MWEs in Universal Dependency treebanks

Koenraad De Smedt*, Victoria Rosén*† and Paul Meurer†
University of Bergen*, Uni Research Computing†

1 Background

One of the goals of **Parseme WG4** is to create annotation guidelines for representing MWEs in treebanks. So far, the collection of examples from treebank annotations has shown that there is considerable variation in how MWEs are annotated. In our poster we demonstrate how treebank search may be helpful in examining the consistency with which annotations are applied.

As a case study, we have used the second release of annotated treebanks in Universal Dependencies (UD) v1.1, which has recently become available. These treebanks represent an effort to create similarly annotated treebanks across many languages (McDonald et al., 2013).

We have imported the UD treebanks for 17 different languages into the INESS treebanking infrastructure (Rosén et al., 2012). Since these treebanks have a common annotation format, they can all be searched simultaneously with INESS-Search (Meurer, 2012). This makes it possible to efficiently study to what degree certain constructions are annotated in a parallel way across different languages in the UD treebanks, and also to what degree this annotation is consistently applied within each treebank.

2 Annotation Guidelines for the UD treebanks

The online annotation guidelines[1] do not provide a separate treatment of MWE annotation in UD, but they do say that there are three dependency relations that may be used for compounding: *compound*, *mwe* and *name*. It is not clear whether all *compound* constructions are to be considered MWEs in UD, but at least one subtype clearly is: phrasal verbs. These are to be annotated with the *compound:prt* dependency relation (where *prt* stands for *particle*). For example, in English *shut down* the *compound:prt* relation holds between the verb and its particle.

The *name* relation is to be used for "proper nouns constituted of multiple nominal elements", for example *Hilary Rodham Clinton*. The structure is flat, with all words modifying the first one using the *name* label. This annotation is only to be used when there is no clear syntactic modification structure, in which case regular syntactic relations are used, as in *the king of Sweden*, where *king* is the head and is modified by *the* through a *det* relation and *Sweden* through an *nmod* relation, while *of* modifies *Sweden* through a *case* relation.

The *mwe* dependency relation is to be used for roughly the category of *fixed expressions* (Sag et al., 2002), with the exception of relations that should be annotated with the *compound* or *name* labels. The annotation is a "flat, head-initial structure, in which all words in the expression modify the first one using the mwe label". An example is *as well as*, where *as* is the head and the other words are dependent on it through *mwe* relations.

3 Searches for MWEs in the UD treebanks

A search for *mwe* dependencies involving only two words resulted in 2752 match types, 11996 match tokens in all UD treebanks. Of the latter, there are 8860 adjacent words which have the correct dependency direction and 2026 which have the incorrect direction according to the guidelines. The remaining 1110 involve non-adjacent words, as in Swedish *för ... skull* "for the sake of …". A *mwe* dependency between non-adjacent words may indicate possible errors, e.g. German *wie auch auf* "as well as on" having a *mwe* dependency between the first and third words, whereas

[1] `http://universaldependencies.github.io/docs/`, consulted on June 30, 2015.

the first and second words are more likely candidates for a fixed expression. In the annotation of Spanish *ya que* "since", *ya* dominated *que* 78 times, whereas the opposite, which does not comply to the guidelines, was found 77 times.

Search results for MWEs consisting of more than two words (i.e. involving at least two *mwe* dependencies) suggest that many of these seem to be annotated according to the guidelines. However, the only German MWE consisting of three words, *nach wie vor* "still", has *vor* as the head.

Proper nouns constituted of multiple nominal elements were searched for with the dependency *name*. A frequent example across treebanks is *New York*, with 50 occurrences where *New* dominates *York* in Danish, Croatian, Indonesian, Italian and Swedish, in agreement with the guidelines, while 20 occurrences with the opposite dependency direction were found in German and Spanish.

The *name* relation does not occur in all treebanks, however. The English treebank uses the *compound* relation for names, with the last element as the head. Thus, there is a *compound* dependency from *York* to *New*. In the Greek treebank, there is an *amod* dependency to *Νέα* "New" from *Υόρκη* "York" and there is an *nmod* dependency to *Χίλαρι* "Hillary" from *Κλίντον* "Clinton". The Italian treebank annotates some names, e.g. *Scènes de la Vie privée*, with a mixture of *mwe* and *name* relations, which does not seem consistent with the guidelines and creates difficulties for searching multiword names as separate from other MWEs.

Modifiers following names, such as titles and appositions, are sometimes treated as part of the name, sometimes not. In the Danish treebank, the last word in the string *Stefan Fryland, formand* "Stefan Fryland, leader" is annotated with a *name* dependency, whereas a similar construction in Spanish is annotated with the *appos* dependency, e.g. *Jerónimo Martín Caro y Cejudo [...], humanista* "…, humanist". German uses *appos* dependencies for modifiers preceding names, e.g. for *Inhaber Michael Walther* "Proprietor …", whereas Swedish uses the *det* dependency for *professor* or for *författaren* "the author" preceding a name. The use of these relations deserves further investigation as it may involve additional distinctions.

Phrasal verbs with particles were searched for by means of the *compound:prt* dependency relation which according to the documentation is only used in the annotation of English, German and Swedish. This dependency relation is, however, not found in the German treebank, whereas it is found in other treebanks (Danish, Finnish, Farsi, Irish). In the German treebank, the dependency relation *mark* is used for such constructions, e.g. *teilte ... mit* "informed". This is not in accordance with the way the relation *mark* is described in the guidelines: "the word introducing a finite clause subordinate to another clause". The Hungarian treebank seems to use *compound:preverb* for the particle verb relation.

Phrasal verb constructions can be discontinuous, e.g. *blow something up*. A search for only discontinuous phrasal verbs in the Swedish treebank showed 137 match types, 162 matches. The latter represents 18% of the total number of phrasal verb occurrences for Swedish, which is an interesting finding in itself. In the Danish treebank, the dependency relation *compound:prt* is not only used for phrasal verbs, but also between the elements of (discontinuous) circumpositions such as the frequent *for ... siden*, as in *for to år siden* "two years ago".

4 Conclusion

We have have reported on a pilot study which addresses the goals of WG4. Our objective has not been to identify all MWEs in the UD treebanks, nor to provide alternative recommendations for the UD treatment of MWEs. We have searched for some specific dependencies according to the UD annotation guidelines. Our findings so far indicate that the annotations of MWEs in the various UD treebanks show important similarities to each other, but also apparent differences with respect to their adherence to the guidelines. We also found seemingly unmotivated discrepancies even within treebanks. We believe that INESS may be useful to treebank authors as an aid to achieving consistency in treebank annotation, and to researchers wishing to consult treebanks in order to find MWEs.

References

McDonald, Ryan, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee (2013). "Universal Dependency Annotation for Multilingual Parsing". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, pp. 92–97.

Meurer, Paul (2012). "INESS-Search: A search system for LFG (and other) treebanks". In: *Proceedings of the LFG '12 Conference*. Ed. by Miriam Butt and Tracy Holloway King. LFG Online Proceedings. Stanford, CA: CSLI Publications, pp. 404–421.

Rosén, Victoria, Koenraad De Smedt, Paul Meurer, and Helge Dyvik (2012). "An Open Infrastructure for Advanced Treebanking". In: *META-RESEARCH Workshop on Advanced Treebanking at LREC2012*. Ed. by Jan Hajič, Koenraad De Smedt, Marko Tadić, and António Branco. Istanbul, Turkey, pp. 22–29.

Sag, Ivan, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan 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*. Vol. 2276. Springer, pp. 189–206.