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

Koenraad De Smedt and Victoria Rosén

University of Bergen, Norway

An objective of WG4 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.

• 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].

The PARSEME Shared Task has developed guidelines for the annotation of MWEs in flat tokenized text.

• These guidelines offer a classification of verbal MWEs (VMWEs) together with linguistic tests for their identification and categorization.

• The tests and decision trees developed for the shared task are valuable aids for the development of guidelines for treebank annotation.

• There are, however, important differences in the annotation of flat text and annotation in treebanks.

The internal syntactic structure of the VMWEs is not annotated for the shared task.

Example: *Delegates are in little doubt…*

1	Delegates	
2	are	1 LVC
3	in	1
4	little	
5	doubt	1

• It is unclear how *little* relates to the VMWE consisting of the words marked by the number 1.

• In fact, *little* is syntactically related and modifies the VMWE.

• A treebank should show this relation.

Only the minimal phrase that cannot be substituted with other lexical items should be considered a MWE.

• The shared task guidelines list *come off with flying colors* as an idiom.

• However, *with flying colors* can occur with many other verbs with the same idiomatic meaning:

- *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.

Idioms, which have meanings that cannot be derived compositionally, should if possible be represented at two levels:

1. one level that reflects the idiomatic meaning, and
2. one that represents the internal syntactic structure.

This can be achieved in different ways depending on the grammar formalism. We show two examples below.

In some dependency/constituency treebanks, secondary edges can be used.

• The example below is from the Eukalyptus treebank.

• The sentence node (S) contains a subject, head and object.

• The head dominates a 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(onstituent)-structure and f(unctional)-structure.

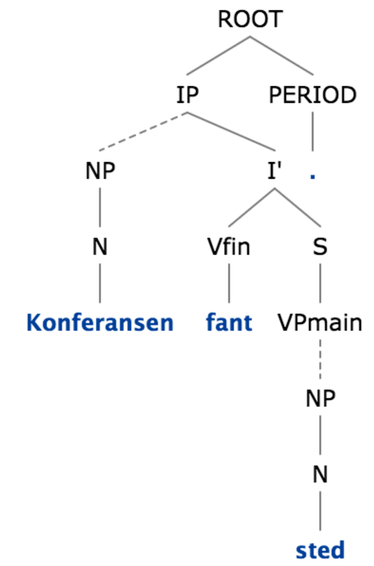
• The figure below shows an example from the Norwegian treebank NorGramBank.

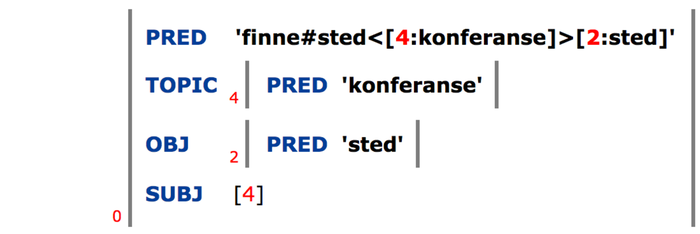
• The verbal idiom *finne sted* “take place / occur” is represented as a combined predicate in the PRED attribute in the f-structure on the right.

• 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 the figure; it reflects the flexibility of the expression by representing each component of the MWE as a separate node.





Eukalyptus:
`http://clarino.uib.no/iness/page?page-id=euk-vpid`

NorGramBank:
`http://clarino.uib.no/iness/page?page-id=iness-vpid`

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 Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2323–2330, 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.