MWEs in Universal Dependencies 1.3 (WG4)

1 Introduction

Universal Dependencies[1] [3] is a project providing harmonized morphological and syntactic annotation in 40 languages. The annotation scheme is based on (universal) Stanford dependencies [2], Google universal part-of-speech tags [4], and the Interset interlingua for morphosyntactic tagsets [5]. Because the treebanks were mostly developed independently, it is not straightforward to come up with one universal style, that is why annotation is not always consistent.

Multiword expressions in UD are associated with one of the three dependency relations: `mwe`, `compound` or `name`, see [1].

From the perspective of natural language understanding, the most interesting MWEs are idioms with non-compositional semantics. The current UD annotation does not delve into semantics so deeply: idiomatic expressions are usually analyzed only at the level of their surface syntax. Nevertheless, the UD relation `mwe` is used for the more frozen expressions, often corresponding to function words in other languages. Here the special annotation increases parallelism across languages, especially in expressions containing nouns, which would be otherwise treated as content words. Unfortunately the current UD guidelines are not very specific about what expressions should be annotated this way.

In [1], the authors presented basic statistics on `mwe` over several selected treebanks from UD 1.2, giving many examples of inconsistency of MWE annotation for different treebanks. We provide a more detailed analysis of the `mwe` relation in the latest version of UD, 1.3, showing the statistics over the most frequent POS patterns that `mwe`-annotated tokens tend to have.

2 Statistics of MWEs in UD 1.3

The statistics were acquired from UD 1.3 using the platform Treex[2] to parse the trees.

Table 1 provides statistics on non-unique POS sequences for MWEs in selected languages.

pos sequence	bg	ca	cs	da	en	es	fr	hr	it	pl	pt	ro	ru	sl	sv
ADP NOUN	8	611	1510	236	26	187	170	2	7	5	3	759	0	16	296
ADV ADP	1	162	330	54	30	874	1067	10	434	17	404	232	79	0	21
ADP NOUN ADP	84	986	811	24	0	49	85	8	5	0	370	212	0	0	175
ADP ADP	66	416	283	0	149	11	128	0	156	0	4	544	0	0	0
ADV SCONJ	79	543	11	0	6	115	539	28	83	0	28	51	0	123	91

Table 1: The five most frequent pos sequences of mwe in selected treebanks

It can be seen that the annotation is not consistent among the languages, it can be due to several reasons. Firstly, it is the difference between the languages themselves, when a MWE in one language corresponds to a single word in another language. Secondly, it is the decision on what should or should not been analyzed using the `mwe` relation: which constructions are fixed enough to be grouped together and which can be treated separately, according to their surface syntax.

In order to illustrate this problem, we have analyzed one POS sequence across various languages: `ADP NOUN ADP` (`ADP` stands for "adposition", i.e. either preposition or postposition). We suppose that related languages are also similar with respect to MWE, so we are especially interested in comparison of languages within the same family.

The largest set of instances of the `ADP NOUN ADP` pattern can be found in the Czech UD treebank (*na rozdíl od* – 'in contrast to, unlike'). It is also annotated in Bulgarian *za razlika ot* or in Croatian *za razliku od*. However, the corresponding expression in Russian *v otlichie ot* is analysed in a different way; in the other Slavic treebanks (Polish, Slovenian) it is not marked with the `mwe` relation at all.

[1] `http://universaldependencies.org/`

[2] `http://ufal.mff.cuni.cz/treex`

Somewhat more consistent is the annotation of this pattern in Romance languages. The multi-word preposition lit. *for reason of* – 'because of' has the same analysis in all the Romance treebanks except Romanian: *por causa de* (Portuguese), *a causa de* (Catalan, Spanish), *à cause de* (French), and *a causa di* (Italian) are all treated as multi-word expressions.

However, in Scandinavian languages the annotation is not that consistent. Even though Swedish, Danish and Norwegian are closely related, Danish has nine times fewer different MWEs than Swedish, and Norwegian has none. Specifically for the `ADP NOUN ADP` pattern, the difference is even more pronounced: Swedish has 49 unique expressions, Danish only 2. Such a large disproportion can hardly be attributed to genre differences alone. Both of the Danish MWE have correlates with Swedish ones (e.g. *på grund af* – 'because of' (da) vs. *på grund av* (sv)). The other multiword prepositions that are marked as 'mwe' in Swedish are connected using different relations in Danish.

3 Conclusion

The UD relation `mwe` is not used consistently in the current release of Universal Dependencies. While part of the issue may be caused by true linguistic differences, we demonstrate on closely related languages that it is not always the case; even literally equivalent expressions do not always receive the same analysis. Obviously, it would be beneficial and in accord with the UD goals if the UD treebanks converge much more. A better cross-linguistic definition of the `mwe` relation would surely help but there is probably no good way of constraining the set only with language-independent rules. Quite likely the annotators of the source treebanks (later converted to UD) had to enumerate the MWEs as lists. We suggest to compare these lists using POS patterns and harmonize the treebanks bottom-up: first try to make sure that similar expressions in related languages are treated the same way, then proceed to more distant languages, as far as possible.

References

[1] Koenraad De Smedt, Victoria Rosén, and Paul Meurer. Studying consistency in ud treebanks with iness-search. In *14th International Workshop on Treebanks and Linguistic Theories (TLT 2015)*, pages 258–267, 2015.

[2] Marie-Catherine De Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. Universal stanford dependencies: a cross-linguistic typology. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavík, Iceland, may 2014. European Language Resources Association (ELRA).

[3] Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666, Portorož, Slovenia, 2016. European Language Resources Association.

[4] Slav Petrov, Dipanjan Das, and Ryan McDonald. A universal part-of-speech tagset. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2089–2096, İstanbul, Turkey, 2012. European Language Resources Association.

[5] Daniel Zeman. Reusable tagset conversion using tagset drivers. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Language Resources and Evaluation Conference, LREC*

2008, pages 28–30, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA).