# Recent Developments within BulTreeBank

**Petya Osenova**
IICT-BAS,
Sofia, Bulgaria
petya@bultreebank.org

**Kiril Simov**
IICT-BAS,
Sofia, Bulgaria
kivs@bultreebank.org

## Abstract

The paper discusses recent developments in BulTreeBank (BTB). First of all, these developments include the preparatory steps for transferring richer linguistic knowledge from the original BTB into BTB-UD in order to for the enhanced dependencies to be added in the next release in May 2018. The new line of research also handles the extension of the BTB valency lexicon with subatom-based embeddings for English. The aim is to check automatically how good they are for detecting the core participants in an event. Since there are not enough resources for Bulgarian, we rely on transferring the embeddings trained on English data but enhanced with mappings to the Bulgarian WordNet and evaluated over BTB as gold standard.

## 1 Introduction

The original BulTreeBank (BTB) is an HPSG-based treebank including constituent annotation that reflects the HPSG hierarchy of phrases, annotation of the head constituent in each phrase, coreference annotation, named entities, ellipsis and discontinuous elements. Later on, the annotated sentences have been transferred into two different dependency formats: (1) CoNLL 2006 format where we used our own list of dependency relations and (2) Universal Dependency (UD) format where we focused rather on universal mappings of our data than on the language specific relations. As a follow-up, all newly annotated sentences adhere directly to the UD format. In addition to the mainly syntactic information, in the last few years we annotated the treebank with senses from the BulTreeBank Bulgarian WordNet (BTB-WN), aligned to Princeton WordNet (Osenova and Simov, 2017), and with valency frames (Osenova et al., 2012).

On the basis of the available rich linguistic information within the original HPSG-based treebank as well as the semantic annotation and valency frames information, new extensions were performed in two directions: (1) transferring linguistic information from the HPSG-based annotation to the UD format with the goal to facilitate the addition of the so-called *enhanced dependencies*; and (2) assigning sense embeddings to valency slots in the valency lexicon for supporting better feature representations that are learned from huge corpora. In this paper we discuss these two developments as well as the preliminary results from them.

The paper is structured as follows: next section presents related works. Section 3 describes the strategies behind the transfer of the linguistic information from the original treebank to the UD one. Section 4 focuses on the syntactic roles transfer from English to Bulgarian with the help of word embeddings. Section 5 concludes the paper.

## 2 Related Work

Many treebanks in Universal Dependencies (UD) initiative have been converted from already existing ones that were not necessarily dependency-based. This is also the case of BulTreeBank. Thus, initially the main focus was put on the mapping and proper transfer of parts-of-speech, grammatical and syntactic information from the existing annotation scheme into the UD one. As described in (Osenova and Simov,

*Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT16)*, pages 129–137,
Prague, Czech Republic, January 23–24, 2018. Distributed under a CC-BY 4.0 licence. 2017.

**Keywords:** enhanced dependencies, grammatical role embeddings, transfer between languages

2015) this transfer was performed by rules of two kinds: (1) lexical head identifier moving up the constituent tree; and (2) relation assignment for a constituent node of the dependent child when all children of the parent node have lexical identifiers. The example, given in that paper, was as follows: Let us have the following constituent, whose lexicalized example might be this one: tvarde visok zelen stol 'too tall green chair' [NPA [APA too tall] [NPA green chair]].

```
NPA -> APAid1 NPAid2
```

where id1 is a lexical head identifier for the adjectival phrase APA and id2 is a lexical head identifier for the noun phrase NPA. Then we establish the relation amod from NPAid2 to APAid1 and the identifier for the child NPA is moved up, because the lexical head of the child NPA is the lexical head for the whole phrase. After the application of these two rules we have the constituent tree annotated with lexical identifiers and dependency relations in this way:

```
NPAid2 -> APAid1 amod NPAid2.
```

However, it became clear that richer annotation in treebanks is needed to capture syntax-semantics-pragmatics interfaces. It should be noted that there already exist a number of semantically and discourse annotated treebanks (for example Prague Dependency Treebank annotated on discourse level — (Zikánová et al., 2015) and Italian Syntactic-Semantic Treebank (Montemagni et al., 2003), among others). However, they are not so many considering the multilinguality dimensions. At the same time, the NLP applications started to require the availability of richer cross-level linguistic knowledge.

Hence, the idea of the enhanced dependencies reflects exactly the linguistic multilevel interfaces (syntax, semantics, discourse). More precisely, it aims "to make implicit relations between content words more explicit by adding relations and augmenting relation names." (Schuster and Manning, 2016). They build on the basic dependencies and include the following phenomena:[1]

- *Null nodes for elided predicates.* This dependency involves the addition of special null nodes in clauses with an elided predicate. An example is: 'I go to Varna, and you [NULL NODE] to Sofia'. With this ellipsis recovery the grammatical relations are maintained also in the clause without an explicit predicate.

- *Propagation of conjuncts.* Apart from attaching the governor and dependents of a conjoined phrase to the first conjunct, dependencies are established between the other conjuncts and the governor, and dependents of the phrase. An example of conjoined subjects is: [The boy and the girl] are walking.

- *Additional subject relations for control and raising constructions.* In the enhanced dependency there is a relation between the embedded verb and the subject of the matrix clause. An example is: *She* intends to *go*. Between 'she' and 'go' there is a relation.

- *Arguments of passives (and other valency-changing constructions).* Here the enhanced dependency assignes a type (passive or agent) to the subject or a complement in a passive sentence. An example is: The vase was broken by the child, where 'vase' is a nominal subject of type *passive*, and 'child' is an oblique of type *agent*.

- *Coreference in relative clause constructions.* The enhances dependencies add a relation between the relative pronoun and its antecedent as well as between this antecedent and the predicate in the relative clause. An example is: The man who came ran away quickly. 'Who' refers to 'man'. Also. between 'man' and 'came'.

- *Modifier labels that contain the preposition or other case-marking information.* This means that some modifier relations, such as nominal and adverbial modification, etc., reflect also the preposition involved either as a case or the preposition itself. An example is: He put the book on the table, where the relation between 'book' and 'table' is oblique and copies also 'on' in the relation label.

[1] http://universaldependencies.org/v2/enhanced.html and http://universaldependencies.org/u/overview/enhanced-syntax.html

Since BTB has been originally annotated with information additional to the grammatical functions on the syntactic level, its annotation can be transferred also into the UD enhanced dependencies. It should be noted, however, that some of these relations have been annotated explicitly in the original treebank, while others stayed implicit, but they might be derived when necessary from the present annotations. Such a case are the arguments of passive predicates. Subjects and obliques are not explicitly marked as passive/agent, but in some cases this information can be derived automatically on the base of the predicate form. Needless to say, not all mappings are straightforward and trivial.

The assignment of sense embeddings to valency slots in the valency lexicon follows our previous work on grammatical role embeddings for English — (Simov et al., 2018). In this work we used two corpora: real text corpora (RTC) and pseudo corpus generated over WordNet (PCWN). The RTC was annotated with POS tags and parsed with Stanford CoreNLP pipeline — (Manning et al., 2014). Then on the basis of syntactic information we substituted the subject, direct object and indirect object lemmas with pseudo words representing the corresponding grammatical roles for the corresponding verb. Then we mixed the RTC with PCWN in order to train sense embeddings for the senses represented in the joint corpus in the same vector space. This allowed us to compare the embeddings for the grammatical roles with the embeddings for noun senses. This approach proved to be successful for English and we evaluated it via an extension of the Princeton WordNet with new syntagmatic relations between synsets which improved the results for Knowledge-based Word Sense Disambiguation — (Simov et al., 2018).

A similar application of the same approach to Bulgarian is justified by the fact that the BTB Bulgarian WordNet (BTB-WN) does not have good coverage on Bulgarian texts — (Osenova and Simov, 2017). Thus, we exploited the mapping from BTB-WN to Princeton English WordNet (PWN) — (Fellbaum, 1998) — in order to transfer the grammatical role embeddings trained for English to Bulgarian and to assign them to valency slots in the Bulgarian valency lexicon. We consider these tasks as part of a bigger task of transferring lexical semantic relations from the English WordNet to the Bulgarian one, but we will not report on this issue here. We performed the training of sense embeddings and grammatical role embedding in a similar way as for English, but first we extended the English WordNet with Bulgarian Synsets that lack the same meanings among the English Synsets. Then we generated a pseudo corpus using the UKB system[2] for knowledge-based word sense disambiguation. The sense embeddings were trained again over a joint corpus real texts and pseudo corpus.

Our work seems similar to the work of (Vulić et al., 2017). In their paper they consider three research questions: (**Q1**) Given their fundamental dependence on the distributional hypothesis, to what extent can unsupervised methods for inducing vector spaces facilitate the automatic induction of VerbNet-style verb classes across different languages? (**Q2**) Can one boost verb classification for lower-resource languages by exploiting general-purpose cross-lingual resources to construct better word vector spaces for these languages? (**Q3**) Based on the stipulated cross-linguistic validity of VerbNet-style classification, can one exploit rich sets of readily available annotations in one language (e.g., the full English VerbNet) to automatically bootstrap the creation of VerbNets for other languages? Our work differs from theirs in the fact that in our case the valency lexicon already existed before the experiment. In this respect more relevant to us are Q2 and Q3 with the modification that our goal is not to construct a VerbNet-like-lexicon for Bulgarian, but to perform a sense-embeddings-transfer from English to Bulgarian. in this way we use the larger availability of data in one language to address the contexts of sentence participants in a language with less data availability. However, in future it will be interesting to apply the approach, decsribed in (Vulić et al., 2017), to our resources in order to transfer additional knowledge from VerbNet to our valency lexicon.

## 3 The BulTreeBank Annotation Scheme with regard to the enhanced dependencies

BulTreeBank in its original format is HPSG-based and it consists of 15 000 sentences (or 214 000 tokens). More information on the annotation strategies can be found in the BulTreeBank Stylebook.[3]

[2] http://ixa2.si.ehu.es/ukb/

[3] http://bultreebank.org/wp-content/uploads/2017/04/BTB-TR05.pdf

### 3.1 Glances at the BTB annotation scheme

The original BTB annotation scheme is constituency-based with indication of head-dependency relations. In spite of the fact that XML was used as a main encoding format, additional graph-forming relations had been also assigned. These include several semantics and discourse-oriented (named entities, intrasentential coreferences, ellipsis, etc.) phenomena. To start with, subject and object control were annotated when one or both elements are pro-drop. This relation was introduced by a co-indexation mechanism that binds the overt subject and the pro-ss element (as in Fig. 1, left part) or two (more) pro-drop elements. For example, the instances of elided subjects (marked as *pro-ss*) always being part a co-referential chain within a sentence, are 6953.

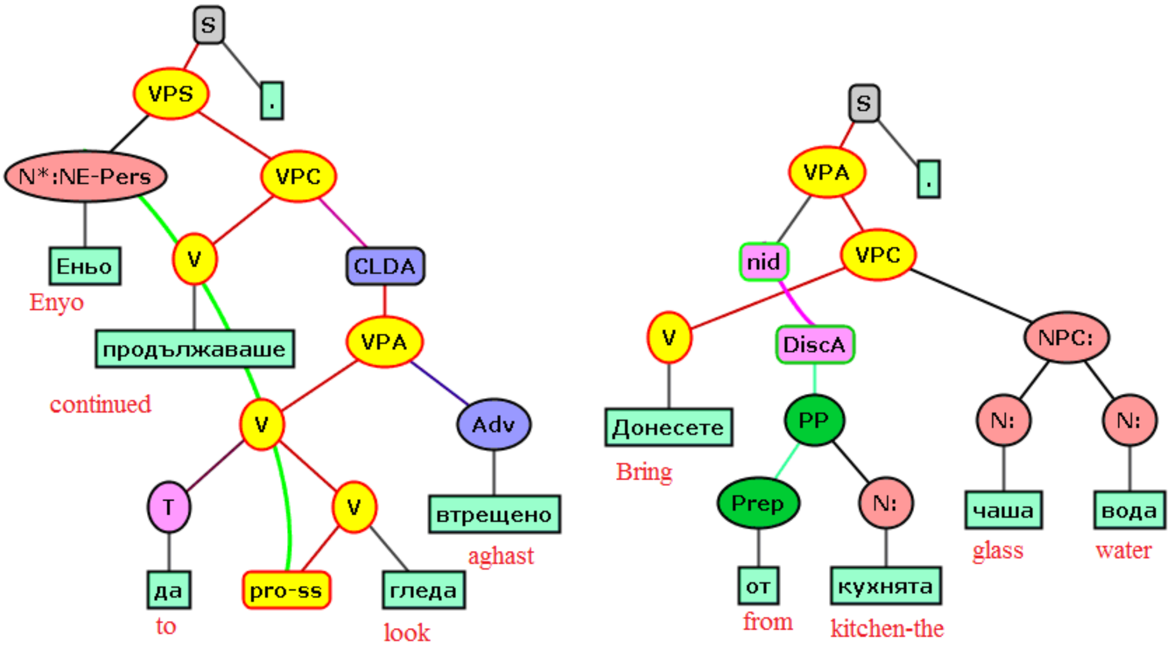

Figure 1: **Left:** Enyo [1] continued to [1] look aghast (Enyo kept looking aghast.) **Right:** Bring from kitchen-the glass water (Bring me from the kitchen a glass of water)

Another relation is the discontinuity one, introduced through three more specific relations: *DiscA*, *DiscM* and *DiscE*. The first one reflects scrambling. The second one — the so-called mixing arguments,[4] and the third one — topicalization. The most frequent type is scrambling *DiscA* (2447 instances), then comes topicalization *DiscE* (932 instances) and the rarest one, as expected, is *DiscM* (8 instances). See Fig. 1, right part, for an example of scrambling.

Further, ellipsis was added as well. It was marked on two levels: syntactic (V-Elip) and discourse (VD-Elip). The syntactic one has 262 instances in the treebank. It marks one verb form that is recoverable from the nearest context. It has subtypes only for marking equality of the missing element, its opposite or a different grammatical form. The discourse one has 255 instances. This type marks not only verb forms that are recoverable in a bigger context or even with the help of our common world knowledge. The subtypes represent existential verbs (to be, there is) with 120 instances, possessive verbs with 10 instances and a discourse element with 35 instances. It also marks whole VPs with a head and a complement. It can be seen that both types are almost equally represented. See Figure 2 for an example of syntactic ellipsis.

[4] By mixing arguments we mean a situation in which two constituents swap their elements. It can be found mainly in folklore and colloquial speech. For example: *Malki* **go** *momi beryaha* ‘Little it-ACC girls picking-were’ instead of *Malki momi* **go** *beryaha* ‘Little girls were picking it’. The accusative clicic comes between the adjective and the head noun in the NP, while belonging to the VP and thus causing an extraction-like process in this VP.

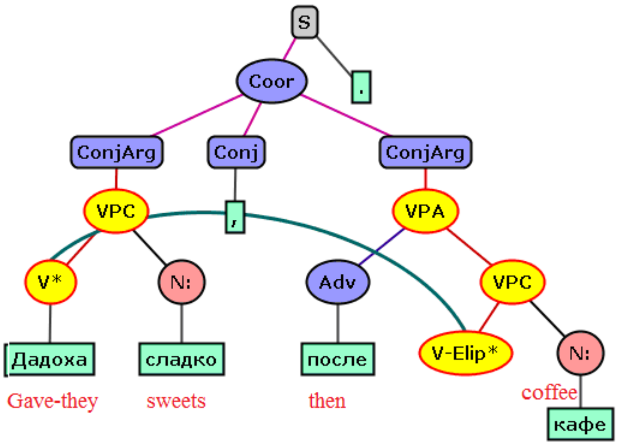

Figure 2: Gave-they sweets, then coffee (They gave sweets and then - coffee)

### 3.2 Towards enhanced dependencies

In this section the UD enhanced dependencies are considered with regard to the transfer procedure.

As it was mentioned above, there are two ways of transferring the information: implicit one in which the needed information is derived from the the linguistic annotation below the syntactic one, and explicit one in which the information can be directly processed through the represented syntactic nodes.

Similarly to our strategy for transferring basic graph information from the original BTB to the UD one, the same two types of rules are used, as described above: assigning a syntactic label to all nodes and then assigning appropriate relations among them.

***Null nodes for elided predicates***. In BTB such predicates are introduced as V-Elip or VD-Elip. Thus, both labels can be mapped directly into the so-called null nodes. V-Elip is the more straightforward one, while VD-Elip considers also cases of VP-ellipsis and copula ellipsis. While the latter is more systematic, the former varies in the length of the involved recovered material, such as different parts of a verb form or a head verb with a complement, etc. Apart from that VD-Elip provides discourse labels with the meaning that it is difficult to identify the type (let alone the form) of the missing element(s). These difficult cases can be processed only manually.

***Propagation of conjuncts***. Here we have to rely on the implicit but straightforward information, since in UD each dependant in a coordinated phrase has to be attached to its head (subject, object, modifier). In BTB the coordination phrases are considered head-less and thus - flat. However, the overall approach with respect to the treatment of conjuntcs is similar to the UD ideology. For example, two modifiers that modify the same head are coordinated as [**NP** [**CoordP** flat-the and lonely] voice]. Thus *amod* relations can be established on the base of the morphosyntactic and lexical information coming from the elements of the coordination phrase. The same holds for the core/non-core arguments. For example, the coordinated subjects can be assigned the *nsubj* relation per each subject with respect to the predicate.

***Additional subject relations for control and raising constructions***. In BTB the subject connects to the predicate in the main clause (i.e. the controller). Then the controller is connected with the unexpressed subject of the embedded verb. Thus, the *nsubj* relation between the subject of the main verb and the embedded verb can be established rather easily. Just the *pro-ss* element has to be substituted with *nsubj* and to be moved on the verb itself (see Fig. 1, left part, for the original tree).

***Arguments of passives (and other valency-changing constructions)***. In BTB there are no special markings of these arguments. Some of them can be derived automatically (such as the participle passive due to its special morphological form), and some of them are not trivial, such as the se-passives (being formed with the originally reflexive accusative clitic ‘se’ attached to the tensed verb form), since they are ambiguous across types of voice as well as markers of intransitive/detranzivised verbs. In the present UD version these labels are already available.

***Coreference in relative clause constructions***. The representation in BTB is similar to the representation in the basic UD graph where the relative is connected to the predicate with a grammatical relation. Thus, the *ref* relation with its antecedent can be established automatically.

***Modifier labels that contain the preposition or other case-marking information***. Since Bulgarian is analytic language, the non-core or nominal dependants (nmod, obl, acl and advcl) would have labels with propagated prepositions. This step can be done automatically.

Thus, it seems that the necessary information for the preliminary list of enhanced dependencies can be covered in BTB-UD almost in a straightforward way, since this kind of information has been already encoded in the original treebank. The main problems would go to some types of ellipsis and some non-typical coordinations.

In our case the only fully non-covered phenomenon were the arguments of passives, but they were (semi)automatically added where needed. Some difficulties in the transfer are expected in the following directions: (a) lack of enough instruction documentation in UD on some very complex examples and interrelated phenomena, such as ellipsis and coordination, (b) attempts to expand the treebank automatically and (c) some errors or problematic cases in the original treebank.

## 4 Syntactic Role Embeddings over the BTB Valency Lexicon

In our view word embeddings have to reflect the relational structure of the corresponding word. Thus, for a verb having a subject, a direct object and an indirect object we expect that its word embedding will be formed by four parts: embedding for the whole verb reflecting the semantics of the event denoted by the verb; then embeddings for the subject, direct and indirect objects. Such embeddings have to represent the selectional restrictions for the corresponding grammatical roles. There are many possible applications of such embeddings such as in the coreference resolution task where embeddings for the used pronouns are provided, also in word sense disambiguation, parsing, etc.

Here we report on the first experiments for learning such vector representations for the verb valency slots in Bulgarian valency lexicon, that correspond to subject, direct objects and indirect objects of verbs. We perform this through a knowledge transfer from English-to-Bulgarian with the help of the WordNet alignments. Our long-term goal is to train such embeddings for all lexical items with a relational structure including adjectives, adverbs, nouns (plus relational nouns), etc. We call such embeddings *subatom embeddings* because they contain features only on some aspects of a given event (or state).

The training of such embeddings for Bulgarian is not so easy because of the lack of sufficient language resources especially with respect to the coverage of BTB-WN. Thus we decided to exploit the available resources and their alignment to English in order to transfer these sense embeddings back to Bulgarian. Hence, we reused most of the work that has been already done for English — (Simov et al., 2018). In this work we learned subatom semantic embeddings on the basis of dependency-parsed corpora. We determined the arguments as wordforms in the text. As an example, let us consider the following sentence:

```
Every dog chases some white cat.
```

The generalization over the various word forms (or lemmas) in the different examples in the corpus has been performed by substituting the word forms for the corresponding argument with a pseudoword form. For example, for the above sentence the following variations have been generated with pseudoword forms for the different arguments of the different predicates:

```
Every SUBJ_chase chases some white cat.
Every dog chases some white DOBJ_chase.
```

Having learned embeddings for these pseudowords, we assume that they represent the selectional features for the corresponding grammatical roles of the verbs.

The corpus for training the embeddings reported in the paper consists of two parts: (1) real text corpora (RTC); and (2) pseudo corpus generated over WordNet (PCWN). RTC is used to represent relevant contexts for learning embeddings of pseudo words for subjects, direct objects and indirect objects. PCWN is used to ensure that the embeddings represent features extracted from the knowledge within the WordNet and also the coverage is extended to all synsets in WordNet.

As RTC we have used WaCkypedia_EN corpus — (Baroni et al., 2009). The WaCkypedia_EN corpus was reparsed with a more recent version of the Stanford CoreNLP dependency parser. The dependency of type "collapsed-cc" was selected, which collapses several dependency relations in order to obtain direct dependencies between content words, and in addition propagates dependencies involving conjuncts. For instance, a parse of the sentence "the dog runs and barks" would result in the relations nsubj(dog, runs) and nsubj(dog, barks). This type of dependency allows for a token to have multiple head words.

The head word of each nominal subject, as well as direct and indirect object, is then replaced by its predicate role and its governing verb's lemma (SUBJ_run, SUBJ_bark — both for the noun 'dog'). When a token has more than one head word suitable for substitution, copies of the sentence are created for each alternative replacement.

For the relation `has-subj` we use the dependency relations 'nsubj' and 'nsubjpass'; for the relation `has-dobj` we use the dependency relation 'dobj'; and for the relation `has-iobj` we use the dependency relation 'iobj'. In order to minimize some errors we enforced a condition that the dependency word should be a noun.

Here is a real example from RTC that was processed:

```
few high-quality SUBJ_address address long-term DOBJ_address
```

In the example both subject and direct object are substituted with pseudo words. All of the word forms are substituted with lemmas because our goal is getting sense embeddings.

The PCWN consists of pseudo texts that are the output from the Random Walk algorithm, when it is set to the mode of selecting sequences of nodes from a knowledge graph (KG) — see (Goikoetxea et al., 2015) for generation of pseudo corpora from a WordNet knowledge graph and (Ristoski and Paulheim, 2016) for generation of pseudo corpora from RDF knowledge graphs such as DBPedia, GeoNames, FreeBase. Here we report results only for knowledge graphs based on WordNet and its extensions. The pseudo corpus is generated using the UKB system[5] for knowledge-based word sense disambiguation.

Here is an example of a pseudo sentence from PCWN:

```
unfit function use undertake disposal
```

The pseudo sentences in PCWN represent sequences of related words on the basis of relations within WordNet. Such pseudo corpora provide good basis for learning lemma embeddings — see (Goikoetxea et al., 2015) and (Simov et al., 2017).

The union of both corpora is used in the experiments. As said before, in RTC all the words were substituted with their lemmas. Punctuation marks and numbers were deleted. We used the Word2Vec tool[6] in order to train the embeddings. From the various models we selected the one with the best score on the similarity task. This model was trained with the following settings: context window of 5 words; 7 iterations; negative examples set to 5; and frequency cut sampling set to 7. This approach worked for English and we used it to extend Princeton WordNet which improved the results for Knowledge-based Word Sense Disambiguation — (Simov et al., 2018). As it was described already, the resulting embeddings are related to lemmas, but not to senses, which is actually our goal. In order to have sense embeddings we performed some additional processing. Thus, for each synset, we obtained its vector by averaging the vectors for all lemmas it can be expressed by (this information is retrieved from WordNet). For grammatical roles, we averaged the corresponding grammatical role vectors per each lemma in the particular verb synset.

For the transfer from English to Bulgarian we extended the corpus of English senses with Bulgarian senses. In the BTB-WN an alignment to the Princeton WordNet has been maintained. We have supported three main relations of mapping Bulgarian-to-English synsets: *equality*, *subsumption*, and *generalization*. Here are some examples: vertolet = helicopter; chicho[7] is-subsumed-by uncle; mafia[8] generalized-over Cosa Nostra and Sicilian Mafia. A new PCWN was generated using this extended knowledge graph. The new corpus includes enough examples of Bulgarian synsets. The sense embeddings were trained over

[5] http://ixa2.si.ehu.es/ukb/

[6] https://code.google.com/archive/p/word2vec/

[7] Brother of the father.

[8] Organized crime group using the mechanisms of power.

the new PCWN and the RTC from English WikiPedia. Here we assume that the new trained embeddings represent well enough the Bulgarian senses.

The evaluation of the approach was done over the sense annotation of BulTreeBank. From it we extracted 285 instances of the subject–verb relation (subj(NounSynset, VerbSynset)), 207 instances of the direct object–verb relation (dobj(NounSynset, VerbSynset)), and 98 instances of the indirect object–verb relation where VerbSynset is presented in the training corpus and also there are embeddings for the related grammatical roles. Thus we were able to calculate the cosine similarity between the NounSynset embedding vector and the embedding vector for the corresponding grammatical role. If there is an instance of the relation subj(NounSynset, VerbSynset) we calculated the cosine similarity between the embedding for NounSynset and the embedding for SUBJ_VerbSynset. The results from this evaluation are presented in Table 1. The threshold for a good relation is set to 0.40. This value was selected by empirical evaluation on the impact of adding new syntagmatic relations to WordNet. The results showed that the embeddings selected the correct relations: in one third of the Subject–Verb cases, almost half of the Direct Object–Verb cases and one third of the Indirect Object–Verb cases.

| Grammatical Relation | Minimum | Maximum | Mean | Number over 0.40 |
|---|---|---|---|---|
| **Subject–Verb** | 0.2304 | 0.7463 | 0.3798 | 91 |
| **Direct Object–Verb** | 0.2387 | 0.5924 | 0.3947 | 96 |
| **Indirect Object–Verb** | 0.2199 | 0.5202 | 0.3698 | 28 |

Table 1: Evaluations for Grammatical Roles Embeddings.

Although the results are not very impressive we believe that they show the utility of aligning the slots of the frames in a valency lexicon with embeddings that generalize over the concrete words in real texts. In our future work we plan to extend the BTB-WordNet in order to create such embeddings directly from Bulgarian resources. The proposed evaluation approach needs to be made more precise with respect to the quality of the embeddings. We also plan to incorporate these embeddings in some mainstream NLP applications like parsing, coreference resolution and word sense disambiguation.

## 5 Conclusions

The paper presented two recent developments in BTB. The first one is the preparation work for transferring the knowledge from the original BTB in order to add enhanced dependencies into BTB-UD for the next release in May 2018. Our expectation is that the transfer will be done relatively smoothly, since the linguistic information in the original treebank covers the list of proposed UD enhanced dependencies.

The second one is the extension of the BTB valency lexicon with subatom-based embeddings for English with the aim to check automatically how good they are for detecting the core participants in an event. Due to the scarce Bulgarian resources for this task, we relied on transferring the embeddings trained on English data but enhanced with mappings to the Bulgarian WordNet. The evaluation was performed against BTB as gold standard. Our preliminary results showed the feasibility of the approach. There are many directions of future work, such as: better transfer from English to Bulgarian, exploiting of more Bulgarian resources, using approaches like retrofitting with respect to human created resources for tuning the initially assigned embeddings.

## Acknowledgements

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.

## References

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation* 43(3):209–226.

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

Josu Goikoetxea, Aitor Soroa, and Eneko Agirre. 2015. Random walks and neural network language models on knowledge bases. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Denver, Colorado, pages 1434–1439. http://www.aclweb.org/anthology/N15-1165.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*. pages 55–60. http://www.aclweb.org/anthology/P/P14/P14-5010.

Simonetta Montemagni, Francesco Barsotti, Marco Battista, Nicoletta Calzolari, Ornella Corazzari, Alessandro Lenci, Antonio Zampolli, Francesca Fanciulli, Maria Massetani, Remo Raffaelli, Roberto Basili, Maria Teresa Pazienza, Dario Saracino, Fabio Zanzotto, Nadia Mana, Fabio Pianesi, and Rodolfo Delmonte. 2003. *Building the Italian Syntactic-Semantic Treebank*, Springer Netherlands, Dordrecht, pages 189–210.

Petya Osenova and Kiril Simov. 2015. Universalizing bultreebank: a linguistic tale about glocalization. In *The 5th Workshop on Balto-Slavic Natural Language Processing*. INCOMA Ltd. Shoumen, BULGARIA, Hissar, Bulgaria, pages 81–89. http://www.aclweb.org/anthology/W15-5313.

Petya Osenova and Kiril Simov. 2017. Challenges behind the Data-driven Bulgarian Wordnet (Bultreebank Bulgarian Wordnet). In *John P. McCrae, Francis Bond, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Jorge Gracia, Ilan Kernerman, Elena Montiel Ponsoda, Noam Ordan and Maciej Piasecki (eds): Proceedings of the LDK 2017 Workshops*. pages 152–163.

Petya Osenova, Kiril Simov, Laska Laskova, and Stanislava Kancheva. 2012. A treebank-driven creation of an ontovalence verb lexicon for bulgarian. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet UÄ§ur DoÄ§an, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), Istanbul, Turkey.

Petar Ristoski and Heiko Paulheim. 2016. *RDF2Vec: RDF Graph Embeddings for Data Mining*, Springer International Publishing, Cham, pages 498–514.

Sebastian Schuster and Christopher D. Manning. 2016. Enhanced english universal dependencies: An improved representation for natural language understanding tasks. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), Paris, France.

Kiril Simov, Petya Osenova, and Alexander Popov. 2017. *Comparison of Word Embeddings from Different Knowledge Graphs*, Springer International Publishing, Cham, pages 213–221.

Kiril Simov, Alexander Popov, Iliana Simova, and Petya Osenova. 2018. Grammatical Role Embeddings for Enhancements of Relation Density in the Princeton WordNet. In *Proceedings of the 9th Global Wordnet Conference*.

Ivan Vulić, Nikola Mrkšić, and Anna Korhonen. 2017. Cross-lingual induction and transfer of verb classes based on word vector space specialisation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, pages 2546–2558. https://www.aclweb.org/anthology/D17-1270.

Šárka Zikánová, Eva Hajičová, Barbora Hladká, Pavlína Jínová, Jiří Mírovský, Anna Nedoluzhko, Lucie Poláková, Kateřina Rysová, Magdaléna Rysová, and Jan Václ. 2015. *Discourse and Coherence. From the Sentence Structure to Relations in Text*. ÚFAL, Praha, Czechia.