Processing the MWEs that are not “words-with-spaces” in BulTreeBank – WG4

Petya Osenova and Kiril Simov, IICT-BAS

In this work we present a classification of examples of MultiWord Expressions (MWEs) from BulTreeBank - an HPSG-based treebank of Bulgarian (Simov, Popova and Osenova 2002). The main distinction between MWEs is as follows: the ones that cannot be considered a phrase with syntactic structure (words-with-spaces), and the ones that exhibit syntactic structure, but have non-compositional semantics. Thus, in the original treebank only the so-called words-with-spaces MWEs, such as the complex conjunctions, adverbs, prepositions, and similar have been annotated explicitly. They have neither compositional semantics, nor syntax. The number of their occurrences amounts to1800. The other MWEs (such as, light verb constructions, idioms, names, etc.) have been analyzed compositionally at the syntactic level.

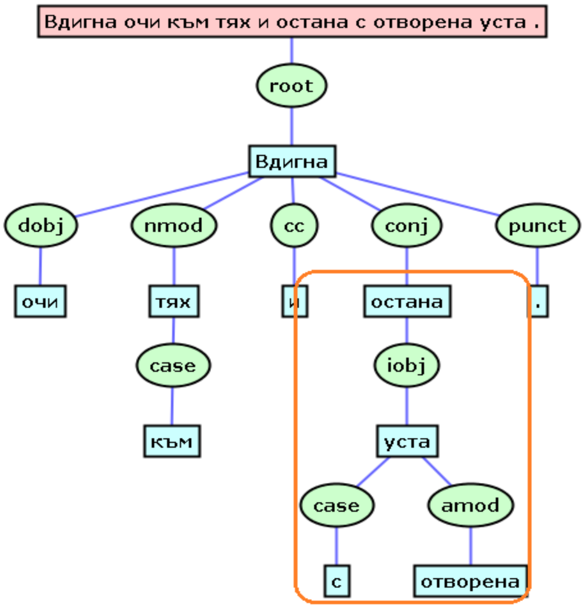
In this abstract we report our work on the processing of these other MWEs, which are “non-words-with-spaces”. It includes the following steps: detection, extraction and modeling of the MWEs as catenae. The catenae approach generally means that (1) The MWE is represented as a dependence subtree; (2) The points of possible insertion of modifiers have been indicated; and (3) The semantics is represented for the participation specific words as well as for the whole MWEs. The semantics of the specific words is necessary for recording their literal meaning. More information on modeling Bulgarian MWEs can be found in Osenova and Simov (2015a) and Simov and Osenova (2015b).

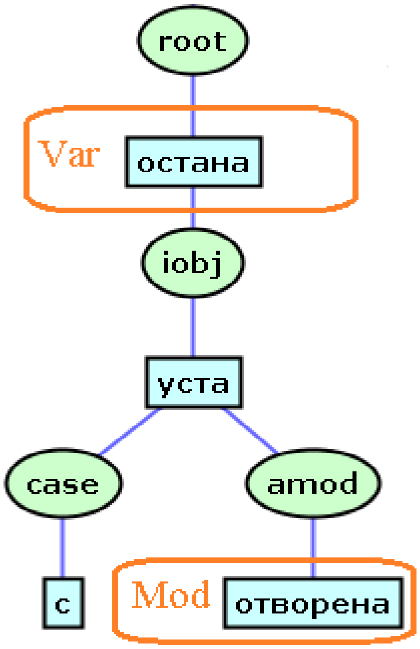
The identification of MWEs in the treebank has been done manually during the creation of the valency lexicon, which builds on the syntactic annotation in the treebank and the semantic annotation of the open class words. We have identified 1858 cases (occurences) of MWEs. During the first phase of annotation the senses of the individual words within MWEs have been determined. At this stage, the recognizable elements of the MWEs have been mapped to their senses. Sometimes, these senses had already reflected the meaning of the MWE expression. For example, книжен плъх (literally: book/ADJ rat/NOUN, ‘book worm’) has been mapped to the sense of a man that reads too much, but the adjective *книжен* was mapped also to the meaning ‘related to books’. Another example is *гореща линия* (literally: hot line, ‘hot line’) with the meaning of a direct telephone line between two officials. In these cases the noun *линия* was already mapped to ‘a telephone connection’. However, most of the examples have not received any related senses for their participating parts. The exact meaning of the whole MWEs was annotated at the second phase of the process. In this case, an experienced lexicographer determined the final main form of the MWE, its meaning and variability. For example, *мръсни пари* (literally: dirty money, 'dirty money') got the sense for dirty as non-clean, but the meaning of illegally earned income had to be added. Another example is *дребна риба* (literally: small fish, 'small fry') with the meaning of "someone who is small and insignificant". The two parts of the MWE were mapped respectively to the following senses: ‘limited or below average in number or quantity or magnitude or extent’ and ‘any of various mostly cold-blooded aquatic vertebrates’, thus the corresponding meaning of the whole MWE has been added during the second phase of processing.

The MWEs with verbal heads have been mapped also to their valency frames. For example, the verb *остана* (remain) in the MWE *остана с отворена уста* (literally: stay with open mouth, 'remain open-mouthed') has more than 15 senses, associated with the relevant frames. The one that goes to the MWE-related meaning is ‘to be amazed’. Here is the glossed and translated example sentence:

Вдигна очи към тях и остана с отворена уста.
Raised.3PersonSG eyes to them and stayed with opened mouth.
She/he looked up at them and stayed with her/his mouth open.

Below the same sentence in the treebank and as lexicon catena model of the MWE are given:





The catena, presented graphically on the right side, is in its base form (as given in the lexicon), but in the sentence the head verb is in past tense, third person, singular (the graphics on the left side). In the lexicon representation the verb root was marked as an item that could vary in its forms (Var). Additionally, the adjective *отворена* (open/FEM) could be modified (Mod) as in the example: *остана с широко отворена уста* (literally: stayed with *widely* open mouth). The semantics of the MWEs is stated in the lexicon as part of the lexical entry of the catena. The communication between the lexicon and the text is operated through the realizations of the lexicon catena in the context. The lexicon catena subtree encodes the possibilities of realization with respect to insertion of possible modifiers and variability of the participating parts.

The annotation scheme generally follows the one elaborated within WP4. It considers the following factors: (1) The POS of the head (noun, verb, etc.) (2) The normalized form of the MWE as well as the syntactic, lexical and word order variants; (3) MWE as showing no variation (Fixed) or showing variation (Nofixed).

References

Simov, Popova and Osenova 2002: K. Simov, G. Popova and P. Osenova. HPSG-based syntactic treebank of Bulgarian (BulTreeBank). In: *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World*, edited by Andrew Wilson, Paul Rayson, and Tony McEnery; Lincom-Europa, Munich, pp. 135-142.

Osenova and Simov 2015a: P. Osenova and K. Simov 2015: Modeling Lexicon-Syntax Interaction with Catenae. In: Journal of Cognitive Science, vol. 16/3, pp. 287-322. Seoul National University, College of Humanities. ISSN: 1598-2327.

Simov and Osenova 2015b: K. Simov and P. Osenova 2015. Catena Operations for Unified Dependency Analysis. In: Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015), pages 320–329, Uppsala, Sweden, August 24–26 2015.