Towards principles for the annotation of MWEs in treebanks – WG 4

Koenraad De Smedt and Victoria Rosén

University of Bergen, Norway

An important goal for WG4 in PARSEME is to provide guidelines for the annotation of MWEs in treebanks.

• Treebanks are valuable sources of information on MWEs.

• Few treebanks explicitly address the range of MWEs that could be annotated.

• Annotation guidelines may improve the consistency of MWE annotations within and across treebanks.

• Guidelines may also improve the ease of retrieving and studying MWEs in their syntactic context.

Previous work in WG4 has resulted in:

• An overview of existing MWE annotations in various treebanks [3].

• An exploration of the consistency of MWE annotations in UD treebanks [1].

• A preliminary proposal for general principles for MWE annotations in treebanks [2].

Proposed general principles for MWE annotation:

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

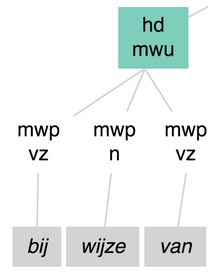
B. The annotation of noncompositional MWEs should distinguish them from homonymous strings with a compositional analysis.

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.

A Dutch example from [3]:



The recursive case of this principle is that MWEs which occur as part of other MWEs should also be annotated as such, so that embeddings of MWEs (e.g. the complex name *Johann Wolfgang Goethe–Universität Frankfurt am Main*) can be discovered.

Principle B is a corollary: ease of identification implies that MWEs should be distinguished from homonymous constructions which are compositional.

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

This is an English idiom meaning “undergoing surgery”.

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

The annotation should distinguish the idiomatic meaning in (1) from the compositional meaning in (2).

The principle of marking the distinction should not prevent a treebank from having different levels, so that on some level one may provide the same ‘regular’ syntactic analysis for (1) and (2).

Principle C will allow identification of non-fixed MWEs irrespective of their surface forms and word orders. For instance, the morphological and word order 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.

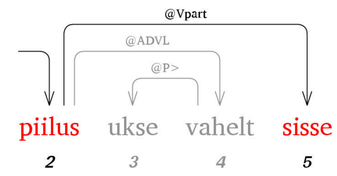
(4) The company has *shut* the power plant *down*.

In order to fulfill principle C, some normalization is recommended, i.e. each MWE occurrence in a corpus should be associated with its canonical form so as conflate different morphosyntactic variants of the same MWE. In the simplest case a canonical form is a MWE lemma, e.g. *man servant* for *men servants*. Linking to a lexicon or similar knowledge base of MWEs (e.g. DUELME) should be considered.

To the extent that a treebank is a parsed corpus, this should normally be achieved by having appropriate MWE entries in the lexicon used in parsing, as is the case in NorGramBank. Automatic lemmatization of MWEs is non-trivial in the general case, since components of a MWE lemma may not be lemmas themselves, as in *to spill the beans* but not *to spill the bean*.

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.

An Estonian example from [3]:



References

[1] Koenraad De Smedt, Victoria Rosén, and Paul Meurer. Studying consistency in UD treebanks with INESS-Search. In Markus Dickinson, Erhard Hinrichs, Agnieszka Patejuk, and Adam Przepiórkowski, editors, *Proceedings of the Fourteenth Workshop on Treebanks and Linguistic Theories (TLT14)*, pages 258–267, Warsaw, Poland, 2015. Institute of Computer Science, Polish Academy of Sciences.

[2] Victoria Rosén, Koenraad De Smedt, Gyri Smørdal Losnegaard, Eduard Bejček, Agata Savary, and Petya Osenova. MWEs in treebanks: From survey to guidelines. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia, 2016. ELRA.

[3] 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. In Markus Dickinson, Erhard Hinrichs, Agnieszka Patejuk, and Adam Przepiórkowski, editors, *Proceedings of the Fourteenth Workshop on Treebanks and Linguistic Theories (TLT14)*, pages 179–193, Warsaw, Poland, 2015. Institute of Computer Science, Polish Academy of Sciences.