MWEs in Universal Dependency treebanks – WG 4

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

One of the goals of PARSEME is to provide guidelines on the annotation of MWEs in treebanks. Some suggestions:

- MWEs should be annotated consistently so as to promote their searchability (also across languges)
- individual MWEs should be searchable even if they are variable in form and discontinuous
- types of MWEs should be searchable based on their characteristics

What are possible tools for checking the annotation of MWEs in existing treebanks for many languages?

INESS (INfrastructure for the Exploration of Syntax and Semantics) offers INESS-Search, a powerful and efficient tool suitable for searching treebanks with LFG, HPSG, constituency and dependency annotations.

As a case study, we have used the second release of annotated treebanks in Universal Dependencies (UD) v1.1. These are similarly annotated across many languages [1].

We have imported the UD treebanks into INESS [3]. They can all be searched simultaneously with INESS-Search [2]. This makes it possible to study to what degree they are annotated in a parallel way.

Annotation guidelines for UD include the following relations:

- **compound:prt** for particle verbs: *shut down*. Can be discontinuous (*shut it down*).
- **name** for proper nouns with multiple elements (flat, head-initial): *Hillary Rodham Clinton*. When there is a syntactic relation with a name (*the king of Sweden*), regular syntactic relations can be used.
- **mwe** for fixed expressions (flat, head-initial) not covered by the previous: *as well as*

To what extent have these guidelines been consistently applied? With INESS-Search we have had a first look at the UD treebanks.

Searching for "mwe" (fixed expressions)

Searching simple dependencies (or dominance) is easy in INESS:

```
#x >mwe #y
```

Searching mwes consisting of two words only:

```
#x >mwe #y &!(#x >mwe #z & #z != #y)
```

Searching head-initial binary dependency relations:

```
#x >mwe #y & #x . #y &!(#x >mwe #z & #z != #y)
```

The previous expression produces 8860 matches (1813 types) in all UD treebanks which are in accordance with the guidelines.

Searching binary dependency relations where the second word dominates the first:

```
#x >mwe #y & #y . #x &!(#x >mwe #z & #z != #y)
```

This expression produces 2026 matches (308 types) which are not head initial, and are therefore not in accordance with the guidelines.

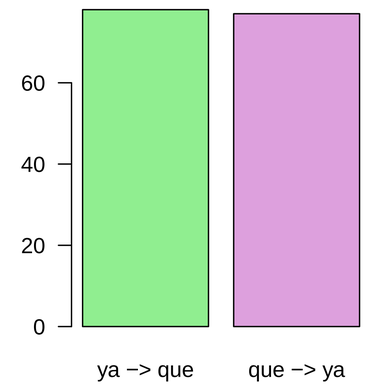
Ya que "since", a frequent fixed expression in the Spanish treebank, is annotated inconsistently with respect to its head:

```
#x:[word="ya"] >mwe #y:[word="que"] & #x . #y
```

produces 78 matches.

```
#y:[word="que"] >mwe #x:[word="ya"] & #x . #y
```

produces 77 matches.



Searching for "name" (names with multiple elements)

INESS gives nice tabular overviews across languages, where *lang* is a metadata parameter:

```
#x:[word="New|York"] >name #y:[word="New|York"]::lang
```

Count	#x: word	#y: word	globals: lang
34	New	York	ita
19	York	New	deu
12	New	York	ind
2	New	York	dan
1	York	New	spa
1	New	York	hrv
1	New	York	swe

Other observations:

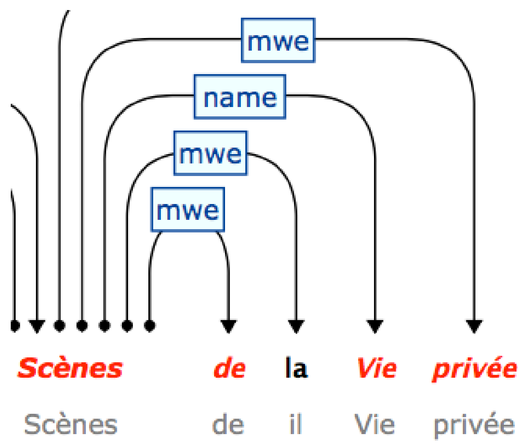
- The English treebank uses *compound* for names
- The Greek treebank has only dependencies based on regular syntactic relations:

Νέα "New" ← amod ← *Υόρκη* "York"
Χίλαρι "Hillary" ← nmod ← *Κλίντον* "Clinton"

- The Spanish treebank sometimes has *name*, sometimes regular syntactic relations:

Les "The" ← name ← *Pieux* "Pieux" (Les Pieux)
los *Países* "Countries" → amod → *Bajos* "Low" (The Low Countries)

- The Italian treebank has combinations of different relations, e.g. for the title *Scènes de la Vie privée*:

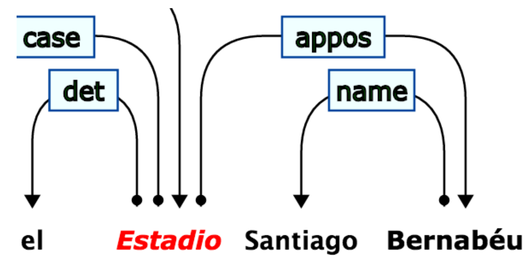


Modifiers of names, such as titles, professions, occupations, and other descriptions, are sometimes treated as part of the name, sometimes not. The following relations are examples which were found for modifiers following names:

- Danish treebank: *name*
- Spanish treebank: *appos* (apposition)

The following relations are examples which were found for modifiers preceding names:

- Swedish treebank: *det* (determiner)
- German, Spanish treebanks: *appos* (apposition)



Searching for "compound:prt" (phrasal verbs) can be done with the following expression:

```
#x >compound:prt #y
```

UD guidelines say: this relation is used in the annotation of English, German and Swedish.

But it is found in the Danish, English, Finnish, Farsi, Irish and Swedish treebanks.

The German treebank has instead *mark*, e.g. for *teilte … mit* "informed"

The Hungarian treebank seems to use *compound:preverb* instead.

In the Danish treebank, *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".

Discontinuous phrasal verbs can be searched with the following expression:

```
#x >compound:prt #y & !(#x . #y)::lang
```

In the Swedish treebank, for instance, there are 162 matches (137 types). These matches represent 18% of the total number of phrasal verb occurrences for Swedish (the remainder being continuous).

In conclusion, INESS-Search, as a part of the INESS infrastructure, is a useful tool for searching many treebanks at the same time.

It can be used for searching MWEs if they are properly annotated as such, and for checking the consistency of annotations both within and across treebanks.

Visit us at http://clarino.uib.no/iness

References

[1] Ryan McDonald, 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. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

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

[3] Victoria Rosén, Koenraad De Smedt, Paul Meurer, and Helge Dyvik. An open infrastructure for advanced treebanking. In Jan Hajič, Koenraad De Smedt, Marko Tadić, and António Branco, editors, *META-RESEARCH Workshop on Advanced Treebanking at LREC2012*, pages 22–29, Istanbul, Turkey, 2012.

UD treebank documentation:
http://universaldependencies.github.io/docs/
The treebanks and their documentation were consulted on June 30, 2015.