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

Petya Osenova and Kiril Simov (IICT, Bulgarian Academy of Sciences)
WG4: Annotating MWEs in Treebanks (related also to WG1)

1. Overview

• Here 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 (not *words-with-spaces*)
• 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 (1800)
• The other MWEs (such as, light verb constructions, idioms, names, etc.) have been analyzed compositionally at the syntactic level

2. “non-words-with-spaces” MWEs

The **processing of these MWEs** includes the following steps: *detection*, *extraction* and *modeling of the MWEs as catenae*

The catenae approach generally means that
• The MWE is represented as a dependence subtree;
• The points of possible insertion of modifiers have been indicated; and
• The semantics is represented for the participating specific words as well as for the whole MWEs.

The semantics of the specific words is necessary for recording their literal meaning.

3. Identification

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.

During the first phase of annotation the senses of the individual words within MWEs have been determined.

Examples with Adj N patterns:

книжен плъх (literally: *book rat*, ‘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’

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’

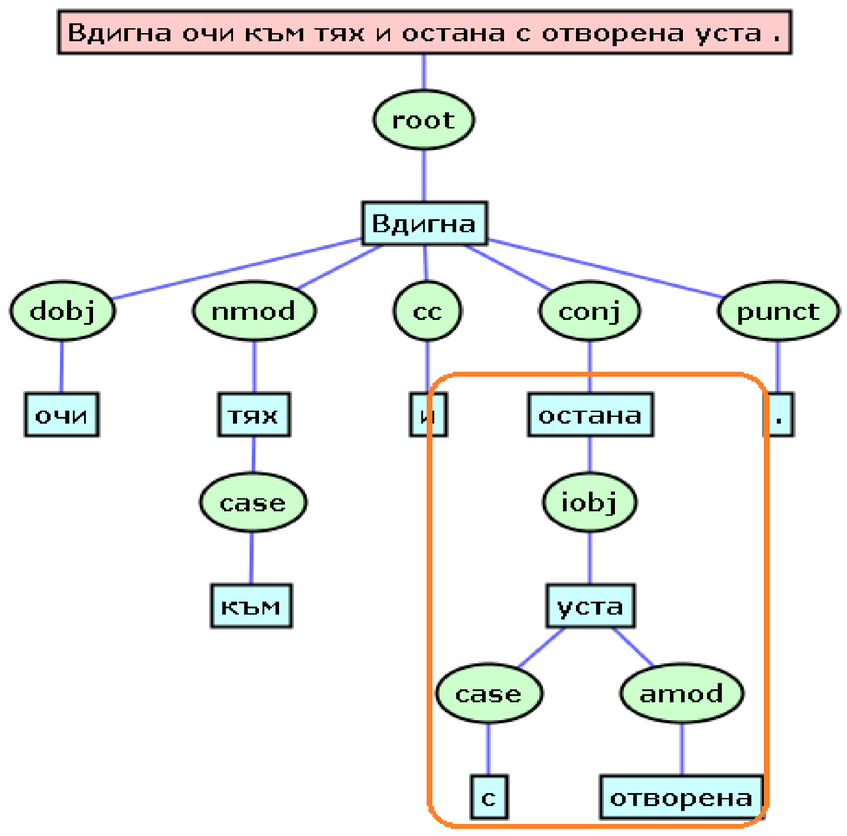
The exact meaning of the whole MWEs was annotated at the second phase of the process.

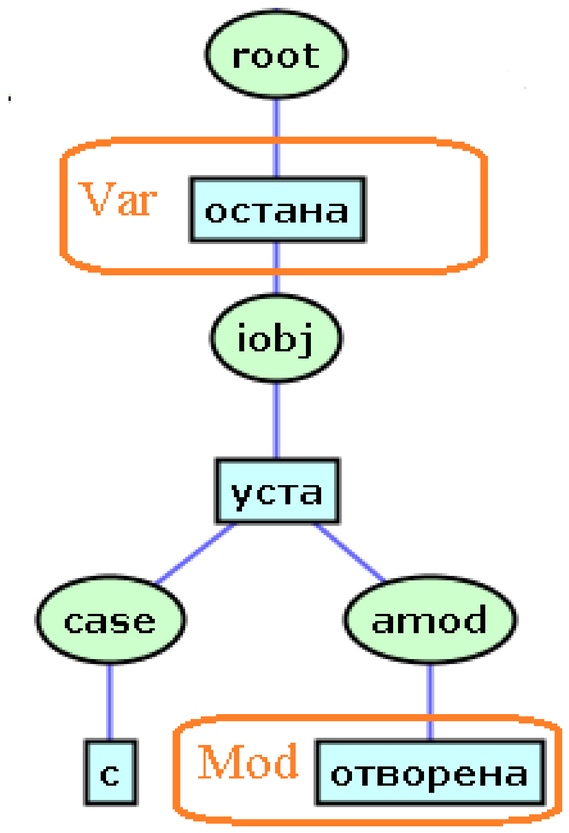
мръсни пари (literally: *dirty money*, 'dirty money') got the sense for dirty as non-clean, but the meaning of illegally earned income had to be added

4. Modeling of Frame and Variation

The MWEs with verbal heads have been mapped also to their valency frames.

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





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 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*).

This research has received support by the EC's FP7 under grant agreement number 610516:
“QTLeap: Quality Translation by Deep Language Engineering Approaches”