Norwegian Universal Dependencies

Lilja Øvrelid
Department of Informatics, University of Oslo
liljao@ifi.uio.no

1. Introduction

With the increasing popularity of dependency-based representations of syntactic structure in recent years, a wealth of different dependency annotation schemes have surfaced. It has been shown that the choice of dependency scheme influences parsing results (Schwartz et al., 2012) as well as down-stream applications (Elming et al., 2013) and even though attempts have been made to contrast different schemes theoretically (Ivanova et al., 2012), it is clear that the diversity of representation makes comparisons difficult. Cross-linguistically even more so, and it can often be difficult to tease apart aspects of annotation scheme from typological differences in cross-lingual learning (Søgaard, 2011; Skjærholt and Øvrelid, 2012).

Universal Dependencies (UD) (de Marneffe et al., 2014; Nivre, 2015) is a recent community-driven effort to create cross-linguistically consistent syntactic annotation. UD is based on the Stanford dependency scheme (de Marneffe et al., 2006) which has become a widely used dependency scheme for English in recent years. A number of existing dependency treebanks have been converted to UD (Pyysalo et al., 2015; Nivre, 2014) and new data has also been annotated from scratch in order to enable multilingual parser development, cross-lingual learning and typological studies of syntactic structure. Treebanks involved in this effort represent a diverse range of languages such as English, German, Swedish, Spanish, Italian, Persian, Japanese, and the UD release 1.1 contains treebanks for as many as 34 different languages of varying sizes.

This paper describes a fully automatic conversion procedure for the Norwegian Dependency Treebank (NDT) to UD. Due to differences both in the tag set, as well as structural analyses, the conversion requires non-trivial transformations of the dependency trees, in addition to mappings of tags and labels that make reference to a combination of various kinds of linguistic information. This paper details the mapping of PoS tags, morphological features and dependency relations and provides a description of the structural changes made to NDT analyses in order to make it compliant with the UD guidelines. The full converted treebank will be made available with the next release of the UD treebanks scheduled for November 2015.

2. NDT and UD

Universal Dependencies (UD) builds on several previous initiatives for universally common morphological (Zeman, 2008; Petrov et al., 2012) and syntactic dependency (McDonald et al., 2013; Rosa et al., 2014) annotation. Among its main tenets are the primacy of content-words, i.e. content words, as opposed to function words, are syntactic heads wherever possible. It is intended to be a universal annotation scheme, i.e. applicable to any language, however also offers some possibilities for language-specific information.

The Norwegian Dependency Treebank (NDT) (Solberg et al., 2014) contains morphological and syntactic annotation for both varieties of written Norwegian (Bokmål and Nynorsk). The morphological annotation follows the Oslo-Bergen Tagger scheme (Hagen et al., 2000). The syntactic annotation scheme is, to a large extent, based on the Norwegian Reference Grammar (Faarlund et al., 1997). and the dependency representations are inspired by choices made in comparable treebanks, in particular the Swedish treebank Talbanken (Nivre et al., 2006).

3. Parts-of-speech

The part-of-speech tag set used in the UD scheme is based on the Universal PoS tag set of Petrov et al. (2012) and contains 17 tags. The NDT tag set contains 19 tags. The conversion of the part-of-speech information in NDT to the UD pos tag set is fairly straightforward and largely relies on a direct mapping presented in Table 1. A few parts-of-speech require conversion rules which make reference to additional information in the treebank, represented by disjunction in the mapping. Below we will discuss a few of these cases.

The universal scheme makes a distinction between proper and common nouns at the part-of-speech level.

NDT	UD
adj	ADJ
adv	ADV
clb	PUNCT, SYM
det	DET, NUM
konj	CONJ
interj	INTJ
inf-merke	PART
prep	ADP, ADV
pron	PRON
<komma>	PUNCT
sbu	SCONJ
<strek>	PUNCT
subst	NOUN, PROPN
<anf>	PUNCT
<parentes-slutt>	PUNCT
<parentes-beg>	PUNCT
symb	SYM
ukjent	X
verb	AUX, VERB

Table 1: Mapping between NDT and UD parts-of-speech

This information can be found among the morphological features in NDT (prop), hence the mapping is straightforward.

For verbs UD distinguishes auxiliaries (AUX) from main verbs (VERB). This distinction is not explicitly made in NDT, hence our conversion procedure must make use of the syntactic structure of the verbs in order to implement this distinction. Verbs that have a direct, non-finite dependent (a dependent with the NDT dependency relation INF) are marked as auxiliaries and all other verbs as regular verbs.

4. Morphological information

In addition to part-of-speech information, NDT contains a rich inventory of morphological features, e.g. information about properties like gender, definiteness, tense, voice, etc. The UD guidelines specify a universal set of morphological features and the conversion between the two does not require reference to information in addition to the feature information. The feature mapping is described in Table 2. Note that since the number of UD features is larger than the NDT features, some of the NDT features correspond to a set of UD features, e.g. the NDT features for verbs (pres, pret) which instantiate both the Mood, Tense and VerbForm features.

NDT	UD
mask,fem,nøyt	Gender=Masc,Fem,Neut
ent,fl	Number=Sing,Plur
be,ub	Definite=Def,Ind

pres,pret	Mood=Ind,
	Tense=Pres,Past,
	VerbForm=Fin

perf-part	VerbForm=Part

imp	Mood=Imp,
	VerbForm=Fin

pass	Voice=Pass
inf	VerbForm=Inf
1,2,3	Person=1,2,3
nom,akk,gen	Case=Nom,Acc,Gen
pos,komp,sup	Degree=Pos,Cmp,Sup
hum	Animacy=Anim
pers	PronType=Prs
dem	PronType=Dem
sp	PronType=Int
res	PronType=Rcp
poss	Poss=Yes
refl	Refl=Yes

Table 2: Mapping between NDT and UD morphological features

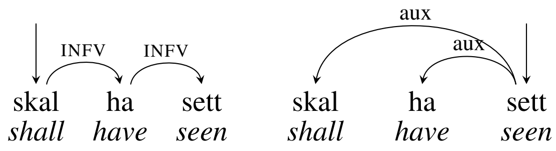
5. Structural conversion

The NDT annotation scheme differs structurally from the UD scheme in a number of important ways. The conversion is therefore non-trivial and requires a set of structural rules which operate on the dependency graphs in addition to a mapping procedure over the dependency labels. The conversion is implemented as a cascade of structural rules followed by a relation conversion procedure over the modified graph structures. The structural rules employ a small set of graph operations that reverse, reattach, delete and add arcs.

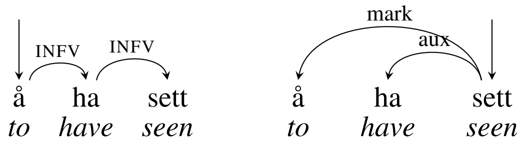
5.1. Verbal groups

NDT consistently marks the finite verb as head of a clause, with other non-finite verbs as dependents (INFV), see example (1a). In a parallel manner, infinitival markers are also annotated as heads with the infinitival verb as its dependent, see (2a). UD on the other hand annotates the lexical, main verb as head of the verbal group and various finite and non-finite auxiliaries receive an auxiliary relation (aux, auxpass) , see (1b) and (2b) below. The conversion rule locates the main verb within the chain of nonfinite dependents of the finite verb and makes this node the head of the other verbs in the chain.

(1) a) b)



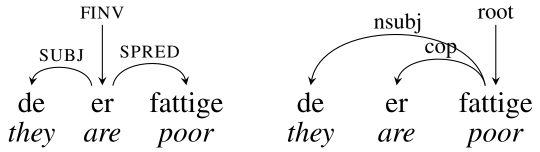
(2) a) b)



5.2. Copula constructions

The treatment of copula constructions within the UD scheme differs markedly from that of the NDT by appointing the predicative element as head of the entire construction and attaching the copula verb with a special relation `cop`, see (3b). Our conversion thus reverses the arc between the copula and its complement and reattaches all its dependents to the predicative element.

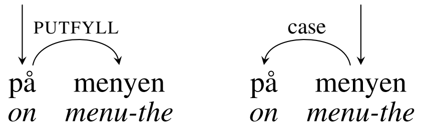
(3) a) b)



5.3. Prepositions and their complements

In NDT, prepositions are heads of their prepositional complements which receive the `PUTFYLL` label, see (4a). Seeing that languages differ greatly in their use of pre/postpositions, the UD scheme annotates the prepositional complement as head and attaches the preposition using the `case` relation, see (4b). In conversion this is once again a matter of reversing the arc and reattaching dependents to the new head.

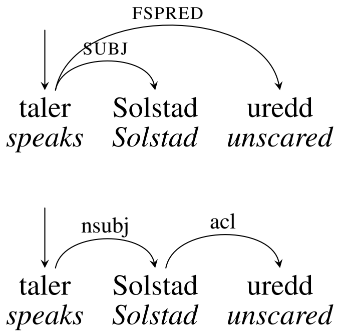
(4) a) b)



5.4. Predicatives

NDT distinguishes several types of predicatives, both predicatives that are arguments of verbs (subject predicative `SPRED` and object predicative `OPRED`) and "free predicatives" which are not arguments of the verb, but nonetheless characterize either a subject or an object in the preceding context (free subject predicative `FSPRED` and free object predicative `FOPRED`). Both of these are attached to the finite verb in NDT, as we can see in (5a). In a similar manner, UD distinguishes between obligatory and optional predicatives, where the former are analyzed using the `xcomp` relation and attached to the main predicate, whereas the optional predicatives are attached as adverbial clauses (`acl`) modifying the argument they characterize, see (5b). Our conversion thus attaches the `FSPRED` argument to its sibling subject argument, `FOPRED` to an object sibling.

(5) (a) (b)

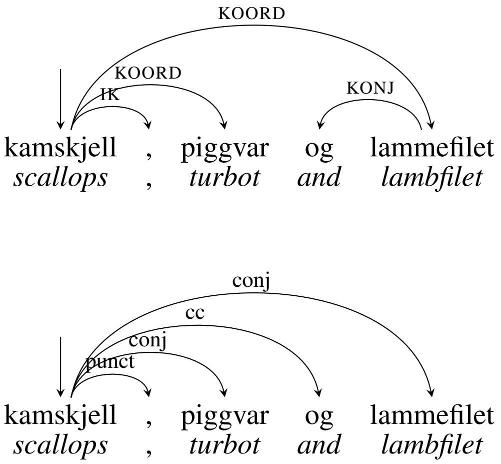


5.5. Coordination

Coordination is a phenomenon which exhibits considerable variation in terms of dependency representation across various annotation schemes (Popel et al., 2013). As we can see from the example in (6), the analyses chosen in the NDT and UD schemes are fairly similar in their choice of the first conjunct as head of the coordinate structure. They differ mainly in the attachment of the conjunction and the relation names.

(6) (a)

(b)


KOORD
KOORD
IK
KONJ
kamskjell , piggvar og lammefilet
scallops , turbot and lambfilet
conj
cc
conj
punct
kamskjell , piggvar og lammefilet
scallops , turbot and lambfilet

6. Conversion of dependency relations

A minority of the dependency relations in NDT may be converted directly, based on the mapping described in Table 3, the rest require the formulation of mapping constraints which make reference to information in addition to the dependency relation itself, i.e. part-of-speech tag, morphological features, dependency structure or even a combination of these. An overview of

NDT	UD
APP	appos
FSUBJ	expl
FOBJ	expl
FSPRED	acl
FOPRED	acl
FRAG	root
IOBJ	iobj
OPRED	xcomp
INTERJ	discourse
KONJ	cc
KOORD	conj
KOORD-ELL	remnant
IP	punct
IK	punct
PAR	parataxis
SPRED	xcomp
UKJENT	goeswith

Table 3: Direct mapping between NDT and UD dependency relations.

the non-direct mapping of dependency relations that require additional information from the linguistic context is provided in Table 4. These mappings often also require heuristics which approximate some syntactic property which is not explicitly annotated in NDT. Below we will present these heuristics and their usage in the conversion.

Active/Passive There are two ways of expressing passive voice in Norwegian: a morphological passive expressed by an addition of a *-s* to the verb, e.g. *danses* 'to be danced' or a periphrastic passive which is composed of the auxiliary *bli* 'to become' and a participle form, e.g. *danset* 'danced'. Only the morphological passive is marked explicitly as being in the passive voice. We therefore define a heuristic which counts a lexical main verb as passive, if it is (i) a morphological passive, or (ii) is a participle headed by a form of the auxiliary *bli*. This heuristic is used in the conversion of passive auxiliaries `auxpass` and passive subjects `nsubjpass, csubjpass`.

Nominal/Clausal Several of the UD relations assume a distinction between nominal and clausal elements. This distinction is complicated somewhat by the fact that in copula constructions, as described above, the complement of the copula is head of the construction as a whole. This means that adjectives or even nouns may be counted as clausal in contexts where they have a copula dependent, as in (3b). In the conversion we introduce a notion of a *predicate*, which may be either verbal (`AUX, VERB`) or the complement in a copula construction. This notion is used to distinguish nominal and clausal subjects (`nsubj, nsubjpass` vs. `csubj, csubjpass`), objects (`dobj` vs. `ccomp, xcomp`), various modifiers (`nmod` vs. `acl`) and adverbials (`nmod` vs. `advcl`).

Control UD is inspired by the syntactic framework of Lexical Functional Grammar (Kaplan and Bresnan, 1982) and adopts its distinction between complement clauses with obligatory subject control (`xcomp`) and those without (`ccomp`). The notion of control is not native to NDT, hence we approximate it by requiring the presence of an explicit subject dependent of the head verb of the clause.

Negation UD distinguishes negation modifiers which modify either a noun (*no problem*) or a predicate (in the aforementioned sense) (*is not a problem, doesn't argue*). Our conversion explicitly marks the negative determiner (*ingen* 'no') and the negative adverb (*ikke* 'not') in Norwegian.

Particles NDT distinguishes between transitive and intransitive prepositions, or so-called particles, in the annotation. In order to account for the relation between the verb and its particle we introduce the language-specific relation `compound:prt`, for prepositions which are attached to a verb and furthermore does not have an explicit prepositional complement, e.g. *si opp* 'discontinue' (lit. 'say up').

Relative clauses In many languages, relative markers are pronouns. In Norwegian, there are good reasons for choosing to treat the relative marker *som* 'that' as a subjunction (Faarlund et al., 1997). This relative marker, unlike those found in many other languages does not inflect (as in English *who-whom*, or German *der-die-das*) and exclusively occurs initially in a subordinate clause. In our conversion we introduce a language-specific variant of the clausal relation `acl`, `acl:relcl` which signals that this is a relative clause.

7. References

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 449–454.

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 International Conference on Language Resources and Evaluation (LREC)*, pages 4585–4592.

NDT	UD
ADV	advcl, advmod, compound:prt, neg, nmod
ATR	acl:relcl, amod, nmod
DET	nmod, nummod, det
FINV	aux, auxpass, root
FLAT	foreign, name
OBJ, POBJ	dobj, ccomp, xcomp
SBU	nsubj, nsubjpass, dobj, iobj, mark
SUBJ, PSUBJ	nsubj, nsubjpass, csubj, csubjpass

Table 4: Non-direct mapping between NDT and UD dependency relations; requires additional constraints with reference to PoS, morphological features or dependency context.

Jacob Elming, Anders Johannsen, Sigrid Klerke, Emanuele Lapponi, Hector Martinez, and Anders Søgaard. 2013. Downstream effects of tree-to-dependency conversions. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 617–626, Atlanta, Georgia, USA.

Jan Terje Faarlund, Svein Lie, and Kjell Ivar Vannebo. 1997. *Norsk referansegrammatikk*. Universitetsforlaget, Oslo, Norway.

Kristin Hagen, Janne Bondi Johannessen, and Anders Nøklestad. 2000. A constraint-based tagger for norwegian. In *17th Scandinavian Conference in Linguistics*, pages 31–48.

Angelina Ivanova, Stephan Oepen, Lilja Øvrelid, and Dan Flickinger. 2012. Who Did What to Whom? A contrastive study of syntacto-semantic dependencies. In *Proceedings of the Sixth Linguistic Annotation Workshop*, pages 2–11, Jeju, Republic of Korea.

Ronald Kaplan and Joan Bresnan. 1982. Lexical Functional Grammar. A formal system for grammatical representation. In Joan Bresnan, editor, *The Mental Representation of Grammatical Relations*, pages 173–281. MIT Press, Cambridge, MA, USA.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, and Oscar Täckström. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of Association for Computational Linguistics (ACL)*, pages 92–97.

Joakim Nivre, Jens Nilsson, and Johan Hall. 2006. Talbanken05: A swedish treebank with phrase structure and dependency annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy.

Joakim Nivre. 2014. Universal Dependencies for Swedish.

Joakim Nivre. 2015. Towards a universal grammar of natural language processing. In *Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics (CICLING)*.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 2089–2096.

Martin Popel, David Mareček, Jan Štěpánek, Daniel Zeman, and Zdeněk Žabokrtský. 2013. Coordination Structures in Dependency Treebanks. In *Proceedings of Association for Computational Linguistics (ACL)*, pages 517–527.

Sampo Pyysalo, Jenna Kanerva, Anna Missilä, Veronika Laippala, and Filip Ginter. 2015. Universal dependencies for finnish. In *Proceedings of the Nordic Conference of Computational Linguistics (NODALIDA)*.

Rudolf Rosa, Jan Mašek, David Mareček, Martin Popel, Daniel Zeman, and Zdeněk Žabokrtskỳ. 2014. Hamledt 2.0: Thirty dependency treebanks stanfordized. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 2334–2341.

Roy Schwartz, Omri Abend, and Ari Rappoport. 2012. Learnability-based syntactic annotation design. In *Proceedings of the International Conference on Computational Linguistics COLING*, Mumbai, India.

Arne Skjærholt and Lilja Øvrelid. 2012. Impact of treebank characteristics on cross-lingual parser adaptation. In *Proceedings of Treebanks and Linguistic Theories (TLT11)*.

Anders Søgaard. 2011. Data point selection for cross-language adaptation of dependency parsers. In *Proccedings of the meeting of the Association for Computational Linguistics*.

Per Erik Solberg, Arne Skjærholt, Lilja Øvrelid, Kristin Hagen, and Janne Bondi Johannessen. 2014. The Norwegian Dependency Treebank.

Daniel Zeman. 2008. Reusable tagset conversion using tagset drivers. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.