Guidelines for Universal Dependency Annotation

Joakim Nivre and Ryan McDonald

This document describes the annotation guidelines used in the Universal Dependency Treebank Project, Version 2.0. The aim of the project is to create dependency treebanks with cross-linguistically consistent annotation by adapting and harmonizing variants of the Stanford typed dependencies (de Marneffe et al., 2006; de Marneffe and Manning, 2008). This scheme was originally developed for English but has subsequently been adapted and applied to a number of other languages including Chinese (Chang et al., 2009), Finnish (Haverinen et al., 2013), Persian (Seraji et al., 2012), and Modern Hebrew (Tsarfaty, 2013). We first give an overview of the modifications to the original Stanford scheme and then provide a detailed description of each dependency relation and its relation to the original scheme(s). Besides a syntactic dependency annotation, the treebanks also contain part-of-speech annotation using the Google Universal Part-of-Speech Tags (Petrov et al., 2012).[1]

1 Overview of the Annotation Scheme

We assume the Stanford *basic* dependencies (with punctuation included), where every dependency structure is a tree spanning all the input tokens, because this is the kind of representation that most available dependency parsers require.[2] A sample dependency tree from the French treebank is shown in Figure 1.

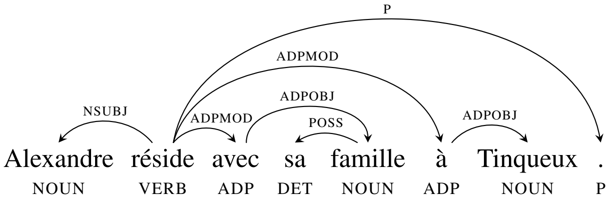
P
ADPMOD
ADPOBJ
NSUBJ
ADPMOD
POSS
ADPOBJ
Alexandre réside avec sa famille à Tinqueux .
NOUN VERB ADP DET NOUN ADP NOUN P

Figure 1: A sample French sentence.

The universal annotation scheme was created by harmonizing available treebanks in slightly different variants of Stanford dependencies, some developed through manual annotation, some produced through automatic conversion from other schemes.[3] In the harmonization step, we have eliminated cases where the same label was used for different linguistic relations in different languages and, conversely, where one and the same relation was annotated with different labels, both of which could happen accidentally when the original Stanford scheme was adapted to specific languages. Secondly, we have avoided, as far as possible, labels that are only used in one or two languages.

In order to satisfy these requirements, a number of language-specific labels have been merged into more general labels. For example, in analogy with the *nn* label for (element of a) noun-noun compound, the German scheme had a label *aa* for compound adjectives, and the Korean scheme had a label *vv* for compound verbs. In the universal scheme, these three labels have been merged into a single label *compmod* for modifier in compound. For Korean, the annotation scheme distinguished four different subtypes of nominal subjects, which have all been merged to the single relation *nsubj* in the universal annotation.

In addition to harmonizing language-specific labels, we have also renamed relations where the name would be misleading in the universal context (although quite appropriate for English). For example, the label *prep* (for a modifier headed by a preposition) has been renamed to *adpmod*, to make clear the relation to other modifier labels and to allow postpositions as well as prepositions. Consequently, *pobj* and *pcomp* have been changed to *adpobj* and *adpcomp*. Similarly, *npadvmod* has been replaced by *nmod* (in analogy with *amod* and *advmod*). We have also eliminated a few distinctions in the original Stanford

[1] In addition to the universal tags, we also provide language-specific tags when available.

[2] This is in contrast to the *collapsed* dependencies, where multiple heads are allowed and where some tokens may not correspond to nodes in the dependency structure.

[3] For a more detailed description of this process, see McDonald et al. (2013).

scheme that were not annotated consistently across languages, for example, merging *complm* with *mark*, *number* with *num*, and *purpcl* with *advcl*.

Although the ultimate goal is to arrive at a single universal annotation for all languages, there are still two types of constructions where the annotation may vary across languages. The Stanford basic dependencies in general favor content words over function words as syntactic heads, but make an exception for copula constructions (optionally) and adpositional phrases (always). In some of the language-specific adaptations, notably for Finnish (Haverinen et al., 2013), this has been changed enforce the content-head principle also in these constructions, making both copulas and adpositions dependents of their complements in the dependency structure. For some languages, the annotation permits accurate conversion between these two representations, but for others it is difficult to perform the conversion without introducing too much noise.

In Version 2.0, we therefore maintain two versions of the annotation scheme: the *standard* version, which treats copulas and adpositions as heads of their complements, and the *content-head* version, which consistently treats content words as syntactic heads. Currently, English, Portuguese, and Indonesian are only available in the standard version, while Finnish is only available in the content-head version. For Japanese and Korean, where the syntactic annotation is at the chunk (bunsetsu) level, the distinction is neutralized, and for the remaining languages we provide both versions (although the content-head version should be regarded as tentative and experimental at this point). For illustration, Figure 2 shows a German sentence annotated in the standard version (left) and the content-head version (right).

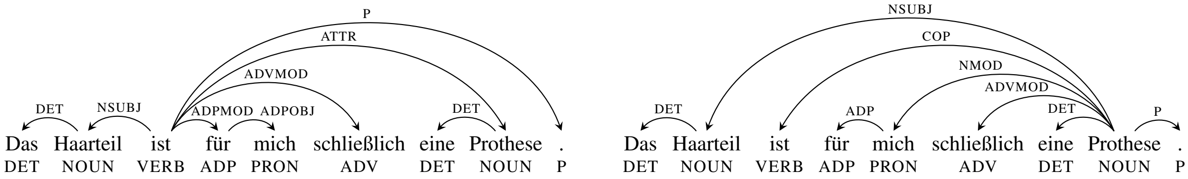
P
ATTR
ADVMOD
DET
NSUBJ
ADPMOD
ADPOBJ
DET
Das Haarteil ist für mich schließlich eine Prothese .
DET NOUN VERB ADP PRON ADV DET NOUN P
NSUBJ
COP
NMOD
ADVMOD
DET
ADP
DET
P
Das Haarteil ist für mich schließlich eine Prothese .
DET NOUN VERB ADP PRON ADV DET NOUN P

Figure 2: A sample German sentence with standard (left) and content-head (right) annotation.

In addition to the two annotation versions, there are a few known inconsistencies across languages, notably in the annotation of multiword expressions and in particular multiword names. Most treebanks follow the practice from English to annotate name parts as components of (nominal) compounds (which is questionable in languages like German where real nominal compounds are normally realized as single orthographic words), while some treebanks instead annotate them as parts of multiword expressions. In the future, it might be desirable to instead add a new relation *name* for this type of expression.[4] In addition to the inconsistency in name annotation, the internal structure of multiword expressions varies between treebanks, being sometimes head-initial, sometimes head-final, and sometimes with no consistent headedness direction.

2 Dependency Relations

Below we give a brief description of each dependency relation used in the universal annotation. For each relation, we also list the language-specific relation(s) that it replaces or subsumes. We talk about replacement when it is a simple renaming and about subsumption when a more specific relation is merged with a more general one.

[4] This relation already exists in the native version of the Finnish treebank, but has been eliminated in the cross-linguistic harmonization process.

ROOT Root of the dependency tree, normally a verb (or a predicative complement in the content-head version). Replaces: none. Subsumes: none.

acomp Adjectival complement (including predicative complements in the standard version). Replaces: none. Subsumes: none.

adp Adposition analyzed as dependent of noun (case marker). Replaces: *adpos*. Subsumes: *objp*, *ps*.

adpcomp Clausal complement of adposition. Replaces: *pcomp*. Subsumes: none.

adpmod Adpositional modifier (with the adposition taken as the head of the adpositional phrase). Replaces: *prep*, *postp*. Subsumes: *agent*, *comp*.

adpobj Nominal complement of adposition. Replaces: *pobj*. Subsumes: none.

advcl Adverbial clause modifier. Replaces: none. Subsumes: *compar*, *purpcl*.

advmod Adverbial modifier. Replaces: none. Subsumes: *quantmod*, *tmod*.

amod Adjectival modifier. Replaces: none. Subsumes: none.

appos Appositional modifier. Replaces: none. Subsumes: *abbrev*.

attr Nominal predicative complement (dependent on a copula verb). Replaces: none. Subsumes: none.

aux Auxiliary verb (dependent on main verb), including infinitive marker. Replaces: none. Subsumes: *infumzu*.

auxpass Auxiliary verb in passive construction. Replaces: none. Subsumes: none.

cc Coordinating conjunction (dependent on conjunct). Replaces: none. Subsumes: *preconj*.

ccomp Clausal complement. Replaces: none. Subsumes: *iccomp*.

compmod Compound modifier (non-head part of compound), sometimes including multiword names. Replaces: none. Subsumes: *aa*, *nn*, *vv*.

conj Conjunct (dependent on first conjunct in coordination). Replaces: none. Subsumes: none.

cop Copula verb (dependent on predicative complement, primarily in content-head version). Replaces: none. Subsumes: none.

csubj Clausal subject. Replaces: none. Subsumes: *csubj-cop*.

csubjpass Clausal subject in passive construction. Replaces: none. Subsumes: none.

dep Unclassifiable dependent. Replaces: none. Subsumes: *emot*, *emoticon*, *intj*, *interj*, *voc*.

det Determiner. Replaces: none. Subsumes: *predet*, *postdet*.

dobj Direct object (with or without dependent case marker). Replaces: none. Subsumes: *dobj1*, *dobj2*, *dobj3*, *pronl*.

expl Expletive subject. Replaces: none. Subsumes: none.

infmod Infinitival modifier. Replaces: none. Subsumes: none.

iobj Indirect object (with or without dependent case marker). Replaces: none. Subsumes: *gobj*.

mark Subordinating conjunction and equivalent expressions. Replaces: none. Subsumes: *complm*, *comparator*.

mwe Multiword expression (non-head part), sometimes including multiword names. Replaces: none. Subsumes: *name*.

neg Negation. Replaces: none. Subsumes: *postneg*.

nmod Nominal modifer (including adpositional phrases headed by nominal in content-head version). Replaces: *nommod*, *nommod-own*, *npadvmod*. Subsumes: none.

nsubj Nominal subject. Replaces: none. Subsumes: *gsubj*, *nsubj1*, *nsubj2*, *nsubj3*, *nsubj-cop*.

nsubjpass Nominal subject in passive construction. Replaces: none. Subsumes: *nsubjpass1*, *nsubjpass2*.

num Numeral. Replaces: none. Subsumes: *cln*, *nnumber*, *number*.

p Punctuation. Replaces: *punct*. Subsumes: none.

parataxis Clause-like structure loosely dependent on preceding clause. Replaces: none. Subsumes: none.

partmod Participial modifier. Replaces: none. Subsumes: none.

poss Possessive (or genitive) modifier. Replaces: none. Subsumes: *gmod*.

prt Verb particle. Replaces: none. Subsumes: *dvp*.

rcmod Relative clause modifier. Replaces: none. Subsumes: none.

rel Relative pronoun/adverb with unidentifiable grammatical function. Replaces: none. Subsumes: none.

vmod Verbal modifier (underspecified label used only in content-head version). Replaces: none. Subsumes: none.

xcomp Non-finite clause-like complement. Replaces: none. Subsumes: *vinf*.

For reference, Table 1 shows the distribution of different dependency relations across languages. A plus (+) indicates that the relation is present in all versions available for that language, and a minus (−) that it is absent in all versions. The symbols STD and CH indicate that the relation occurs in the standard but not the content-head version, and vice versa.

References

Chang, P.-C., Tseng, H., Jurafsky, D., and Manning, C. D. 2009. Discriminative reordering with Chinese grammatical relations features. In *Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation (SSST-3) at NAACL HLT 2009*, 51–59.

de Marneffe, M.-C. and Manning, C. D. 2008. The Stanford typed dependencies representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, 1–8.

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

Haverinen, K., Nyblom, J., Viljanen, T., Laippala, V., Kohonen, S., Missilä, A., Ojala, S., Salakoski, T., and Ginter, F. 2013. Building the essential resources for Finnish: the Turku Dependency Treebank. *Language Resources and Evaluation*.

McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., Bedini, C., Bertomeu Castelló, N., and Lee, J. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 92–97.

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

Seraji, M., Megyesi, B., and Joakim, N. 2012. Bootstrapping a Persian dependency treebank. *Linguistic Issues in Language Technology*, 7:1–10.

Tsarfaty, R. 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)*, 578–584.

Relation	DE	EN	ES	FI	FR	ID	IT	JA	KO	PT	SV
ROOT	+	+	+	+	+	+	+	+	+	+	+
acomp	+	+	+	+	+	+	+	−	−	+	+
adp	+	+	+	+	CH	−	−	−	−	+	CH
adpcomp	STD	+	+	−	+	+	+	−	−	+	STD
adpmod	+	+	+	−	+	+	+	+	+	+	+
adpobj	+	+	+	−	+	+	+	−	+	+	+
advcl	+	+	+	+	+	+	+	+	+	+	+
advmod	+	+	+	+	+	+	+	+	+	+	+
amod	+	+	+	+	+	+	+	+	+	+	+
appos	+	+	+	+	+	+	+	+	+	+	+
attr	+	+	+	−	+	+	+	−	−	+	+
aux	+	+	+	+	+	+	+	−	−	+	+
auxpass	+	+	+	+	+	−	−	−	−	+	−
cc	+	+	+	+	+	+	+	+	+	+	+
ccomp	+	+	+	+	+	+	+	+	+	+	+
compmod	+	+	+	+	+	+	−	+	+	+	−
conj	+	+	+	+	+	+	+	+	+	+	+
cop	+	+	+	+	+	−	−	−	−	−	CH
csubj	+	+	+	+	+	+	+	−	+	+	+
csubjpass	+	+	+	−	−	+	+	−	−	+	−
dep	+	+	+	+	+	+	+	+	+	+	+
det	+	+	+	+	+	+	+	−	+	+	+
dobj	+	+	+	+	+	+	+	−	+	+	+
expl	+	+	−	−	+	−	−	−	−	−	+
infmod	+	+	+	+	+	+	+	−	−	+	+
iobj	+	+	+	−	+	+	−	−	+	+	+
mark	+	+	+	+	+	+	+	−	+	+	+
mwe	+	+	+	+	+	+	+	−	+	+	+
neg	+	+	+	+	+	+	+	−	+	+	+
nmod	+	+	+	+	+	+	+	+	+	+	+
nsubj	+	+	+	+	+	+	+	−	+	+	+
nsubjpass	+	+	+	−	+	+	+	−	+	+	−
num	+	+	+	+	+	+	+	+	+	+	+
p	+	+	+	+	+	+	+	+	+	+	+
parataxis	+	+	+	+	+	+	−	−	−	+	+
partmod	+	+	+	+	+	+	+	−	−	+	+
poss	+	+	+	+	+	+	−	−	+	+	+
prt	+	+	+	+	+	−	−	−	−	+	+
rcmod	+	+	+	+	+	+	+	+	+	+	+
rel	+	+	+	+	+	−	+	−	−	−	−
vmod	CH	CH	CH	−	CH	−	−	−	−	−	CH
xcomp	+	+	+	+	+	+	+	−	+	+	+

Table 1: Dependency relations across languages.