Universal Dependencies for Finnish

Sampo Pyysalo[1] **Jenna Kanerva**[1,2] **Anna Missilä**[4] **Veronika Laippala**[3,4] **Filip Ginter**[1]
[1]Department of Information Technology [2]University of Turku Graduate School (UTUGS)
[3]Turku Institute for Advanced Studies (TIAS) [4]School of Languages and Translation Studies,
University of Turku, Finland
first.last@utu.fi

Abstract

There has been substantial recent interest in annotation schemes that can be applied consistently to many languages. Building on several recent efforts to unify morphological and syntactic annotation, the Universal Dependencies (UD) project seeks to introduce a cross-linguistically applicable part-of-speech tagset, feature inventory, and set of dependency relations as well as a large number of uniformly annotated treebanks. We present Universal Dependencies for Finnish, one of the ten languages in the recent first release of UD project treebank data. We detail the mapping of previously introduced annotation to the UD standard, describing specific challenges and their resolution. We additionally present parsing experiments comparing the performance of a state-of-the-art parser trained on a language-specific annotation schema to performance on the corresponding UD annotation. The results show improvement compared to the source annotation, indicating that the conversion is accurate and supporting the feasibility of UD as a parsing target. The introduced tools and resources are available under open licenses from http://bionlp.utu.fi/ud-finnish.html.

1 Introduction

The Universal Dependencies (UD) initiative seeks to develop cross-linguistically consistent annotation guidelines and apply them to many languages to create treebank annotations that are uniform in e.g. their theoretical basis, label sets, and structural aspects. Such resources could substantially advance cross-lingual learning, improve comparability of evaluation results, and facilitate new approaches to automatic syntactic analysis.

UD builds on the Google Universal part-of-speech (POS) tagset (Petrov et al., 2012), the Interset interlingua of morphosyntactic features (Zeman, 2008), and Stanford Dependencies (de Marneffe et al., 2006; Tsarfaty, 2013; de Marneffe et al., 2014). In addition to the abstract annotation scheme, UD defines also a treebank storage format, CoNLL-U. A first version of UD treebank data, building on the Google Universal Dependency Treebanks (McDonald et al., 2013) and many other previously released resources (Bosco et al., 2013; Haverinen et al., 2013b), was recently released[1] (Nivre et al., 2015).

In this paper, we present the adaptation of the UD guidelines to Finnish and the creation of the UD Finnish treebank by conversion of the previously introduced Turku Dependency Treebank (TDT) (Haverinen et al., 2013b). We also provide a first set of experiments comparing the parsing scores of language-specific treebank annotation to that of a UD treebank, providing an evaluation of both the conversion quality and the feasibility of UD annotation as a parsing target. In a related but separate effort within the UD initiative, the FinnTreeBank 1[2] (ftb-1) (Voutilainen, 2011) is also being converted into the UD format. The *ftb-1* is a treebank based on all grammatical examples from the VISK[3] Finnish grammar reference (Hakulinen et al., 2004), and will thus complement the TDT-based UD Finnish treebank in the set of UD treebanks.

2 Treebank conversion

The conversion of TDT into the UD Finnish treebank was implemented following the UD specification (Nivre et al., 2014) (version 1, Oct 2014),

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

[2]http://www.ling.helsinki.fi/kieliteknologia/tutkimus/treebank/sources/

[3]http://scripta.kotus.fi/visk

the Finnish grammar of Hakulinen et al. (2004) and the TDT annotation guidelines (Haverinen et al., 2013b) as the primary references. The initial stages of the work involved identifying similarities and differences between the TDT and UD annotation guidelines, adapting the general UD guidelines to Finnish, and planning the implementation of the conversion. Technically, the conversion was implemented as a pipeline of processing components, each of which consumed and produced CoNLL-U-formatted data. The following sections present the source data and primary stages of processing in detail.

2.1 Turku Dependency Treebank

As the source data for the conversion, we selected the most recent published distribution of TDT.[4] The source treebank contains 15,000 sentences (200,000 words) drawn from a variety of sources and annotated in a Finnish-specific version of the Stanford Dependencies (SD) scheme, and it has previously been demonstrated to be applicable e.g. for training broad-coverage dependency parsers for Finnish (Kanerva et al., 2014).

In addition to converting the annotation to UD standards, we also addressed a number of instances where tokenization differed from UD specifications, corrected a small number of sentence-splitting errors, and updated the lemmas to improve both treebank-internal consistency and conformance with the UD specification. We further introduced a fully manually annotated morphology layer, replacing the automatically generated morphological annotation of the initial data. This modified TDT not only serves as the basis for conversion but is also made available as a separate contribution.

2.2 Part-of-speech annotation

The UD specification defines 17 POS tags, and requires that all conforming treebanks use only these tags.[5] The TDT annotation uses a comparatively coarse-grained set of 12 POS tags, of which approximately half correspond straightforwardly to one of the 17 UD POS tags (Table 1). Several other TDT tags could be assigned the appropriate UD tag based on the value of the SUBCAT feature, which distinguishes e.g. coordinating conjunctions from subordinating conjunctions (CONJ and SCONJ in UD, respectively). Just four TDT tags, marking pronouns, punctuation, symbols and verbs, required further information to resolve correctly.

TDT	UD	TDT type
A	ADJ	adjective
Adp	ADP	adposition
Adv	ADV	adverb
C[SUBCAT=CC]	CONJ	coord. conj.
C[SUBCAT=CS]	SCONJ	subord. conj.
Foreign	X	foreign word
Interj	INTJ	interjection
N[SUBCAT=Prop]	PROPN	proper noun
N[!SUBCAT=Prop]	NOUN	common noun
Num[SUBCAT=Card]	NUM	cardinal number
Num[SUBCAT=Ord]	ADJ	ordinal number
Pron	PRON or ADJ	pronoun
Punct	PUNCT or SYM	punctuation
Symb	PUNCT or SYM	symbol
V	VERB or AUX	verb

Table 1: Part-of-speech tag mapping from TDT to UD. TAG[FEATURE=VALUE] specifies a mapping that applies only in cases where a word has both the given tag and the feature value, TAG[!FEATURE=VALUE] in cases where the feature is absent or has a different value.

Punctuation and symbols The guidelines covering the use of the Punct and Sym tags in the TDT annotation differed to such an extent from the UD specification of PUNCT and SYM that the Punct/Sym distinction in the original treebank was ignored in creating the mapping. Instead, words assigned either of these tags in TDT were assigned UD POS based on newly implemented surface form-based heuristics, with e.g. currency symbols, mathematical operators, URLs and emoticons assigned SYM and other non-alphabetical character sequences PUNCT.

Verbs All verbs that can serve as auxiliaries were assigned AUX or VERB based on the presence of an aux dependency. This is the only rule concerning the morphological annotation layer that refers to the syntactic annotation. It should be noted that this rule cannot be applied deterministically in a standard syntactic analysis pipeline where morphological analysis precedes dependency analysis, but will instead require these verbs to be assigned both a VERB and AUX reading.

Pronouns The TDT POS tag Pron maps to PRON for UD Finnish in most cases, but pro-

[4] Available from http://bionlp.utu.fi/

[5] While no language-specific POS tags can thus be defined in the primary POS annotation, the CoNLL-U format allows a secondary POS tag to be assigned to each word to preserve treebank-specific information.

adjectives such as *millainen* "like-what" are analyzed as Pron in TDT but assigned to ADJ in UD Finnish following the reference grammar and the UD specification. The annotation of related cases such as pro-adverbs was already consistent with the reference resources and could thus be processed using the general mapping rules.

Finally, we note that UD Finnish excludes by design two of the UD POS tags, DET (determiner) and PART (particle). As Finnish has no true articles (Sulkala and Karjalainen, 1992) and words (primarily pronouns) that play a determiner role syntactically can be identified using the dependency annotation layer (namely, the det relation), we opted not to apply DET in UD Finnish annotation. Similarly, although various words have been categorized as particles in different descriptions of Finnish, the reference grammar (Hakulinen et al., 2004) does not assign any Finnish words to the category covered by PART in the UD specification. This POS tag is correspondingly excluded from use in UD Finnish.

2.3 Morphological features

The UD specification defines a set of 17 widely attested morphological features such as Case, Person, Number, Voice and Mood. However, by contrast to the POS tag annotation, the specification allows conforming treebanks to introduce language-specific features that are not included in this universal inventory, suggesting that such features be drawn when possible from the extended Interset compilation of morphological feature names and labels (Zeman, 2008).

The morphological annotation of TDT draws directly on the rich features provided by the OMorFi morphological analyzer (Pirinen, 2008), and many of the generally applicable UD features can be generated by direct mapping from TDT POS tags and features (Table 2). For brevity, we refer to UD documentation for descriptions of UD standard features, focusing in the following on UD Finnish features not among the basic 17.

To minimize information loss from the conversion, we made liberal use of the possibility to introduce language-specific features to mark aspects of the TDT morphological annotation that were not captured by the basic 17 UD features. We aimed to primarily apply extended Interset features, drawing from this inventory the features Abbr (abbreviation or acronym), Style (colloquial or archaic style), Typo (typographic error), Foreign (foreign word or script) and AdpType (adposition type: pre- or postposition). Finally, we added features to capture aspects of TDT annotation that did not have representation in Interset: InfForm (differentiates between Finnish infinitives), PartForm (similar for participles), Connegative (verb in connegative form) and Clitic and Derivation, identifying steps in the morphological derivation and modification processes to create the wordform.

TDT	UD
CASE	Case
CLIT	Clitic
CMP	Degree
DRV	Derivation
INF	InfForm and VerbForm=Inf
MOOD	Mood
NEG=ConNeg	Connegative=Yes
OTHER=Coll	Style=Coll
OTHER=Arch	Style=Arch
OTHER=Err	Typo=Yes
PCP	PartForm and VerbForm=Part
POSS	Person[psor] and Number[psor]
V[SUBCAT=Neg]	Negative=Yes
SUBCAT=Pfx	-
Pron[SUBCAT]	PronType or Reflex
Adp[SUBCAT]	AdpType
SUBCAT=Card|Ord	NumType
NUM	Number
TENSE	Tense
VOICE	Voice
PRS	Person and Number
ABBR	Abbr
ACRO	Abbr
not INF and not PCP	VerbForm=Fin
FOREIGN[...]	Foreign

Table 2: Morphological feature mapping. FEATURE denotes a mapping that applies for all features with the given name, FEATURE=VALUE for a specific name-value pair, and TAG[FEATURE=VALUE] also for a specific POS tag. Person[psor] and Number[psor] are layered UD features for Person and Number of possessor, respectively.

While the great majority of UD Finnish features could be deterministically generated by reference only to the TDT POS tag and features, there were a few cases that required more complex heuristics to meet UD requirements. For example, the value of the Person feature is assigned to personal pronouns based on a lemma lookup table as OMorFi does not generate it, and the value of the Foreign value is assigned based on comparison of characters in the surface form against Unicode script tables.

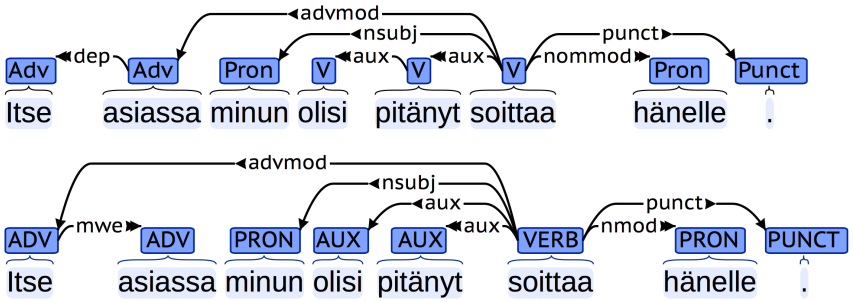

Figure 1: Top: TDT-style syntax and part-of-speech annotation for a Finnish sentence. Bottom: The same sentence converted to the UD Finnish scheme. Analyses visualized using BRAT (Stenetorp et al., 2012).

2.4 Dependency annotation

UD defines as set of 40 broadly applicable dependency relations, further allowing language-specific subtypes of these to be defined to meet the needs of specific resources. Unlike the fairly straightforward mappings for morphological annotations, the conversion from TDT dependency annotation to UD often required not only relabeling types, but also changes to the tree structure. This mapping is summarized in Table 3 and presented in detail below.

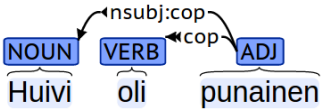

Figure 2: Annotation of *Huivi oli punainen* "The scarf was red".

The UD syntactic annotation is based on the universal Stanford Dependencies (SD) scheme (de Marneffe et al., 2014). One of the key properties of these schemes is that they emphasize direct relations between content words, treating function words as dependents of content words rather than as their heads. For example, this leads to a structure where a copula subject is attached directly to the predicative with the copular verb also becoming a dependent of the predicative (Figure 2). Furthermore, function words can only have a very limited set of dependents, with strong preference given to attachment of function words to content words rather than to other function words. This will tend to produce relatively flat tree structures.

The UD emphasis on content words is not universally shared with other dependency annotation schemes, many of which mediate connections between content words through function words. However, TDT is originally annotated using a language-specific variant of the SD scheme, and thus already applies an annotation scheme with predicatives as heads in copular expressions and content-word heads in prepositional phrases. The conversion of the syntactic annotation to UD thus involved fewer challenges than might be encountered for other treebanks.

During the conversion, relatively few structural reconfigurations were required. In the original TDT annotation, function words were allowed to have dependents of their own, permitting e.g. chains of auxiliary verbs (see Figure 1). These modifiers were reattached to the upper-level content words. Additionally, multi-word expressions and names were annotated with head-final structures in TDT, but UD specifies head-initial annotation for all expressions that do not have internal structure of their own. For UD Finnish, multi-word expressions were revised to follow the UD head-initial approach. However, the head-final structure was kept for names. This decision reflects the fact that in Finnish multi-word names, only the last word carries the morphological inflections, providing evidence that it is the head of the phrase. By contrast, fixed multi-word expressions (UD `mwe`) do not typically inflect, and thus do not provide sufficient cause to diverge from the UD guideline of head-initial annotation.

One problematic issue arose from the fact that UD makes a systematic distinction between core arguments and other modifiers, which are only partly distinguished in TDT annotation. For example, participial modifiers of predicates, which usually include also secondary predication, were annotated simply as participial modifiers in TDT, while in UD these are seen as clausal dependents and a distinction must thus be made between com-

Unchanged types
advcl, amod, appos, aux, auxpass, cc, conj, cop, csubj, det, dobj, mark, name, nsubj, neg, root, parataxis, xcomp
Simple mapping
acomp → xcomp, adpos → case, compar → advcl, comparator → mark, complm → mark, csubj-cop → csubj:cop, gobj → nmod:gobj, gsubj → nmod:gsubj, iccomp → xcomp:ds, infmod → acl, intj → discourse, nommod-own → nmod:own, nsubj-cop → nsubj:cop, num → nummod, number → compound, poss → nmod:poss, preconj → cc:preconj, prt → compound:prt, quantmod → advmod, rcmod → acl:relcl, voc → vocative, xsubj → nsubj, xsubj-cop → nsubj:cop
More complex mapping
advmod → advmod, cc, mark ccomp → ccomp, xcomp:ds dep → dep, mwe nommod → nmod, xcomp, xcomp:ds nn → compound:nn, goeswith partmod → acl, advcl, ccomp, xcomp, xcomp:ds punct → discourse, punct ∅ → remnant
Unmapped TDT types (removed)
ellipsis, rel
Unused UD types
csubjpass, dislocated, foreign, expl, iobj, list, nsubjpass, reparandum

Table 3: Dependency type mapping from TDT to UD Finnish.

plements and adjuncts. To implement the conversion for cases like these, we made reference to the manually annotated predicate-argument structures of the Finnish Propbank (Haverinen et al., 2013a). Since the Finnish Propbank and the Turku Dependency Treebank are built on top of the same texts, we had access to semantic information where each argument is marked to identify whether it serves as a core argument or a modifier.

In some cases the original TDT annotation is more fine-grained than the relation types defined in the UD guidelines. We use two approaches to resolve this issue in UD Finnish. First, most of the more specific dependency types not defined in UD are simply dropped from UD Finnish, replacing occurrences of the types with their more general UD types. This is done in particular for TDT types that are not specific to Finnish and encode distinctions not targeted in UD syntactic relations, such as the difference between finite and non-finite clauses (cf. SD `partmod` and `infmod`). However, some fine-grained dependencies were defined in the TDT variant of the SD scheme to capture properties that are unique or especially important to the Finnish language. We introduce some of these relations also in UD Finnish as subtypes of UD relations. This allows us to preserve the information while allowing a fully comparable UD analysis to be generated by simply replacing detailed types with those that they are subtypes of. For example, Finnish does not have a specific verb expressing ownership (such as *to have* in English), and typically the verb *olla* "to be" is used instead with the owner expressed with a nominal modifier. The surface forms of possessive clauses and existential clauses are similar (*Minulla on koira* "I have a dog", lit. *At_me is a dog* and *Pihalla on koira* "These is a dog in the yard"), and using the standard nominal modifier type `nmod` for both would fail to distinguish these constructions. Thus, UD Finnish carries over the original TDT distinction and defines a language-specific subtype `nmod:own` to address this issue. `nmod:own` can then trivially be mapped to `nmod` when the distinction is not required.

The total number of dependency relation types defined in UD Finnish is 43, consisting of 32 universal relations and 11 language-specific subtypes. In the original TDT annotation, 46 dependency types are used, with an additional 4 types to mark non-tree structures used in the second annotation layer of TDT. In UD Finnish, the second annotation layer does not expand the set of dependency types. Although not currently formalized in UD, the *extended* layer of annotation from TDT (Haverinen et al., 2013b) was converted as well and is included in the UD version of TDT. This extended TDT layer includes (1) conjunct propagation, where dependencies of the head of a coordination structure are propagated where applicable also to the other coordinated elements, (2) external subjects (`xsubj`) of open clausal complements,

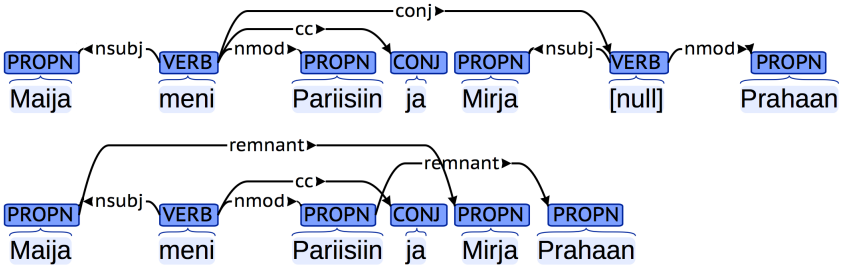

Figure 3: TDT-style (top) and UD-style (bottom) analysis for the sentence *Maija meni Pariisiin ja Mirja Prahaan* "Maija went to Paris and Mirja to Prague".

(3) `name` dependencies marking named entities spanning several words and having some internal syntactic structure, (4) dependencies marking the syntactic function of relativizers, and (5) the `ellipsis` dependency marking constructions involving ellipsis. Of these, conjunct propagation is converted using the same rules as the base syntax dependencies, external subjects are renamed to the standard subject relation `nsubj` or the language-specific `nsubj:cop` copula subject relation, the `name` dependencies are preserved, dependencies marking the syntactic function of relativizers are converted and placed into the base layer, replacing the `rel` dependency type (which is eliminated) and `ellipsis` dependencies are removed together with the `null` nodes they marked. (We refer to the UD Finnish documentation for further details.)

2.4.1 Implementation

While POS tags and morphological features could be mapped with rules affecting a single word and only referencing properties of that word, the dependency annotation mapping requires changes to the tree structure and the ability to refer to a wider syntactic context in mapping rules. The conversion is implemented using the *dep2dep* tool which allows rules that produce dependencies in the output tree based on an input tree context that can be specified in considerable detail: it can match subtree structures, specify negations (e.g. *does not have a property, dependent, or subtree*), refer to the morphological layer, the linear order of tokens, and to additional meta-data such as PropBank argument roles. The tool is implemented as a compiler that converts the source expressions into predicates in Prolog, which is then used to apply the rules.

As an illustration, consider the rule below, which specifies that an `advcl` UD dependency is to be produced between a verb and its participial modifier `partmod` in the transitive case, providing that the participle is not a core argument of the verb in the PropBank.

```
[v p ('advcl')] : [
  @[v-"POS_V" p-"CASE=Tra" ("partmod")]
  ![v p ("Arg_.*")]
]
```

In total, the conversion consists of 116 such rules, of which 22 are simple direct dependency renamings, and the remaining refer to a broader context. We note that these rules did not aim to be universal or exhaustive: a small number of dependencies, on the order of 250, were not covered by the rules and were edited manually upon conversion. This was more efficient than writing rules that only apply to generate very few or only single dependencies.

2.4.2 Null tokens

In many situations sentences can be incomplete and elements obvious from the context can be omitted. In *gapping*, an elliptic sentence element is omitted to avoid unnecessary repetition, whereas in *sentence fragments* the main predicate is absent. The analysis of fragments and sentences including gapping is difficult, and many different approaches have been proposed. In TDT the omitted token, most commonly a verb, is replaced with a *null* token, which is given a full morphological analysis and which acts as a normal token in the syntactic analysis.

UD takes a different approach to analyzing omitted sentence elements. UD aims in general to avoid representing things that are absent, and does not define a way to introduce null tokens. Instead, for example to address coordination with ellipsis, UD introduces a special dependency type `remnant`. Thus, e.g. *Maija meni Pariisiin ja Mirja Prahaan* "Maija went to Paris and Mirja to Prague" is analysed with an empty token representing *meni* "went" in the second constituent in

Language	Tokens	Source treebank
Czech	1,506,490	Prague Dependency Treebank 3.0 (PDT) (Bejček et al., 2012)
Spanish	432,651	Universal Dependency Treebank v2.0 (UDT) (McDonald et al., 2013)
French	400,620	Universal Dependency Treebank v2.0 (UDT) (McDonald et al., 2013)
German	298,614	Universal Dependency Treebank v2.0 (UDT) (McDonald et al., 2013)
English	254,830	English Web Treebank v1.0 (EWT) (Silveira et al., 2014)
Italian	214,748	Italian Stanford Dependency Treebank (ISDT) (Simi et al., 2014)
Finnish	**202,085**	Turku Dependency Treebank (TDT) (Haverinen et al., 2013b)
Swedish	96,819	Talbanken (Nivre, 2014)
Hungarian	26,538	Szeged Treebank (Farkas et al., 2012)
Irish	23,686	Irish Dependency Treebank (IDT) (Lynn et al., 2014)

Table 4: Statistics of the UD Finnish treebank in comparison to the other treebanks included in the first UD data release.

TDT, but with `remnant` relations between *Maija* and *Mirja* and between *Pariisiin* and *Prahaan* in UD Finnish (see Figure 3). We applied a combination of custom scripts and manual reannotation to resolve empty nodes in the conversion of TDT to UD Finnish.

2.5 Annotation statistics

Table 4 shows token statistics for the 10 languages for which treebanks were included in the initial UD data release. With over 200,000 tokens, the UD Finnish treebank is in a mid-size cluster among the UD version 1 languages together with German, English and Italian. This is a relatively prominent position for Finnish, which until recently had no publicly available treebanks. We hope that the availability of this corpus will encourage further interest in Finnish dependency parsing.

3 Experiments

As discussed by de Marneffe et al. (2014) in the context of the Universal Stanford Dependencies which formed the basis on which UD was built, parsing accuracy has not been a major consideration in the definition of the scheme. In fact, a number of the design choices taken, such as the attachment of auxiliaries and prepositions as dependents rather than governors of their semantic head is known to result in a numerically worse parsing accuracy. Additionally, as the conversion is an automatic process, the resulting noise may have a detrimental effect on parsing accuracy as well. To quantify these effects, we carry out several parsing experiments, comparing the Stanford Dependencies annotation in TDT with its conversion to the UD format. Further, since TDT now contains also fully manually annotated morphology, we will pay extra attention to morphological processing in the evaluation.

We base the experiments on the publicly available Finnish parsing pipeline.[6] The pipeline uses the CRF-based tagger Marmot (Müller et al., 2013), in conjunction with the two-level morphological analyzer OMorFi (Pirinen, 2008; Lindén et al., 2009). The morphological analyzer is used to provide the set of possible morphological readings (lemma, POS, and features) of every recognized word, which are subsequently given as features to the Marmot tagger. We initially apply a *hard* constraint approach, where the output of the tagger is used to select one of these readings (the reading with the highest overlap of tags and a priority for readings matching the main POS), effectively disambiguating OMorFi output. For words not recognized by OMorFi, the reading produced by Marmot is used as-is, and the wordform itself is used in place of the lemma. This has so far been the strategy taken when learning to parse Finnish (Bohnet et al., 2013). The tagged text is then parsed with the Mate tools graph-based dependency parser (Bohnet, 2010).[7]

As baseline, we consider the most recent Finnish dependency parser trained and evaluated on the original distribution of TDT. Note that the test sets differ: the baseline is evaluated on a test set matching the data it was trained on, which differs from the new test set in several aspects such as the treatment of named entities. The results are thus broadly comparable, but not directly so.

[6] http://turkunlp.github.io/Finnish-dep-parser/

[7] https://code.google.com/p/mate-tools

	POS	PM	FM	LAS	UAS
Baseline (Haverinen et al., 2013b)	94.3	90.5	89.0	81.4	85.2
Stanford Dependencies (SD)	96.3	93.4	90.3	80.1	84.1
Universal Dependencies (UD)	96.0	93.1	90.5	81.0	85.0
Pure Universal Dependencies (Pure UD)	96.0	93.1	90.5	81.5	84.7

Table 5: Results of the parsing experiments. *SD* refers to the morphological tagset and dependency relations as defined in TDT, *UD* to the universal tagset and relations, and *pure UD* to UD relations with no language-specific extensions. *POS* is the POS tagging accuracy, *PM* the accuracy of POS and all features, *FM* the accuracy of full morphology (including the lemma), and *LAS* and *UAS* are the standard labeled and unlabeled attachment score metrics.

	POS	PM	FM	LAS	UAS
Universal Dependencies (soft)	97.0	93.0	89.3	81.5	85.4
Universal Dependencies (hard-pos)	97.0	94.0	90.7	82.1	85.8
Pure Universal Dependencies (soft)	97.0	93.0	89.3	82.0	84.9
Pure Universal Dependencies (hard-pos)	97.0	94.0	90.7	82.7	85.4

Table 6: Results of the UD parsing experiments with the *soft* and *hard-pos* morphological tagging strategies.

The results are summarized in Table 5. Firstly, we see that all results are roughly comparable, meaning that the conversion to UD has had no major effect on the parsing accuracy. However, the attachment scores are somewhat lower compared to the baseline, likely due at least in part to the different treatment of named entities in the previously published baseline parser as opposed to both the newly introduced SD and UD versions of TDT. Unsurprisingly, the labeled attachment score is slightly higher for the pure UD scheme with no language-specific relations.

We additionally focused on morphological tagging. As TDT now contains manual morphological annotation, the analyses are no longer tightly bound to OMorFi as they were in the original release of TDT. We therefore consider also a *soft* constraint approach, where the tags given by Marmot are preserved, and OMorFi is only used to select the lemma (from the reading with the highest overlap of tags). This results in morphological analyses superior in POS accuracy but inferior in the prediction of full features. To address this issue, we implemented a new tagging strategy that applies the hard constraint only in cases where the predicted POS can be found among the analyses given by OMorFi (referred to as *hard-pos*). The results show an across-the-board improvement for this strategy as well as numerically the best scores for Finnish with the graph-based parser of Bohnet (2010) (Table 6).

4 Conclusions

We have presented Universal Dependencies (UD) for Finnish, detailing the application of general UD guidelines to the annotation of parts-of-speech, morphological features, and dependency relations in Finnish and introducing a conversion from the previously released Turku Dependency Treebank corpus into the UD Finnish treebank released in the first UD data release. We also performed experiments evaluating a state-of-the-art parser on both the source treebank, TDT, and the target UD Finnish treebank, finding that performance is slightly improved in the conversion, which supports both the accuracy of the conversion and the feasibility of UD as a parsing target.

All of the tools and resources described in this work are available under open licenses from `http://bionlp.utu.fi/ud-finnish.html`.

Acknowledgments

This work was supported by the Kone Foundation and the Emil Aaltonen Foundation. Computational resources were provided by CSC - IT Center for Science. This paper builds on joint work with Jinho Choi, Marie-Catherine de Marneffe, Tim Dozat, Yoav Goldberg, Jan Hajič, Christopher Manning, Ryan McDonald, Joakim Nivre, Slav Petrov, Natalia Silveira, Reut Tsarfaty, and Dan Zeman.

References

Bejček, E., Panevová, J., Popelka, J., Straňák, P., Ševčíková, M., Štěpánek, J., and Žabokrtskỳ, Z. (2012). Prague dependency treebank 2.5 – a revisited version of pdt 2.0. In *Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012)*, pages 231–246.

Bohnet, B. (2010). Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of COLING'10*, pages 89–97.

Bohnet, B., Nivre, J., Boguslavsky, I., Farkas, R., Ginter, F., and Hajič, J. (2013). Joint morphological and syntactic analysis for richly inflected languages. *Transactions of the Association for Computational Linguistics*, 1:415–428.

Bosco, C., Montemagni, S., and Simi, M. (2013). Converting italian treebanks: Towards an italian stanford dependency treebank. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 61–69.

de Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., and Manning, C. D. (2014). Universal Stanford Dependencies: A cross-linguistic typology. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, volume 14, pages 4585–4592.

de Marneffe, M.-C., MacCartney, B., and Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, volume 6, pages 449–454.

Farkas, R., Vincze, V., and Schmid, H. (2012). Dependency parsing of hungarian: Baseline results and challenges. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 55–65.

Hakulinen, A., Korhonen, R., Vilkuna, M., and Koivisto, V. (2004). *Iso suomen kielioppi*. Suomalaisen kirjallisuuden seura.

Haverinen, K., Laippala, V., Kohonen, S., Missil, A., Nyblom, J., Ojala, S., Viljanen, T., Salakoski, T., and Ginter, F. (2013a). Towards a dependency-based propbank of general finnish. In *Proceedings of the 19th Nordic Conference on Computational Linguistics (NoDaLiDa'13)*, pages 41–57.

Haverinen, K., Nyblom, J., Viljanen, T., Laippala, V., Kohonen, S., Missilä, A., Ojala, S., Salakoski, T., and Ginter, F. (2013b). Building the essential resources for finnish: the Turku Dependency Treebank. *Language Resources and Evaluation*, pages 1–39.

Kanerva, J., Luotolahti, J., Laippala, V., and Ginter, F. (2014). Syntactic n-gram collection from a large-scale corpus of internet finnish. In *Proceedings of the Sixth International Conference Baltic HLT*, pages 184–191.

Lindén, K., Silfverberg, M., and Pirinen, T. (2009). HFST tools for morphology — an efficient open-source package for construction of morphological analyzers. In *State of the Art in Computational Morphology*, volume 41 of *Communications in Computer and Information Science*, pages 28–47.

Lynn, T., Foster, J., Dras, M., and Tounsi, L. (2014). Cross-lingual transfer parsing for low-resourced languages: An Irish case study. In *Proceedings of the First Celtic Language Technology Workshop*, pages 41–49.

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 (ACL)*, pages 92–97.

Müller, T., Schmid, H., and Schütze, H. (2013). Efficient higher-order CRFs for morphological tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*.

Nivre, J. (2014). Universal Dependencies for Swedish. In *SLTC 2014*.

Nivre, J., Bosco, C., Choi, J., de Marneffe, M.-C., Dozat, T., Farkas, R., Foster, J., Ginter, F., Goldberg, Y., Hajič, J., Kanerva, J., Laippala, V., Lenci, A., Lynn, T., Manning, C., McDonald, R., Missilä, A., Montemagni, S., Petrov, S., Pyysalo, S., Silveira, N., Simi, M., Smith, A., Tsarfaty, R., Vincze, V., and Zeman, D. (2015). Universal dependencies 1.0.

Nivre, J., Choi, J., de Marneffe, M.-C., Dozat, T.,

Ginter, F., Goldberg, Y., Hajič, J., Manning, C., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2014). Universal dependencies documentation 1.0.

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

Pirinen, T. (2008). Suomen kielen äärellistilainen automaattinen morfologinen jäsennin avoimen lähdekoodin resurssein. Master's thesis, University of Helsinki.

Silveira, N., Dozat, T., de Marneffe, M.-C., Bowman, S., Connor, M., Bauer, J., and Manning, C. D. (2014). A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

Simi, M., Bosco, C., and Montemagni, S. (2014). Less is more? towards a reduced inventory of categories for training a parser for the italian stanford dependencies. In *Proceedings of LREC 2014*.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.

Sulkala, H. and Karjalainen, M. (1992). *Finnish*. Descriptive Grammar Series. Routledge, London.

Tsarfaty, R. (2013). A unified morpho-syntactic scheme of stanford dependencies. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 578–584.

Voutilainen, A. (2011). FinnTreeBank: Creating a research resource and service for language researchers with Constraint Grammar. In *Proceedings of the NODALIDA 2011 workshop Constraint Grammar Applications*.

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