Verbal Multiword Expressions in Treebanks

Veronika Vincze and István Nagy T.
University of Szeged

WG4: Annotating MWEs in Treebanks

In this poster, we show how verbal multiword expressions or multiword verbs (MWVs) are annotated in treebanks. We focus on light verb constructions (LVCs) and verb-particle constructions (VPCs), with regard to English, Hungarian, French and German data.

Some treebanks contain MWE-aware annotations, i.e. there are MWE-specific morphological or syntactic tags in them. Here, we take our examples from the French Treebank, which contains explicit annotations for MWEs (Abeillé et al. 2003), the Penn Treebank (Marcus et al., 1993), which contains annotation for VPCs, the TIGER corpus (Brants et al. 2004) and the Szeged Dependency Treebank (Vincze et al. 2013), both of which contain annotation for LVCs. In these treebanks, the special relation of the two components of the MWE is distinctively marked by a certain syntactic or morphologic label.

English MWVs

The special relation of the verb and particle within a VPC is distinctively marked in the Penn Treebank, the particle is assigned a specific part of speech tag (RP) and it also has a specific syntactic label (PRT). Thus, it is marked both at the level of morphological and syntactic annotations. For instance, in the imperative *Turn the light off* the POS tag of *off* is RP and the sentence is syntactically annotated in the following way:

```
(S (NP-SBJ *)
        (VP turn
                (NP the light)
                (PRT off)))
```

Hungarian MWVs

The Szeged Constituency Treebank has also been manually annotated for light verb constructions (Vincze and Csirik, 2010). Later, in the dependency version of the corpus, dependency relations were enhanced with LVC-specific relations that can be found between the two members of the constructions. Among the two versions, the annotation found in the dependency treebank is more detailed since in the constituency treebank, it was only the syntactic boundaries of the phrases that were marked but in the dependency treebank the inner syntactic relation of the MWE (i.e. which word is the head and which is the dependent) is also encoded. For instance, *a támogatásról* ***hozott döntés*** "the **decision made** on the support" contains a participle form of a LVC and its dependency annotation is the following:

1	a	DET	4
2	támogatásról	OBL	3
3	hozott	ATT-LVC	4
4	döntés	ROOT	0

German MWVs

In the TIGER corpus, LVCs that consist of a verb and a prepositional phrase are annotated with the relation CVC (collocational verb construction). However, the annotators excluded from the annotation verb-object pairs, so *Abschied nehmen* "to take leave" was not considered here as LVC but *zur Diskussion bringen* "to discuss" has the following syntactic structure:

(zur Diskussion)$_{\text{CVC}}$ bringen

French MWVs

The French Treebank contains explicit annotations for MWEs (Abeillé et al. 2003). Here, we focus on the verbal MWEs, which are grouped in the treebank according to their POS patterns (like V N, V P N etc.). Verbal MWEs also include light verb constructions like *avoir lieu* "to take place" or *entrer en vigueur* "to enter into force", which is annotated as follows:

entrer en vigueur – VW-VPN

In the poster presentation, we illustrate the annotation principles of these treebanks regarding MWVs with examples and also make interlingual comparisons on the annotation practice of MWEs. We believe that a comparative approach to MWV annotations may shed more light on how future treebanks should deal with MWEs in general.

References

Abeillé, Anne; Clément, Lionel; Toussenel, François 2003: Building a treebank for French. In Abeillé, Anne (ed.): Treebanks: building and using parsed corpora. Kluwer, Chapter 10.

Brants, Sabine; Dipper, Stefanie; Eisenberg, Peter; Hansen, Silvia; König, Esther; Lezius, Wolfgang; Rohrer, Christian; Smith, George; Uszkoreit, Hans 2004: TIGER: Linguistic Interpretation of a German Corpus. *Journal of Language and Computation* 2, 597–620.

Marcus, Mitchell P.; Santorini, Beatrice; Marcinkiewicz, Mary Ann 1993: Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics* 19(2): 313–330.

Vincze, Veronika; Csirik, János 2010: Hungarian Corpus of Light Verb Constructions. In: *Proceedings of COLING 2010*, Beijing, China, pp. 1110-1118.

Vincze, Veronika; Szauter, Dóra; Almási, Attila; Móra, György; Alexin, Zoltán; Csirik, János 2010: Hungarian Dependency Treebank. In: *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

Vincze, Veronika; Zsibrita, János; Nagy T., István 2013: Dependency Parsing for Identifying Hungarian Light Verb Constructions. In: *Proceedings of IJCNLP 2013*.