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
joakim.nivre@lingfil.uu.se

[○]The Ohio State University
mcdm@ling.ohio-state.edu

[•]University of Turku
ginter@cs.utu.fi

[†]Bar-Ilan University
yoav.goldberg@gmail.com

[‡]Charles University in Prague
{hajic,zeman}@ufal.mff.cuni.cz

[◇]Stanford University
{manning,natalias}@stanford.edu

[◁]Google Inc.
{ryanmcd,slav}@google.com

[▷]University of Cambridge
sampo@pyysalo.net

[⋆]The Open University of Israel
reutts@openu.ac.il

Abstract

Cross-linguistically consistent annotation is necessary for sound comparative evaluation and cross-lingual learning experiments. It is also useful for multilingual system development and comparative linguistic studies. Universal Dependencies is an open community effort to create cross-linguistically consistent treebank annotation for many languages within a dependency-based lexicalist framework. In this paper, we describe v1 of the universal guidelines, the underlying design principles, and the currently available treebanks for 33 languages.

Keywords: treebanks, dependency, annotation, multilingual, cross-linguistic, universal.

1. Introduction

Multilingual research on syntax and parsing has for a long time been hampered by the fact that annotation schemes vary enormously across languages, which has made it virtually impossible to perform sound comparative evaluations and cross-lingual learning experiments. A striking illustration of this problem can be found in Figure 1, which shows three parallel sentences in Swedish, Danish and English, 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.

The Universal Dependencies (UD) project seeks to tackle this problem by developing cross-linguistically consistent treebank annotation for many languages, aiming to capture similarities as well as idiosyncracies among typologically different languages (e.g., morphologically rich languages, pro-drop languages, and languages featuring clitic doubling). In this way, we hope to be able not only to support comparative evaluation and cross-lingual learning but also to facilitate multilingual natural language processing and enable comparative linguistic studies. To serve all these purposes, the framework needs to have a solid linguistic foundation and at the same time be transparent and accessible to non-specialists.

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

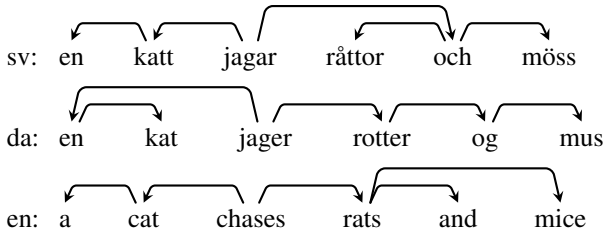


Figure 1: Divergent annotation of parallel structures

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.

In this paper, we present version 1 of the universal guidelines and explain the underlying design principles. We give an overview of the 37 treebanks that constitute the latest release (v1.2), representing 33 different languages, and conclude with a few words about the future of the project. Guidelines for specific languages can be found at `http://universaldependencies.org`.

[1]The UDT project has been deprecated and redirects to UD. HamleDT still exists as an independent project but uses the UD standard from version 3.0.

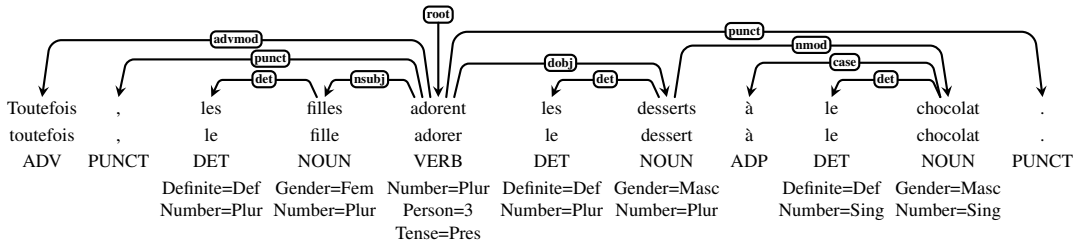


Figure 2: UD annotation for a French sentence. (Translation: However, girls love chocolate desserts.)

2. History

UD comprises two layers of annotation with diverse origins. The Google universal tag set used in the morphological layer 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 standard for mapping diverse tag sets to a common standard. The morphological layer also builds on Interset (Zeman, 2008), which 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, used in the syntactic layer, were developed for English in 2005 and eventually emerged as the de facto standard for dependency analysis of English. They 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 were released for 6 languages in 2013 (English, French, German, Spanish, Swedish and Korean) 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 with different native annotations. These efforts were followed by the development of the universal Stanford dependencies (USD), revising 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. There have been three releases of treebanks, for 10 languages (January 2015), 18 languages (May 2015), and 33 languages (November 2015), respectively.

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 another 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 parallelism 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 they exist 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 the French sentence *Toutefois, les filles adorent les desserts au chocolat* (However, the girls love chocolate desserts) to exemplify the different UD annotation layers, which are described in more detail in the following sections.

3.1. 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) when this is necessitated by the syntactic analysis, 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.

3.2. Morphology

The morphological specification of words in UD consists of three levels of information: a lemma, a part-of-speech

Open class words		Closed class words		Other	
ADJ	adjective	ADP	preposition/postposition	PUNCT	punctuation
ADV	adverb	**AUX**	auxiliary	**SYM**	symbol
INTJ	interjection	CONJ	coordinating conjunction	X	unspecified POS
NOUN	noun	DET	determiner		
PROPN	proper noun	NUM	numeral		
VERB	verb	PART	particle		
		PRON	pronoun		
		SCONJ	subordinating conjunction		

Table 1: 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.)

Lexical	Inflectional	
	(Nominal)	(Verbal)
PronType	Gender	VerbForm
NumType	Animacy	Mood
Poss	Number	Tense
Reflex	Case	Aspect
	Definite	Voice
	Degree	Person
		Negative

Table 2: Morphological features in UD v1.

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 17 part-of-speech tags, which come from a revised version of the Google universal POS, divided into open class words, closed class words, and other symbols. This tag inventory is meant to be fixed and used for all languages, but not all categories have to be used in all languages. For example, the distinction between common nouns (NOUN) and proper nouns (PROPN) is not grammaticalized in all languages.

Table 2 lists the current inventory of morphological features, based on the Interset system. Each feature is associated with a set of values (e.g., Number can take the values Sing[ular], Plur[al], Dual, Ptan [plurale tantum], and Coll[ective]). Languages select the subset of features and values that are relevant, but it is also possible to add new features and values if needed.

3.3. Syntax

In v1, UD contains 40 grammatical relations between words, listed in Table 3.

Grammatical Relations

The organization of the relations distinguishes between three types of structure: nominals, clauses and modifier words. The scheme also makes a distinction between core arguments (e.g., subject and object) and other dependents, but does not attempt to distinguish complements vs. adjuncts. By design, UD indicates in the dependency labels whether dependents are phrases or clauses, thus distin-

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

guishing *nsubj* and *csubj*, *dobj* and *ccomp*, *advmod* and *advcl*. It also recognizes a non-canonical voice subject (where the proto-agent argument is not subject, e.g., in passives). Following Lexical-Functional Grammar (Bresnan, 2001), UD includes a distinction between *ccomp* and *xcomp* for clausal complements that are standalone (have an internal subject) versus those having obligatory control (omission) of the dependent predicate's subject (have an external subject). The non-core clausal dependents are all modifiers. UD does not attempt to differentiate finite from non-finite clauses, but differentiates attachment to predicates from attachment to nominals: an adverbial clause *advcl* modifies a predicate whereas an *acl* ("clausal modifier of noun") modifies a nominal.

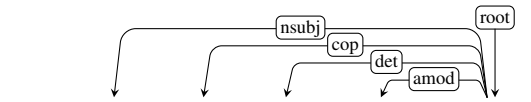
The UD scheme has a rich taxonomy of noun dependents inherited from the Stanford dependencies, as well as relations to capture phenomena appearing in non-edited or informal texts (such as *goeswith* to connect multiple tokens that correspond to a single standard word (e.g., "hand some" for "handsome") or *reparandum* to indicate disfluencies overridden in a speech repair).

UD differentiates compounding from modification or complementation, and there are three relations for compounding. We use *mwe* for fixed grammaticized expressions with function words, left-headed (e.g., *instead of*: *mwe*(instead, of), *de facto*: *mwe*(de, facto)). We use *name* for names constituted of multiple proper nouns, left-headed. That is, *name* would be used between the words of *Hillary Rodham Clinton* but not to replace the usual relations in a phrasal or clausal name like *The Lord of the Rings*. And we use *compound* to label other types of multiword lexemes, with headedness according to the language and/or compound type. Thus, *compound* is used for any kind of X^0 compounding: noun compounds (e.g., *phone book*), but also verb and adjective compounds that are more common in other languages (such as Persian or Japanese light verb constructions); for numbers (e.g., *three thousand books* gives *compound*(thousand, three)); for particles of phrasal verbs (e.g., *put up*: *compound*(put, up)).

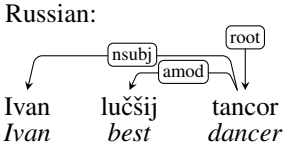
Relations between Content Words

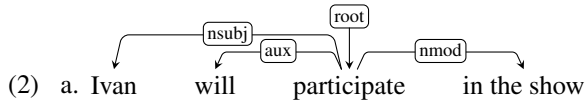
Each word depends either on another word in the sentence or on a notional "root" of the sentence, following three principles: content words are related by dependency relations; function words attach to the content word they further specify; and punctuation attaches to the head of the phrase or clause 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.

These principles lead to the following treatment of copula and auxiliaries: they are not the head of a clause, but depend on a lexical predicate, as in (1a) and (2a). Such treatment maximizes the parallelism between dependency trees in different languages: compare (1a) and (1b) where Russian does not have an overt copula. Similarly, compare (2a) and (2b) where the future tense in French can be marked morphologically.

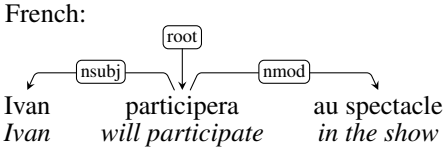


b. Russian:

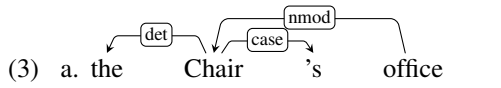


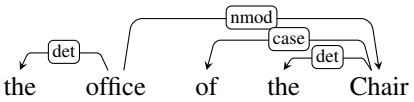


b. French:

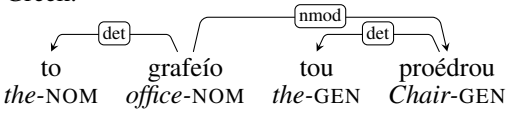


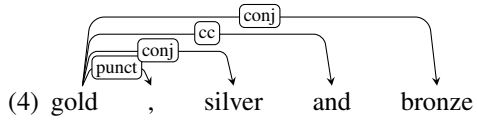
To have relations between content words, any case-marking element (including prepositions, postpositions, and clitic case markers) is treated as a dependent of the noun it attaches to or introduces. As can be seen in (3a) and (3b), *nmod* labels the relation between the two content words *office* and *Chair*, whereas the preposition or the possessive marker is a *case* depending on its complement. In general, *nmod* expresses some form of oblique or adjunct relation which can be further specified by the *case* or be morphologically marked as in (3c). Coordination follows a similar treatment: the leftmost conjunct is the head, and other conjuncts as well as the coordinating conjunction depend on it, as in (4).



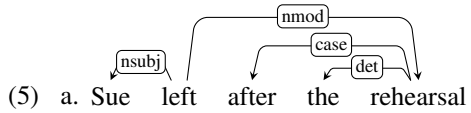


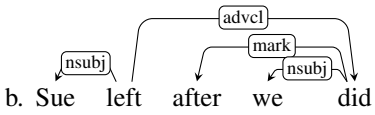
c. Greek:





These principles provide parallelism between different constructions across and within languages, as emphasized in (3) where the different constructions of possessive (possessive clitic, preposition or morphologically marked) are all parallel. For instance in English, we also obtain parallel representations between prepositional phrases and subordinate clauses, which are in practice often introduced by a preposition, as in (5).

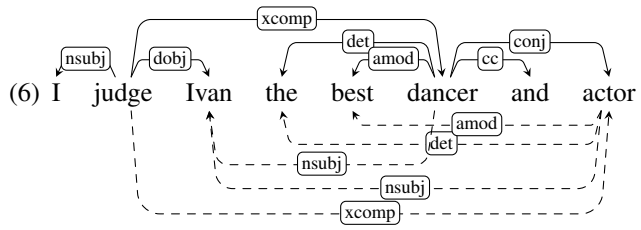




The choice to make content words the backbone of the syntactic representations may seem to be at odds with the strong tendency in modern syntactic theory to give priority to functional heads, a tendency that is found in both constituency-based and dependency-based approaches to syntax (Brugé et al., 2012; Osborne and Maxwell, 2015). We believe, however, that this conflict is more apparent than real. The UD view is that we need to recognize both lexical and functional heads, but in order to maximize parallelism across languages, only lexical heads are inferable from the topology of our tree structures. Functional heads are instead represented as specifying features of content words, using dedicated relation labels, features which can alternatively be specified through morphological processes. In the dependency grammar tradition, this is very close to the view of Tesnière (1959), according to whom dependencies hold between nuclei that always contain a content word, and where function words combine with content words to form dissociated nuclei. Moreover, it seems highly compatible with the view dominant in typologically grounded syntactic theories such as that of Dixon (2009).

Enhanced Representation

The basic dependency structure is assumed to form a (possibly non-projective) tree but UD also allows additional dependencies in an *enhanced* dependency representation.[2] The idea behind the enhanced dependency representation is to explicitly mark external subjects and the external role in relative clauses as well as to propagate relations over conjunctions, as shown in (6) where additional dependencies are indicated with dashed arrows below the sentence.



Language-specific Relations

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.

[2]Complete guidelines for the enhanced representations have not been worked out yet, and only one treebank (Finnish) uses them so far, but see Schuster and Manning (2016) for a concrete proposal for English.

Language	Sentence	Token	Word	Lemma	PoS	Feat	Dep	LDep
Ancient Greek	16221	244993	244993	15721	13	33	26	0
Ancient Greek PROIEL	16633	206966	206966	9256	13	41	31	0
Arabic	7664	282384	282384	–	16	36	30	1
Basque	8993	121443	121443	11085	16	69	30	0
Bulgarian	11138	156319	156319	14900	16	44	31	0
Croatian	3957	87765	87765	8884	14	38	39	0
Czech	87913	1503738	1506490	59008	17	82	35	5
Danish	5512	100733	100733	13355	17	46	36	5
Dutch	13735	200654	200654	21505	16	59	29	2
English	16622	254830	254830	17784	17	34	40	7
Estonian	1315	9491	9491	3634	15	59	23	3
Finnish	13581	181022	181022	24177	15	84	33	11
Finnish FTB	18792	159531	159829	21573	14	64	23	2
French	16446	389764	401491	–	17	–	35	2
German	15894	293088	298242	–	15	–	32	1
Gothic	5450	56128	56128	3353	13	36	30	0
Greek	2411	59156	59156	6201	11	30	27	1
Hebrew	6216	115535	158855	–	16	39	29	14
Hindi	16647	351704	351704	15586	16	43	26	1
Hungarian	1299	26538	26538	6477	16	70	32	22
Indonesian	5593	121923	121923	–	16	–	30	0
Irish	1020	23686	23686	3964	16	63	28	10
Italian	12677	252967	271180	18576	17	36	35	4
Japanese KTC	9995	267631	267631	5241	16	–	30	0
Latin	3269	47303	47303	7457	12	35	26	0
Latin ITT	15295	259684	259684	3374	14	50	29	2
Latin PROIEL	14982	165201	165201	7059	13	40	32	0
Norwegian	20045	311277	311277	23651	17	31	34	2
Old Church Slavonic	6346	57507	57507	2964	13	41	29	0
Persian	5997	151624	152871	–	15	30	31	7
Polish	8227	83571	83571	12904	13	46	27	0
Portuguese	9359	212545	212545	19499	17	46	29	2
Romanian	633	12094	12094	3917	17	53	38	11
Slovenian	7996	140418	140418	16946	16	60	31	1
Spanish	16013	423346	431587	35923	16	45	31	1
Swedish	6026	96819	96819	10260	15	27	35	4
Tamil	600	9581	9581	2023	14	41	24	2
TOTAL	430512	7438959	7529911					

Table 4: Statistics on treebanks released in UD v1.2. Sentence: number of sentences. Token: number of unsegmented tokens. Word: number of segmented tokens (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 not all treebanks have lemmas and features.)

3.4. 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 10 columns on a word/token line are used to specify a unique id (integer for words, ranges for multiword tokens), word form, lemma, universal part-of-speech tag, optional language-specific part-of-speech tag, morphological features, head, dependency relation, additional dependencies in the enhanced representation and miscellaneous information. The format is illustrated in Figure 3, with the French sentence from Figure 2.

To support work on treebanks in this format, we have introduced Python and JavaScript libraries for reading and validating CoNLL-U.[3] The UD documentation efforts are supported by the Annodoc system[4] (Pyysalo and Ginter, 2014), with annotation visualizations generated using brat

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

[4]http://spyysalo.github.io/annodoc/

```
ID  FORM      LEMMA     UPOSTAG XPOSTAG 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=Def|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.

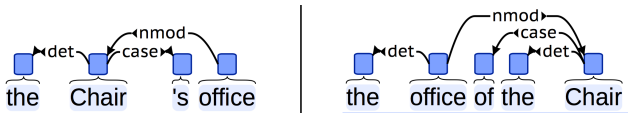


Figure 4: Examples of annotation visualization from UD documentation.

(Stenetorp et al., 2012).[5] Figure 4 shows an example of the visualization. The treebanks can also be queried online using the SETS[6] and PML TreeQuery[7] tools (Luotolahti et al., 2015; Štěpánek and Pajas, 2010). These tools allow querying any of the UD treebanks, freely combining restrictions on the existence or absence of wordforms, lemmas, POS tags, morphological features, dependency labels, and subtrees. The results are shown in a graphical form in the browser, with the relevant tokens highlighted, or as a summary table with frequency counts.

4. Existing Treebanks

The release of UD treebanks in November 2015 (v1.2) comprises 37 treebanks representing 33 languages, listed in Table 4. All treebanks contain annotation of parts-of-speech and dependency relations. Most treebanks in addition provide lemmas and morphological features. There is variation in the treebank sizes. The treebanks are listed with descriptive statistics in Table 4, which gives the number of sentences (ranging from 600 sentences to almost 90,000), unsegmented tokens (ranging from about 9,000 tokens to well over 1.5 million tokens), segmented tokens (syntactic words), unique lemmas, unique part-of-speech tags, unique morphological Feature=Value pairs, unique dependency relations, and unique language-specific dependency relations. Zeros in the table indicate annotation layers that are still missing rather than language-specific properties. Similarly, some treebanks have not yet produced the syntactic word segmentation, resulting in identical numbers of unsegmented and segmented tokens.

Figure 5 is a screenshot from the web documentation of the UD treebanks in March 2016 (treebanks included in v1.2, as indicated by the check mark in the 7th column, as well as in progress and scheduled to be released in the future). Table 5 gives the mapping for genre icons used in Figure 5. Most treebanks are constituted of different genres. While newswire is quite present, there are other genres well represented in several languages such as web data (reviews, blogs), fiction and legal documents. As indicated in Figure 5, the extent to which the data has been manually annotated or automatically converted from existing treebanks varies, and there is a continuing effort to further improve the consistency of the annotation across languages.

Icon in Figure 5	Genre
	bible
	blog
	fiction
	grammar examples
	legal text
	medical text
	news
	non-fiction
	reviews
	spoken
	social (other user-generated content)
	web
W	wikipedia

Table 5: Genres present in the UD treebanks.

5. Conclusion

The UD project aims at developing cross-linguistically consistent treebank annotation for many languages in order to support multilingual parsing research, as well as practical development of multilingual NLP systems and comparative linguistic studies of syntax. To date, we have produced a first version of the universal guidelines and released 37 treebanks where the guidelines have been applied to 33 different languages. According to Wikipedia,[8] these languages cover almost 35% of native speakers in the world (adding Chinese would bring us up to almost 60%). Although there is still a strong bias towards contemporary Indo-European

[5] http://brat.nlplab.org/
[6] http://bionlp-www.utu.fi/dep_search
[7] http://lindat.mff.cuni.cz/services/pmltq/
[8] https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers

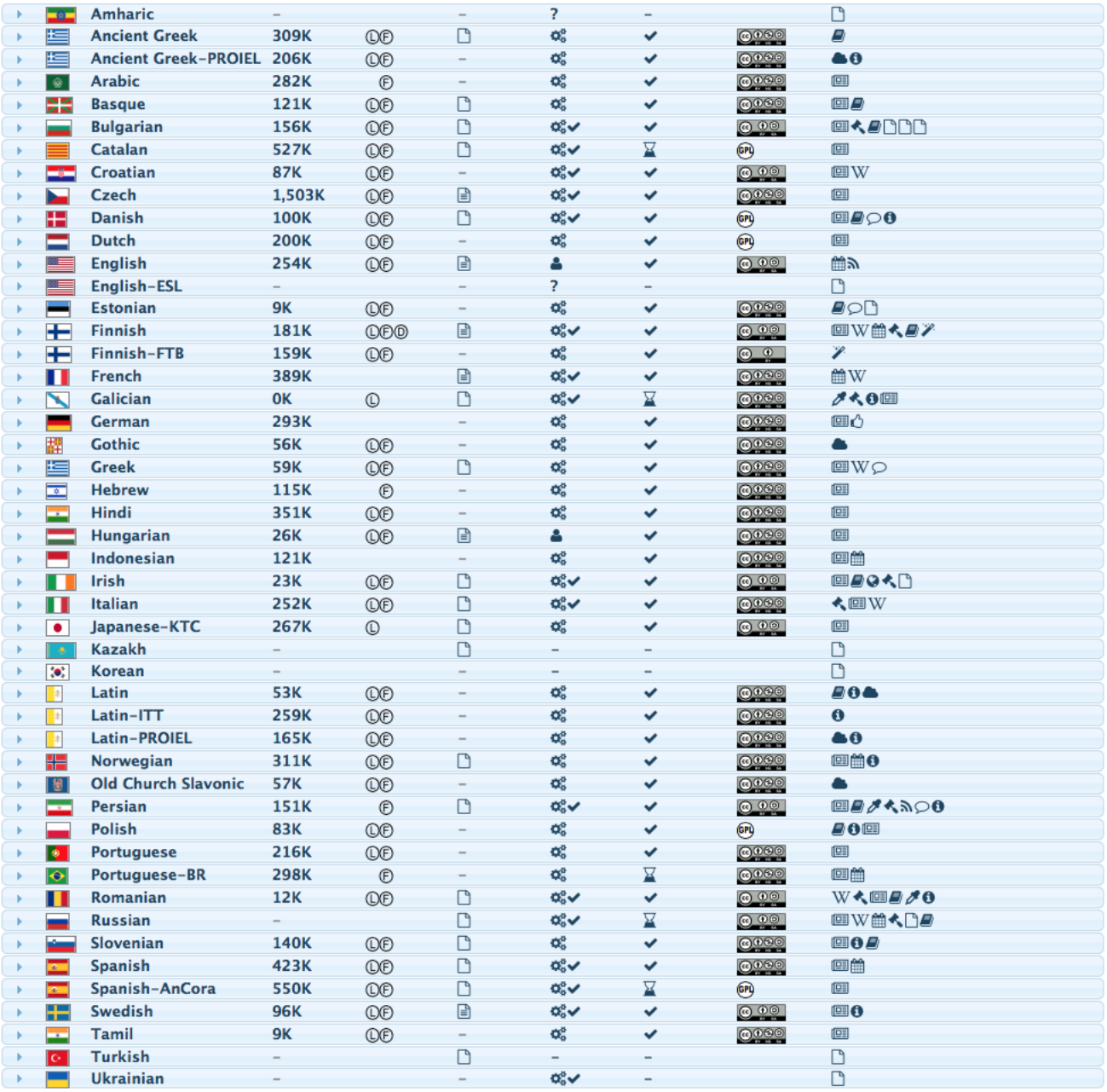


Figure 5: UD treebanks at a glance. Columns show the number of unsegmented tokens in each treebank, whether the treebank contains features (F), lemmas (L) and secondary dependencies (D), the status of the online documentation (partial 🗋 or complete 🗎), the type of conversion to the UD scheme (automatic ⚙, automatic with some manual corrections ⚙✔, or fully checked manually 👤), the release status, the treebank license type, and the data genres.

languages in the sample, we are starting to see the emergence of treebanks for other language families as well as treebanks for classical languages.

We plan to continue with treebank releases twice a year to keep up the momentum of the project. In the near future, our main priority is to improve the consistency and completeness of annotations for all languages, but we are also eager to expand the sample of languages and welcome all new contributors to the project. As a medium-term goal we envisage an improved version of the universal guidelines, based on an analysis of issues that have arisen in the work on improving consistency across languages. Ideally, the next version of the guidelines should also cover the enhanced dependencies. In parallel to the development of guidelines and annotated corpora, finally, we hope to be able to release tools for tokenization, morphological analysis and syntactic parsing for all languages, as well as large-scale parsebanks (automatically parsed corpora).

Acknowledgements

We thank all contributors working on annotated corpora under the UD guidelines, to whom we owe a substantial part of the success and momentum achieved within the UD project so far: Željko Agić, Riyaz Ahmad, Maria Jesus Aranzabe, Masayuki Asahara, Aitziber Atutxa, Miguel Ballesteros, Cristina Bosco, Giuseppe G. A. Celano, Jinho

Choi, Çağrı Çöltekin, Kaja Dobrovoljc, Timothy Dozat, Binyam Ephrem, Tomaž Erjavec, Richárd Farkas, Jennifer Foster, Iakes Goenaga, Koldo Gojenola, Bruno Guillaume, Nizar Habash, Dag Haug, Hiroshi Kanayama, Jenna Kanerva, Simon Krek, Juha Kuokkala, Veronika Laippala, Alessandro Lenci, Krister Lindén, Nikola Ljubešić, Olga Lyashevskaya, Teresa Lynn, Aibek Makazhanov, Catalina Mărănduc, Héctor Martínez Alonso, Anna Missilä, Simonetta Montemagni, Verginica Mititelu, Yusuke Miyao, Shinsuke Mori, Hanna Nurmi, Petya Osenova, Lilja Øvrelid, Petr Pajas, Elena Pascual, Marco Passarotti, Jussi Piitulainen, Barbara Plank, Prokopis Prokopidis, Loganathan Ramasamy, Sebastian Schuster, Wolfgang Seeker, Mojgan Seraji, Maria Simi, Kiril Simov, Aaron Smith, Jan Štěpánek, Alane Suhr, Takaaki Tanaka, Anders Trærup Johannsen, Francis Tyers, Sumire Uematsu, Veronika Vincze, Rob Voigt, and Jonathan Washington. The work has been partially funded by the Czech Science Foundation grant GA15-10472S, Czech MEYS grant LM2015071, and SWE-CLARIN.

References

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

Bresnan, J. (2001). *Lexical-Functional Syntax*. Blackwell, Oxford.

Laura Brugé, et al., editors. (2012). *Functional Heads. The Cartography of Syntactic Structures, Volume 7*. Oxford University Press.

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

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

Das, D. and Petrov, S. (2011). Unsupervised part-of-speech tagging with bilingual graph-based projections. In *ACL*, pages 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*, pages 4585–4592.

Dixon, R. M. (2009). *Basic Linguistic Theory. Volume 1: Methodology*. Oxford University Press.

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*, pages 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*, pages 51–55.

McDonald, R. and Nivre, J. (2007). Characterizing the errors of data-driven dependency parsing models. In *Proceedings of EMNLP-CoNLL*, pages 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*, pages 97–102.

Osborne, T. and Maxwell, D. (2015). A historical overview of the status of function words in dependency grammar. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 241–250.

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 Žabokrtský, Z. (2014). HamleDT 2.0: Thirty dependency treebanks stanfordized. In *LREC*, pages 2334–2341.

Schuster, S. and Manning, C. D. (2016). Enhanced english universal dependencies: An improved representation for natural language understanding tasks. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*.

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*, pages 102–107.

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

Tesnière, L. (1959). *Éléments de syntaxe structurale*. Editions Klincksieck.

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*, pages 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*, pages 2735–2741.

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