An English-Bulgarian Comparable Corpus Annotated with FrameNet Valence Patterns

Ivelina Stoyanova
Department of Computational Linguistics
Institute for Bulgarian Language – Bulgarian Academy of Sciences
iva@dcl.bas.bg

Relevant UniDive working groups: WG2

This abstract presents ongoing work on developing a bilingual corpus that demonstrates the syntactic realisation of the conceptual description of verbs in English and Bulgarian, and in a broader context, contributes to cross-lingual studies through enabling theoretical and practical data-driven observations for the two languages involved. The work relies on the universal aspects of conceptual description and syntactic realisation through exploiting the underlying organisational principles of the two main resources used (WordNet and FrameNet), which facilitate cross-language linking and transfer of information across languages (in this instance from English to Bulgarian) and resources, in particular transfer of the semantic description in FrameNet to the verbs in WordNet. The resource developed links the semantic level of the frame elements to the syntactic level of patterns representing the syntactic realisation of frame elements in terms of syntactic categories and grammatical function.

1 Motivation

It has long been acknowledged that combining WordNet with conceptual resources such as FrameNet produces a more complete semantic and syntactic representation of the lexical entries (Baker and Fellbaum, 2009; Schneider et al., 2012; Das et al., 2014), thus expanding the possible applications of the resources for the purposes of syntactic and semantic parsing.

WordNet ensures vast lexical coverage of the English lexicon structured and enriched with lexical and semantic information in the form of synset glosses, usage examples, notes on the usage or grammatical specificities, and a rich network of semantic relations. However, WordNet encodes no explicit semantic information about the participants in the situations described by the predicates and only limited information about their syntactic behaviour.

FrameNet provides a rich semantic description of the predicates using schematic representations (frames) of the configurations of the participants and circumstances that define the situation described. The corpus of examples in FrameNet annotated with explicit and implicit frame elements supplies empirical evidence about the syntactic realisations of semantic frames that is particularly valuable not only for linguistic generalisations about the target language (English) but as a point of departure for making observations cross-linguistically.

The granularity of frame elements in FrameNet is handled by involving them into a shallow hierarchy based on the hierarchy and inheritance relations between the frames (Litkowski, 2014)[1]. Consider for instance the taxonomy of frame elements AIR > FLUID > THEME derived from the frame hierarchy (built on the frame-to-frame relation of Inheritance) Breathing > Fluidic motion > Motion.

The description of verb semantics and the grouping of verbs into semantically homogeneous classes in WordNet and FrameNet reflects complementary aspects of verb semantics. Enriching WordNet synonym sets with conceptual information from FrameNet, and vice versa, populating the FrameNet frames with new lexical units coming from WordNet provides a more comprehensive semantic and syntactic description.

2 Creating a Comparable English-Bulgarian corpus with annotated examples

The corpus compilation relies on the mapping between WordNet synsets and FrameNet frames, through which each verb in WordNet is associated with a number of possible syntactic patterns defined for the relevant frames in FrameNet. Although the syntactic component of the description is more language specific than the semantic component, the generalised patterns are applicable, at least to a certain degree, to other languages.

I take as a point of departure the lattices of the

[1] urlhttps://www.clres.com/clr/fetax.php

frame elements and their syntactic realisations for certain verbs and the valence patterns of frame elements as described in the annotated FrameNet examples[2] (Burchardt and Pennacchiotti, 2008). The dataset for English is supplemented with examples from SemCor (Miller et al., 1993) in order to illustrate the usage of particular verb senses and verb literals.

The dataset for Bulgarian consists of examples excerpted from BulSemCor (Koeva et al., 2011), additionally supplemented with examples from other corpora. In the process of collecting examples in English and Bulgarian, I consider both the shared syntactic patterns (the ones valid for both languages) as well as patterns that are specific to each of the languages. An annotated example is shown on Fig. 1.

After preprocessing (morphosyntactic description and lemmatisation for English and Bulgarian (Koeva et al., 2020)), the syntactic components are identified: (a) the verb; (b) noun phrases – subject NPs (marked as NP.Ext) or direct object NPs (NP.Obj); (c) prepositional phrases (PP); (d) subordinate clauses marked with different conjunctions and other lexical elements; etc.

Sentence components corresponding to core frame elements are manually annotated and marked with both the name of the frame element (e.g., AGENT, THEME, INSTRUMENT, etc.) and the syntactic category that realises it – NP.EXT, NP.OBJ, PP, ADVP, CLAUSE, etc.

There may be mismatches in the syntactic category across languages, e.g. a certain frame element may be a direct object in one language and a prepositional object in another. Languages may also differ in terms of the overtness of syntactic information, i.e. the possibility to leave an obligatory element non-explicit (null instantiations retrievable from the context or the grammatical construction); language-specific diatheses, constructions, word order, morphosyntactic features, etc. We consider indefinite null instantiations (INI), constructional null instantiations (CNI) and definite null instantiations (DNI) and annotate them in the sentence. We do not annotate the cases of incorporated frame elements (INC).

The inventory of means that introduce certain frame elements such as prepositions, conjunctions, wh-words, etc. may also vary across languages. For this reason, the original patterns from FrameNet are generalised in order to allow cross-language match with the Bulgarian data. For example, patterns with `Sinterrog`, `Sfin`, `VPing` are clustered together and considered as subclasses of `Clause` so as to be matched to their different realisation in Bulgarian where some of these subclasses are not present (e.g., `VPing`).

Prepositional phrases realising the same frame element but headed by different prepositions (e.g., PP[of], PP[from] introducing the frame element COMPONENTS in the frame `Building`), are also grouped together in a PP-phrase.

Particular attention is paid to examples which are not matched to a pattern in order to identify patterns characteristic for Bulgarian that do not appear in FrameNet or for English in general. However, these cases are very rare.

3 Results

The dataset is compiled from the resources outlined above with focus on several semantic verb classes: verbs of communication, verbs of motion and verbs of change.

The English dataset covers 211 verbs (lexical units in FrameNet) with their assigned frame. The verbs are aligned to 135 WordNet synsets using WordNet-to-FrameNet mappings. For each verb (lexical unit in FrameNet) there is a number of examples in the FrameNet dataset illustrating its valence patterns, and the dataset contains a total of 13,295 illustration examples representing 3,577 different patterns. The annotation of each sentence shows the verb and the sentence components marking the realisation of core frame elements.

The Bulgarian dataset is considerably smaller and represnet work in progress. It covers 146 verbs across 125 WordNet synsets of the semantic classes under study. There are 2,050 annotated example sentences representing 272 different patterns. Similarly, the annotation included labelling the sentence components with respect to their syntactic category and the frame elements they realise.

A cross-lingual analysis is performed, aiming to match the FrameNet lexical units to WordNet synsets, and thus to obtain verb pairs in Bulgarian and English that exhibit the same set of frame element lattices and syntactic patterns. The pairs that have the same syntactic realisation are considered to be closer translation equivalents than verbs that share only part of the valence patterns or differ significantly in their syntactic realisation.

[2] http://framenet.icsi.berkeley.edu/

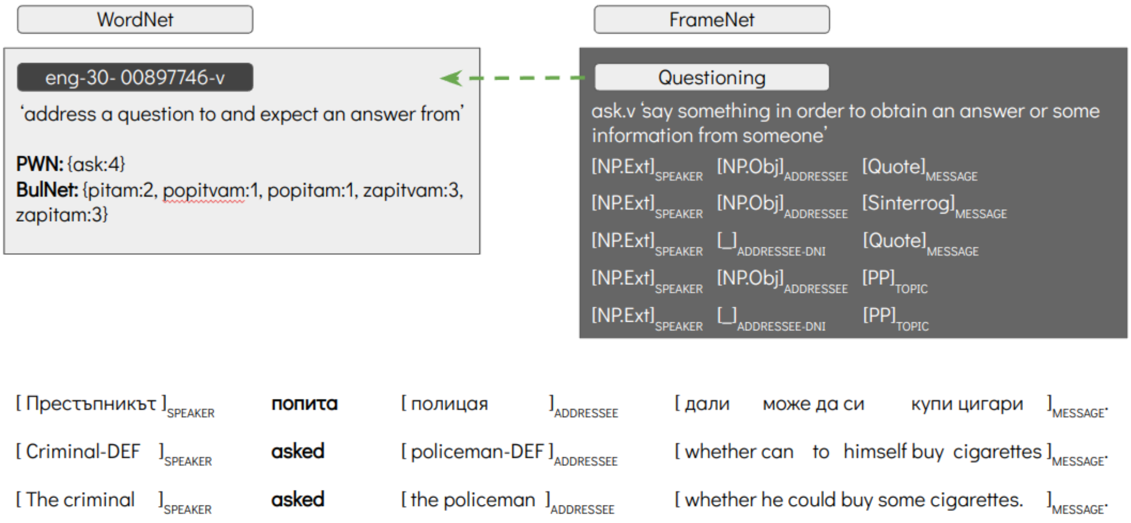
WordNet
eng-30- 00897746-v
'address a question to and expect an answer from'
PWN: {ask:4}
BulNet: {pitam:2, popitvam:1, popitam:1, zapitvam:3, zapitam:3}
FrameNet
Questioning
ask.v 'say something in order to obtain an answer or some information from someone'
[NP.Ext]SPEAKER [NP.Obj]ADDRESSEE [Quote]MESSAGE
[NP.Ext]SPEAKER [NP.Obj]ADDRESSEE [Sinterrog]MESSAGE
[NP.Ext]SPEAKER [_]ADDRESSEE-DNI [Quote]MESSAGE
[NP.Ext]SPEAKER [NP.Obj]ADDRESSEE [PP]TOPIC
[NP.Ext]SPEAKER [_]ADDRESSEE-DNI [PP]TOPIC
[Престъпникът]SPEAKER попита [полицая]ADDRESSEE [дали може да си купи цигари]MESSAGE
[Criminal-DEF]SPEAKER asked [policeman-DEF]ADDRESSEE [whether can to himself buy cigarettes]MESSAGE
[The criminal]SPEAKER asked [the policeman]ADDRESSEE [whether he could buy some cigarettes.]MESSAGE

Figure 1: An annotated example

However, it should be taken into account that the Bulgarian dataset is not sufficiently large at this stage to draw reliable conclusions on the pattern correspondences.

4 Future work

The work on describing the conceptual and syntactic properties of Bulgarian verbs shows that the conceptual description encoded in the FrameNet frames is largely language-independent and transferrable cross-linguistically. By employing the potential of wordnets aligned by means of shared identification numbers with the original Princeton WordNet, we can map conceptual description from FrameNet to less-resourced languages and thus facilitate cross-linguistic analysis.

Acknowledgements

This research is carried out as part of the project *Enriching the Semantic Network WordNet with Conceptual Frames* funded by the Bulgarian National Science Fund, Grant Agreement No. KP-06-H50/1 from 2020.

References

C. F. Baker and C. Fellbaum. 2009. WordNet and FrameNet as Complementary Resources for Annotation. In *Proceedings of the Third Linguistic Annotation Workshop (ACL-IJCNLP '09), Association for Computational Linguistics, Stroudsburg, PA, USA*, pages 125–129.

Aljoscha Burchardt and Marco Pennacchiotti. 2008. FATE: a FrameNet-Annotated Corpus for Textual Entailment. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. European Language Resources Association (ELRA), Marrakech, Morocco.

Dipanjan Das, Desai Chen, André F. T. Martins, Nathan Schneider, and Noah A. Smith. 2014. Frame-Semantic Parsing. *Computational Linguistics*, 40(1):9–56.

Svetla Koeva, Svetlozara Leseva, Borislav Rizov, Ekaterina Tarpomanova, Tsvetana Dimitrova, Hristina Kukova, and Maria Todorova. 2011. Design and development of the Bulgarian sense-annotated corpus. In *Information and communications technologies: present and future in corpus analysis: Proceedings of the III International Congress of Corpus Linguistics*, pages 143 – 150.

Svetla Koeva, Nikola Obreshkov, and Martin Yalamov. 2020. Natural language processing pipeline to annotate Bulgarian legislative documents. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6988–6994, Marseille, France. European Language Resources Association.

Ken Litkowski. 2014. The FrameNet Frame Element Taxonomy. https://www.clres.com/online-papers/FETaxonomy.pdf.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A Semantic Concordance. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

Nathan Schneider, Behrang Mohit, Kemal Oflazer, and Noah A. Smith. 2012. Coarse Lexical Semantic Annotation with Supersenses: An Arabic Case Study. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 253–258. Association for Computational Linguistics.