Distribution and specificities of MWEs in BulTreeBank

Petya Osenova and Kiril Simov

WG4

As reported in our poster at PARSEME meeting in Struga, 1858 occurrences of non-words-with-spaces have been identified in BulTreeBank. We showed a way to model the connection between the lexicon and the text via catenae[1].

Our initial annotation approach was compositional. Thus, each content part of an MWE was connected to its meaning in WordNet. The next step was to annotate the MWEs with syntactic subtypes. For this purpose we selected the following labels from the WG4 classification scheme (Rosén et al. 2015; Rosén et al. 2016): *adjectival MWEs* (AMWE), *nominal MWEs* (NE, NMWE), *prepositional MWEs* (PMWE), *verbal MWEs* (LVC, VI) and *Other*. As it can be seen, the nominal and verbal MWEs have two subtypes. The nominal one includes *names* (NE) and *nominal MWEs* (NMWE). The verbal ones include *light verb constructions* (LVC) and *verbal idioms* (VI). For the MWE examples that do not belong to any of these categories, the label *Other* was provided.

The statistics over the annotated MWEs is presented in Table 1 below.

MWE Types	**Percentage**
NE	40 %
NMWE	28 %
VI	11 %
Other	10 %
PMWE	7 %
LVC	3 %
AMWE	1 %

Table 1. Percentage of different MWE types in BulTreeBank.

It can be observed that the subtypes with lowest frequency are the adjectival ones (AMWE – 1 %) and the light verb constructions (LVC – 3 %). Unsurprisingly, the most frequent subtype is the named entities (NE – 40 %). Next come the nominal MWEs (NMWE – 28 %). Somewhat in the middle of the scale are the verbal idioms (VI – 11 %), other (Other – 10 %) as well as the prepositional subtypes (PMWE – 7 %). It should be noted that the *Other* subtype includes newly identified complex prepositions (most cases), proverbs or adverbial MWEs.

Let us consider in more detail the groups NMWE and VI, since the former is the most frequent MWE nominal type after the NE group, and the latter is the most frequent one among verbal types.

Nominal MWEs (NMWE)

These include predominantly the syntactic type A + N. Semantically, most of them are terms *полска мишка* (‘field mouse’, field mouse), *глобална мрежа* (‘global net’, world wide

[1] See more details at: http://typo.uni-konstanz.de/parseme/images/Meeting/2016-04-07-Struga-meeting/WG1-4-OSENOVA-SIMOV-abstract.pdf

web), *дафинов лист* (‘bay leaf’, bay leaf) or collocations *текуща сметка* (‘current bill’, current account), *земеделска партия* (‘agricultural party’, country party), *резервна скамейка* (‘reserve bench’, substitute bench). Another syntactic type is N + p + N *средство за масово осведомяване* (‘means of mass information’, mass media), *мярка за неотклонение* (‘measures for diversion’, detention measures); *ябълка на раздора* (‘apple of disagreement-the’, apple of discord); *сняг на парцали* (‘snow in rags’, snow in big flakes).

Some of the phrases are only literal *пръстов отпечатък* (‘finger-ADJ print’, finger print); *средство за масово осведомяване* (‘means of mass information’, mass media), some are only figurative *дървена ваканция* (‘wooden vacation’, unexpected vacation due to epidemic); *ябълка на раздора* (‘apple of disagreement-the’, apple of discord), while others are ambiguous *задънена улица* (‘no through street (e.g. for a car to pass)’, blind alley or dead-end (e.g. for a person in crisis)). This is reflected in the semantics part of the lexical entry. But the majority of them are non-fixed, i.e. open to external modifiers.

Verbal idioms (VI)

The syntactic types are: a) V + PP *отивам на вятъра* (‘go to wind’, be wasted); *встъпя в длъжност* (‘enter in duty’, take office); *изпадам в несъстоятелност* (‘fall into bankruptcy’, bankrupt); b) Aux + p + N *съм обеца на ухото* (‘to be ear-ring on the ear’, to be a good warning); *съм на хартия* (‘to be on paper’, to be on paper); *съм вързан в кърпа* (‘to be tied in napkin’, to be dead sure); c) V + NP + PP *изпратя някого на улицата* (‘send someone to the street’, throw into the street); *подведа някого под отговорност* (‘bring someone under responsibility’, prosecute); d) V + NP *бера душа* (‘pick soul’, be on one’s death-bed); *развързвам кесията* (‘untie wallet’, loosen one’s purse-strings); *разрушавам стената* (‘destroy the wall’, destroy the wall or break the constraints); *пробвам си късмета* (‘try one’s luck’, chance one’s luck). The cases a) and d) are the most frequent ones. The a) type refers to intransitive verbs and the MWEs are mostly fixed. This means that modifiers within the MWEs are rare. The same holds for the b) type. Type c) shows a possibility of internal modification preferably in the NP complement part. In type d) the internal modification concerns mainly the verbal head.

Conclusions

The general observations are as follows: the nominal MWEs prevail in comparison to other types. From the other types the most frequent ones are the verbal idioms, type *Other* and then – prepositional MWEs. The low frequency of adjectival MWEs is not surprising. However, it is interesting to see the infrequent usages of the LVCs. Partly, this result might be due to the fact that Bulgarian prefers content synonym verbs to the decomposed LVC, and partly it might be due to the fact that some LVC have been identified as verbal idioms.

References

Rosén et al. 2015: Victoria Rosén, Gyri Smørdal Losnegaard, Koenraad De Smedt, Eduard Bejček, Agata Savary, Adam Przepiórkowski, Petya Osenova and Verginica Barbu Mititelu. A Survey of Multiword Expressions in Treebanks. *Proceedings of the Fourteenth International Workshop on Treebanks and Linguistic Theories (TLT14)*, 11–12 December 2015 Warsaw, Poland, pp. 179–193.

Rosén et al. 2016: Victoria Rosén and Koenraad De Smedt and Gyri Smørdal Losnegaard and Eduard Bejček and Agata Savary and Petya Osenova 2016: MWEs in Treebanks: From Survey to Guidelines. Proceedings of LREC 2016. 23-28 May, Portorož, Slovenia, pp. 2323-2330.