Towards guidelines 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. Earlier work in WG4 has been aimed towards creating an overview of current treebank annotations as well as towards general principles that assure retrieval of MWEs and MWE types (Rosén et al., 2016). For the PARSEME Shared Task, guidelines have been produced for the annotation of MWEs in flat tokenized text.[1] We take these previous efforts as a starting point for formulating more specific guidelines for the annotation of MWEs in treebanks. Only a few of these can be discussed in this abstract.

Whereas the shared task guidelines "do not annotate the internal syntactic structure", such structure is normally annotated in treebanks, with the possible exception of fixed expressions as defined in Sag et al. (2002). Treebanks thereby do not only provide a richer annotation, but also a less ambiguous one. In the examples provided for the shared task guidelines,[2] the string *There is little doubt* is annotated such that *There*, *is* and *doubt* make up the (minimal) MWE, whereas *little* is left unannotated. This leaves it unclear whether *little* is syntactically unrelated, making the MWE truly discontinuous, or whether it is syntactically integrated and modifies the MWE. In a treebank, the latter would be indicated in the annotation of the example's grammatical structure.

Only the minimal phrase that cannot be substituted with other lexical items should be considered a MWE. The shared task guidelines list *to come off with flying colors* as an idiom. However, *with flying colors* can occur with many other verbs with the same idiomatic meaning, e.g. *He passed the exam with flying colors*, *The team won with flying colors*, *The bill passed the Senate with flying colors*, etc. The suggested guideline to annotate the minimal phrase as a MWE is relevant for annotation of flat text as well as grammatical structures. This does not prevent the annotation of MWEs in which other MWEs are embedded.

To the extent that treebanks can have multiple levels of structure, we suggest that idioms (which have meanings that cannot be derived compositionally) ideally be represented at two levels, one that reflects the idiomatic meaning, and one that represents the internal syntactic structure. This can be achieved in different ways depending on the grammar formalism. We show two examples. In some versions of dependency/constituency annotation, secondary edges can be used, as exemplified in the description of the Eukalyptus treebank.[3] As shown in Figure 1, a sentence node (S) contains a subject, head and object. In addition, the head dominates a

[1] http://typo.uni-konstanz.de/parseme/images/shared-task/guidelines/PARSEME-ST-annotation-guidelines-v5.pdf

[2] http://typo.uni-konstanz.de/parseme/images/shared-task/pilot-annotation/PARSEME-shared-task-pilot-annotation-format-sample.txt

[3] http://clarino.uib.no/iness/page?page-id=euk-vpid

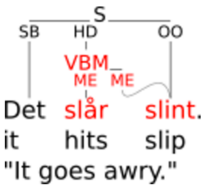

Figure 1: Analysis of *Det slår slint.* “It goes awry.” in the Swedish Eukalyptus treebank

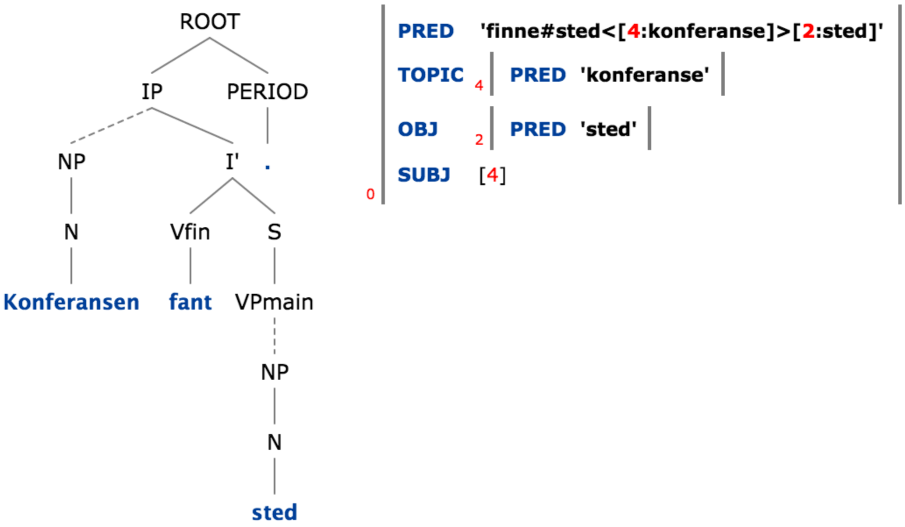

Figure 2: Analysis of *Konferansen fant sted.* “The conference took place.” in NorGramBank

multiword verb node (VBM), and secondary edges (labeled “ME”) are used to connect the remaining MWE parts to this node. Associated with the multiword node is a semantic identifier for the idiomatic sense.

In the LFG formalism, there are two levels of syntactic structure, c-structure and f-structure. Figure 2 shows an example from the Norwegian treebank NorGramBank,[4] in which the verbal idiom *finne sted* “take place / occur” is represented as a combined predicate in the PRED attribute in the f-structure on the right. The new predicate name “finne#sted” is built by concatenating the verb predicate and the object predicate. This predicate has only the subject as a semantic argument; the object argument is outside the angled brackets, indicating that it is only a syntactic and not a semantic argument of the predicate. The c-structure represents the internal constituent structure of the MWE as shown on the left of Figure 2. It reflects the flexibility of the expression by representing each component of the MWE as a separate node.

References

Rosén, V., K. De Smedt, G. S. Losnegaard, E. Bejček, A. Savary, and P. Osenova (2016). MWEs in treebanks: From survey to guidelines. In N. Calzolari, K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, Portorož, Slovenia, pp. 2323–2330. ELRA.

Sag, I., T. Baldwin, F. Bond, A. Copestake, and D. Flickinger (2002). Multiword expressions: A pain in the neck for NLP. In *Lecture Notes in Computer Science. Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, Volume 2276, pp. 189–206. Springer.

[4] `http://clarino.uib.no/iness/page?page-id=iness-vpid`