Incorporating Prepositions in the BulTreeBank WordNet

Zara Kancheva
Institute of Information and Communication Technologies,
Bulgarian Academy of Sciences / Sofia, Bulgaria
zara@bultreebank.org

Abstract

A model for preposition incorporation in the BulTreeBank WordNet is presented which follows the model for presenting open class words in wordnets. An adapted semantic classification of prepositions is done on the base of Bulgarian grammars and the classes are used for synset categories. The good coverage of prepositions in the wordnet will be used for the aim of neural language models creation for Bulgarian. This extension of the wordnet improves its utility for semantic annotation.

1 Introduction

The paper aims at presenting a model for preposition incorporation in the BulTreeBank WordNet (BTB-WN) (Osenova and Simov, 2018). Prepositions are considered a beneficial extension of the part of speech coverage of BTB-WN, because they would improve its utility for semantic annotation, word sense disambiguation, machine translation, etc. Additionally, they would provide a better quality of neural language models for Bulgarian, which is the long term purpose of this task.

BTB-WN was created on the base of the Core WordNet subset[1] of Princeton WordNet (PWN) (Fellbaum, 1998) that contains the 5000 most frequent English senses. After that it was expanded with content words from the BulTreeBank and a Bulgarian frequency list, as well as with senses from the Bulgarian versions of Wikipedia and Wiktionary. Initially it was mapped to the PWN, but since 2020 it shifted to the Open English WordNet (McCrae et al., 2020), because it is being developed and updated, in contrast to the PWN. The current version of BTB-WN[2] – 4.0 – contains more than 33 000 synsets and could be browsed online[3]. BTB-WN will be soon freely available for downloading in the WordNet LMF format.

Prepositions like any other closed class words usually are not presented in wordnets, including BTB-WN, which contains the four most common parts of speech – nouns, verbs, adjectives and adverbs. Prepositions are one of the most frequent and at the same time ambiguous parts of speech and additionally the independence of their semantics is often argued (Baldwin et al., 2009). Here a semantic categorisation of Bulgarian prepositions is done and the model for presenting prepositions in BTB-WN follows the model for the open class words.

Prepositions serve to establish different relations between words: on one hand grammatical – they indicate the syntactic position of words in phrases (object, adverbial, modifier), and on the other hand semantic – they reveal their sense relations (local, temporal, causal, etc.). In this research only the semantic function of prepositions is considered. Section 2 gives an overview of different preposition research, Section 3 presents the semantic preposition classification that is used, Section 4 contains the synset model for prepositions in BTB-WN, and Section 5 concludes the paper.

2 Related Works

Bulgarian grammars classify prepositions by their origin, morphological composition and semantics. A detailed review of the history of preposition classifications is presented in Konstantinova (1982). For the aim of preposition incorporation in BTB-WN two classifications are considered and a new more compact categorization is compiled (Boyadzhiev et al. (1998), Stoyanov (1983)).

The role of prepositions in NLP and the variety of approaches towards their processing are thor-

[1] http://wordnetcode.princeton.edu/standoff-files/core-wordnet.txt

[2] https://clada-bg.eu/bg/centers-and-services/language-technologies/btb-wordnet.html

[3] https://concordance.webclark.org/

Semantic Class	Synset category	Prepositions
locative	prep.location	в, всред, върху, въз, връз, до, за, зад, из, низ, изпод, измежду, извън, край, към, между, на, над, насред, о, около, от, откъм, отвъд, отсам, оттам, оттатък, по, под, подир, подире, помежду, посред, пред, през, при, против, след, сред, срещу, спроти, у
temporal	prep.time	в, всред, до, за, между, край, към, на, насред, около, от, по, подир, подире, помежду, посред, пред, през, при, с, след, спроти, сред, срещу
manner and instrument of action	prep.manner	без, в, като, на, по, под, посредством, при, с (със), според, чрез
cause	prep.cause	за, заради, от, по, поради, пред
purpose	prep.purpose	върху, до, за, заради, към, по, поради
possession	prep.possession	на, от, с, у
origin and part of a whole	prep.origin	в, от
quantitative, degree and exceeding of a limit	prep.quantity	до, за, към, между, на, над, около, от, с (със), около, свръх
exchange	prep.obj.exchange	вместо, за, заради, наместо, срещу, спроти
exclusion	prep.obj.exclusion	без, освен
opinion	prep.obj.opinion	за, по, според, спрямо
thought	prep.obj.thought	върху, връз, въз, за, заради, около, по, спрямо
transition	prep.transition	в, на, от
comparison	prep.comparison	като
opposition	prep.opposition	въпреки, против, пряко, спроти, срещу

Table 1: Semantic classes and synset categories of prepositions

oughly presented in Baldwin et al. (2009) and here I will outline only some of the most relevant research on the topic. Schneider et al. (2015) introduce a taxonomy of preposition functions called supersenses for classification of prepositions. The work is directed towards automatic word sense disambiguation and the classification is aimed to be suitable for manual annotation. 73 preposition supersenses are determined and mapped to other resources such as VerbNet[4].

There are two resources particularly dedicated to prepositions: PrepNet (Saint-Dizier, 2008) and the Preposition Project (Litkowski and Hargraves, 2005). The Preposition Project is a semantic database for English prepositions which has been used for word sense disambiguation. It combines data for prepositions from a dictionary and from FrameNet[5], where English prepositions are functionally tagged. PrepNet was originally build for French, but later extended for several languages. The approach in PrepNet is inspired by thematic role classifications and it also uses data from FrameNet.

O'Hara and Wiebe (2009) approached preposition disambiguation using the semantic roles from Penn Treebank, the semantic network Factotum and FrameNet, and hypernyms from the Princeton WordNet[6].

Preposition classifications particularly directed towards wordnets are done by Amaro (2018) and Harabagiu (1996). Harabagiu (1996) shows an approach where by using information from Princeton WordNet and applying inferential heuristics two types of phrases are analysed – *noun+preposition+noun* and *verb+preposition+noun*. The phrases are organized in classes if there are *hyperonymy*, *hyponymy* or *synonymy* relations between the verbs and the nouns in the phrases or if they have common hypernym/hyponym.

Amaro (2018) presents a very interesting approach towards preposition integration in wordnet, including visual description by using typical wordnet relations for Portuguese prepositions. The integration is not large scale, only prepositions for movement are processed and the following relations are introduced: *synonymy*, *antonymy*, *hyponymy/hyperonymy* and *causes/is caused by*.

As far as I am concerned BulNet[7] is the only wordnet that has prepositions but the incorporation in it is not beneficial enough for many NLP tasks – they are presented with definitions, synonyms,

[4] https://verbs.colorado.edu/verbnet/

[5] https://framenet.icsi.berkeley.edu/fndrupal/

[6] https://wordnet.princeton.edu/

[7] http://dcl.bas.bg/en/resursi/wordnet/

examples and English translations, but they do not have any relations.

Another relevant work is the research of Da Costa and Bond (2016) on the incorporation of non referential words in wordnet. They expand the Open Multilingual Wordnet[8] with interjections and numeral classifiers motivated by the task of semantic annotation, similarly to the case of BTB-WN. For interjections two relations are used: *exemplifies* (with other parts of speech) and *see also* (with interjections). For the classifiers *exemplifies* is used, but also two new relations are introduced: *classifies* and *classified by*. The Penn Discourse Treebank also contains annotations of closed class words including some prepositions Prasad et al. (2019).

My approach is similar to that of Amaro (2018) and Harabagiu (1996), because the presentation of prepositions in BTB-WN is following the model for the open class words in wordnets.

3 Semantic Classification of Prepositions

For their integration in BTB-WN prepositions have been semantically classified in 15 groups: `location`, `time`, `transition`, `manner and instrument of action`, `possession`, `quantity`, `degree and exceeding of limit`, `purpose`, `origin and part of a whole`, `opposition`, `comparison`, `cause` and `object class`: `exchange`, `exclusion`, `opinion` and `thought`.

There are several differences from the grammars in the adapted classification mainly motivated by the aim of having a more compact and general-purpose system. For example, the classes `manner of action` (слушам с внимание 'listen with attention') and `instrument of action` (пиша с молив 'write with a pencil') here are united in one class, because they are very closely related. The same applies for the `origin` (тя е от града 'she is from the city') and `part of a whole` (яж от този хляб 'eat from this bread') classes. The `approximation of time` (към 9 часа 'around 9 o'clock') and `approximation of quantity` (около 3 килограма 'about 3 kilograms') classes from Stoyanov (1983) here are included respectively in the `time` and `quantity` classes. The `exceeding of limit` sense (това е свръх възможностите ми 'this is beyond my abilities') is included in the `quantity` category. An `object` superclass is outlined to unite the expression of relations for `exchange` (отиди вместо мен 'go instead of me'), `exclusion` (няма други гости освен семейството 'there are no other guests except the family'), `thought` (разкажи ми за пътешествието 'tell me about the journey') and `opinion` (според мен това е добра идея 'to me this is a good idea'). The *prep.obj.thought* class includes expression of object of thought, speech and writing. The metaphorical usages of a given class are considered part of it, not a separate class. For instance, usages like "Тия неща са врязани в паметта ми." ('These things are etched in my memory') are considered as examples of the `location` class.

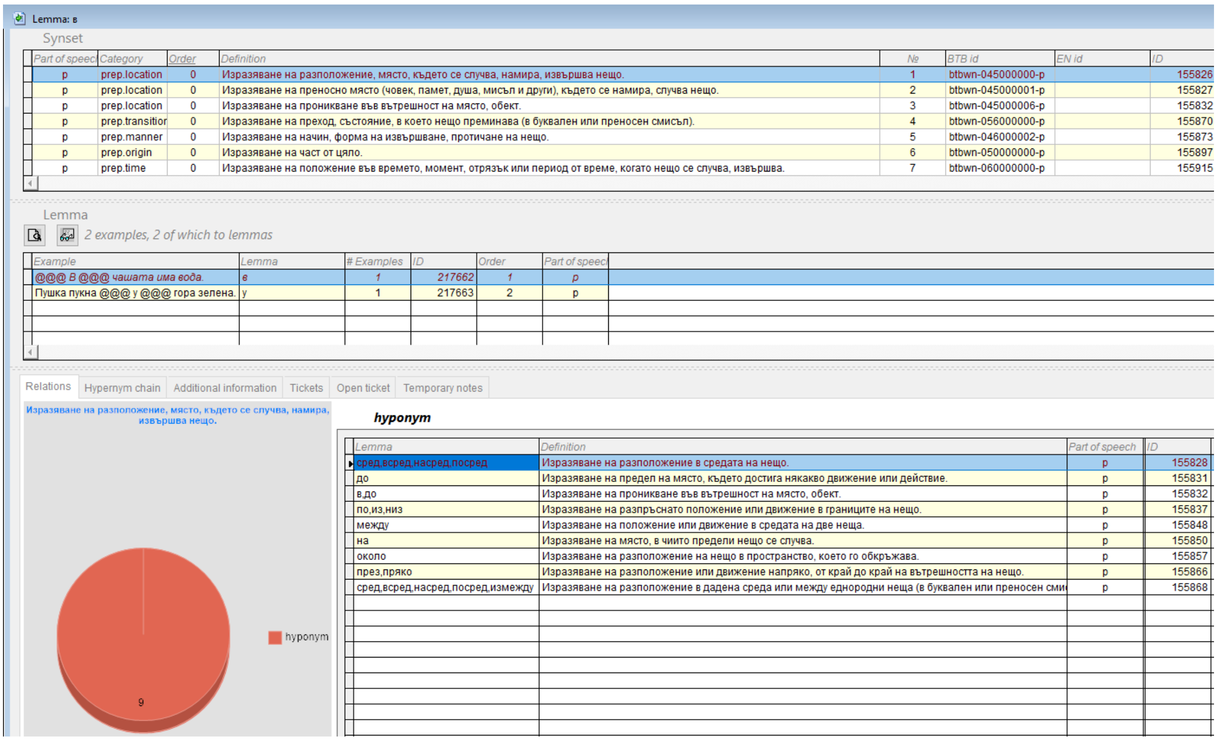
4 Preposition Synset Model

Preposition synsets have synset category (based on their semantic class), detailed definition, examples, synonyms if available and as much as possible relations. The part of speech value for prepositions in BTB-WN is `p`, following the format of the Global WordNet Association[9]. An example is shown in Figure 1 with the synsets for preposition в ('in').

The main intention for the preposition relations is that they follow the relations model of any other part of speech in wordnets. Two types of relations are used: between preposition synsets and between a preposition synset and other parts of speech. Examples for the first type are: *synonymy* (the prepositions върху, въз, връз, на 'over, on' all express position or motion over some surface, something or someone), *antonymy* (върху 'over' is antonym of под 'under'), *hyperonymy* and *hyponymy* (в 'in' in its most general meaning for 'position or action in the limits of something, somewhere' is hypernym of several prepositions which express more specific location relations, such as сред, всред, насред, посред 'in the middle', из, низ, по 'through', между 'between', през, пряко 'across'), *similar* (върху 'over' is similar with над 'above'). The second type is intended to link combinations of verbs and prepositions (and as a plan for future work – nouns and prepositions) which tend to express a particular meaning together (such as the combination of the verb превръщам се 'turn into' and the preposition в 'in, into' express `transition in new state`). The *sem-derived-from* relation can be

[8] compling.hss.ntu.edu.sg/omw/

[9] https://globalwordnet.github.io/schemas/



Part of speecl	Category	Order	Definition	№	BTB id	EN id	ID
p	prep.location	0	Изразяване на разположение, място, където се случва, намира, извършва нещо.	1	btbwn-045000000-p		155826
p	prep.location	0	Изразяване на преносно място (човек, памет, душа, мисъл и други), където се намира, случва нещо.	2	btbwn-045000001-p		155827
p	prep.location	0	Изразяване на проникване във вътрешност на място, обект.	3	btbwn-045000006-p		155832
p	prep.transitior	0	Изразяване на преход, състояние, в което нещо преминава (в буквален или преносен смисъл).	4	btbwn-056000000-p		155870
p	prep.manner	0	Изразяване на начин, форма на извършване, протичане на нещо.	5	btbwn-046000002-p		155873
p	prep.origin	0	Изразяване на част от цяло.	6	btbwn-050000000-p		155897
p	prep.time	0	Изразяване на положение във времето, момент, отрязък или период от време, когато нещо се случва, извършва.	7	btbwn-060000000-p		155915

2 examples, 2 of which to lemmas

Example	Lemma	# Examples	ID	Order	Part of speecl
@@@ В @@@ чашата има вода.	в	1	217662	1	p
Пушка пукна @@@ у @@@ гора зелена.	у	1	217663	2	p

Изразяване на разположение, място, където се случва, намира, извършва нещо.

hyponym

Lemma	Definition	Part of speech	ID
сред,всред,насред,посред	Изразяване на разположение в средата на нещо.	p	155828
до	Изразяване на предел на място, където достига някакво движение или действие.	p	155831
в,до	Изразяване на проникване във вътрешност на място, обект.	p	155832
по,из,низ	Изразяване на разпръснато положение или движение в границите на нещо.	p	155837
между	Изразяване на положение или движение в средата на две неща.	p	155848
на	Изразяване на място, в чиито предели нещо се случва.	p	155850
около	Изразяване на разположение на нещо в пространство, което го обкръжава.	p	155857
през,пряко	Изразяване на разположение или движение напряко, от край до край на вътрешността на нещо.	p	155866
сред,всред,насред,посред,измежду	Изразяване на разположение в дадена среда или между еднородни неща (в буквален или преносен смис	p	155868

Figure 1: Preposition synsets in the CLaDA-BG Dict – the editing system for BTB-WN

used both between prepositions and between prepositions and other parts of speech (върху 'over' is derived from the noun връх 'top' and so does the preposition свръх 'above', so they are also linked with this relation). More relations applicable to prepositions are planned to be considered.

Currently 62 preposition lemmas are available in BTB-WN with 105 synsets. The most polysemous prepositions prove to be на (most frequently could be translated as 'on', 'of', 'in', etc.) with 12 synsets, followed by по ('over', 'in', 'on', etc.) with 11 synsets. The prepositions за ('for', 'to', 'about', etc.) and от ('from') are part of nine synsets each; с ('with') is in eight synsets and до ('to', 'until', etc.) and в ('in', 'at') are found in seven. As Table 1 shows the `locative` class has the most prepositions – 42, followed by `time` with 24 and `manner` and `quantity` with 11 prepositions.

5 Conclusion and Future Work

An attempt for preposition incorporation in the BTB-WN is presented. A semantic classification of prepositions is adapted on the base of Bulgarian grammars. The preposition synsets follow the structure and relations model of the nouns, verbs, adjectives and adverbs in wordnets and currently six semantic relations are introduced for prepositions: *synonymy*, *antonymy*, *hyperonymy*, *hyponymy*, *similarity* and *semantic derivation*. There are several directions in which this research would be elaborated: the hierarchy inheritance and categorization of the verbs and nouns in wordnet will be used and also features from a valency lexicon for Bulgarian – it will provide data about the types of prepositions which occur in the verbs' frames and about the semantic roles of their arguments. A classification based on semantic roles could be applied, given the good results that it provides for different languages.

Recent research (Amaro (2018), Da Costa and Bond (2016), etc.) show that closed class words have a place in wordnets and contribute for different NLP tasks if integrated. The good coverage of prepositions in BTB-WN will benefit its utility for semantic annotation and generation of pseudo corpora, which to be used for creation of neural language models in Bulgarian.

Acknowledgements

This work was supported by *the Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH – CLaDA-BG*, Grant number DO01-377/18.12.2020. and the Grant No. BG05M2OP001-1.001-0003, financed by *the Science and Education for Smart Growth Operational Program (2014-2020)* and co-financed by *the European Union through the European structural and Investment funds*.

References

Raquel Amaro. 2018. Integrating prepositions in wordnets: Relations, glosses and visual description. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Timothy Baldwin, Valia Kordoni, and Aline Villavicencio. 2009. Prepositions in applications: A survey and introduction to the special issue. *Computational Linguistics*, 35(2):119–149.

Todor Boyadzhiev, Ivan Kutsarov, and Yordan Penchev. 1998. *Contemporary Bulgarian language. Phonetics, lexicology, word formation, morphology, syntax.* Petar Beron, Sofia, Bulgaria.

Luis Morgado Da Costa and Francis Bond. 2016. Wow! what a useful extension! introducing non-referential concepts to Wordnet. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4323–4328, Portorož, Slovenia. European Language Resources Association (ELRA).

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Sanda M. Harabagiu. 1996. An application of wordnet to prepositional attachment. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics (ACL 1996)*, page 360–362, Santa Cruz, USA. Association for Computational Linguistics.

Violeta Konstantinova. 1982. *Prepositions in Bulgarian grammar literature*. Publishing house of BAS, Sofia, Bulgaria.

Ken Litkowski and Orin Hargraves. 2005. The preposition project. In *ACL-SIGSEM Work- shop on "The Linguistic Dimensions of Prepositions and Their Use in Computational Linguistic Formalisms and Applications"*, pages 171–179.

John P. McCrae, Ewa Rudnicka, and Francis Bond. 2020. English WordNet: A new open-source WordNet for English. *K Lexical News*, (28):37–44.

Tom O'Hara and Janyce Wiebe. 2009. Exploiting semantic role resources for preposition disambiguation. *Computational Linguistics*, 35:151–184.

Petya Osenova and Kiril Simov. 2018. The data-driven Bulgarian Wordnet: BTBWN. *Cognitive Studies – Études cognitives", 2018(18).*

Rashmi Prasad, Bonnie Webber, Alan Lee, and Aravind Joshi. 2019. The Penn Discourse Treebank 3.0 Annotation Manual.

Patrick Saint-Dizier. 2008. Syntactic and semantic frames in PrepNet. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*.

Nathan Schneider, Vivek Srikumar, Jena D. Hwang, and Martha Palmer. 2015. A hierarchy with, of, and for preposition supersenses. In *Proceedings of the 9th Linguistic Annotation Workshop*, pages 112–123, Denver, Colorado, USA. Association for Computational Linguistics.

Stoyan Stoyanov. 1983. *Grammar of contemporary Bulgarian standard language. Morphology.* Publishing house of BAS, Sofia, Bulgaria.