Annotation of Multiwords in BulTreeBank
(current status and perspectives)

Petya Osenova and Kiril Simov
Bulgarian Academy of Sciences

BulTreeBank exists in two formats: HPSG-based (original - constituent-based with head annotation and grammatical relations) and Dependency-based (converted from the HPSG-based format). In both of them the representations of the various kinds of MWE is a challenge.

Since there is no broadly accepted standard for MWE (see the various classifications in Villavicenncio and Kordoni 2012), in this presentation we adopt the MWE classification, presented in (Sag et. al 2001). They divide the MWE into two groups: lexicalized phrases and institutionalized phrases. The former are further subdivided into: *fixed-expressions*, *semi-fixed expressions* and *syntactically-flexible expressions*. Fixed expressions are said to be fully lexicalized and undergoing neither morphosyntactic variation nor internal modification. Semi-fixed expressions have fixed word order, but "undergo some degree of lexical variation, e.g. in the form of inflection, variation in reflexive form, and determiner selection" (non-decomposable idioms, proper names). Syntactically-flexible expressions show more varieties in their word order (light verbs, decomposable idioms).

In its inception and development phase, the HPSG-based Treebank adopted the following principles: When the expression is fixed, thus inseparable, with fixed order and can be viewed as a part-of-speech, it receives lexical treatment. Thus, this group concerns the running words with complex POS: complex prepositions, conjunctions, pronouns, adverbs. There are 1081 occurrences of running words with such complex POS (compared to 214 000 running words with POS in the Treebank). Of course, there are also exceptions. For example, one of the complex indefinite pronouns in Bulgarian shows variation in its ending part: *каквито и да* ***е/са/било*** (whatever).The varying part is either a 3-person-singular-present-tense-auxiliary, 3-person-plural-present-tense-auxiliary or its 3-person-neuter-singular past participle.

The semi-fixed expressions (mainly proper names) have been interpreted as multiwords. However, all the idioms, light verb constructions, etc. have been treated syntactically. This means that in the annotations there is no difference between: *kick the door* and *kick the bucket*. In both cases we indicated that the verb *kick* takes its nominal complement.

After some exploration of the treebank, such as the extraction of the valency frames from it and training of statistical parsers on it, we discovered that the present ways of MWE annotations are not the most useful ones. In both cases the corresponding generalizations are overloaded with specific cases which are not easy to incorporate in more general classifications. Needless to say, the group of lexically treated POS remained stable. However, the other two groups were reconsidered. Proper names, as semi-fixed, are treated separately, i.e. as non-MWE, since we need coreferencing the single occurrence of the name with the occurrence of two or more parts of the name of the same person in the text. Also, light constructions have to be marked as such explicitly. The same holds for the idioms.

At the moment we are considering a number of possible approaches for handling idioms and light constructions. The approaches are not necessarily conflicting with each other.

The first one is selection-based. This approach is appropriate for MWE in which there is a word that can play the role of a head. For example, a verb subcategorizes for only one lexical item[1] or a very constrained set of lexical items. When combined with nouns, such as *time*, *shape*, *hope*, the verb *lose* forms idioms (lose time= 'to be in a hurry'; lose shape = 'to be not well'; lose hope = 'to be unhappy'). However, when combined with other nouns, such as *wallet* or *relative*, the verb takes canonical complements. In former cases, verbs like *обръщам* (*pay*) take only noun *внимание* (*attention*) for making an idiom. Another example is the verb *вземам* (*take*), which combines in such cases with *дума* (*word*). However, light expressions with desemantisized verbs, such as *има* (*have*) or *става* (*happens*) – *I have the word = I am allowed to speak* or *it happens word = refers to* can take more semantic classes as complements. In this case we mark just the head of MWE, since according to HPSG in head-complements phrases the head incorporates the information from its dependants.

In this approach the assumption is that the verb posits its requirements on the complements. However, such an analysis needs a very detailed valency lexicon. One problem with this approach is when the dependant elements allow modifications.

The second approach is construction-based. In this case there is no head. In this case MWEs are with fixed order and inseparable parts. They are annotated via brackets on lexical level. One example is the idiom: *from needle to thread = from the beginning to the end*. This approach is problematic for syntactically flexible MWE.

The third approach marks all the parts of the MWE. It is based on the notion of catena[2] which could be considered as a continuous subtree in the syntactic representation. It operates on both representations – constituency and dependency. Here is an example of this annotation, where the syntactically treated idiom in our previous analysis is now presented as a catena sequence:

(**VPS** *Той* (**VPC-C** (**V-C** *ритна*) (**N-C** *камбаната*))) = He kicked the bucket.

where the suffix "**-C**" marked the catena. Maybe this approach would add some spurious compositionality to the idioms, but it would be indispensable for handling idiosyncratic cases, such as separable MWE, ellipsis, coordination, etc. Also, this approach would facilitate the processing of MWE, since - as far as we observed the data, two catena chains (i.e. MWE) cannot overlap.

References

Bejček and Straňák 2010: Eduard Bejček and Pavel Straňák. Annotation of multiword expressions in the Prague dependency Treebank. In: *Language Resources and Evaluation* (2010) 44, pp. 7–21.

Sag et. al 2001: Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, Dan Flickinger. Multiword Expressions: A Pain in the Neck for NLP. In: *Proc. of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pp. 1-15.

Villavicencio and Kordoni 2012: Aline Villavicencio and Valia Kordoni. Multiword Expressions and Collocations in theory and practice. *Course materials for European Masters Program in Language and Communication Technologies*.

[1] In the sense of the notion 'monosemic lexeme', adopted in the Prague Dependency Treebank (Bejček and Straňák 2010).

[2] http://en.wikipedia.org/wiki/Catena_(linguistics)