COGNITIVE STUDIES | ÉTUDES COGNITIVES, 18
Warsaw 2018

Article No.: 1713
DOI: 10.11649/cs.1713

Citation: Osenova, P., & Simov, K. (2018). The data-driven Bulgarian WordNet: BTBWN. *Cognitive Studies | Études cognitives*, *2018*(18). https://doi.org/10.11649/cs.1713

PETYA OSENOVA[A] & **KIRIL SIMOV**[B]

Institute of Information and Communication Technologies, BAS,
Akad. G. Bonchev. 25A, 1113 Sofia, Bulgaria

[A]petya@bultreebank.org ; [B]kivs@bultreebank.org

THE DATA-DRIVEN BULGARIAN WORDNET: BTBWN

Abstract

The paper presents our work towards the simultaneous creation of a data-driven WordNet for Bulgarian and a manually annotated treebank with semantic information. Such an approach requires synchronization of the word senses in both — syntactic and lexical resources, without limiting the WordNet senses to the corpus or vice versa. Our strategy focuses on the identification of senses used in BulTreeBank, but the missing senses of a lemma also have been covered through exploration of bigger corpora. The identified senses have been organized in synsets for the Bulgarian WordNet. Then they have been aligned to the Princeton WordNet synsets. Various types of mappings are considered between both resources in a cross-lingual aspect and with respect to ensuring maximum connectivity and potential for incorporating the language specific concepts. The mapping between the two WordNets (English and Bulgarian) is a basis for applications such as machine translation and multilingual information retrieval.
Keywords: Bulgarian WordNet; WordNet mappings; data-driven WordNet construction

1 Introduction

There have been two prominent trends in language resources creation — compiling syntactically annotated resources (treebanks), on the one hand, and building lexical resources (WordNets), on the other. The former resources reflect the syntagmatic connectedness of the words, while the latter encode primarily the paradigmatic relations among words (via hierarchies). There are also works focused on the semantic annotation of corpora/treebanks, which apply the lexical knowledge onto real texts. Here we report on the challenges behind the construction of the BulTreeBank WordNet for Bulgarian (BTBWN). BTBWN has been created in three different ways: (1) by manual translation of English synsets from Core WordNet subset of Princeton WordNet (PWN — Fellbaum, 1998)[1] into Bulgarian. This step ensures comparable coverage between the two WordNets on the most frequent senses; (2) by identification of senses used in Bulgarian Treebank BulTreeBank (BTB). The identified senses have been organized in synsets for the BulTreeBank

[1]The Core WordNet contains the 5000 most frequent synsets of PWN. http://wordnetcode.princeton.edu/standoff-files/core-wordnet.txt

WordNet. The newly created Bulgarian synsets are being mapped onto the conceptual structure of PWN. In this way, the BTBWN was extended with real usages of the word meanings in texts. Also, the coverage of the core and base concepts for Princeton WordNet has been evaluated over a Bulgarian syntactic corpus; (3) by sense extension, which includes two activities: a) detection of the missing senses of processed lemmas in BulTreeBank and adding them to the BTBWN, and b) a semi-automatic extraction of information from the Bulgarian Wiktionary mapped to synsets from PWN and then manually checked[2] In this paper we present the second step of creating the BTBWN — simultaneous annotation of BTB with senses, the extension of BTBWN with these new synsets and their mapping to PWN.

The structure of the papers is as follows. Section 2 briefly discusses related work. The construction of BTBWN is presented in Section 3. Section 4 introduces the general principles of mapping. Section 5 presents two extensions of BTBWN in progress and the future direction of development. The last section concludes the paper.

2 Related work

Concerning WordNets, many of them for the European languages have been created within EuroWordNet and BalkaNet projects (including BulNet for Bulgarian). However, some of these WordNets are not publicly available (including BulNet). This motivated us to start our own WordNet creation endeavor, since we needed the lexico-semantic information in our work on Machine Translation, eLearning and Word Sense Disambiguation.

There are two main methods for building a WordNet as pointed out in Raffaelli, Tadić, Bekavac, and Agić (2008) the *expand* method and the *merge* method. The former relies on the translation of the synsets from the source into the target language, thus complying initially to the source hierarchy of concepts. The latter takes (also) into account the language specific resources. Different WordNet projects used the above mentioned methods alone or in combination with other strategies. For example, translation of English PWN into another language; data-driven approaches via identification of synsets within real texts; automatic extraction from existing lexical resources; various combinations of these. In our WordNet project we exploit all of these approaches at different stages of the resource development. They will be explained in more detail in the next sections.

Let us introduce briefly some best practices in the WordNet creation for specific languages. Most of them seem to go for the expand approach first. Some WordNets were created on the basis of publicly available resources. For example, the Open Dutch WordNet[3] (see Postma, Miltenburg, Segers, Schoen, & Vossen, 2016) was created by "removing the proprietary content from Cornetto[4], and by using open source resources to replace this proprietary content." For the Basque language (Pociello, Agirre, & Aldezabal, 2011) the construction approach relies on the joint development of WordNets and annotated corpora. The Basque WordNet was developed within the EuroWordNet framework. First, a quick core Basque WordNet was developed through semi-automatic methods. The quality control included a concept-to-concept manual review. Afterwards, an additional word-to-word review was performed as a higher-level quality check. The Slovenian WordNet started as automatic translation from a closely-related language resource, namely — the Serbian WordNet with the help of a bilingual dictionary. Later on, manual correction has been performed Erjavec and Fišer (2006). The Croatian WordNet Raffaelli et al. (2008) also used the expand method, but it additionally explored monolingual dictionaries for incorporating language-specific relations into the resource. One of the few endevours for constructing a language-specific WordNet first, and then mapping it to some already existing one, such as the Princeton WordNet, is the Polish WordNet Rudnicka, Maziarz, Piasecki, and Szpakowicz (2012).

[2]We would like to thank Antoni Oliver Gonzalez who provided the automatic mapping from Bulgarian Wiktionary to PWN.

[3]`http://wordpress.let.vupr.nl/odwn/`

[4]`http://www2.let.vu.nl/oz/cltl/cornetto`

To sum up, there is no easy way to achieve typological consistency in building WordNets — if the expand method is chosen, the language resource suffers from lack of nativeness in the hierarchy and relations. If the merge method is followed, the language resource differs too much from other similar resources for other languages, and thus it is time-consuming to map it back to them.

Now let us turn to the accompanying sense corpora. The usual way of annotating senses in treebanks is the following: there is a WordNet for the language in question, and then the treebank is annotated with senses from it. This is the case in the German Tuba/DZ treebank, the Italian treebank and the Polish treebank Hajnicz (2014). All of them use the WordNets they created in the EuroWordNet Project for sense annotation of the treebanks. Thus, they bear also the restrictions that are presented in the so-called static lexical resources. This means the following: if we want to annotate our texts with senses, but some sense is missing in the lexical database, and we cannot control the WordNet resource to add it, then the sparseness of the sense coverage would be really problematic.

Our work differs from the above mentioned approaches in the fact that we first annotated the treebank with senses from an explanatory dictionary of Bulgarian Popov et al. (2014) and then started the formation of synsets. They were mapped to the PWN while keeping track of the various sense discrepancies by differently marked mappings. We explain our motivation for such a decision below in a more contextually-bound manner. Here it can be only mentioned that in this way a wider sense coverage was achieved quickly for the purposes of Machine Translation, since our initial WordNet covered only the core concepts from PWN.

3 Building the BTBWN

In this section we present the steps of building the BTBWN including also the retrospective point of view. The creation of this resource started as an attempt to construct terminological vocabularies for two domain ontologies: the domain of *Information Technology for End Users*, and the domain of *Home Textile* — see Simov and Osenova (2008) and Simov (2009). In both cases the domain ontologies were aligned to an upper ontology for the reasons of consistency and inheritance of general knowledge. The ontology and the aligned lexicons were used for several tasks: (1) semantic annotation of domain documents; (2) multilingual search; (3) common conceptualization; and (4) interaction with the end users. Thus, the lexicon communicated the concepts from the ontology to the lexical knowledge used by the grammar in order to recognize the realizations of the concepts in the text; the lexicon represented the main interface between the user and the ontology. For achieving such an interface, the need of general lexica became apparent. Thus our next goal was to extend the domain lexicons to cover (at least) the most frequent senses in Bulgarian. We could not find any evaluation on the distribution of word senses in Bulgarian. Thus we decided to solve this problem in two steps: (1) by transferring the most frequent senses from another language to Bulgarian, assuming that European languages share substantial number of most frequent senses; and (2) by annotation of Bulgarian texts where we believed that the most frequent senses would be present. For the purposes of applications, such as word sense disambiguation, annotated texts were needed. So we decided to annotate the senses for all open class words in the texts.

Concerning the first step — transfer of most frequent senses from another language — we translated manually the English synsets from the Core WordNet subset of the Princeton WordNet into Bulgarian. The translation was done by two people with excellent knowledge of English. First, they formulated a Bulgarian definition reflecting the content of the concept represented by its correspondence to the English synset. Then they formed the Bulgarian synset recording the Bulgarian lemmas that have this meaning. Some of the lemmas might be multiword expressions. After this first phase a lexicographer checked both — the definition and the lemmas. The result from this work was published as part of the Open Multilingual WordNet[5] under CC BY 3.0

[5] `http://compling.hss.ntu.edu.sg/omw/`

license[6].

Our next step for extending the BTBWN was the manual annotation of running Bulgarian texts. Here our goals were: (1) to extend the coverage of BTBWN to really frequent Bulgarian words; (2) to have a corpus of semantically annotated texts which to be used for experiments within tasks like Word Sense Disambiguation; and (3) to check how many of the English most frequent senses are frequent also in Bulgarian. The actual annotation of the treebank was done in the following way: (1) for each lemma of the open class word forms in the treebank a concordance was created; (2) each lemma in the concordance was annotated with all possible senses from the Core WordNet version of BTBWN as well as from an explanatory dictionary of Bulgarian; (3) the annotators selected the appropriate sense for each example, if available. If there was no appropriate sense, or there was no available senses for a given lemma, the annotator had the possibility to create a new sense (definition). After the completion of this initial annotation the result was turned into lexical entries which contain the lemmas, selected in the text, the chosen definitions and the examples.

The next step was to manually map each new lexical entry to an appropriate synset in PWN. Thus we achieved several goals: (1) different lemmas with similar senses were grouped together and in this way the lexical entries for synonyms were recorded in the corresponding synsets; (2) the mapping to PWN allowed the execution of various bilingual applications; (3) mediated by PWN mapping to WordNets of other languages. The annotation was checked by a second person and validated by a judicator. After the completion of the annotation, BTBWN contained about 11000 synsets. From them about 1800 synsets are from the Core WordNet version of BTBWN. In this way we empirically showed that the most frequent senses in the texts of BulTreeBank correspond roughly to one third of the English Core WordNet.

The next extension of BTBWN was performed by a semi-automatic addition of Bulgarian Wiktionary mapped to synsets from PWN and then manually checked. Behind this extension we added new senses for the words that had already been included in BTBWN synsets. The idea is that each word is represented with all its senses.

The extensions on the basis of text annotation and the existing lexicons exhibit however the sparseness problem: not all synonyms appear in the annotated texts and the lexical entries. For that reason, we performed checks on the completeness of the synsets with respect to the missing synonyms. The checks have been performed with respect to the available monolingual synonymic dictionaries of Bulgarian. Special attention was paid to the aspect variation of verbs (perfective and imperfective). In many of the synsets it turned out that for one of the verbs in the aspect pair there were only few real examples or no examples at all in the data. Thus, we started searching for examples in bigger corpora or on the web. Our goal is to have at least five examples for each synset. Ideally, examples are expected to be included for each lemma in the synsets.

4 Mapping to other resources

In the process of creation of BTBWN we have been performing mappings to several other resources in order to enrich its structure with new word senses and new relations. We thus map the BTBWN to PWN, to the Bulgarian Wikipedia and to the Predicate Matrix. For the English resources the mappings are performed through the PWN. The synchronization of word senses in BTBWN and the word senses in PWN appeared to be complicated due to the fact that many senses in BTBWN originate from real data, namely BulTreeBank. In this corpus the words reflect Bulgarian lexicalizations of concepts that might differ from PWN being the provider of the English-specific view on the lexical relations. Also, since BulTreeBank comprises mainly texts from media, for many lemmas only the figurative senses are registered, not the most common ones.

[6] `https://creativecommons.org/licenses/by/3.0/`

4.1 Mapping to PWN

From the annotator perspective, the mapping of the word from text starts with its translation into English and is followed by a search through the corresponding lemmas in the PWN. The factors of importance for the adequate mapping are the following: the Bulgarian definition and the matching examples from the treebank. The provision of examples plays a crucial rule for the specification of the correct definition as well as the English description and accompanying examples. The PWN examples themselves can also help in indicating the matching concept, since the Bulgarian definition and the English one can be phrased in different ways and thus might reflect various degrees of granularity within conceptualizations.

Several types of correspondences have been attested during the mapping process: full correspondence (one-to-one); partial correspondence (one-to-many or many-to-one); forced connectivity (re-design of the Bulgarian definition); common general meaning; resolving metonymies; incorrect and extended correspondences. Needless to say, these types of correspondences are not novel at all. However, they are quite relevant, because they provide feedback for typologically-based and resource-oriented similarities and differences between Bulgarian and English, thus opening the path to comparisons with other languages.

Full Correspondence

The ideal case in the mapping is when equal concepts are encountered, i.e. the concepts in the two languages map one-to-one. That is, the Bulgarian concept matches equally to the one in the Princeton WordNet. For example, the Bulgarian noun "сигурност", "sigurnost" and the English noun "safety" both mean in short 'lack of danger'. If a Bulgarian definition corresponds well to more than one definition in the Princeton WordNet, then all these definitions are mapped to the Bulgarian one, through a special separator. For example, the English nouns "answer" and "response" map to the Bulgarian noun "отговор", "otgovor".

Partial Correspondence

In many cases, however, the concepts differ in terms of specificity in both language directions. In the first case, the Bulgarian definition is more specific than the English one. In this case, it is mapped to a more general English one, but it is also marked with a specificity label. The most frequent cases here is the regular polysemy — for example, in Bulgarian the noun "прокуратура", "prokuratura", is given also with the meaning of the building, while in English it is meant as the institution, the group of people and the act.

A second scenario is possible, where the Bulgarian definition is more general and subsumes one or more synsets from PWN. In this case, the following approach has been taken — the common definition in Bulgarian is mapped once to the more specific English definitions (with the *specificity*) relation and a second time — to their hypernyms (with the *subsumption* relation). For instance, in Bulgarian "режисьор", "rezhisyor", "director" has only one definition: *The lead person in the making of a theater play, film, TV program, etc.* However, in PWN there exist two synsets that can be related to it: *director as someone who supervises the actors and directs the action in the production of a show* and *director as the person who directs the making of a film.*

Ensuring a One-to-One Mapping. In some cases of mismatch the one-to-one mapping can be achieved through re-working the Bulgarian definitions. This often means dividing the Bulgarian definition into two separate ones. For example, the word "седмица", "sedmitsa", "week", has the following definition: 'seven consecutive days, usually counted from Monday to Sunday'. All examples correspond to this definition. There are two synsets in English: "week" as any period of seven consecutive days, and "week" as a period of seven consecutive days starting on Sunday.

Searching for a More General Meaning. There is another group of examples to which no equivalent sense can be detected. In this case the strategy is to find a more general one. Usually this applies to the cases of regular polysemy:

For example, Words like "'цар", "tsar", "king", etc. in Bulgarian refer to both concepts — a person and a title. In PWN there is a definition only for a person and there is no word sense corresponding to the title meaning, as in "The title of the Bulgarian and Russian monarchs."

Therefore this Bulgarian definition is mapped to the more general concept for title. On the other hand, a second definition in Bulgarian is added to reflect the dimension of persons.

Figurative Usages. If a word is used with its figurative sense, then this sense is mapped to the appropriate sense in the PWN. For example, in Bulgarian the word "армеец", "armeets", can be used in two senses: literal and rare (soldier), or figurative and more frequent (member of a specific football team). When used in the latter sense, it cannot be mapped to the concept of 'soldier', but to the concept of 'footballer'. In this way, the specific features of the figurative language are kept in the lexicon.

4.2 Mapping to DBpedia/Wikipedia

Some time ago we have annotated the Named Entities in Bulgarian Treebank BulTreeBank with URIs from DBpedia (Wikipedia) — Popov et al. (2014). The annotation with DBpedia[7] URIs covers 10 885 named entities — 2877 organizations, 2938 locations, 4195 people, the rest are from different categories: events, books, others. Unfortunately, the coverage of the Bulgarian DBpedia is rather small. For that reason, the Bulgarian Wikipedia was used for adding the respective links into the data. As expected, many Named Entities are not present in Wikipedia and consequently — in DBpedia. For that reason such entities have been annotated with classes from DBpedia ontology[8]. We have selected classes on the basis of the context of usage. We aimed to cover as many classes as possible. For example, Romano Prodi could be annotated with classes like `dbo:Person`,[9] `dbo:President`, `dbo:Politician`, and other if such are available for the particular Named Entity. In order to facilitate the further use of the BTBWN we mapped the DBpedia ontology to it.

The integration of WordNets for other languages with additional semantic resources has a long history. Here we will mention only some of them: BabelNet — Navigli and Ponzetto (2012), UBY — Gurevych et al. (2012). However, all the mentioned approaches are automatic and despite having quite high level of precision they still do not provide gold level of mapping as it was stated in (McCrae, 2018). Our goal is to extend BTBWN in two directions: (1) an extension of the number of senses for lemmas that are already in BTBWN; and (2) an extension of BTBWN with instances.

The first extension was done manually for the lemmas that are already in BTBWN. For each lemma we collected all the senses that are currently in the wordnet, and from Wikipedia we collected all the articles that started with the same lemma while cleaning the modification suffixes. For example, for the lemma 'маса' in Wikipedia by using the disambiguation page we found five senses from which three were already in BTBWN and two were missing. One is a name of a language, and we consider it as a named entity that we leave for later elaboration. The other missing sense is used in electrical engineering for Ground.[10]. Thus, the discovered missing sense is added to BTBWN together with the definition and link to the Wikipedia article. This work is not completed yet. Additionally to the extension of BTBWN with new senses this enrichment contributes also new encyclopedic knowledge to BTBWN. The second extension is done through a mapping from the DBpedia ontology to the synsets in BTBWN. In this way we automatically classify the instances of the classes as instances of the corresponding synsets. We have performed these extensions for the moment for the three more frequent types of Named Entities annotated in the BulTreeBank treebank — persons, locations, and organizations. For these mappings we used the corresponding classes in the ontology and all their sub-classes. We do not try to map these new instances to the instances that are already in PWN, but we plan to use the interlanguage links from Wikipedia in future applications. The addition of these Named Entities to BTBWN allows us also to maintain extensions of the coverage of the resource with aliases that are not in Wikipedia.

[7] `https://wiki.dbpedia.org/`
[8] `https://wiki.dbpedia.org/services-resources/ontology`
[9] `dbo:` is used as a name space for DBpedia ontology.
[10] `https://en.wikipedia.org/wiki/Ground_(electricity)`

Although this work is very time consuming we consider it very important for the creation of a lexical resource for Bulgarian that has a good coverage in terms of senses and relations between them.

4.3 Sense annotation with BTBWN

In the treebank 79 703 tokens have been annotated with senses from BTBWN. More specifically, 37 330 nouns; 14 341 verbs, 17 304 adjectives and 10 728 adverbs.

We performed an experiment to test the coverage of the wordnet senses from the perspective of the mappings of BTBWN to PWN. For this experiment we used texts from the parallel English-Bulgarian parts of the Setimes corpus. The English part was first annotated with Predicate Matrix senses that combine information from wordnet, VerbNet and FrameNet. Then the senses have been transferred to the Bulgarian texts. We have not performed any quantitative evaluation, because the coverage of BTBWN was not that good, but we performed some qualitative analysis.

Thus, the transferred senses reflect the following cases: the sense differs from the one in Predicate Matrix, but still is valid; several senses matched, and the correct one was there; one sense matched from BTBWN and it was the correct one; wrong sense was mapped due to the missing corresponding sense of the lemma in Bulgarian.

The non-transferred senses were due to: errors in the part-of-speech tagger (English or Bulgarian) or a missing lemma in BTBWN.

5 Current and future developments

Here we discuss briefly: the treatment of MultiWord Expressions (MWEs) as specific cases of lexicalization; and extensions of the current version with new words. We also present some directions of future developments.

5.1 Treatment of MultiWord Expressions in BTBWN

Currently, we include MWEs as strings of several words separated by spaces. They are represented in their standard form: lemmatized (where possible) and reflecting the canonical word order. However, in this representation we lose information about possible word order variations of the MWE elements and their potential for morphosyntactic variation and modification. In order to add this information we rely on the notion of *catena*.

The notion of catena (chain) was introduced in O'Grady (1998) as a mechanism for representing the syntactic structure of idioms. He shows that for this task a definition of syntactic patterns is needed that does not coincide with constituents. He defines the catena in the following way: *The words A, B, and C (order irrelevant) form a chain if and only if A immediately dominates B and C, or if and only if A immediately dominates B and B immediately dominates C.* In our work on BTBWN we convert MWEs into a representation defined in Simov and Osenova (2014) and Simov and Osenova (2015) in which the catena is depicted as a dependency tree fragment with appropriate grammatical and semantic information. Here we demonstrate the model by just one lexical entry for the Bulgarian MWE: затварям си очите, "zatvaryam si ochite", "I close my eyes".

The lexical entry uses the following format: a **lexicon-catena** (LC), **semantics** (SM) and **valency** (Frame). The lexicon-catena for the MWEs is stored in its canonical form. The realization of the catena in a sentence has to obey the rules of the grammar. In this way the possible word order is managed. The semantics of a lexical entry specifies the list of elementary predicates contributed by the lexical item. When the MWE allows for some modification (including adjunction) of its elements, i.e. modifiers of a noun, the lexical entry in the lexicon needs to specify the role of these modifiers. For example, the MWE represented in Fig. 1.[11]

[11] The grammatical features are: 'poss' for possessive pronoun, 'plur' for plural number and 'def' for definite noun.

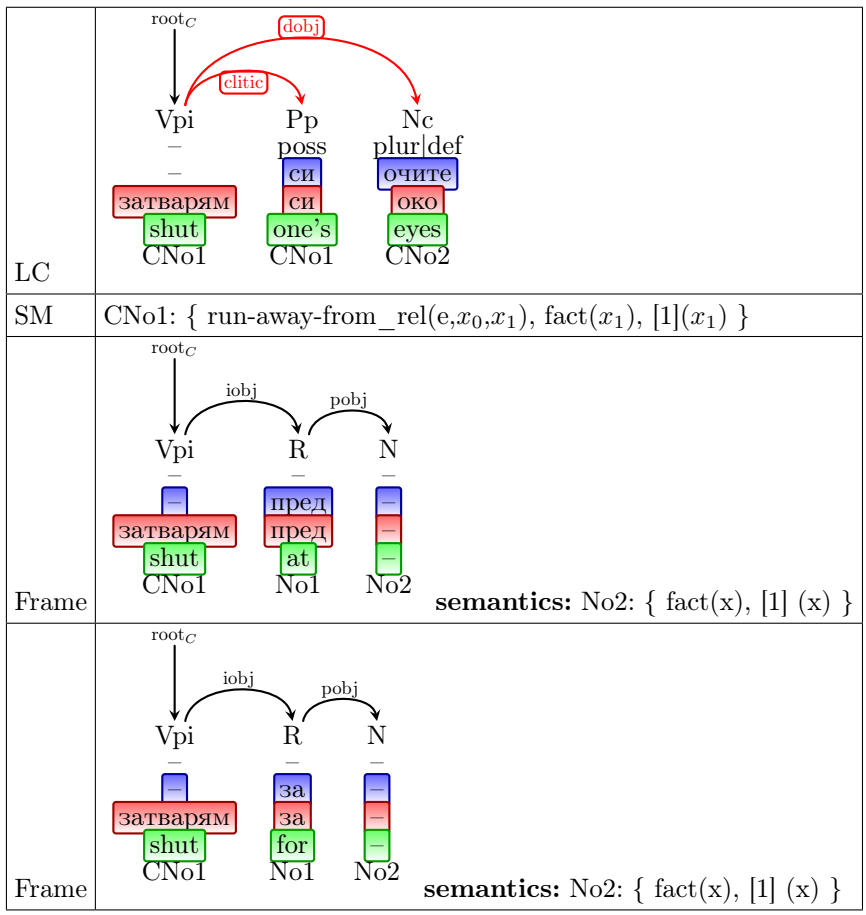

Figure 1: Lexical entry for затварям си очите, "zatvaryam si ochite", 'I close my eyes'.

The valency frame contains two alternative elements for indirect object introduced by two different prepositions. The situation that the two descriptions are alternatives follows from the fact that the verb has no more than one indirect object. If there is also a direct object then the valency set will contain elements for it as well. The semantic contribution of the indirect object is specified for each valency element. This semantic contribution is added to the semantic contribution of the lexical entry when the valency element is realized. In the dependency tree fragments also grammatical features and lemmas are represented. The catena for the frame and for the whole lexical entry are unified on the basis of nodes with the same names. Within BTBWN the semantic contribution will depend also on the corresponding synsets to which MWEs belong.

5.2 Extensions of BTBWN

The extension of the BTBWN coverage is a constant task. The selection of new lexical entries, including new synonyms, new senses for words that are already in synsets in BTBWN, new words with corresponding new senses and synsets is an on-going activity. To perform this task we use the following approaches: (1) a task-based approach; (2) a dictionary-based approach; (3) a corpus-based approach; and (4) a linguistically-based approach. All of these approaches are used by different language groups for the construction of the corresponding wordnets. We briefly present each of them.

The *task-based approach* handles the coverage of BTBWN for a concrete application. For

example, the construction of lexicons for domain ontologies. In this case we identify concepts for the task to be performed and construct an aligned lexicon. The synsets in the domain lexicon are also aligned to the rest of BTBWN. The identification of the domain concepts is ususally based on the annotation of appropriate domain texts.

The *dictionary-based approach* is performed by comparing senses for each word that is already in BTBWN with senses of the same word in a given dictionary. We performed this in two ways: comparing with senses registered in the Bulgarian Wiktionary, and with senses in the Bulgarian explanatory dictionary. In many cases we reformulated the identified non-covered senses on the basis of the existing senses in BTBWN and with the requirement for a better mapping to the English PWN.

The *corpus-based approach* uses several mechanisms for identification of new words and senses to be added to BTBWN. They include at least the following ones: (1) annotation of new texts; (2) clustering of word forms in a large corpus on the basis of their contexts (Polish WordNet and some others); and (3) checking the coverage of BTBWN over a frequency list compiled from a large corpus. Currently, we perform (3) on a frequency list compiled over a 7-million-word corpus covering different types of text. Our goal is to include all the words that appear in the corpus at least 100 times. Any word that is not presented in BTBWN is lemmatized in all possible ways and then included by each possible lemma and each possible sense. For example, the Bulgarian word form "поет", "poet", is lemmatized as a noun "поет", "poet", and a verb "поема", "poema", "take". In this way we cover all frequent word forms in the corpus with their relevant senses, independently of the context. The people who add the new words and senses are free to search for usages of the corresponding lemmas not only within the corpus, but also on the web. Currently 3,708 new synsets have been added to BTBWN in three months by one person. Each new synset includes a list of lemmas, a definition and at least five examples. From these new synsets 3,302 have been mapped to the English synsets with the equivalent meanings relation and 406 have been mapped through some other relation. The percentage of the non-equivalent mappings was expectedly similar to the percentage for the other part of BTBWN. The coverage over the whole corpus (excluding the stop words) increased from 70.58 % to 80.39 %. We consider this as a good sign that the data-driven approach to construction of wordnets provides good coverage with relatively small number of synsets. Keep in mind that many of the uncovered words are names.

The *linguistically-based approach* exploits productive phenomena within the language. We mainly exploit derivation patterns with clear new semantics. For example, the names of citizens of a given location is such a case: from "New York" to form "New Yorker". This pattern is easy to recognize in the corpus, and the definition and mapping to the rest of BTBWN is thus predictable.

Besides these two current activities we plan to perform also the following two tasks: (1) addition of relational structure over BTBWN; and (2) including the synsets that are not mapped exactly to synsets in PWN to the Collaborative Interlingual Index (CILI) — Vossen, Bond, McCrae, and Fellbaum (2016). For the latter task it appropriate definitions in English have to be added. For the former task we will exploit the mapping to the English WordNet and additionally the mapping from the English WordNet to the Polish Wordnet. In this way we will be able to transfer relations between the synsets in English and Polish WordNets to Bulgarian. Thus, we expect to impose a reliable relation structure over BTBWN. We manually will check the cases of lexical relations like *antonymy* and *derivation* as well as the cases where English and Polish Wordnets disagree with each other.

6 Conclusion

The paper discusses our strategy for mapping the word senses in a treebank to the WordNet ones from the perspective of the overall construction of the BTBWN. In the presented approach the resource annotation does not rely on pre-created WordNet, but rather on an explanatory dictionary of Bulgarian. Later on, these senses have been mapped to the PWN 3.0, while keeping

the language specific concepts through the introduction of special markings. The adopted strategy allowed for dense connectivity between the resources, and at the same time it leaves room for the further creation of a language-specific hierarchy. Currently BTBWN contains more than 20 000 synsets equivalent to the synsets in PWN, and about 2900 additional synsets mapped in a special way as described in the paper.

These mappings have been exploited actively for knowledge-based word sense disambiguation of Bulgarian. This was done by using the English WordNet as a knowledge graph that transfers the linguistic relations to Bulgarian lemmas. Also, these mappings will determine the language-specific hierarchy of concepts over the Bulgarian definitions.

References

Erjavec, T., & Fišer, D. (2006). Building Slovene WordNet. In *Proceedings of the 5th International Conference on Language Resources and Evaluations (LREC 2006)* (pp. 1678–1683). Retrieved from http://www.lrec-conf.org/proceedings/lrec2006/

Fellbaum, C. (Ed.). (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Gurevych, I., Eckle-Kohler, J., Hartmann, S., Matuschek, M., Meyer, C. M., & Wirth, C. (2012). UBY — A large-scale unified lexical-semantic resource based on LMF. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 580–590). Avignon: The Association for Computer Linguistics. Retrieved from http://www.aclweb.org/anthology/E12-1059

Hajnicz, E. (2014). The procedure of lexico-semantic annotation of *Składnica* Treebank. In *Proceedings of LREC-2014* (pp. 2290–2297). Retrieved from http://www.lrec-conf.org/proceedings/lrec2014/pdf/444_Paper.pdf

McCrae, J. P. (2018). Mapping WordNet instances to Wikipedia. In *Proceedings of the 9th Global WordNet Conference (GWC 2018)* (pp. 62–69). Singapore: Global WordNet Association. Retrieved from http://compling.hss.ntu.edu.sg/events/2018-gwc/pdfs/gwc-2018-proceedings.pdf

Navigli, R., & Ponzetto, S. P. (2012). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, *193*, 217–250. https://doi.org/10.1016/j.artint.2012.07.001

O'Grady, W. (1998). The syntax of idioms. *Natural Language and Linguistic Theory*, *16*(2), 279–312. https://doi.org/10.1023/A:1005932710202

Pociello, E., Agirre, E., & Aldezabal, W. (2011). Methodology and construction of the Basque WordNet. *Language Resources and Evaluation*, *45*(2), 121–142. https://doi.org/10.1007/s10579-010-9131-y

Popov, A., Kancheva, S., Manova, S., Radev, I., Simov, K., & Osenova, P. (2014). The sense annotation of BulTreeBank. In V. Henrich, E. Hinrichs, D. de Kok, P. Osenova, & A. Przepiórkowski (Eds.), *Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories (TLT13)* (pp. 127–136). Retrieved from http://tlt13.sfs.uni-tuebingen.de/tlt13-proceedings.pdf

Postma, M., Miltenburg, E. van, Segers, R., Schoen, A., & Vossen, P. (2016). Open Dutch WordNet. In V. Barbu Mititelu, C. Forascu, C. Fellbaum, & P. Vossen (Eds.), *Proceedings of the Eighth Global WordNet Conference* (pp. 300—308). Retrieved from http://jiangbian.me/papers/2016/gwc2016.pdf

Raffaelli, I., Tadić, M., Bekavac, B., & Agić, Z. (2008). Building Croatian WordNet. In A. Tanács, D. Csendes, V. Vincze, C. Fellbaum, & P. Vossen (Eds.), *GWC 2008: The Fourth Global WordNet Conference, Szeged, Hungary, January 22–25, 2008: Proceedings* (pp. 349–359). Retrieved from http://www.inf.u-szeged.hu/projectdirs/gwc2008/GWC2008_Proceedings_Final.pdf

Rudnicka, E., Maziarz, M., Piasecki, M., & Szpakowicz, S. (2012). A strategy of mapping Polish WordNet onto Princeton WordNet. In M. Kay & C. Boitet (Eds.), *Proceedings of COLING 2012: Poster* (pp. 1039—1048). Retrieved from http://anthology.aclweb.org/C/C12/C12-2.pdf

Simov, K. (2009). Ontology-based lexicon of Bulgarian. *Journal for Language Technology and Computational Linguistics*, *24*(2), 40–55.

Simov, K., & Osenova, P. (2008). Language resources and tools for ontology-based semantic annotation. In *Proceeding of OntoLex 2008 Workshop at LREC 2008* (pp. 9–13). Retrieved from http://www.lrec-conf.org/proceedings/lrec2008/

Simov, K., & Osenova, P. (2014). Formalizing multiwords as catenae in a treebank and in a lexicon. In V. Henrich, E. Hinrichs, D. de Kok, P. Osenova, & A. Przepiórkowski (Eds.), *Proceedings of the*

Thirteenth International Workshop on Treebanks and Linguistic Theories (TLT13) (pp. 198–207). Retrieved from http://tlt13.sfs.uni-tuebingen.de/tlt13-proceedings.pdf

Simov, K., & Osenova, P. (2015). Catena operations for unified dependency analysis. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)* (pp. 320– 329). Uppsala: Uppsala University. Retrieved from http://www.aclweb.org/anthology/W15-2135

Vossen, P., Bond, F., McCrae, J. P., & Fellbaum, C. (2016). CILI: The Collaborative Interlingual Index. In V. Barbu Mititelu, C. Forascu, C. Fellbaum, & P. Vossen (Eds.), *Proceedings of the Eighth Global WordNet Conference* (pp. 50–57). Retrieved from http://jiangbian.me/papers/2016/gwc2016.pdf

Acknowledgment

This research has received partial support by the grant 02/12 — *Deep Models of Semantic Knowledge (DemoSem)*, funded by the Bulgarian National Science Fund in 2017–2019. We are grateful to the anonymous reviewers for their remarks, comments, and suggestions. All errors remain our own responsibility.

The authors declare that they have no competing interests.

The authors' contribution was as follows: concept of the study: Kiril Simov; data analyses: Petya Osenova (sense annotation and mappings to PWN), Kiril Simov (catena approach and mappings to other resources); the writing: Petya Osenova (Introduction, Related Work, Building the BTBWN, Mapping to Other Resources – 4.1, 4.3), Kiril Simov (Mapping to Other Resources – 4.2, Current and Future Developments, Conclusion).

Publisher: Institute of Slavic Studies, Polish Academy of Sciences, University of Silesia & The Slavic Foundation