Universal Dependencies for Swedish

Joakim Nivre

Uppsala University
Department of Linguistics and Philology
joakim.nivre@lingfil.uu.se

1. Introduction

Natural language parsing has made tremendous progress during the last twenty years thanks to the availability of syntactically annotated corpora, or treebanks. Treebanks can be used for statistical learning as well as evaluation and are available for an increasing number of languages. However, the annotation schemes used for different languages show considerable variation, mainly due to the existence of language-specific grammatical traditions but also to different theoretical inclinations among treebank developers. This makes it hard to port existing parsers to new languages and undermines the comparability of parsing results across languages (Nivre et al., 2007). This problem has been highlighted in particular in recent work on cross-lingual learning (McDonald et al., 2011; Naseem et al., 2012).

There have been several initiatives recently that try to come to terms with these problems by creating data sets with cross-linguistically consistent syntactic annotation, either by converting existing treebanks or by annotating new data, or both (Zeman et al., 2012; McDonald et al., 2013; Tsarfaty, 2013). In this paper, we report work in progress within the project Universal Dependencies,[1] which seeks to unify several recent proposals into a single coherent framework based on the universal Stanford dependencies (de Marneffe et al., 2006; de Marneffe et al., 2014), the Google universal part-of-speech tags (Petrov et al., 2012), and the Interset inventory of morphosyntactic features used in the HamleDT treebanks (Zeman, 2008; Zeman et al., 2012). More specifically, we discuss how the dependency version of the Swedish Treebank (Nivre and Megyesi, 2007) can be converted to the new standard.

2. Morphological Annotation

The morphological annotation consists of a coarse-grained part-of-speech tag and a set of morphological features. The part-of-speech tag set is a revised version of the universal Google tags (UGT) with five new categories: auxiliary verb (AUX), interjection (INTJ), proper noun (PROPN), subordinating conjunction (SCONJ), and symbol (SYM). The inventory of morphological features currently comprises a subset of the Interset inventory used in the HamleDT treebanks, but it is likely to grow as data from more languages get annotated and converted. In addition, it is possible to add language-specific features if needed. Figure 1 (bottom) illustrates the morphological annotation for a sentence from the Swedish Treebank.

[1] http://universaldependencies.github.io/docs/

SUC	UGT	SUC	UGT
AB	ADV	PC	ADJ
DL	PUNCT	PL	{ADP, ADV, ...}
DT	DET	PM	PROPN
HA	ADV	PN	PRON
HD	DET	PP	ADP
HP	PRON	PS	{DET, PRON}
HS	DET	RG	NUM
IE	PART/SCONJ	RO	ADJ
IN	INTJ	SN	SCONJ
JJ	ADJ	UO	X
KN	CONJ	VB	{AUX, VERB}
NN	NOUN		

Table 1: Mapping from SUC to UGT

The Swedish treebank uses the part-of-speech tag set from the Stockholm-Umeå Corpus (SUC) (Ejerhed et al., 1992), which consists of 23 base tags and a rich set of morphosyntactic features. Mapping the base tags to Google tags is mostly straightforward, but there are a few tricky cases. First of all, infinitive markers (IE) can conceivably be treated either as subordinating conjunctions (SCONJ) or grammatical particles (PART). Secondly, there are a few cases where SUC categories need to be split into multiple UGT categories depending on lemma and/or syntactic function. This includes possessives (PS), which need to be split into determiners (DET) and pronouns (PRON), and verbs (VERB), which have to be divided into auxiliary verbs (AUX) and lexical verbs (VERB). Finally, UGT does not recognize verb particles as a part of speech (only as a syntactic function), so the SUC category PL needs to be split into ADP, ADV, etc. When converting the Swedish Treebank, these cases can be handled adequately by taking the syntactic annotation into account. However, if we want to use the universal representation in practical applications, we also need to find a way of mapping the output of a tagger trained on SUC to the universal morphological representation. Table 1 illustrates the basic correspondences between the two tag sets.

3. Syntactic Annotation

The syntactic annotation is based on the universal Stanford dependencies (USD), consisting of 42 grammatical relations that are widely attested across typologically different languages. A basic assumption in this scheme is that dependency relations hold primarily between content words, while function words are pushed to the bottom of the trees and attached in a flat structure to the content word with which they are most closely associated. This principle is

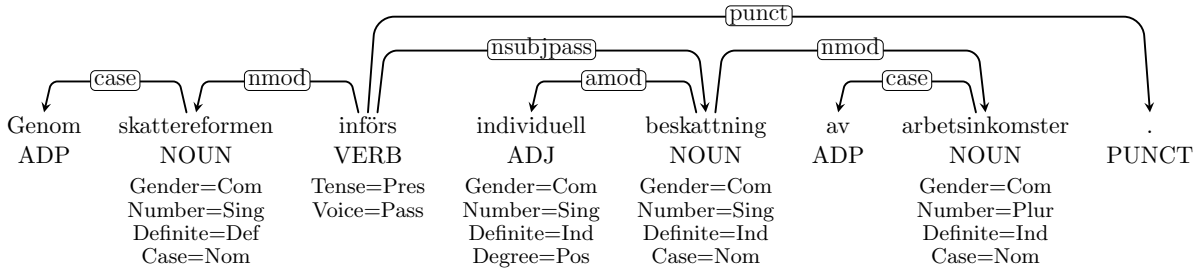

Figure 1: Swedish sentence with morphological and syntactic annotation

Label Mapping
*A → {advmod, advcl}
AN → appos
AT, XT → amod
DT → det
EO → {dobj, xcomp, ccomp}
ES → {nsubj, csubj}
ET → {nmod, acl, . . . }
FO, FS → expl
FP, PT → acl
I*, J* → punct
IM → mark
IO → iobj
NA → neg
OO → {dobj, xcomp, ccomp}
OP, VO, VS → xcomp
PL → prt
SS → {nsubj, csubj, nsubjpass, csubjpass}
TT → vocative
UK → mark
XF → dislocated
XX → {foreign, . . . }
YY → dislocated
Structural Conversion
Coordination: conjunction → first conjunct
Prepositional phrases: preposition → noun
Verb groups: finite auxiliary → main verb
Copula constructions: copula → predicate

Table 2: Conversion from MAMBA to USD

enforced to maximize parallelism across languages, since content words and their relations are more likely to be similar across languages, while function words in one language often correspond to morphological inflection (or nothing at all) in other languages. Figure 1 (top) illustrates the syntactic annotation for a Swedish sentence.

The Swedish Treebank uses a dependency annotation where most relations are inherited from the MAMBA annotation scheme used in the original version of Talbanken (Teleman, 1974). For the majority of grammatical constructions, the two schemes assume the same structure, which means that converting from MAMBA to USD is just a matter of mapping from one label set to the other. The mapping may be one-to-one or many-to-one, which is unproblematic, but there are also a number of cases where USD makes finer distinctions and where the mapping therefore has to be context-sensitive. In most cases, this is due to the conscious design decision in USD to use different labels depending on whether a function is filled by a phrase or a clause. Thus, the single subject relation (SS) in Talbanken corresponds to nominal subject (nsubj) or clausal subject (csubj) in USD. In the case of objects (OO), there is even a three-way distinction between phrasal objects (dobj) and clausal objects with (xcomp) and without (ccomp) obligatory control.

In addition, there are four types of constructions where the two schemes make different assumptions about headedness: coordination, prepositional phrases, verb groups, and copula constructions. For coordination structures, the Swedish Treebank treats the coordinating conjunction as the head, while USD takes the first conjunct to be the head. For prepositional phrases, verb groups, and copula constructions, the Swedish Treebank takes the function word (preposition, auxiliary verb, copula) to be the head, while USD consistently prefers content words as heads. In these cases, the conversion requires changing both the structure and the labels. Interestingly, however, the original MAMBA annotation can be used to facilitate this process. For prepositional phrases, MAMBA agrees with USD in assuming that the head is a noun rather than a preposition. For coordination, verb groups, and copula constructions, MAMBA basically assumes a flat structure with no designated head. By going back to the original annotation, we can therefore hope to convert these constructions more accurately than if we had to rely only on the more recent dependency version in the Swedish treebank. Table 2 gives an overview of the conversions that need to be performed with respect to the syntactic annotation.

4. Conclusion

The conversion of the Swedish Treebank is currently work in progress, but we hope to have a first rough version ready in time for SLTC 2014. The treebank will be made freely available in the Universal Dependencies repository as part of the first data release. This release is scheduled for January 2015 and meant to contain treebanks for 10–15 languages, but we welcome contributions to extend the coverage in later releases. The guidelines for the Universal Dependencies were released October 1, 2014 and will be kept stable for at least a year, but may subsequently be revised as the empirical basis for generalization increases.

Acknowledgements

This paper builds on joint work with Ryan McDonald, Slav Petrov, Marie de Marneffe, Chris Manning, Natalia Silveira, Tim Dozat, Yoav Goldberg, Reut Tsarfaty, Filip Ginter, Sampo Pyysalo, Jan Hajič, Dan Zeman and Jinho Choi.

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)*.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. Universal Stanford Dependencies: A cross-linguistic typology. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, pages 4585–4592.

Eva Ejerhed, Gunnel Källgren, Ola Wennstedt, and Magnus Åström. 1992. The linguistic annotation system of the Stockholm-Umeå corpus project. Technical Report 33, University of Umeå: Department of Linguistics.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 62–72.

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. 2013. 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.

Tahira Naseem, Regina Barzilay, and Amir Globerson. 2012. Selective sharing for multilingual dependency parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 629–637.

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, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the CoNLL Shared Task of EMNLP-CoNLL 2007*, pages 915–932.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*.

Ulf Teleman. 1974. *Manual för grammatisk beskrivning av talad och skriven svenska*. Studentlitteratur.

Reut Tsarfaty. 2013. A unified morpho-syntactic scheme of Stanford dependencies. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 578–584.

Daniel Zeman, David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Zdeněk Žabokrtský, and Jan Hajič. 2012. HamleDT: To parse or not to parse? In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, pages 2735–2741.

Daniel Zeman. 2008. Reusable tagset conversion using tagset drivers. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, pages 213–218.