A Treebank-driven Creation of an OntoValence Verb Lexicon for Bulgarian

Petya Osenova, Kiril Simov, Laska Laskova, Stanislava Kancheva

Linguistic Modelling Department, IICT-BAS
Acad. G.Bonchev 25A, 1113 Sofia, Bulgaria
{petya, kivs, laska, stanislava}@bultreebank.org

Abstract

The paper presents a treebank-driven approach to the construction of a Bulgarian valence lexicon with ontological restrictions over the inner participants of the event. First, the underlying ideas behind the Bulgarian Ontology-based lexicon are outlined. Then, the extraction and manipulation of the valence frames is discussed with respect to the BulTreeBank annotation scheme and DOLCE ontology. Also, the most frequent types of syntactic frames are specified as well as the most frequent types of ontological restrictions over the verb arguments. The envisaged application of such a lexicon would be: in assigning ontological labels to syntactically parsed corpora, and expanding the lexicon and lexical information in the Bulgarian Resource Grammar.

Keywords: Ontology-based Lexicon, Valence Lexicon, Treebank

1. Introduction

There exist various strategies towards constructing valence dictionaries for a specific language. Some of these resources are being created together with the extension of a Treebank (Hinrichs and Telljohann, 2009) or comprise the most frequent verbs, evaluated on a large corpus (Zabokrtsky and Lopatkova, 2007). Also, valence dictionaries keep closer to the surface syntactic level (Hinrichs and Telljohann 2009) or elaborate on more semantically oriented representations (FrameNet, PropBank, PDT-VALLEX, etc.).

The Prague strategy for building valence lexicons has been adapted also to Arabic (Bielický and Smrž, 2008) and Croatian (Agic et al, 2010) or for parallel lexicons, such as the English-Czech valency lexicon attempt, reported in (Šindlerová and Bojar, 2009). (Agic et al, 2010) claim that 1923 verb valency frames for 594 different lemmas have been extracted. Although the data in Croatian is smaller, the ratio between the lemma and the frames is comparable to our data, excluding in both cases the valencies of the verb "to be". Another approach is taken for Danish – a combined representation of the valency information is presented, which collapses the LFG functions with HPSG categories (Amussen and Ørsnes, 2005). An approach similar to ours with respect to a Treebank-driven lexicon has been taken for Latin (McGillivray and Passarotti, 2009). They follow the notation of the Treebank itself, and thus the lexicon is more data-dependent. In our case, the extracted frames have been post-edited, if necessary, and also a new layer of ontological abstraction over the arguments has been added.

We aim at constructing a valency lexicon, which covers the verbs in the syntactically analyzed corpus of Bulgarian – BulTreeBank (www.bultreebank.org). Although the surface syntactic structure is adopted, the lexicon consists of ontological constraints. In this respect our strategy resembles the approach in (Hinrichs and Telljohann, 2009). However, it also adds semantic information, but rather based on a formal ontology than on thematic roles or topic frames. Thus, our mechanism is more in lines with the KYOTO Project (Vossen and Rigau, 2010) which also uses DOLCE as a top ontology and OntoWordNet as a mapping layer between the lexicon and the ontology.

This paper reports our observations on 3283 verb lemmas in BulTreeBank. The number of distinct valence frames for these lemmas is 6469. This means that the average is almost 3 valence frames per lemma.

2. Bulgarian Ontology-based Lexicon

The valence lexicon presented in this paper is a part the Bulgarian Ontology-based Lexicon (BOL) – (Simov and Osenova, 2010). BOL is constructed on the basis of *ontology-to-text* relation which interrelates an ontology, a lexicon, an annotation grammar and text. The lexicon exploits the relation between the ontology and the text. We assume that the ontology represents the conceptual meaning of the lexical units. The lexicon is mapped to the ontology in order the lexical units to be connected to their conceptual meaning. Additionally, the lexicon contains linguistic knowledge, such as phonological, morphological, and syntactic one. The mapping from the lexicon to the ontology is done by two relations *equality* and *subsumption*. The first relation is used when the concept (relation) in the ontology corresponds exactly to the conceptual meaning of the lexical unit. The second relation is used when the meaning of the lexical unit is a subconcept (subrelation) of the corresponding element in the ontology. The usage of two relations allows us to construct the ontology and the lexicon relatively independent of each other in the sense that the coverage of the lexicon can be extended more easily than the coverage of the ontology.

The current version of BOL is based on DOLCE ontology (Masolo et al, 2003) extended with concepts from OntoWordNet (Gangemi et al, 2003) - a version 1.6 of WordNet aligned to DOLCE. We selected DOLCE Ontology as upper ontology for several reasons: (1) it is constructed on rigorous basis which reflects the OntoClean methodology (Guarino and Welty, 2002); (2) it is represented in OWL-DL. We assume that the middle layer of OntoWordNet contains concepts that are better

understood by people. Thus the alignment between the two ontologies facilitates the understanding of the concepts in the upper ontology. The concepts from OntoWordNet are additionally restricted by synsets presented in EuroWordNet Base Concepts (BC) - (Vossen, 1999) and Core WordNet[1] (CWN contains 5000 synsets from WordNet on the basis of analysis of British National Corpus). We first used the intersection of both selections of WordNet synsets. It contains 1504 synsets. The corresponding concepts from OntoWordNet were extracted and their alignment to DOLCE was used as a first version of the ontology on which BOL has been constructed. After we completed the Bulgarian lexicon for the concepts from DOLCE and the intersection between CWN and BC, we proceeded with rest of the concepts from CWN. In addition, the lexicon was extended with lexical units extracted from the Bulgarian National Referent Corpus. These lexical units were ranked on the basis of their frequency in the corpus and the number of the documents in which they occurred. The last step ensures that BOL will reflect in a better way the Bulgarian-specific conceptualization. Thus, we assume that an ontology, which consists of concepts and relations from DOLCE and which is extended with concepts from CWN, is abstract enough to provide a language independent conceptual model for the construction of ontology-based lexicons. We might envisage that most of the lexical items in BOL would require more language dependent concepts. Keeping a larger language independent ontology might require more sophisticated mapping relations between the lexicon and the ontology.

In (Simov and Osenova, 2010) we suggested that the appropriate information for verb lexical units is represented in two ways: (1) in the ontology each verb is connected to an event concept related to the meaning of the verb. In the ontology all the participants (irrespectivly of whether they are considered to be arguments, adjuncts, etc.) are represented as such via appropriate relations; (2) the linguistic behavior is encoded in the lexicon as a set of frames. These frames determine the role of each participant in the given event. Unfortunately, there is no available ontology with such detailed information on events. Thus the presented work is in the direction of constructing a verb lexicon according to the above guidelines of a mapping between the lexicon and the ontology. Later on, it will be used for the construction of an event ontology.

To sum up, we consider ontologies and lexicons as artifacts reflecting the human abilities for representing, processing and managing linguistic and conceptual knowledge. As such, they are not complete and exhausting. This requires we to proceed in different ways depending on the available information. In the next sections we present how we use the information from the syntactic treebank of Bulgarian - BulTreeBank in order to construct a verbal lexicon in accordance with the model, outlined above.

[1]http://wordnet.princeton.edu/wordnet/download/standoff/

3. OntoValence Lexicon Extraction and Manipulation

The valence lexicon has been acquired from the syntactically annotated corpus BulTreeBank. BulTreeBank consists of 15 000 sentences (214 000 tokens). It includes predominantly newspaper texts, but also prose and administrative documents. All the verbs have been extracted together with the sentences they have been used in. Then they have been lemmatized and sorted by the lemma marker. The extracted valence frames present the predefined decisions of the annotators, who followed the BulTreeBank annotation scheme. Also, a default valence frame has been inserted, which presents a predicate with a SUBJ, DIROBJ and INDOBJ (NP predicate N PP). In the BulTreeBank annotation it is presented in the following way: (**VPS** NP (**VPC** V NP PP))), where VPS stands for a *head-subject phrase* and VPC stands for a *head-complement phrase*[2]. Thus, the human expert received a list of lemmas, accompanied by the respective sentence, which contains the annotated frame, as well as the inserted default frame. The reasons for choosing such an approach are as follows:

1. The pre-annotated frames in BulTreeBank might differ syntactically from our present postulations of constructing valence frames due to an error or different view;
2. The pure copying of the annotated frame, which might be considered a trivial step, has been abandoned, since our aim is to add also ontological constraints.

Thus, the default frame is viewed as a medium level to the ontological abstraction.

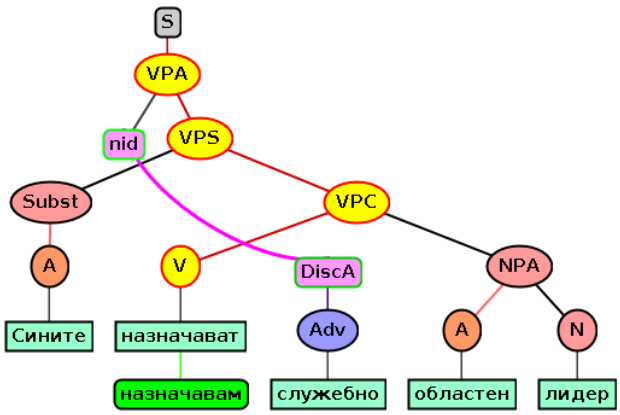

Fig. 1: Original representation of a sentence tree in BulTreeBank. The green rounded rectangle depicts the lemma node of the verbal head of the sentence.

In Fig. 1 one of the usages of the verb *назначавам* 'appoint - imperfective' is presented in the BulTreeBank visualization format. The gloss of the sentence is: *Blue-the appoint officially area leader*. The translation is: *The blue team ex officio appoints an area leader*. In Fig. 2 the

[2] Note that the canonical order is *Subject – Verb – Direct Object – Indirect Object*.

inserted default frame is presented, which says: [SOMEBODY appoints SOMEONE for SOMETHING]. Then the expert can modify the frame according to the adopted principles (more about the principles, see next section). After modifying the frame (in this case – deleting the PP), the expert adds the semantic constraints over the arguments in accordance with the DOLCE ontology. Such a modified and ontologically constrained frame for the sentence, presented in Fig. 1, is showed in Fig. 3. There the syntactic labels within the frame are adapted to the scheme: SUBJ predicate DIROBJ, but the lexical values of the arguments are replaced by ontological concepts. Thus, the ontologically constrained scheme is: ORGANIZATION [appoint - lemma] PERSON.

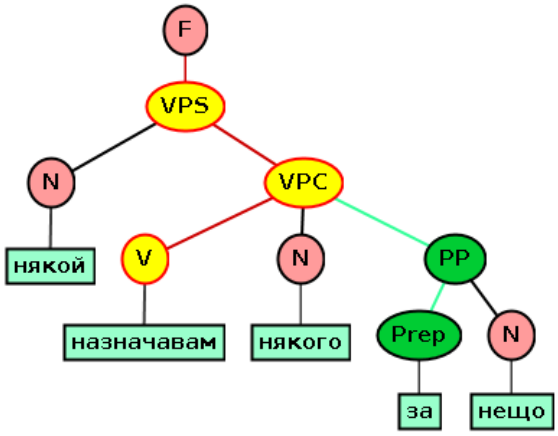

Fig. 2. The default frame included for each instance of a verb. In this way some manual work is saved.

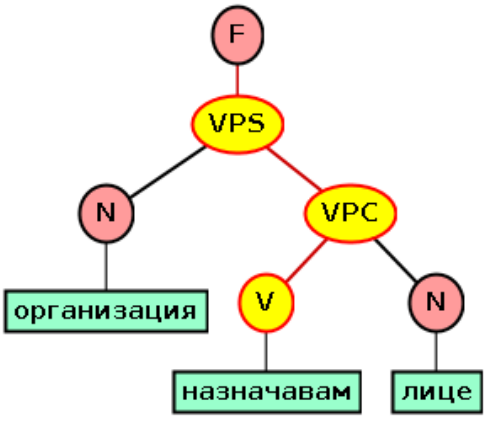

Fig. 3: The resulting frame

The extracted annotated frames from BulTreeBank are 18081. However, since our aim is to cover as many syntactic and ontological variations per verb meaning as possible, some additional example material has been extracted also from the Bulgarian National Reference Corpus. For the moment this step was executed only when the examples per verb meaning are less than 5. Its importance is justified by the fact that 920 verb lemmas out of the processed 2180 have occurred in BulTreeBank only once; 313 lemmas have occurred 2 times; 200 lemmas – 3 times; 115 – 4 times and 94 – 5 times. Thus, such cases constitute approximately 75 %, although the most frequent verbs are well covered by the other 25 % (for example, the verb *гледам* 'look at' has 93 frames, extracted from BulTreeBank). We are aware of the fact, that the real picture is much more complex, and for that reason we envisage further investigations and elaborations over the lexicon in the future.

4. OntoValence Lexicon Architecture and Principles

Our approach to the valence lexicon follows the syntactic annotation scheme in BulTreeBank. Here is the table with the labels, which represent the main grammatical relations between the predicate and the participant, and also between other heads and their dependants:

Label	Description
VPA	head (verb)-adjunct
VPC	head(verb)-complement
VPS	head(verb)-subject
NPA	head(noun)-adjunct
NPC	head(noun)-complement
PP	head(preposition)-complement
PPA	head(preposition)-adjunct
APC	head(adjective)-complement
APA	head(adjective)-adjunct
AdvC	head(adverb)-complement
AdvA	head(adverb)-adjunct

Table 1: Description of the syntactic labels in BulTreeBank

Apart from this, the syntactic architecture is more similar to (Hinrichs and Telljohann, 2009) than to (Zabokrtsky and Lopatkova, 2007). This means that we encode the surface behavior of the verb occurrence and thus the valence frame is kept to the surface syntax. In such a framework the pro-drops of any kinds are also presented within the frames. The frame considers the clausal complements as well. We encode the verb usage in active voice, while marking the passive occurrence in the lemmatized wordform within the sentence. Contrary to (Zabokrtsky and Lopatkova, 2007), the verbs in perfective and imperfective aspect are considered separate lemmas. As a rule, the frame includes only the inner participants (semantically obligatory for the event or situation, presented by the predicate, but might be unexpressed on the surface level). Thus, the lexical meaning of the verb is crucial for the frame creation. Our model follows the ideas in (Pustejovsky, 1998: 63), i.e. we include the TRUE ARGUMENTS (syntactically realized), DEFAULT ARGUMENTS (part of the semantics, but not necessarily expressed) and SHADOW ARGUMENTS (semantically incorporated into the lexical item), but not TRUE ADJUNCTS (part of situational interpretation

rather than of lexical one). Since the distinction between complements and adjuncts is still problematic in the theories of argument realization, for some unclear cases an extended frame has been proposed. One problem is whether some non-dative PPs, which seem to be synonymic to the dative PP, are arguments, or not: *говоря на Иван* (speak-I to Ivan 'I speak to Ivan') and *говоря пред публика* (speak-I in front of audience 'I speak to some audience'). Thus, such extended frames have been added for the verb *гледам* 'look at', among other verbs.

As is was already mentioned, our lexicon uses the DOLCE ontology for constraining the inner participants. This process is manual. The annotators have been instructed to use the most abstract notion that keeps the *differentia specifica* of the specific lexical meaning. In cases when there is not such a notion, it has been added. Thus, also the middle and domain layers of the top ontology are extended. Our work is also facilitated by the fact that parts of DOLCE have been integrated into WordNet via OntoWordNet.

5. Some Observations over the current state of the OntoValence Lexicon

Let us give some more specific information on the valence frames of the processed lemmas. The most frequent syntactic frames with the number of their occurrences are presented in Table 2:

N	Syntactic Frame Type	Number of Frame Occurrences
1.	Predicate – Direct Object (NP)	4089
2.	Subject (NP) – Predicate – Direct Object (NP)	3122
3.	Subject (NP) – Predicate	1339
4.	Subject (NP) – Predicate – Indirect Object (PP)	1243
5.	Predicate	1082
6.	Predicate – Direct Object (NP) – Indirect Object (PP)	1013
7.	Predicate – Indirect Object (PP)	888
8.	Predicate – Complement (CLDA)	807
9.	Subject (NP) - Predicate – Direct Object (NP) – Indirect Object (PP)	695
10.	Subject (NP) - Predicate – Complement (CLDA)	643

Table 2: Frequency of syntactic Frames

The most frequent frame is the one with a pro-drop SUBJ and with an explicit Direct Object (NP). This fact proves that the canonical Bulgarian utterance is a subject-null one. The next most frequent frame is the one with a realized SUBJ (NP) and a realized Direct Object (NP). The intransitive frame takes the third position. It is worth mentioning that also intransitive pro-drop frames seem to occur often. In the frequency table they take the 5th position. The most frequent type of clause that takes a complement position is the so called '*da*-clause', which in general is considered an analytical counterpart of the Old Bulgarian infinitive. The subject-null frame outnumbers the subject explicit one.

N	Syntactic Frame	Ontological Label
1.	Predicate	No Ontological Restrictions
2.	Predicate – Complement (CLDA)	EVENT
3.	Subject (NP) – Predicate	PERSON
4.	Predicate – Direct Object (NP)	PERSON
5.	Subject (NP) - Predicate – Complement (CLDA)	PERSON - EVENT
6.	Predicate – Direct Object (NP)	OBJECT
7.	Subject (NP) - Predicate – Direct Object (NP) – Indirect Object (PP)	PERSON – ARTEFACT – (for) OBJECT
8.	Subject (NP) – Predicate – Direct Object (NP)	PERSON - PERSON
9.	Predicate – Direct Object (NP)	SOCIAL OBJECT
10.	Subject (NP) – Predicate – Direct Object (NP)	PERSON - OBJECT

Table 3: Frequency of ontological constraints

As it can be seen in Table 3 above, from an ontological point of view, the most frequent type is the type without any restrictions. It is followed by a clausal complement, which is constrained as an EVENT. Only then the intransitive type with an explicit subject comes, with the restriction PERSON. The same restriction holds for the transitive subject-null frame, which is at the 4th place. Thus, the following restrictions over participants come in a frequency order:

EVENT > PERSON > OBJECT > ARTEFACT > SOCIAL OBJECT

It would be interesting to make also observations on the frequency of participants within the sub-events within the EVENT.

The 10 most frequent verbs with respect to the occurrences in texts are as follows: the modal verb *мога* 'can, be able to'; *имам/има* 'have/there is'; *нямам* 'do not have'; *кажа* 'say'; *искам* 'want'; *направя* 'do-perfective'; *правя* 'do-imperfecive'; *знам* 'know'; *заявя* 'declare, announce-perfective'; *взема* 'take-perfective'; *видя* 'see-perfective'. Excluding the most frequent verb *съм* 'to be', the next to come is another representative of the stopword list, namely the modal verb *мога* 'can'. Closely to it are the verbs of possession and existence, and the modal verb *искам* 'want'. The following verbs show some domain dependency in a corpus with predominantly newspaper texts. This holds especially for the verbs *say* and *announce*.

6. Conclusions

The OntoValency lexicon has been fully processed with respect to its coverage – both on syntactic and ontological layers. However, more efforts are needed for testing the correct level of abstraction for the ontological labeling, and also the handling of the metaphorical usages of a specific verb meaning.

The main problems in inter-annotator agreement seem to be: the granularity of concept mappings as well as the metaphorical and idiomatic usages. Concerning the granularity, since the annotation has been example-based, some annotators preferred to keep the most specific label, while others provided some more abstract concept.

The metaphorical and idiomatic usages raised the question of the adequate concept mapping. In both cases a mapping is required also of whole phrases to concepts.

Presently, the OntoValency lexicon is envisaged to be used for at least two tasks:

1. Annotation of syntactically parsed texts with ontological constraints, and
2. Expanding the lexicon of the HPSG Bulgarian Resource Grammar with valence information and later on – incorporating the ontological mappings.

The completion of the first task would support the various applications of knowledge extraction, while the expanded lexicon would improve massively the performance of the HPSG grammar of Bulgarian.

A validation of the selected ontological restrictions over the participants will be performed in two steps:

1. Mapping of the lexical units in BulTreeBank that denote the participants in the concrete sentences to the ontology of BOL;
2. Annotation of new sentences extracted from the corpus via a clustering on the dependency analyses of the sentences.

The first step will provide evidences whether the conceptual restrictions on the event participants are compatible with the concepts mapped to the lexical units in BOL. The second step will provide evidences for completeness of the frameset for each verb.

7. Acknowledgements

The work reported in this paper has been supported by the EU project EuroMatrixPlus.

References

Željko Agic, Krešimir Šojat, Marko Tadic. 2010. *An Experiment in Verb Valency Frame Extraction from Croatian Dependency Treebank*. In: Proceedings of the 32nd International Conference on Information Technology Interfaces, Zagreb, SRCE University Computer Centre, University of Zagreb, 2010, pp. 55-60.

Jørg Asmussen and Bjarne Ørsnes. 2005. *Valency information for dictionaries and NLP lexicons: Adapting valency frames from The Danish Dictionary to an LFG lexicon*. In: Ferenc Kiefer and Julia Pajzs (eds.): Papers in Computational Lexicography. Proceedings of the 8th Conference on Computational Lexicography, COMPLEX 2005. Budapest: Research Institute for Linguistics. Hungarian Academy of Sciences, 2005, pp. 28–39.

Viktor Bielický and Otakar Smrž. 2008. *Building the Valency Lexicon of Arabic Verbs*. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation. ELRA

Aldo Gangemi, Roberto Navigli, and Paola Velardi. (2003). The *OntoWordNet Project: extension and axiomatization of conceptual relations in WordNet*. Meersman R, et al. (eds.), Proceedings of ODBASE03 Conference, Springer, 2003.

Nicola Guarino and Welty, C. (2002). *Evaluating Ontological Decisions with OntoClean*. Communications of the ACM, 45(2): 61-65.

Erhard Hinrichs and Heike Telljohann. 2009. *Constructing a Valence Lexicon for a Treebank of German*. In: Frank Van Eynde, Anette Frank, Koenraad De Smedt and Gertjan van Noord (eds.) Proceedings of the 7th International Workshop on Treebanks and Linguistic Theories (TLT 7), LOT, Groningen, pp. 41-52.

Claudio Masolo, Stefano Borgo,. Aldo Gangemi, Nicola Guarino, Alessandro Oltramari. 2002. *Ontology Library* (final). WonderWeb Deliverable D18, December 2003. http://www.loa-cnr.it/Publications.html.

Barbara McGillivray and Marco Passarotti. 2009. *The Development of the Index Thomisticus Treebank Valency Lexicon*. In: Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education –LaTeCH – SHELT&R 2009, pages 43–50, Athens, Greece, 30 March 2009.

James Pustejovsky. 1998. *The Generative Lexicon*. The MIT Press.

Kiril Simov and Petya Osenova. 2010. *Constructing of an Ontology-based Lexicon for Bulgarian*. In Proceedings of LREC 2010.

Jana Šindlerová and Ondrej Bojar. 2009. *Towards English-Czech Parallel Valency Lexicon via Treebank Examples*. In: Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories, 4-5 December 2009, Milan, Italy, pp. 185-197.

Piek Vossen and G. Rigau. 2010. *Division of semantic labour in the Global WorNet Grid*. In: Proceedings of the 5th Global WordNet Conference, Mumbai, India.

Piek Vossen 1999. Editor. *EuroWordNet General Document*. Version 3, Final, July 19, 1999. http://www.hum.uva.nl/~ewn

Zdenek Zabokrtsky and Marketa Lopatkova. 2007. *Valency Information in VALLEX 2.0: Logical Structure of the Lexicon*. The Prague Bulletin of Mathematical Linguistics No. 87, pp. 41-60, 2007.