# Universal Dependencies v1: A Multilingual Treebank Collection

**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**[⋆] **Daniel Zeman**[‡]
[*]Uppsala University [○]The Ohio State University [•]University of Turku
[†]Bar-Ilan University [‡]Charles University in Prague [◇]Stanford University
[◁]Google Inc. [▷]University of Cambridge [⋆]The Open University of Israel

## 1. Introduction

This paper presents the resources so far created in the Universal Dependencies (UD) project, which attempts to address a lack of cross-linguistically adequate dependency representations for natural language processing. Multilingual research on parsing has for a long time been hampered by the fact that annotation schemes vary enormously across languages, which makes it virtually impossible to perform sound comparative evaluations and cross-lingual learning experiments, as well as difficult to develop and maintain multilingual NLP systems based on parsing technology. A striking illustration of this problem can be found in Figure 1, which shows three sentences in Swedish, Danish and English, which are 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 the percentage of shared dependency relations across pairs of languages is at most 40% (and 0% across all three languages). As a consequence, a parser trained on one type of annotation and evaluated on another type will be found to have at least a 60% error rate when it functions perfectly. To facilitate multilingual parser development and cross-lingual learning from a language typology perspective, there is a crucial need for a large collection of treebanks that all follow a common representation, which is able to capture the similarities as well as the idiosyncrasies of disparate language families (e.g., morphologically rich languages, pro-drop languages, languages featuring clitic doubling).

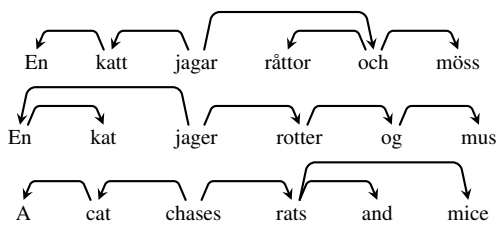


Figure 1: Divergent annotation of parallel structures

Several separate initiatives exist to build consistent resources for many languages, and the UD project is a merger of some of the initiatives. It combines the (universal) Stanford dependencies (de Marneffe et al., 2006; de Marneffe and Manning, 2008; de Marneffe et al., 2014), the universal Google dependency scheme (Universal Dependency Treebanks) (McDonald et al., 2013), the Google universal part-of-speech tags (Petrov et al., 2012), and the Interset interlingua for morphosyntactic tag sets (Zeman, 2008) used in the HamleDT treebanks (a project that transforms existing treebanks under a common annotation scheme, Zeman et al. 2012). UD is thus based on common usage and existing de facto standards, and is intended to replace all the previous versions by a single coherent standard.[1] The general philosophy is to provide a universal inventory of categories and guidelines to facilitate consistent annotation of similar constructions across languages, while allowing language-specific extensions when necessary. We present version 1 of the universal guidelines, the underlying design principles, and the 19 treebanks that constitute the latest release (v1.1). Guidelines for specific languages can be found at `http://universaldependencies.github.io/docs/`.

## 2. History

UD comprises three layers of annotation (parts-of-speech, morphological features, and syntactic dependencies) with diverse origins. The Google universal tag set grew out of the cross-linguistic error analysis based on the CoNLL-X shared task data by McDonald and Nivre (2007). It was initially used for unsupervised part-of-speech tagging by Das and Petrov (2011), and has been adopted as a widely used stan-

[1]HamleDT still exists as an independent project but, from version 3.0., it uses the UD standard.

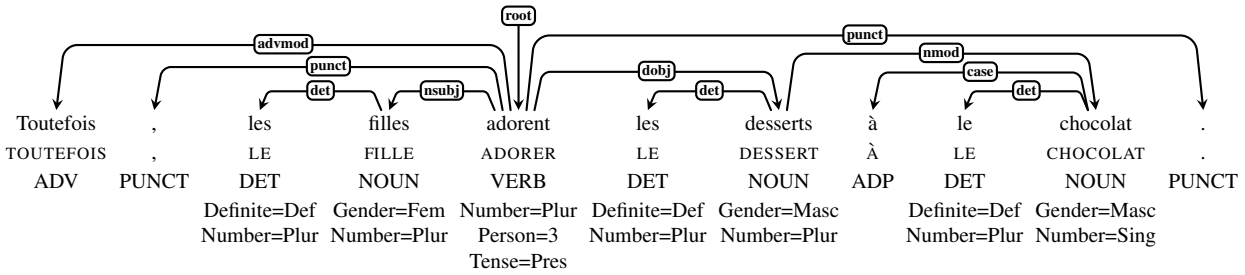


Figure 2: UD annotation for a French sentence. (Translation: However, girls love chocolate desserts.)

| Open class words | | Closed class words | | Other | | Lexical | Inflectional | |
|---|---|---|---|---|---|---|---|---|
| ADJ | adjective | ADP | preposition/postposition | PUNCT | punctuation | | (Nominal) | (Verbal) |
| ADV | adverb | **AUX** | auxiliary | **SYM** | symbol | PronType | Gender | VerbForm |
| **INTJ** | interjection | CONJ | coordinating conjunction | X | unspecified POS | NumType | Animacy | Mood |
| NOUN | noun | DET | determiner | | | Poss | Number | Tense |
| **PROPN** | proper noun | NUM | numeral | | | Reflex | Case | Aspect |
| VERB | verb | PART | particle | | | | Definite | Voice |
| | | PRON | pronoun | | | | Degree | Person |
| | | **SCONJ** | subordinating conjunction | | | | | Negative |

Table 1: **Columns 1–3:** Part-of-speech tags in UD v1. (Bold indicates addition to the original Google tag set, including AUX and PROPN which were added in UDT.) **Column 4:** Morphological features in UD v1.

dard for mapping diverse tag sets to a common standard. Interset (Zeman, 2008) started as a tool for conversion between morphosyntactic tag sets of multiple languages. It dates back to 2006 when it was used in the first experiments with cross-lingual delexicalized parser adaptation (Zeman and Resnik, 2008). The Stanford dependencies, developed for English in 2005, eventually emerged as the de facto standard for dependency analysis of English, and have since been adapted to a number of different languages (Chang et al., 2009; Bosco et al., 2013; Haverinen et al., 2013; Seraji et al., 2013; Lipenkova and Souček, 2014).

These resources have featured in other attempts at universal standards. The Google Universal Dependency Treebank (UDT) project (McDonald et al., 2013)) was the first attempt to combine the Stanford dependencies and the Google universal part-of-speech tags into a universal annotation scheme: treebanks for 6 languages (English, French, German, Spanish, Swedish and Korean) were released in 2013, and for 11 languages in 2014 (Brazilian Portuguese, English, Finnish, French, German, Italian, Indonesian, Japanese, Korean, Spanish and Swedish). The first proposal for incorporating morphology was made by Tsarfaty (2013). The second version of HamleDT (Rosa et al., 2014) provided Stanford/Google annotation for 30 languages, by automatically harmonizing treebanks native to different annotations. These efforts were followed by the development of the universal Stanford dependencies (USD) which revised Stanford Dependencies for cross-linguistic annotations, in light of the Google scheme (de Marneffe et al., 2014).

UD is the result of merging all these initiatives into a single coherent framework, based on the universal Stanford dependencies, an extended version of the Google universal tag set, a revised subset of the Interset feature inventory, and a revised version of the CoNLL-X format (which we call "CoNLL-U"). The first version of the annotation guidelines were released in October 2014. In January 2015, treebanks for 10 languages were released, and in May 2015, 18 languages were released (see Table 3).

## 3. Annotation Guideline Principles

The syntactic annotation in UD is based on *dependency*, which is widely used in contemporary NLP, both for treebank annotation and as a parsing representation. It is also based on *lexicalism*, the idea that words are the basic units of grammatical annotation. Words have morphological properties and enter into syntactic relations, which is what the UD annotation is primarily meant to capture. To arrive at an adequate grammatical representation, it is important to note that syntactic wordhood does not always coincide with whitespace-separated orthographic units, and a final important design consideration is that there should be a transparent relation between the original textual representation and the linguistically motivated word segmentation. We call this the *recoverability* principle.

To obtain a cross-linguistically consistent and transparent annotation, we want to maximize the paral-

| | | |
|---|---|---|
| **Core dependents of clausal predicates** | | |
| *Nominal dep* | *Predicate dep* | |
| nsubj | csubj | |
| nsubjpass | csubjpass | |
| dobj | ccomp | xcomp |
| iobj | | |
| **Non-core dependents of clausal predicates** | | |
| *Nominal dep* | *Predicate dep* | *Modifier word* |
| nmod | advcl | advmod |
| | | neg |
| **Special clausal dependents** | | |
| *Nominal dep* | *Auxiliary* | *Other* |
| vocative | aux | mark |
| discourse | auxpass | punct |
| expl | cop | |
| **Noun dependents** | | |
| *Nominal dep* | *Predicate dep* | *Modifier word* |
| nummod | acl | amod |
| appos | | det |
| nmod | | neg |
| **Case-marking, prepositions, possessive** | | |
| case | | |
| **Coordination** | | |
| conj | cc | punct |
| **Compounding and unanalyzed** | | |
| compound | mwe | goeswith |
| name | foreign | |
| **Loose joining relations** | | |
| list | parataxis | remnant |
| dislocated | | reparandum |
| **Other** | | |
| *Sentence head* | *Unspecified dependency* | |
| root | dep | |

Table 2: The 40 dependency relations in UD. Note: *nmod*, *neg* and *punct* appear in two places.

lelism between languages and make sure that the same construction is annotated in the same way across languages. At the same time, we do not want to go too far and, in particular, we do not want to annotate things that do not exist in a language simply because it exists in other languages. The idea is to use a universal pool of structural and functional categories that languages select from. Moreover, it should be possible to refine the analysis by adding language-specific subtypes of universal categories.

Figure 2 uses a French sentence, "Toutefois les filles adorent les desserts au chocolat" (*However girls love chocolate desserts.*), to exemplify the different UD annotation layers, which are described in more detail in the following sections.

## Word Segmentation

Following the lexicalist view, the basic annotation units in UD are syntactic words (not phonological or orthographic words). Concretely, clitics are split off (e.g., Spanish *dámelo* 'give me it' = *dá me lo*) and contractions are undone (e.g., French *au* = *à le*; see Figure 2), but for recoverability the original tokens are included as well. UD currently does not allow words with spaces, and even though the lexicalist view could be taken to imply that multiword expressions should be treated as single words, multiword expressions are annotated using special dependency relations, rather than by collapsing multiple tokens into one.

## Morphology

The morphological specification of words in UD consists of three levels of information: a lemma, a part-of-speech tag as well as a set of features which encode lexical and grammatical properties associated with the word form (see Figure 2). Table 1 lists the part-of-speech tags, which come from a revised version of the Google universal POS, as well as the morphological features, based on the Interset system. Each feature has defined values (e.g., 'Number' can be 'singular', 'plural', 'dual', 'plurale tantum' or 'collective'). Languages select the features and values that are relevant.

## Syntax

Currently UD contains 40 grammatical relations between words, listed in Table 2. The organization of the relations distinguishes between three types of structure: nominals, clauses and modifiers. The scheme makes a distinction between core arguments (e.g., subject and object) vs. other dependents, but does not attempt to distinguish complements vs. adjuncts.

Each word depends either on another word in the sentence or on a notional "root" of the sentence, following three principles: content words are related by dependency relations; function words attach to the content word they further specify; and punctuation attaches to the head of the phrase or clause in which it appears, as illustrated in Figure 2. Giving priority to dependency relations between content words increases the probability of finding parallel structures across languages, since function words in one language often correspond to morphological inflection (or nothing at all) in other languages.

In addition to universal relations, UD allows the use of language-specific subtypes to capture special phenomena in different languages. For instance, while the universal UD scheme has a single relation *acl* for adnominal clauses, several languages make use of the subtype *acl:relcl* to distinguish relative clauses as an important subtype of adnominal clauses. By design, we can always map back to the core label set by stripping the specific relations that appear after the colon. For a complete list of currently used language-specific relations, we refer to the UD website.

| ID | FORM | LEMMA | CPOSTAG | POSTAG | FEATS | HEAD | DEPREL | DEPS | MISC |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Toutefois | toutefois | ADV | _ | _ | 5 | advmod | _ | _ |
| 2 | , | , | PUNCT | _ | _ | 5 | punct | _ | _ |
| 3 | les | le | DET | _ | Definite=Det\|Number=Plur | 4 | det | _ | _ |
| 4 | filles | fille | NOUN | _ | Gender=Fem\|Number=Plur | 5 | nsubj | _ | _ |
| 5 | adorent | adorer | VERB | _ | Number=Plur\|Person=3\|Tense=Pres | 0 | root | _ | _ |
| 6 | les | le | DET | _ | Definite=Det\|Number=Plur | 7 | det | _ | _ |
| 7 | desserts | dessert | NOUN | _ | Gender=Masc\|Number=Plur | 5 | dobj | _ | _ |
| 8-9 | au | _ | _ | _ | _ | _ | _ | _ | _ |
| 8 | à | à | ADP | _ | _ | 10 | case | _ | _ |
| 9 | le | le | DET | _ | Definite=Def\|Gender=Masc\|Number=Sing | 10 | det | _ | _ |
| 10 | chocolat | chocolat | NOUN | _ | Gender=Masc\|Number=Sing | 7 | nmod | _ | _ |
| 11 | . | . | PUNCT | _ | _ | 5 | punct | _ | _ |

Figure 3: The French sentence from Figure 2 in CoNLL-U format.

| Language | Sentence | Token | Word | Lemma | PoS | Feat | Dep | LDep |
|---|---|---|---|---|---|---|---|---|
| Basque | 5273 | 60563 | 60563 | 7395 | 16 | 69 | 27 | 0 |
| Bulgarian | 9405 | 125592 | 125592 | 13507 | 16 | 42 | 30 | 0 |
| Croatian | 3957 | 87765 | 87765 | 8827 | 14 | 38 | 39 | 0 |
| Czech | 87913 | 1503738 | 1506490 | 57900 | 17 | 82 | 39 | 6 |
| Danish | 5512 | 100238 | 100238 | 13276 | 17 | 47 | 31 | 5 |
| English | 16622 | 254830 | 254830 | 16295 | 17 | 34 | 46 | 7 |
| Finnish | 13581 | 181022 | 181022 | 23932 | 15 | 84 | 43 | 11 |
| Finnish-FTB | 19097 | 161682 | 161984 | 21571 | 14 | 64 | 25 | 2 |
| French | 16468 | 388892 | 400627 | 0 | 17 | 0 | 34 | 2 |
| German | 15918 | 293459 | 298614 | 0 | 15 | 0 | 33 | 1 |
| Greek | 2411 | 59156 | 59156 | 6174 | 10 | 30 | 28 | 1 |
| Hebrew | 6216 | 115535 | 158855 | 0 | 16 | 40 | 43 | 14 |
| Hungarian | 1299 | 26538 | 26538 | 6366 | 16 | 79 | 51 | 22 |
| Indonesian | 5593 | 121923 | 121923 | 0 | 16 | 0 | 30 | 0 |
| Irish | 1020 | 23686 | 23686 | 3916 | 16 | 0 | 36 | 10 |
| Italian | 12330 | 258308 | 277209 | 17852 | 17 | 36 | 38 | 3 |
| Persian | 6000 | 151671 | 152918 | 0 | 15 | 14 | 37 | 7 |
| Spanish | 16006 | 424384 | 432651 | 0 | 16 | 27 | 39 | 4 |
| Swedish | 6026 | 96819 | 96819 | 10252 | 15 | 27 | 39 | 4 |
| **TOTAL** | 250647 | 4435801 | 4527480 | | | | | |

Table 3: Statistics on treebanks released in UD v1.1. Sentence: number of sentences. Token: number of unsegmented tokens. Word: number of segmented (syntactic words). Lemma: number of unique lemmas. PoS: number of unique part-of-speech tags. Feat: number of unique Feature=Value pairs. Dep: number of unique dependency relations. LDep: number of language-specific dependency relations. (Note that most of the zeros are due to missing annotation layers rather than language-specific properties.)

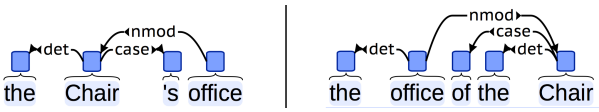


Figure 4: Examples of annotation visualization from UD documentation.

## Format and Tools

The data is encoded in the CoNLL-U format, which is an evolution of the widely used CoNLL-X format (Buchholz and Marsi, 2006), where each word/token is represented in tab-separated columns on one line and sentence boundaries are marked by blank lines. The first change is that the ID column has been redefined to allow the representation of both unsegmented tokens and the syntactic words they correspond to in order to promote recoverability. The second change is that the obsolete PHEAD and PDEPREL columns have been replaced by a DEPS column for additional dependencies in the enhanced representation and a MISC column for annotations that do not fit anywhere else. In addition, the CPOSTAG and FEATS columns have been standardized to hold the universal part-of-speech tags and morphological features, respectively. (The remaining POSTAG column can optionally be used for language-specific part-of-speech tags.) The format is illustrated in Figure 3, with the French sentence from Figure 2.

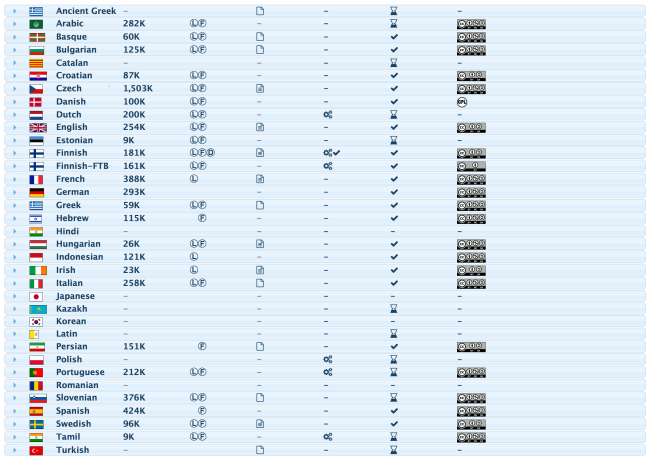


Figure 5: UD treebanks at a glance.

To support work on treebanks in this format, we have introduced Python and JavaScript libraries for reading and validating CoNLL-U.[2] The UD documentation efforts are supported by the Annodoc system (Pyysalo and Ginter, 2014), with annotation visualizations generated using brat (Stenetorp et al., 2012) (see Figure 4). The treebanks can also be queried online using the SETS[3] and PML TreeQuery[4] tools (Luotolahti et al., 2015; Štěpánek and Pajas, 2010).

[2] http://github.com/universaldependencies/.

[3] http://bionlp-www.utu.fi/dep_search

[4] http://lindat.mff.cuni.cz/services/pmltq

## 4. Existing Treebanks

The latest release of UD treebanks (v1.1) comprises 19 treebanks representing 18 languages. They are listed with descriptive statistics in Table 3. All treebanks contain annotation of parts-of-speech and dependency relations. Most treebanks in addition provide lemmas and morphological features. The extent to which the data has been manually annotated or automatically converted from existing treebanks varies per language, and there is a continuing effort to further improve the consistency of the annotation across languages. The plan is to release new versions every six months, gradually adding more languages and improving completeness and consistency of the annotations. Table 3 presents v1.1 but will be updated to reflect the status after the next release in November 2015. The final paper will thus present UD v1.2. Figure 5 shows a screen shot from the web documentation of the UD treebanks currently in progress.

## 5. References


Bosco, C., Montemagni, S., and Simi, M. (2013). Converting Italian treebanks: Towards an Italian Stanford dependency treebank. In *LAW & Interoperability with Discourse*.

Buchholz, S. and Marsi, E. (2006). CoNLL-X shared task on multilingual dependency parsing. In *CoNLL*, 149–164.

Chang, P.-C., Tseng, H., Jurafsky, D., and Manning, C. D. (2009). Discriminative reordering with Chinese grammatical relations features. In *SSST*, 51–59.

Das, D. and Petrov, S. (2011). Unsupervised part-of-speech tagging with bilingual graph-based projections. In *ACL*, 600–609.

de Marneffe, M.-C. and Manning, C. D. (2008). The Stanford typed dependencies representation. In *Workshop on Cross-framework and Cross-domain Parser Evaluation*.

de Marneffe, M.-C., MacCartney, B., and Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. In *LREC*.

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 *LREC*, 4585–4592.

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*. In press. Available online.

Kromann, M. T. (2003). The Danish Dependency Treebank and the DTAG treebank tool. In *TLT*, 217–220.

Lipenkova, J. and Souček, M. (2014). Converting Russian dependency treebank to Stanford typed dependencies representation. In *EACL*.

Luotolahti, J., Kanerva, J., Pyysalo, S., and Ginter, F. (2015). SETS: scalable and efficient tree search in dependency graphs. In *NAACL Demo*, 51–55.

McDonald, R. and Nivre, J. (2007). Characterizing the errors of data-driven dependency parsing models. In *Proceedings of EMNLP-CoNLL*, 122–131.

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 *ACL*.

Nivre, J. and Megyesi, B. (2007). Bootstrapping a Swedish treeebank using cross-corpus harmonization and annotation projection. In *TLT*, 97–102.

Petrov, S., Das, D., and McDonald, R. (2012). A universal part-of-speech tagset. In *LREC*.

Pyysalo, S. and Ginter, F. (2014). Collaborative development of annotation guidelines with application to universal dependencies. In *SLTC*.

Rosa, R., Mašek, J., Mareček, D., Popel, M., Zeman, D., and Žabokrtskỳ, Z. (2014). HamleDT 2.0: Thirty dependency treebanks stanfordized. In *LREC*, 2334–2341.

Seraji, M., Jahani, C., Megyesi, B., and Nivre, J. (2013). Uppsala Persian dependency treebank annotation guidelines. Technical report, Uppsala University.

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 *ACL Demo*, 102–107.

Štěpánek, J. and Pajas, P. (2010). Querying diverse treebanks in a uniform way. In *LREC*.

Tsarfaty, R. (2013). A unified morpho-syntactic scheme of Stanford dependencies. In *ACL*.

Zeman, D. and Resnik, P. (2008). Cross-language parser adaptation between related languages. In *IJCNLP-08 Workshop on NLP for Less Privileged Languages*, 35–42.

Zeman, D., Mareček, D., Popel, M., Ramasamy, L., Štěpánek, J., Žabokrtský, Z., and Hajič, J. (2012). HamleDT: To parse or not to parse? In *LREC*, 2735–2741.

Zeman, D. (2008). Reusable tagset conversion using tagset drivers. In *LREC*, 213–218.