Towards guidelines for the annotation of MWEs in treebanks (WG4)

Koenraad De Smedt and Victoria Rosén

A main goal for WG4 is to develop guidelines for the annotation of MWEs in treebanks. Earlier work in WG4 has resulted in some basic principles for annotation.

The PARSEME Shared Task has developed guidelines for the annotation of verbal MWEs. Their linguistic tests for identification and categorization are useful for us, but there are important differences between annotating flat text and annotation of syntactic structure.

We propose some new treebank annotation guidelines:

- Only the minimal phrase should be annotated as a MWE.
- Syntactic/semantic structure should show how additional elements (such as modifiers) interact with MWEs.
- 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 (often regular) syntactic structure.