Universal Dependencies: A Cross-Linguistic Perspective on Grammar and Lexicon

Joakim Nivre
Department of Linguistics and Philology
Uppsala University

joakim.nivre@lingfil.uu.se

Abstract

Universal Dependencies is an initiative to develop cross-linguistically consistent grammatical annotation for many languages, with the goal of facilitating multilingual parser development, cross-lingual learning and parsing research from a language typology perspective. It assumes a dependency-based approach to syntax and a lexicalist approach to morphology, which together entail that the fundamental units of grammatical annotation are words. Words have properties captured by morphological annotation and enter into relations captured by syntactic annotation. Moreover, priority is given to relations between lexical content words, as opposed to grammatical function words. In this position paper, I discuss how this approach allows us to capture similarities and differences across typologically diverse languages.

1 Introduction

Multilingual research on syntax and parsing has for a long time been hampered by the fact that annotation schemes vary enormously across languages, which has made it very hard to perform sound comparative evaluations and cross-lingual learning experiments. The basic problem is illustrated in Figure 1, which shows three parallel sentences in Swedish, Danish and English, annotated according to the guidelines of the Swedish Treebank (Nivre and Megyesi, 2007), the Danish Dependency Treebank (Kromann, 2003), and Stanford Typed Dependencies (de Marneffe et al., 2006), respectively. The syntactic structure is identical in the three languages, but because of divergent annotation guidelines the structures have very few dependencies in common (in fact, none at all across all three languages). As a result, a parser trained on one type of annotation and evaluated on another type will be found to have a high error rate even when it functions perfectly.

Universal Dependencies (UD) seeks to tackle this problem by developing an annotation scheme that makes sense in a cross-linguistic perspective and can capture similarities as well as idiosyncracies among typologically different languages. However, the aim is not only to support comparative evaluation and cross-lingual learning but also to facilitate multilingual natural language processing and enable comparative linguistic studies. To serve all these purposes, the framework needs to have a solid linguistic foundation and at the same time be transparent and accessible to non-specialists. In this paper, I discuss the basic principles underlying the UD annotation scheme with respect to grammar and lexicon. A more general introduction to UD can be found in Nivre et al. (2016) and on the project website.[1]

2 Grammatical Relations and Content Words

The UD annotation scheme is based on *dependency*, which means that it focuses on grammatical relations between linguistic units, rather than on the internal constituent structure of these units. In this respect, it adheres to the language typology tradition, where concepts like *subject* and *object*, although far from controversial as language universals, have proven more useful than notions of constituency in cross-linguistic investigations.[2]

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

[2] See, for example, the World Atlas of Language Structures (WALS) at http://wals.info.

Proceedings of the Workshop on Grammar and Lexicon: Interactions and Interfaces,
pages 38–40, Osaka, Japan, December 11 2016.

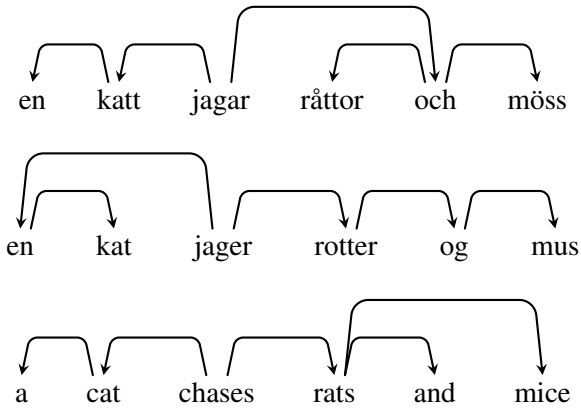

Figure 1: Divergent annotation of Swedish (top), Danish (middle) and English (bottom).

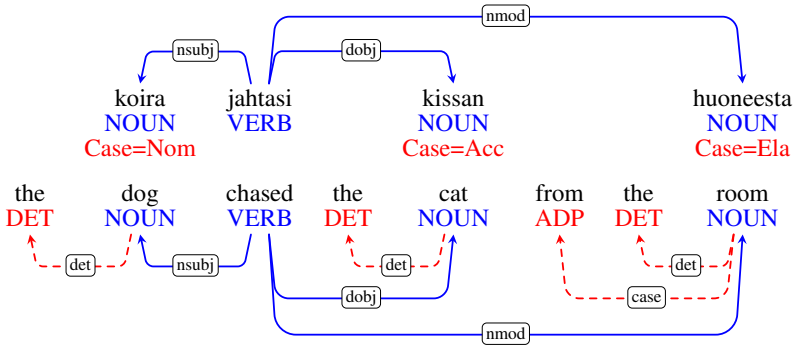

Figure 2: Simplified UD annotation for equivalent sentences in Finnish (top) and English (bottom).

The UD annotation scheme also subscribes to a version of *lexicalism*, which means that the units that enter into grammatical relations are words, more precisely lexical words (or content words), which can be assumed to be more constant across languages. By contrast, function words and bound morphemes are treated as part of the grammatical structure. The former are attached to the lexical word they modify with special functional relations. The latter are captured by morphological features associated with words in a holistic fashion.

The UD annotation scheme is illustrated in Figure 2 with two parallel sentences from Finnish (top) and English (bottom). In both languages, the sentence consists of a single verb and three nouns that act as nominal subject (*nsubj*), direct object (*dobj*) and nominal modifier (*nmod*) of the verb, respectively. What differs is primarily the grammatical encoding of nominals in the two languages. In English, all nouns have a definite article acting as determiner (*det*); *room* in addition is accompanied by the preposition *from*, which is analyzed as a case marker (*case*) indicating that it is not a core argument. In Finnish, no noun is specified by a function word, but all nouns have a morphological case inflection, which shows up as a morphological feature on the noun.[3]

3 Conclusion

The UD project tries to provide cross-linguistically consistent grammatical annotation for typologically diverse languages. To capture similarities and differences across languages, UD uses a representation consisting of three components: (i) dependency relations between lexical words; (ii) function words modifying lexical words; and (iii) morphological features associated with words. This system has so far been applied successfully to over 50 languages.

[3] In both languages, nouns and verbs have additional features that have been suppressed here to highlight the contrast between the two languages.

References

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*.

Matthias Trautner Kromann. 2003. The Danish Dependency Treebank and the DTAG treebank tool. In *Proceedings of the 2nd Workshop on Treebanks and Linguistic Theories (TLT)*, pages 217–220.

Joakim Nivre and Beáta Megyesi. 2007. Bootstrapping a Swedish treeebank using cross-corpus harmonization and annotation projection. In *Proceedings of the 6th Workshop on Treebanks and Linguistic Theories (TLT)*, pages 97–102.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Dan Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*.