Valency as a mediator between grammar and lexicon: integrated annotation of verb valencies in Bulgarian through knowledge transfer from English resources

Petya Osenova
IICT-BAS
petya@bultreebank.org

Kiril Simov
IICT-BAS
kivs@bultreebank.org

Relevant UniDive working groups: WG1, WG2

1 Introduction

In our latest work we consider creating integrated language resources (on monolingual level between different types of resources while on multilingual level between resources from the same type). This step is a key prerequisite for providing more complex language technologies and developing more complex linguistic research on multilevel and multilanguage directions. The main benefit from such an integration effort is the verification of distinct resources, simultaneous usage of complementary knowledge, and transfer of knowledge between resources in different languages. Here we report on our first attempts at integrating a treebank (providing syntactic structure), a valency dictionary (providing syntagmatic potential) and a Wordnet of Bulgarian (providing lexical senses) with a valency dictionary of English (VerbNet) as well as with other English resources related to it. Valency dictionaries can be viewed as minigrammars that connect lexica's potential with full-fledged grammars. For that reason they are a very valuable resource for a language. During the years many such dictionaries for various languages and in multilingual settings have been compiled with respect to differing linguistic approaches. Here we will mention only some of the existing best practices. The interested reader can consult information about Croatian (Birtić et al., 2017), about Czech (Straňáková-Lopatková and Žabokrtský, 2002), about Polish (Przepiórkowski et al., 2014), about a multilingual setting (Di Fabio et al., 2019). At the time, a Valency lexicon for Bulgarian was initially extracted from the original constituency version of BulTreeBank. This version followed the Valency Principle in HPSG. This principle states that 'Unless the rules says otherwise, the mother's value for the VAL features (SPR, COMPS, and MOD) are identical to those of the head daughter (Sag et al., 2003). The work on the Bulgarian Valency Dictionary has been first reported in (Osenova et al., 2012) and consequently in some other works where the lexicographic classes from WordNet were also taken into account for handling valencies. In these works however the semantic roles in valency frames were viewed in a gross way, since only the lexicographic classes were used with their very prototypical roles like Agent, Patient, Experiencer, Theme. The basic XML version of the resource has been submitted to ELEXIS and ELRA repositories. But it respects only the syntactic constraints of valency and only partially includes semantic constraints and semantic information. In this work we would like to discuss an integrated annotation of verb valencies in Bulgarian with the usage of Pustejovsky's ideas on argument structure where he introduces three types of arguments: true (always realized syntactically), default (optional) and shadow (arguments being part of the lexical meaning of the verb)(Pustejovsky, 1991), WordNet lexicographic classes[1], verb senses from BTB-Wordnet and valency frames from VerbNet[2]. Let us consider briefly the role of each of these resources within the valency annotation of Bulgarian verbs. The knowledge transfer from English has been done in two ways: through the mappings between lexica in wordnets (BTB-Wordnet (Osenova and Simov, 2018) and Open English Wordnet (McCrae et al., 2019)) and through incremental localizations from VerbNet. Pustejovsky's ideas were incorporated in the abstracted frames where the role of all types of arguments were taken into account: true, shadow and default ones. The resulted valency resource is planned to be transferred into the verbs in BulTreeBank-UD later on.

2 The approach

Based on previous research – for instance (Osenova, 2022), we got evidence that Bulgarian prefers default arguments in comparison to true arguments and arguments in shadow. Thus, our representation of the Valency frames equals the HPSG representation of the argument structure (ARG-ST) which

[1] https://wordnet.princeton.edu/documentation/lexnames5wn
[2] https://verbs.colorado.edu/verbnet/

covers all the possible argument realizations in texts.

The lexicographic classes were transferred to the verbs in the respective meanings through BTB-Wordnet. The usage of VerbNet resource for English, however, required localization and adaptation to Bulgarian verbs. The fact that VerbNet has been mapped to a large extent to Princeton Wordnet and FrameNet, was an advantage, since as mentioned above, we used the mappings between Bulgarian Wordnet and Princeton wordnet/English Open wordnet. The adaptation went into several directions such as: the number and names of the roles, the verb grammatical behavior, the treatment of metaphorical usages, etc. For the annotation task three annotators from Bulgarian Philology with good knowledge of English were selected as 2022 summer interns. Their task was to check the frames extracted from the treebank with respect to the verb meaning and available examples, as well as to edit, if necessary. Then, they had to assign the semantic roles to the syntactic arguments of each distinct frame having in mind the lexicographic class as a general pointer, and VerbNet with its verb frames for English. This semantic role annotation over the valency frames had and still has back influence for improving the coverage of BTB-Wordnet where the missing meanings and lemmas have been continuously added. Among the challenges during this process are the following: representation of MWEs and of metaphorical usages.

3 Analyses

Let us first briefly introduce our notation of the valency frame. On Fig. 1 the valency frame of the verb 'celebrate' is given. We prefer to present the graphical view to the XML one for the sake of readability.

The frame is as follows:

```
Litse praznuvam sabitie
Person celebrate-1SG event
A person celebrates an event
```

Here the following information can be seen: the lexicographic wordnet class of *verb.social*, the LEMMA 'praznuvam' (celebrate-1P-SG), the definition (DEF) 'Have-I some holiday'.

The ones who celebrate, take the Agent role. The celebrated event is assigned the Theme role. The syntactic structure has been preserved. Here VPS means a phrase of type head-subject, and VPC means a phrase of type head-complement. On the graphics one cannot see the link to VerbNet[3] and the link to the English verb from the Open English Wordnet[4]. At the moment 28 semantic roles have been employed from VerbNet. We opted to have examples and tests when assigning them.

Below come some examples with MWEs. The first one exemplifies an idiomatic expression while the second one – a light verb construction. We skip the metaphorical usage for the sake of length constraints.

```
Igraya na kotka i mishka
Play-1SG on cat and mouse
I play cat and mouse
```

The semantic roles are as follows: the player is Agent and both pseudocomplements are a coordinated Theme.

```
Podlagam nyakogo na stres
Make-1SG someone on stress
I am stressing someone out
```

The semantic roles are as follows: The one who stresses is a Stimulus, the stressed one is an Experiencer and the stress itself is a Theme. The lexicographic category of the construction is *verb.emotion*.

In treating metaphors we follow the strategies in (Bonial et al., 2011) and (Brown and Palmer, 2012). This means to apply the set of roles for literal usages also to their metaphoric usages when possible. For example, the Theme in VerbNet is set to 'a participant that is being literally or metaphorically located, positioned, or moved; this participant may be concrete or abstract.'

Table 1 presents the frequencies of the roles in one file with 688 extracted roles (the number of all valency frames is around 4000). The initial observations show that the most frequent agentive role is Agent and the most frequent patient role is Theme. Also, 22 roles of the whole list of 28 were used.

If we check what verb lexicographic classes have been assigned the Agent role, we can see these frequencies: 1 weather, 2 emotion, 3 competition, 6 consumption, 6 change, 7 motion, 7 body, 8 contact, 14 creation, 15 perception, 17 possession, 22 stative, 36 cognition, 37 social, 62 communication. With these cross-checks, validation can be made

[3] https://verbs.colorado.edu/verb-index/vn/judgment-33.php

[4] celebrate (wn 3; g 2)

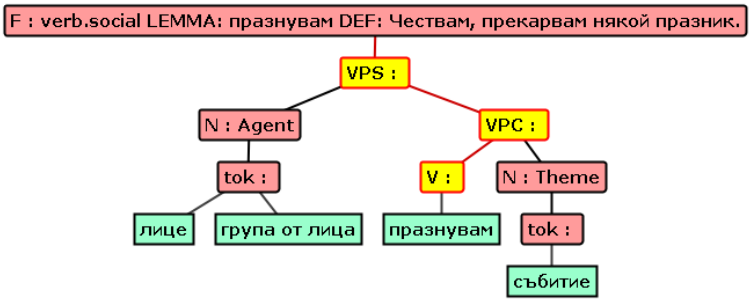

Figure 1: Valency of the verb ‘celebrate’.

Role	Freq	Role	Freq
Agent	243	Source	10
Theme	211	Co-Agent	8
Destination	34	Pivot	8
Patient	29	Recipient	8
Attribute	20	Stimulus	8
Topic	20	Result	7
Beneficiary	17	Co-Patient	6
Experiencer	16	Asset	2
Location	16	Cause	1
Co-Theme	11	Trajectory	1
		Product	1

with respect to which verb groups get differing roles and what the reasons are behind – an error, inconsistencies in VerbNet, wrong localisation to Bulgarian, etc.

4 Conclusions

The main challenges in mapping resources are: diversity in valencies within synsets (due to usage of a different preposition or no preposition); the idiosyncracy of MWEs in wordnets and valency dictionaries; reflexive verbs; various aspectual nuances of verbs; missing meaning and/or frame; differing behaviour of the verb in the two languages; blurred boundaries among some semantic roles; ellipses in the examples.

References

Matea Birtić, Ivana Brač, and Sinisa Runjaic. 2017. The main features of the e-glava online valency dictionary.

C. Bonial, S. Windisch Brown, W. Corvey, M. Palmer, V.V. Petukhova, and H.C. Bunt. 2011. An exploratory comparison of thematic roles in verbnet and lirics. In *Proceedings of the 6th Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation*, pages 39–43. University of Oxford. 978-90-74029-35-3.

Susan Windisch Brown and Martha Palmer. 2012. Semantic annotation of metaphorical verbs: A case study of climb and poison. In *Proceedings of ISA-8:Eighth Joint ACL-ISO Workshop on Interoperable Semantic Annotation*.

Andrea Di Fabio, Simone Conia, and Roberto Navigli. 2019. VerbAtlas: a novel large-scale verbal semantic resource and its application to semantic role labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 627–637, Hong Kong, China. Association for Computational Linguistics.

John P. McCrae, Alexandre Rademaker, Francis Bond, Ewa Rudnicka, and Christiane Fellbaum. 2019. English WordNet 2019 – an open-source WordNet for English. In *Proceedings of the 10th Global Wordnet Conference*, pages 245–252, Wroclaw, Poland. Global Wordnet Association.

Petya Osenova. 2022. The covid-19 pandemic in the valency of its predicates: Observations on a contemporary corpus of parliamentary debates. *Bulgarian Language (Supplement)*, (69):113–121.

Petya Osenova and Kiril Simov. 2018. The data-driven bulgarian wordnet: Btbwn. *Cognitive Studies | Études cognitives*.

Petya Osenova, Kiril Simov, Laska Laskova, and Stanislava Kancheva. 2012. A treebank-driven creation of an ontovalence verb lexicon for bulgarian. In *International Conference on Language Resources and Evaluation*.

Adam Przepiórkowski, Elżbieta Hajnicz, Agnieszka Patejuk, Marcin Woliński, Filip Skwarski, and Marek Świdziński. 2014. Walenty: Towards a comprehensive valence dictionary of Polish. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2785–2792, Reykjavik, Iceland. European Language Resources Association (ELRA).

James Pustejovsky. 1991. The Generative Lexicon. *Computational Linguistics*, 17(4):409–441.

Ivan A. Sag, Thomas Wasow, and Emily M. Bender. 2003. *Syntactic Theory A Formal Introduction*. Center for the Study of Language and Information, Stanford.

Markéta Straňáková-Lopatková and Zdeněk Žabokrtský. 2002. Valency dictionary of Czech verbs: Complex tectogrammatical annotation. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).