

Pre-Processing Terms in Bulgarian from Various Social Sciences and Humanities (SSH) Domains: Status and Challenges

Petya Osenova*, Kiril Simov*, Yura Konstantinova[†]

*Institute of Information and Communication Technologies, Bulgarian Academy of Sciences
Acad. G. Bonchev bl. 2, 1113 Sofia
{petya, kivs}@bultreebank.org

[†]Institute of Balkan Studies and Centre of Tracology, Bulgarian Academy of Sciences
Moskovska St 45, 1000 Sofia
yura.konstantinova@balkanstudies.bg

Abstract

1. Introduction

There exists a great number of focused initiatives, projects and conferences that tackle deeply various topics related to terminology construction, understanding, processing and usage. We will mention only a small part of them here rather as initiatives than as distinct publications. These are, among others, ENeL COST Action on e-Lexicography,¹ related activities in the NexusLinguarum COST Action,² related activities in the ELEXIS project,³ globaLEX organization. There is also ongoing work on providing language technology help to colleagues in SSH within CLARIN-ERIC and DARIAH.^{4,5}

Within the CLaDA-BG infrastructure,⁶ which combines the goals of CLARIN and DARIAH in Bulgaria, there are two types of partners – technological ones and colleagues also from SSH. The latter are historians, ethnographers, specialists in the deeds and lives of Cyril and Methodius, museum and library workers. This combination of complementary partners allows us to construct the necessary resources and immediately to verify their utility for SSH partners.

In the task of creating the Bulgarian-centric Knowledge Graph (BGKG) within CLaDA-BG (Simov and Osenova, 2020) we requested data from our SSH partners in order to perform linguistic pre-processing and to enhance the creation of terminological dictionaries that cover the SSH subdomains based on these data.

The size of the corpus is nearly half a million – 484,815 tokens. The selected words and phrases for pre-processing and creation of entries towards terminological dictionaries were about 5,000 within nearly 26,000 usages annotated within the corpus. From them the rejected candidates, or the false positives, were 542 candidate phrases. Out of them 328 candidates were completely rejected either because they were named entities or free compositional phrases.

Thus, our colleagues from SSH would facilitate their own work with only checking and validating the previously pre-processed data. The data consists of selected texts from various sources such as: scientific texts authored by our SSH colleagues and related to Bulgarian history and society; Linked Open Data like Wikipedia; available textbooks, specialised dictionaries etc.

Here we give a brief outline of our pre-processing strategy towards handling the data-driven terminology in these domains.

2. The Task Overview

The work flow that is discussed here is related to the SSH data (publications, autobiographies, archive documents, newspaper articles from past periods, descriptions of artefacts, etc.) that were collected from partners, and annotated within the INCEption platform⁷ with named entities, events and roles. Thus, while annotating linguistically the texts, the annotators were additionally asked to mark candidate terms with the label *term*. This task was set in the view of the subsequent creation of specialised terminological dictionaries

¹ <https://www.cost.eu/actions/IS1305/>

² <https://nexuslinguarum.eu/>

³ <https://elex.is/>

⁴ <https://www.dariah.eu/>

⁵ <https://www.clarin.eu/>

⁶ <https://clada-bg.eu/en/>

⁷ <https://inception-project.github.io/>

in each participating SSH domain – history, ethnography, biographical studies, etc. The annotators were instructed to view as candidate terms the keywords that are specific for the domain.

Later on, these candidate terms were extracted and transferred to a huge excel table in Google Drive. The table consists of three main areas: a) the candidate term, b) the term in its context of occurrence, and c) the source that delimits the domain of usage. In Figure 1 an excerpt from the excel view is presented:

	Зимно време конете, заедно с другите хайвѐни, държат в айр или в дама за работния добитък /инф. б/. На @@@ ездитния кон @@@ на гърба слагали само едно черджѐ , седлѐ не е имало . Конския юлар е бил от въжета, но в града ги правели – мешинови /инф. б/; „моят баща е	
ЕЗДИТНИЯ КОН	запомнил в селото 60 коня“ /инф. б/.	Etnographs-text01-04
ЕЗДИТЕН КОН	кон, който се използва за езда	етно

Figure 1: An example from the excel table.

In the first row the following information is given: the term as it occurred in the text (riding-the horse, ‘the riding horse’), the text excerpt with the term placed among the symbols @@@, and the name of the source text. In the second row the following information is given: the normalised term (*riding horse*), the definition (*a horse that is used for riding*) and the domain – ethnography.

All the one-word terms got initial definitions from the digitised version of the Explanatory dictionary of Bulgarian (Popov et al., 1994). This step was performed automatically through a rule-base matching method. First, the word forms in the texts were lemmatized with our in-house Inflectional Bulgarian dictionary. Then the coinciding lemmas in the dictionary and in the texts were matched. The terms with more than one meaning also received all the possible definitions automatically. Afterwards, these candidate terms were processed manually by the team that previously worked on the event and roles annotations. The core team engaged with the terminology pre-processing consisted of 4 members as a subpart of the whole annotating team that consisted of 8 people.

The tasks related to the terminology processing were organized as follows: one person (outside the 4 working colleagues) performed the automatic construction of the table and the assignment of the existing definitions and sources. Initially the candidate terms were assigned in an alphabetical order to workers, i.e. each colleague was responsible for the candidate terms that began with certain letters. However, after having completed some letters, a decision was taken to go by domain source instead. This approach allowed us to observe the terms in their domain contexts and interrelations. Then, once more the terms were checked in their alphabetical appearance.

The workflow was generally divided into two phases that respects the competences of the experts. In *Phase 1* the corpus linguists (who were also annotators) pre-processed the candidate terms while in *Phase 2* the specialists in SSH areas are supposed to check and validate these terms against their own area.

3. The Workflow

The respective annotated data was uploaded in advance including the annotated candidate term. The workflow consisted of the following steps:

3.1 Deciding which candidate terms are true terms

Here the main task of the corpus linguists was to try to reduce the list of the obvious non-terms or the common words and expressions from the specialised terms. Sometimes the boundaries were not very clear, especially with respect to the multiword expressions (MWE) and the nested terms. See more about this issue in point 3 below. The annotators had at their disposal three options to select from: *a sure term*, *a maybe term* and *a non-term*.

3.2 Checking the availability of the definition and its relevance

In case there was a definition, the annotator had to: accept it as it is, reject it or modify it. If there was no definition, the annotator had to create one. When the term was one-word, the task was to check the definition that came from the Explanatory dictionary of Bulgarian. In case of lexical ambiguity the annotator had to select the correct definition among the available ones, or again - to provide their own, if no appropriate is present. Then the selected definition was marked as the right one. Please note that the other definitions were not deleted for the sake of completeness and future addition into BTB-Wordnet.

3.3 Handling multiword expressions

Here the prevailing part of terms consisted of a head noun and a pre-positioned modifier. For example, *демократија* (democracy) and *пряка демократија* (direct democracy). The problems might go into two directions: to accept a MWE as a domain term or not, and to provide a definition of it since it is usually not available in the consulted sources. We decided to be inclusive in accepting what had to be a term. This means that all the expressions that were considered specific for the domain, were approved. The annotator could also add the definitions about the parts of compositional MWEs. For example, *невалиден глас* (invalid vote) can have a definition as a phrase, while its two elements *невалиден* (invalid) and *глас* (vote) might also be added below with their own definitions.

3.4 Re-checking the domain/genre.

This step relies on the domain/genre classification that has already been used. Thus, an initial pre-defined schema was explored that in the process of work was further expanded and hierarchized. At the moment the list of the applied domains amounts to 76 categories (for example, architecture with a subdomain of construction; geography with a subdomain of geology; philosophy with subdomains of ethics, rhetorics and logic) and the registers to 15 (for example, dialectal, metaphorical, colloquial, etc.). The initial schema came from the classifications used in the Explanatory dictionary and had 36 domains and 4 registers. At the beginning, we tried to keep the terms in separate groups that do not overlap: history, ethnography, etc. These areas however are highly interdisciplinary and they inter-cross with each other. For that reason this approach was abandoned at a very early stage in our work. In this way one and the same term could be put in more than one domain with the same or different meaning.

Other tasks that were part of the workflow – although with a lower priority were:

3.5 Adding other senses of the lemma of the term, and

3.6 Adding examples to these additional senses.

The idea behind tasks 5 and 6 was that we aim at reaching better coverage also in other language resources like BTB-Wordnet (Osenova and Simov, 2018) and at compiling a sense corpus per lemma and usage.

The results of this preparatory work was a classification of the initially selected about 5000 candidate terms and keywords with respect to the hierarchy of domains. This allows the further processing to be done by different experts in the corresponding domains. Their tasks are the following:

Final Sorting of lexical items within true terms and keywords.

As it was mentioned above, the examples annotated within the domain documents were classified within two main categories: general lexica and compositional phrases, on the one hand, and terms, on the other. The second group sometimes contains keywords that happen to be true terms in the respective domain. Thus, the first task for the experts was to sort out the true terms.

Addition of missing terms.

Despite the wide range of documents selected for annotation, they do not contain all the relevant terms in the domain. For example, from the set of all genres of Old Bulgarian literature, only three were identified within the annotated documents. The rest of terms for these identified genres were added by the experts of Old Bulgarian literature. Thus, by completing the missing slots, we expect that each domain will have a relatively complete list of terms.

Extension of the definitions.

In the original list of candidate terms we had to also add definitions from online sources or to construct our own definitions. Since the pre-processing group included linguists, but not experts in the domains, very often the definitions were not complete and/or precise enough. Thus, the domain experts extended the definitions with encyclopaedic information. In some cases also appropriate images were added. The resulting encyclopaedic entries were cross linked on the basis of the included terms. Here is an example of such an entry from the area of architecture:


<p>АЖУР техника при <i>резбарското, златарското, плетаческото</i> и други изкуства, при която между декоративните елементи има отвори</p>	
--	--

Figure 2: One example from the terminology lexicon in the area of architecture. On the left, the term *openwork* and the definition (ornamental work such as embroidery or latticework having a pattern of openings) are given, and on the right there is an image illustrating it. The links to other terms are represented via italicising the corresponding words/phrases in the definition. This example is only illustrative. The actual entries might contain longer texts, references to relevant literature, more images and links to external resources.

The resulting terminological lexicons are further processed by the team working on the Bulgarian Bultreebank Wordnet (BTB-WN). This work has been done in cooperation with the domain experts. Such alignment of the terminological lexicons and wordnet allows for a joint usage of both lexical resources for the main use cases – explanation of the specific knowledge in the domains and indexing of various types of domain documents. Figure 3 depicts a part of the hierarchy of Bulgarian folk units of measurement. They are linked with a hyponymy relation to the concept for *Bulgarian folk units* and the concept for *linear units*.

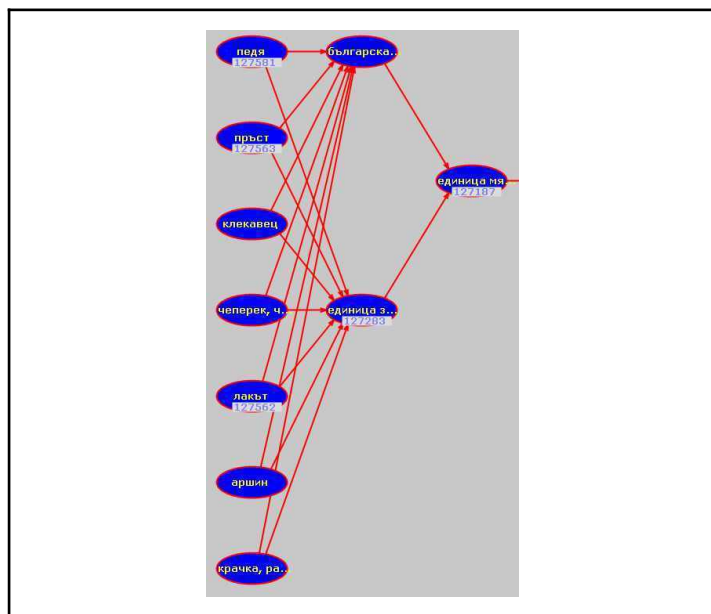


Figure 3: In this figure a graphical view on Bulgarian folk units of measurement is presented. Each term is classified into two ways - as a unit of measurement for distance (*linear units*) and that its domain is Bulgarian folk units. The hierarchy of terms could interleave with synsets that are not terms in the domain. The mapping to synsets in the English WordNet are given with identification (IDs) at the lower part of the graphical representation of each Bulgarian synset. Here measures are given such as *пе́дя* (span), *пръст* (finger), *лакът* (elbow), etc.

Our idea is BTB-WN to be the main resource within CLaDA-BG for representation of lexical data related to general language, terminology and to be aligned to the ontologies on which BGKG is constructed.⁸ In this way we hope to be able to provide access to these data by different types of users with different knowledge about the domains, with different goals in mind, etc.

In addition to the standard wordnet relations (hypernymy, meronymy, etc.), we envisage other semantic relations that represent various aspects of knowledge within the corresponding domains. In this way, we will ensure the representation of encyclopaedic information and will facilitate the representation of Named Entities (NEs) classified with respect to the corresponding concepts. This approach relies on specially created templates based on the domain relations as well as their domain and range restrictions. We already defined about 20 such templates for main classes of NEs like geopolitical entities, historical events (wars, uprisings, etc.), artefacts (icons, stamps, ect), political parties and regimes, and so on.

4. Conclusions

In this extended abstract we described the main steps that were followed in the creation of terminological lexicons in a bottom-up approach starting from real texts within SSH domains. After the domain texts were annotated with named entities, events, roles and candidate terms, a concordance of the latter from different documents was performed where they were grouped together and linguistically processed. As a result, they had the representation of the term in its basic form, listings of related words for MWEs, the existing potential

⁸ This approach is similar to the lexeme assignment in Wikidata.
https://www.wikidata.org/wiki/Wikidata:Lexicographical_data/Documentation

senses from different sources (available locally to the annotators and on the web). The appropriate senses for the given context were selected or created. Then the result was further processed by the domain experts in order to make the definitions more precise and complete. Also, an addition of missing terms was performed. Then the terminological lexicons were aligned to the BTB-WN in order to be used for navigation, annotation of more documents (manually or automatically) and to establish links to the necessary ontologies.

The main challenges can be divided as either technical or theoretical. In the first group we can mention the insufficient context, lack of enough sources for terms related to previous historical times; approaching the task in the best way - alphabetically or by source, etc. In the second group we can outline: the difficulty to differentiate between a term and a non-term; aiming at the most informative definition when there are too many found in the sources; finding and/or constructing a definition when it lacks or is wrong with the help of other available resources; handling close definitions for some lemma; construction of a definition for multiword terms; handling multi-domain inclusion of terms.

5. References

- Petya Osenova and Kiril Simov. 2018. The data-driven Bulgarian WordNet: BTBWN. In: *Cognitive Studies | Études cognitives*, vol. 18, <https://doi.org/10.11649/cs.1713>.
(freely available at: <https://ispan.waw.pl/journals/index.php/cs-ec/article/view/cs.1713/4458>)
- Kiril Simov and Petya Osenova. 2020. Integrated Language and Knowledge Resources for CLaDA-BG. In: *Selected Papers from the CLARIN Annual Conference 2019*, 172 (2020), LiU Electronic Press: Linköping Electronic Conference Proceedings 172 (2020), 2020.
- Dimitar Popov et al. 1994. D. Popov, L. Andreychin, L. Georgiev, St. Ilchev, N. Kostov, Iv. Lekov, St. Stoykov and Tsv. Todorov 1994. *Bulgarian Explanatory Dictionary*. Nauka i izkustvo. Sofia. (in Bulgarian)