Towards principles for the annotation of MWEs in treebanks [WG4]

Koenraad De Smedt and Victoria Rosén
University of Bergen, Norway

One of the objectives of Working Group 4 in PARSEME is the enhancement of MWE-aware methodologies of treebank construction, and among the expected outcomes are annotation guidelines for representing MWEs in treebanks. Creating such guidelines is, however, no simple task. There are many different types of treebank annotation, and the annotations are to some degree rooted in different theoretical frameworks. The two most common types of treebanks are dependency and constituency treebanks, while some treebanks combine these two types of annotation. There are also some treebanks that are based on frameworks such as HPSG and LFG. This has led to many different types of annotation of MWEs (Rosén et al., 2015).

Given this situation, and the lack of full agreement on what constitutes a MWE, how might it be possible to make guidelines for the annotation of MWEs? Although specific guidelines will need to be tuned to the treebank annotation type, we would like to formulate some general principles that might hold for all treebanks. For linguistic research, as well as for the development of some language technology applications, it is important to be able to perform targeted searches for MWEs in treebanks. We argue that the following desiderata are beneficial to effective treebank search:

A. MWEs should be annotated as such, so that treebank queries can directly target them.

B. The annotation of noncompositional MWEs should distinguish them from homonymous strings with a compositional analysis.

C. Individual MWEs should be searchable even if they are discontinuous or variable in form.

D. It should be possible to search for various *types* of MWEs based on their characteristics.

Principle A is a general principle that aims at improving the ease with which MWEs can be identified in treebanks, without the need to be detected by heuristics.

Principle B is a corollary: ease of identification implies that MWEs should be distinguished from homonymous constructions which are compositional. For example, *under the knife* is an English idiom meaning "undergoing surgery". This idiom, illustrated in example (1), should be annotated in a way which distinguishes it from the compositional meaning in (2).

(1) The patient is *under the knife*.

(2) The napkin is *under the knife*.

Consider the problematic treatment of multiword names in the UD treebanks. The *name* relation is to be used for "proper nouns constituted of multiple nominal elements", e.g. *New York*, whereas "regular syntactic relations are used: (i) for a modifying determiner or (ii) to connect

together the words of a description or name which involve embedded prepositional phrases, sentences, etc.", e.g. *Río de la Plata*.[1] This implies that some geographical names will be retrievable by searches for *name* relations, while others will be indistinguishable from regular syntactic dependencies. Neither treebank annotators nor users wishing to retrieve names can be expected to be aware of prepositions, determiners, etc. in all foreign names that may occur in their treebank.

Principle C will allow identification of non-fixed MWEs irrespective of their surface forms and word orders. For instance, the variants of the particle verb *shut down* in examples (3) and (4) should be searchable with a single query.

(3) The company is *shutting down* the power plant.

(4) The company has *shut* the power plant *down*.

Principle D implies that, to the extent possible and depending on the MWE ontology, all MWEs belonging to certain types will be retrievable as a set, for instance, all fixed expressions, all verb-particle constructions or all verbal idioms. Some examples of good practice are the following:

- Fixed expressions (also called 'words with spaces'): In UD treebanks these are annotated with the *mwe* dependency relation. Thus, all of them can be retrieved by searching for this relation. Likewise, in BulTreeBank these can be searched for with the expression *//mw*. In NorGramBank, the same is obtained by searching for all lexical items containing spaces.
- Verb-particle constructions: In UD treebanks these are annotated with the *compound:prt* relation. In DeepBank, one can search for `[type="v_p-.*_l"]`. In NorGramBank, one can search for PRT, which is only used with particle verbs.
- Verbal idioms (including light verb constructions) tend to have regular syntax but a non-compositional meaning. It is therefore an advantage if they are semantically marked in a special way, while their syntactic buildup is also annotated. Light verb constructions are sometimes explicitly annotated with dependency relations, such as OBJ-LVC in the Szeged Dependency Treebank. The label CVC (collocational verb construction) is used in the TIGER treebank for verbal idioms, making it possible to search for all such constructions.

In conclusion, we propose some general principles for the annotation of MWEs in treebanks as a starting point for further discussion in WG4. Our proposals are aimed at allowing MWEs to be easily identified in treebanks, and also allowing types of MWEs to be retrieved.

References

Rosén, Victoria, Gyri Smørdal Losnegaard, Koenraad De Smedt, Eduard Bejček, Agata Savary, Adam Przepiórkowski, Petya Osenova, and Verginica Barbu Mititelu (2015). "A Survey of Multiword Expressions in Treebanks". In: *Proceedings of the Fourteenth Workshop on Treebanks and Linguistic Theories (TLT14)*. Ed. by Markus Dickinson, Erhard Hinrichs, Agnieszka Patejuk, and Adam Przepiórkowski. Institute of Computer Science, Polish Academy of Sciences. Warsaw, Poland, pp. 179–193.

[1] `http://universaldependencies.github.io/docs/`