Petya Osenova
Sofia University "St. Kliment Ohridski"

MODELING VALENCE FRAMES IN BULGARIAN: CORPUS VS. GRAMMAR APPROACH

Abstract: The paper focuses on the construction of valence lexicons for Bulgarian. It presents two approaches: a corpus-based one, which uses a syntactically annotated corpus for Bulgarian, and a grammar-based one, in which the resource is created together with the grammar language rules. BulTreeBank was used as a syntactically annotated corpus. The HPSG-oriented Bulgarian Resource Grammar was used for the grammar-based valency lexicon. Also, the interaction between the two approaches is discussed.

Keywords: valency frames, Bulgarian, resource grammar, treebank

1. Introduction

Valence lexicons have been created now for many languages. These lexicons are often based on existing treebanks – resources, which present syntactically analyzed sentences with respect to a specific linguistic theory. Some valence lexicons have been compiled while extending a treebank (Hinrichs and Telljohann 2009). Thus, a big coverage over verb behavior is ensured. Other resources rely on the most frequent verbs, evaluated on a large corpus (Žabokrtský and Lopatková 2007). Also, valence dictionaries might either reflect the surface syntactic level (Hinrichs and Telljohann 2009), or build also semantically oriented representations (such as lexicons in FrameNet[1] style, PropBank[2] style or Czech lexicon – PDT-VALLEX[3], etc.).

Very often the newly created resources follow best practices. For example, the Prague strategy for building valence lexicons has been applied also to Arabic (Bielický and Smrž 2008) and Croatian (Agić et al. 2010) as well as for parallel lexicons, such as the English-Czech valency lexicon attempt. (Agić 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 lemmas 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 specificities of two constraint-based theories, namely LFG[4] functions with HPSG[5] categories (Asmussen and Ørsnes 2005). An approach similar to ours with respect to a Treebank-driven lexicon has

[1] https://framenet.icsi.berkeley.edu/fndrupal/
[2] http://verbs.colorado.edu/~mpalmer/projects/ace.html
[3] https://ufal.mff.cuni.cz/PDT-Vallex/
[4] Lexical Functional Grammar
[5] Head-driven Phrase Structure Grammar

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.

On the other hand, grammar engineering efforts, which have become very popular in recent years, also provide various valence lexicons for a number of languages together with the language grammars. Such lexicons have been made within the DELPH-IN[6] community, which follows the HPSG-based formalism; GF[7] community, which follows GF formalism, among others. The author has developed a version of Bulgarian grammar[8] in the lines of DELPH-IN grammars (Bender et. al 2010). They follow a core grammar, called Matrix grammar, while developing further their own language-specific versions. Also, the grammar components (types, hierarchy, lexicon, rules) are being developed with respect to the dataset they cover. Sometimes there is as minimum one common dataset to be covered by all grammars thus reflecting the similar and differing phenomena from a typological perspective.

The paper aims at highlighting the peculiarities of valence modeling from a corpus and grammar engineering perspective.

The paper is structured as follows: next section presents BulTreeBank-based valence frames. Section 3 reports on the valence lexicon, developed within Bulgarian Resource Grammar (BURGER). Section 4 discusses the similarities and differences in both approaches. Section 5 concludes the paper.

2. BulTreeBank Valence Frames

The BulTreeBank-based valence lexicon (Osenova et. al 2012) covers the verbs in an existing syntactically analyzed corpus of Bulgarian – BulTreeBank (www.bultreebank.org). The corpus contains 214 000 tokens and around 15 000 sentences. The underlying principles are: 1. keeping the surface syntactic structure, and 2. adding ontological constraints. The first criterion means that the frames are projected directly from corpus occurrences. The second one means that we aim at mapping the grammatical roles into semantic constraints, based on ontology. The second criterion will not be considered in the paper. It is mentioned just for completeness. In **Table 1** the syntactic labels and dependences are presented as used in BulTreeBank annotations. Some specificities come in order:

– the valence frame is kept to the surface syntax
– the pro-drops of any kinds are also presented within the frames
– the frame considers the clausal complements as well
– the verb usage is encoded in active voice

[6] http://www.delph-in.net/wiki/index.php/Home
[7] Grammatical Framework
[8] http://www.bultreebank.org/BURGER/index.html

– the verbs in perfective and imperfective aspect are considered separate lemmas.

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). **Table 1** presents the dependences within BulTreeBank:

Table 1. Description of the syntactic labels in BulTreeBank

LABEL	DESCRIPTION
VPA	head[verb] - adjunct
VPC	head[verb] - complement
VPS	head[verb] - subject
NPA	head[noun] - adjunct
NPC	head[noun] - complement
PP	head[prep] - complement
PPA	head[prep] - adjunct
APC	head[adj] - complement
APA	head[adj] - adjunct
AdvC	head[adv] - complement
AdvA	head[adv] - adjunct

The extracted annotated frames from BulTreeBank are 18 081. After normalization, the verb lemmas in BulTreeBank turned out to be 3283, out of which 2180 have been processed. The number of distinct valence frames for these lemmas is 6469. This means that on average there are approximately 3 valence frames per lemma. Interestingly, 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 "rare occurrence" 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).

The process of valence lexicon compilation had the following steps:

– all the verbs have been extracted together with the sentences they have been used in

– then they have been lemmatized and sorted by their lemmas

– a default valence frame has been inserted, which presents a predicate with a SUBJ, DIROBJ and INDOBJ.

Let us give some illustration. 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*. Thus, the default frame: SUBJ VERB COMPLEMENT [direct object] COMPLEMENT [indirect object] is kept almost without any modifications (deletions, additions, etc.).

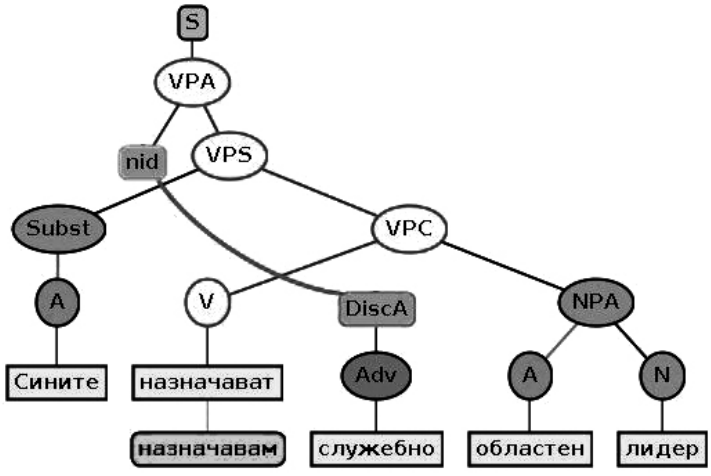

Fig. 1. Original representation of a sentence tree in Bultreebank. The rounded rectangle depicts the lemma node of the verbal head of the sentence

Let us explain the above example in more detail. The default frame has the following structure: (**VPS** N (**VPC** V N PP)), where the generalized model is: (**VPS** SOMEBODY (**VPC** VERB SOMEBODY ABOUT SOMETHING)). In the sentence in Fig. 1 SOMEBODY within VPS is substituted by a substantivized adjective (Subst); VERB is substituted by 'appoint' (wordform and lemma); SOMEBODY within VPC is substituted by 'area leader'. The PP does not match anything in the sentence. The adverb 'officially' does not belong to the valence frame.

3. Grammar-based Valence Frames

BURGER is the Bulgarian Resource Grammar (Osenova 2011), developed in the lines of HPSG theory within DELPH-IN community.

The multilingual testset (originally in English) comprises 100 sentences, each of which exemplifying a different linguistic phenomenon. When translated into Bulgarian, the number of sentences went up to 178 sentences, since some more translational equivalents have been provided per sentence. Then, some additional sentences have been added for illustration of language specific phenomena, such as clitic doubling, double negation, pro-dropness, etc. Thus, the whole set became 193 positive sentences. Since it is a parsing grammar, 20 negative sentences were incorporated for test purposes. The covered cross-language phenomena are: complementation, modification, coordination, agreement, control, quantification, negation, illocutionary force, passivization, nominalization, relative clauses, light verb constructions, etc.

All valence types[9] in the current version of the grammar are 268 (including optional subject, impersonal verbs). All complement valence types are 41.

The valence types have been developed while extending the grammar to cover a predefined dataset with sentences, each of which reflecting some language phenomenon.

In **Tables 2**[10] and **3** the complement-taking frames are considered. The frequencies in the right column are estimated over the treebank, since the related dataset is too small. Also, the valence types in **Table 2** present the frame frequencies with *no overt subjects*, while the valence types in **Table 3** present the frame frequencies *with overt subjects*.

Table 2. The syntactic variety within head-complement dependencies without explicit subject

V_NP	**1307**
V_-	**874**
V_PP	**661**
V_NP-PP	**546**
V_che	**158**
V_da	**143**
V_ADVP	**82**
V_ques	**50**
V_PP-PP	**48**

When comparing the frames *with* and *without* overt subject, some conclusions can be drawn. Pro-drop phenomenon is very strong in Bulgarian valence frames.

It can be noticed that in both tables the most frequent type is verb with nominal complement (V NP) irrespectively of whether the subject is explicit or not. The next valence types differ in their frequencies when both tables are compared. For example, the intransitive verb frame with overt subject (NP V) is less frequent with respect to the other frames than the same frame without overt subject (V). Within the group of clausal complements, in **Table 2** the most frequent one is the *che* clause (=that), while in **Table 3** it is the *da* clause (=to).

[9] Please note that 'valence type' has the same meaning as 'valence frame'. In the grammar-based valence lexicon, the notation 'valence type' is preferred.

[10] The notations present the type of the complement after the underscore. Thus, the defice means intransitive usage. Some of the mentioned complement types are language specific, such as: *che-type* and *da-type*, which correspond roughly to English sentences, introduced with *that* and *to*, respectively.

Table 3. The syntactic variety within head-complement dependencies with explicit subject

V_NP	**52**
V_PP	**40**
V_NP-PP	**29**
V_da	**18**
V_-	**14**
V_che	**12**
V_ADVP	**11**
V_PP-PP	**9**
V_cl	**8**
V_PP-da	**6**

4. Discussion

The differences of the grammar-enhanced lexicon to the treebank-based approach are: bottom-up approach (vs. top-down one); phenomenon-based coverage (vs. text-based one); deep approach (vs. surface one). The first specificity means that each example is modeled separately, and then with the compilation of examples, some general view on verb valence behavior can be generated. In the treebank it is the opposite – all the frames are extracted and then – checked and processed. The second specificity means that the valence depends on the phenomenon to be covered next by the grammar. This might cause either too little or too much variety in the frames. Contrary to this approach, the treebank-based one processes texts, and thus respects the aggregate nature of connected texts. The third specificity means that the valence frames incorporate some generalizations beyond the surface presentations. For example, they respect the optionality of the arguments and in this way one frame would cover more surface syntactic variations.

Within the differences it should be stressed that the grammar-based approach is semantically oriented, while the treebank-based one relies on the syntactic and syntactic-semantic relations only. Thus, in the treebank there is only one *head-complement* type with the copula being the head, while in the grammar there are 4 types. The first distinction is made between the referential index possessed by the nouns as complements, and all other parts-of-speech (adjectives, adverbs, PPs) as being events. The next distinction considers whether there is agreement between the subject and the complement (adjectives), or not (adverbs, PPs).

In both valence approaches the clitics are viewed as lexical projections of the head, while the regular forms are treated as head arguments (complements). However, in the grammar lexicon the semantics is taken into account. The clitic does not contribute its separate semantics, because it is not a full-fledged complement.

Instead, the verb incorporates clitic's contribution in its own semantics. Thus, the personal pronoun clitic lexemes have an empty relation list, while the regular pronoun forms have a pronoun relation.

We started a process of incorporating the treebank-based valence lexicon into the grammar-based one, because the former has a better coverage than the latter. Since the representations and incorporated knowledge have some differences, the following steps are being preformed:

– the verbs have been sorted by frames

– the frames have been automatically transformed into partial syntactic types (v_-; v_pp; v_np…..)

– the information about the value of the aspect has been derived from the morphological dictionary for each verb lemma

– types have been tuned with respect to granularity of valence frame specification

– missing types in the grammar have been detected.

This approach is consistent. However, it is fast only on coarse level of granularity when mapping the frames. For the detailed mapping, the process is accompanied by manual checks and corrections. The types, missing in the grammar, are mainly of the following matter: a) verbs with formally reflexive particles 'se' and 'si'. In the grammar they are considered complements and thus need to be part of the valence frame. This is not the case in the treebank model; b) addition of a perfective or an imperfective verb type to an existing valence frame; c) types with specific prepositions. In the grammar both - prepositions which form complements to verbs and modifying prepositions - have separate entries. Thus, valence types with specific prepositions (dative, locative, predicative, etc.) proliferate.

5. Conclusions

The paper presented two different, but complementing each other approaches to digital valence dictionary compilation – a corpus-based one and a grammar-development-based one. The first one is with high coverage, but oriented to surface syntax, although with a possibility to be generalized on syntax-semantic level. The second one is more semantically oriented and elaborated, but with low verb coverage in the grammar lexicon. Thus, the mapping from the treebank-derived one to the grammar-compiled one is not trivial. On the coarse level it is quick and straightforward. On the deep granularity level, however, it needs human intervention and reconsideration of the frames.

ACKNOWLEDGEMENTS

The work, presented in this paper, was partially supported by the FP7 Capacity Programme AComIn: Advanced Computing for Innovation, hosted at IICT-BAS. Grant Agreement: 316087.

REFERENCES

Agić et al. 2010: Agić, Ž., Šojat, K., Tadić, M. 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, p. 55 – 60.

Asmussen and Ørsnes 2005: Asmussen, J. and Ørsnes, B. Valency information for dictionaries and NLP lexicons: Adapting valency frames from The Danish Dictionary to an LFG lexicon. – In: F. Kiefer and J. 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, p. 28 – 39.

Bender et al. 2010: Bender, E. M., Drellishak, S., Fokkens, A., Poulson, L., Saleem, S. Grammar Customization. – In: *Research on Language and Computation*, 2010, vol. 8(1), p. 23 – 72.

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

Hinrichs and Telljohann 2009: Hinrichs, E. and Telljohann, H. Constructing a Valence Lexicon for a Treebank of German. – In: F. Van Eynde, A. Frank, K. De Smedt and G. van Noord (eds.). *Proceedings of the 7th International Workshop on Treebanks and Linguistic Theories (TLT 7), LOT*. Groningen, 2009, p. 41 – 52.

McGillivray and Passarotti 2009: McGillivray, B. and Passarotti, M. 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, p. 43 – 50.

Osenova 2011: Osenova, P. Localizing a Core HPSG-based Grammar for Bulgarian. – In: H. Hedeland, T. Schmidt, K. Worner (eds.). *Multilingual Resources and Multilingual Applications, Proceedings of GSCL 2011*. Hamburg, 2011, p. 175 – 180.

Osenova et al. 2012: Osenova, P., Simov, K., Laskova, L. and Kancheva, S. A Treebank-driven Creation of an OntoValence Verb lexicon for Bulgarian. – In: N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis (eds.). *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul: ELRA, 2012, p. 2636 – 2640.

Žabokrtský and Lopatková 2007: Žabokrtský, Z. and Lopatková, M. Valency Information in VALLEX 2.0: Logical Structure of the Lexicon. – In: *The Prague Bulletin of Mathematical Linguistics*, 2007, No. 87, p. 41 – 60.