always capital and corresponds to the POS. Each following letter encodes its specific grammatical features, that are listed in a fixed order. Each tag corresponds to a paradigm cell of a determined class of lexical items. In general the signs for the POS and their grammatical categories follow the abbreviations adopted in the BulTreeBank corpus, emended (or replaced) with new ones according to the morpho-syntactic specifics of Old Bulgarian, which unlike the Modern Bulgarian expresses the syntactic relations not analytically, but synthetically through the case endings of nominal forms.

The tagset should contain sufficient morpho-syntactic information for a coherent description of a certain language. Detailed and exhaustive morphological description of the

lexical items is a warranty for its full and wide practical application and for a higher degree of form recognition.

The size of the tagset depends on the richness of the morpho-syntactic system in a given language, f. e. in Spanish the number of tags is about 475, in the modern Bulgarian - between 600 and 700, while in Old Bulgarian - 2 200 (only the tags for the nouns, adjectives, pronouns and pronominal adjectives are 1150).

The tagset, which determines what forms and grammatical features will be automatically recognizable, is an artificial construct and it should be clear that it does not fully reflect our understanding of the Old Bulgarian language system, the hierarchy of categories and oppositions. In many cases, decisions about tagset structure are made to the purpose and depending on the software capability and the needs of future users of the system.

A specific issue that had to be solved in the design of the tagset was the large number of homonymous (syncretic) word forms within the Old Bulgarian paradigm, which explains why the number of tags is different from the number of word forms. Sometimes 2, 3, 5, 6 and even 7 tags can correspond to a single word form in Old Bulgarian (**гради** = gen sg = nom-acc dual), because one case ending might express different syntactic relations. In addition the graphic homonyms might belong to different POS and their grammatical categories, for instance the form **водимъ** is a form for the 1p pl of present and imperative as well as a simple form of the present passive participle.

The Old Bulgarian tagset contains tags only for the homonymous forms that are result of the Protoslavic phonetic tendency for increasing sonority that afflicted the word endings and cuased coincidences between initially different forms, for instance nom. *gord-o-s > gord -ŭ-s > gord -ŭ > **градъ** = acc. *gord-o-n > gord -ŭ-n > gord -ŭ > **градъ**, 2 p. sg. aorist *rek-e-s > **рѣ** = 3 p. ед.ч. aorist *rek-e-t > **рѣ**, nom sg masc of the present active participle *nes-ont-s > *nesons > nesuns> nesūs > **несъı** = nom sg neutr of the same participle *nes-ont-ø > *nesonø > nesūn > **несъı**) etc.

The secondary homonymy, that occurs in the process of phonetic and paradigmatic levelling during the historical development of Bulgarian is not reflected by the tagset. Yet it is taken into account in the grammatical dictionary as an important prerequisite for achieving a higher degree of form recognition. Each wordform (including the homonymous ones) is given a special tag. The only exception is made for the paradigm of the nouns, pronouns and participles in dual, in which nom and acc, gen and loc and dat and instrumental always coincide. (these cases are marked accordings with O (nominative/accusative case), род. п. = мест. п. → G (genitive/locative case), дат. п. = тв. п. → D (dative/instrumental case). This syncretism is inherited from the protolanguage and is typical of all Indo-European languages.

The protoslavic homonymy of nom and acc masculine and neutral did not cause any problems when the nominal forms were referring to objects but created some difficulties in differentiating the subject and the object when referring to persons and animated objects. That is why in the paradigm of the animated nouns the accusative started differentiating from the Nominative through receiving the ending of the Genitive. Thus in ProtoSlavic and in Old Bulgarian in sg a masculine sub gender occurred that was typical of the nominal forms

designating or relating to male creatures. In the tagset this subgender is marked with E , though the wordform is not compulsory for the names of male persons and animals. The nouns of er- and u long-stems also possess a genitive-accusative case.

The finite verbforms are described with 200 tags, each of them corresponding to a single microwordform i.e. a glossemic word (for instance *will be sitting* includes three microwordfroms). In other words, the automatic annotator identifies as words (tokens) the strings of symbols between two spaces. The latter means that the tagger recognizes only synthetics but not composed verbal forms like ¬смь бьралъ, распѧтъ бѫдетъ.

It was foreseen that the automatic morphological tagger should make the difference between full and auxiliary verbs. It was also decided the tagger to provide information about the verbal aspect – perfective or imperfective as well as about whether the verb is transitive or intransitive. The opposition transitivity/intransitivity does not form a grammatical category, dividing the verbal lexis in lexico-grammatical classes. In this case we were driven by the purpose the tagger is meant for: it is a software instrument that should help both specialists and non-specialists, students and all those who are interested in the Bulgarian literary heritage. Keeping that in mind we reached the conclusion that the tagger should provide as much information as possible.

The Old Bulgarian numerals do not form a coherent morphological category: the cardinal numbers are either prononus (ѥдинъ, дъва, ѳба), or adjectives тріѭ/три, чєтъірє/чєтъіри, or nouns of different grammatical gender пѧть, шєсть сєдмь, ѳсмь, дєвѧть, тъісѧштн/тъісѫштн и тьма are feminine, дєсѧть is both masculine and femimime, while сътѳ is neutral, whereas the ordinal numbers that derive from them are compounded adjectives. For this reason, this POS is not described by special tags.

First of all we had to identify the grammatical features for each part of speech, to assign to them the respective codes and to define the order they will appear in the tag: It turned out that

nouns have 49 tags.

adjectives 180 tags

pronouns 620 tags

pronominal adjectives and adverbs 300 tags

finite verbal forms 200

The non-finite verbal forms – the participles, the infinitive and the supine are described with 875 tags in total. The tags for the particles, because of their hybrid grammatical nature turned out to be the most complicated and the grammatical features are displayed in 12 positions.

Working on the tagset we made another important decision and agreed on including in the grammatical dictionary all possible orthographic variations of a certain word form that occurred either due to the phonetic or morphological changes in Bulgarian, or derived from different orthographic versions, in which Old Bulgarian texts are preserved – Middle Bulgarian, Russian or Serbian.

The Old Bulgarian tagset is published on Cyrillomethodiana portal and could be downloaded as a pdf document from https://cyrillomethodiana.uni-sofia.bg/mdocs/category/8-archive.

The tagset helped us to allocate 16 cells of the system to nouns, 129 to adjectives and 33 to verbs. Each cell should be filled with a number of the wordforms, given the linguistic changes and different orthographic versions, in which the writings of Old Bulgarian men of letters came to us. Our main objective was to identify all possible formal types (rules for generating forms) of the Old Bulgarian language. Very soon we realized that the number of patterns is not constant and gradually increases with the inclusion of the new texts in the diachronic corpus. The rules reflect rather a formal word analysis than a traditional morphemic description, since the computer is not able to identify the morphemes but only the common parts of the wordforms. Our ICT specialist and Gergana Ganeva call this *modus operandi* cutting/pasting principle. Thus for each formal type we identified the basic immutable parts (in many cases consisting of a single letter) and changeable parts that could be pasted after them in order to generate the forms. In case of the verbs, the existence of the generated forms was even verified in the corpus.

So far 163 formal types for nouns, 22 for adjectives, and 230 for verbs have been included in the grammatical dictionary. As a result each inflectable word in the historical dictionary was assigned a specific rule according to which its forms are being automatically generated and recognized. Unfortunately, for the time being we have not solved the problem with the participles because of the complexity of their hybrid nature, that combines verbal and nominal features. In its present form the Grammatical Dictionary is published on Cyrillomethodiana portal and could be downloaded from https://cyrillomethodiana.uni-sofia.bg/mdocs/category/8-archive.

Its electronic version is installed in the Historical dictionary and is accessible in two ways: by clicking the sign plus located next to inflectable words or choosing the option Словоформи/Wordforms from the menu. In the first case the computer displays the whole paradigm of the word, while in the second – only the grammatical characteristics of the introduced wordform. Given the widely spread morphological homonymy the user is supposed to decide which characteristics fit into the morpho-syntactic context. In fact, what we have now is a semi-automatic tagger that recognizes the forms of the nouns, adjectives and conjugated verbs. A virtual keyboard is supporting the introduction of the wordforms into the semi-automatic annotator. But as you have already seen it is used also for any kind of searching in the violable electronic resources. The main flaw of the electronic grammatical dictionary is not the lack of rules for participles and pronouns - sooner or later they will be created and verified, but the fact that it could not be edited online. The latter means that the new entries in the historical dictionary should be extracted manually and assigned the grammatical rules. That is why the new entries coming from the dictionary of patriarch Euthymius, though inflectable, do not have sign plus next to their headwords. We hope to solve this problem very son and, in the meantime, do keep dreaming of a fully automatic morphological annotator.