Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4034–4043
Marseille, 11–16 May 2020

Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection

Joakim Nivre[*] **Marie-Catherine de Marneffe**[◦] **Filip Ginter**[•] **Jan Hajič**[†]
Christopher D. Manning[‡] **Sampo Pyysalo**[•] **Sebastian Schuster**[‡]
Francis Tyers[◇] **Daniel Zeman**[†]

[*]Uppsala University [◦]The Ohio State University [•]University of Turku
[†]Charles University in Prague [‡]Stanford University [◇]Indiana University

[*]joakim.nivre@lingfil.uu.se [◦]demarneffe.1@osu.edu [•]{figint,sampo.pyysalo}@utu.fi
[†]{hajic,zeman}@ufal.mff.cuni.cz [‡]{manning,sebschu}@stanford.edu [◇]ftyers@iu.edu

Abstract

Universal Dependencies is an open community effort to create cross-linguistically consistent treebank annotation for many languages within a dependency-based lexicalist framework. The annotation consists in a linguistically motivated word segmentation; a morphological layer comprising lemmas, universal part-of-speech tags, and standardized morphological features; and a syntactic layer focusing on syntactic relations between predicates, arguments and modifiers. In this paper, we describe version 2 of the guidelines (UD v2), discuss the major changes from UD v1 to UD v2, and give an overview of the currently available treebanks for 90 languages.

Keywords: treebanks, annotation, multilingual, universal dependencies.

1. Introduction

Universal Dependencies (UD) is a project that is developing cross-linguistically consistent treebank annotation for many languages, with the goal of facilitating multilingual parser development and research on parsing and cross-lingual learning. The annotation scheme is based on an evolution of (universal) Stanford dependencies (de Marneffe et al., 2006; de Marneffe and Manning, 2008; de Marneffe et al., 2014), Google universal part-of-speech tags (Petrov et al., 2012), and the Interset interlingua for morphosyntactic tagsets (Zeman, 2008). 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.

The project started in 2014 and has developed into an open community effort with a very rapid growth, both in terms of the number of researchers contributing to the project, which now exceeds 300, and in terms of the number of languages represented by treebanks, which is approaching 100. An early snapshot of this development can be found in Nivre et al. (2016), which describes version 1 of the UD guidelines (UD v1) and the treebank resources available in UD v1.2. Since then, there has been one major change of the guidelines, from UD v1 to UD v2, and the number of treebanks has more than quadrupled. Figure 1 shows the growth in number of languages, treebanks and annotated words from UD v1.0 to UD v2.5. During the same period, the number of downloads or accesses at the official repository at `https://lindat.cz` has grown to 46439.[1] The UD resources have also made a significant impact on NLP research, most notably for multilingual dependency parsing through two editions of CoNLL shared tasks (Zeman et al., 2017; Zeman et al., 2018), which have created a new generation of parsers that handle a large number of languages and that parse from raw text rather than relying on pre-tokenized input. Figure 2 visualizes the increase in available data resources and parsing scores for all languages involved in both tasks.

This paper provides an up-to-date description of the project, focusing on the annotation guidelines, especially on the major changes from UD v1 to v2, and on the existing treebank resources. For more information on the project motivation and history, we refer to Nivre et al. (2016). For more information about UD treebanks and applications of these resources, we refer to the proceedings of the UD workshops held annually since 2017 (de Marneffe et al., 2017; de Marneffe et al., 2018; Rademaker and Tyers, 2019).

2. Annotation Scheme

In this section, we give a brief introduction to the UD annotation scheme. For more details, we refer to the documentation on the UD website.[2]

[1]November 25, 2019.

[2]https://universaldependencies.org/guidelines.html

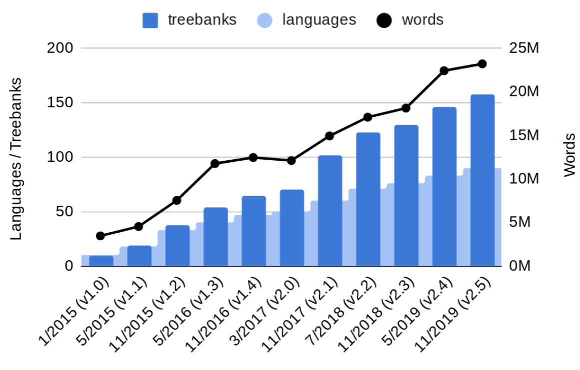


Figure 1: Number of languages, treebanks and words in UD from v1.0 to v2.5.

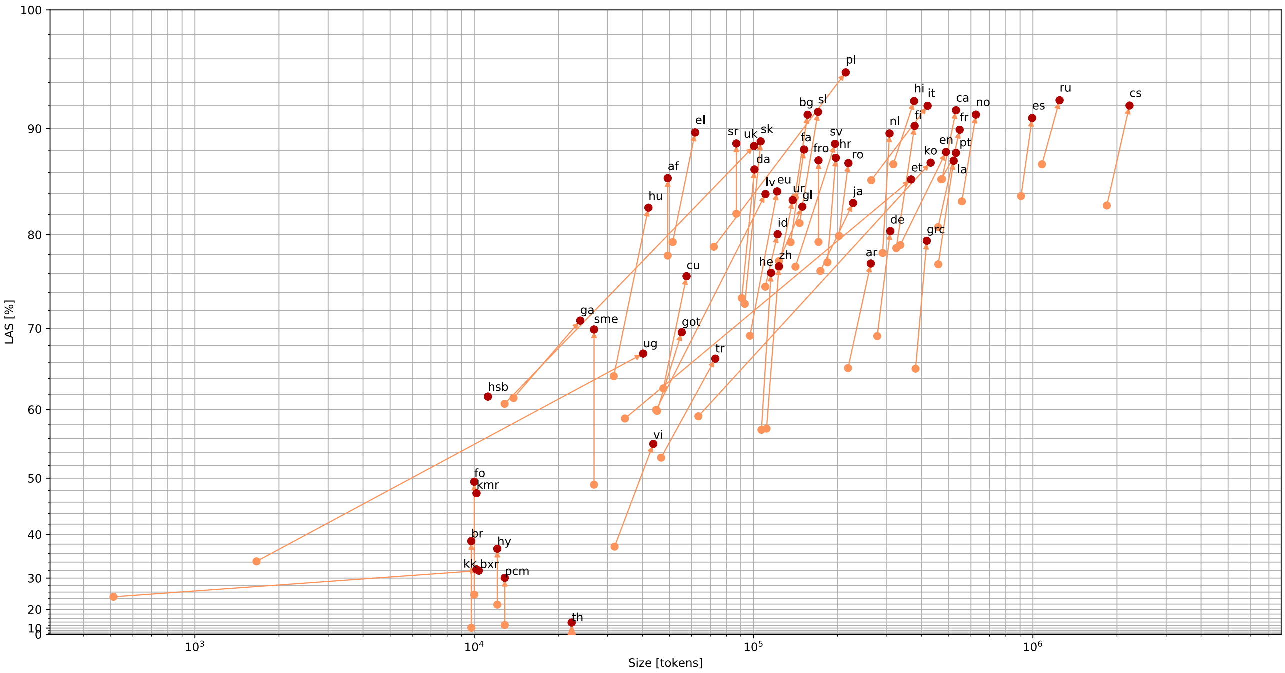


Figure 2: Increase in available data (x-axis) and labeled attachment score (y-axis) from the baseline of the CoNLL 2017 shared task (orange) to the best result of the CoNLL 2018 shared task (red); pairs labeled by ISO language codes.

2.1. Tokenization and Word Segmentation

UD is based on a lexicalist view of syntax, which means that dependency relations hold between words, and that morphological features are encoded as properties of words with no attempt at segmenting words into morphemes. However, it is important to note that the basic units of annotation are syntactic words (not phonological or orthographic words), which means that it is often necessary to split off clitics, as in Spanish *dámelo* = *da me lo*, and undo contractions, as in French *au* = *à le*. We refer to such cases as *multiword tokens* because a single orthographic token corresponds to multiple (syntactic) words. In exceptional cases, it may be necessary to go in the other direction, and combine several orthographic tokens into a single syntactic word (see Section 3.1.).

2.2. Morphological Annotation

The morphological specification of a (syntactic) word in the UD scheme consists of three levels of representation:

1. A lemma representing the base form of the word.
2. A part-of-speech tag representing the grammatical category of the word.
3. A set of features representing lexical and grammatical properties associated with the particular word form.

The lemma is the canonical form of the word, which is the form typically found in dictionaries. In agglutinative languages, this is typically the form with no inflectional affixes; in fusional languages, the lemma is usually the result of a language-particular convention. The list of universal part-of-speech tags is a fixed list containing 17 tags, shown in Table 1. Languages are not required to use all tags, but the list cannot be extended to cover language-specific categories. Instead, more fine-grained classification of words can be achieved via the use of features, which specify additional information about morphosyntactic properties. We provide an inventory of features that are attested in multiple languages and need to be encoded in a uniform way, listed in Table 1. Users can extend this set of universal features and add language-specific features when necessary.

2.3. Syntactic Annotation

Syntactic annotation in the UD scheme consists of typed dependency relations between words. The *basic* syntactic representation forms a tree rooted in one word, normally the main clause predicate, on which all other words of the sentence are dependent. In addition to the basic representation, which is obligatory for all UD treebanks, it is possible to give an *enhanced* dependency representation, which adds (and in a few cases changes) relations in order to give a more complete basis for semantic interpretation. We will focus here on the basic representation and return to the enhanced representation when discussing changes in UD v2.

The syntactic analysis in UD gives priority to predicate-argument and modifier relations that hold directly between content words, as opposed to being mediated by function words. The rationale is that this makes more transparent what grammatical relations are shared across languages, even when the languages differ in the way that they use word order, function words or morphological inflection to encode these relations. This is illustrated in Figure 3, which shows three parallel sentences in Czech, English and Swedish. In all three cases, there is a passive predicate with a subject and an oblique modifier (the relations marked in solid blue), but the languages differ in how they encode certain grammatical categories (marked in dashed red): definiteness is indicated by a separate function word (the article *the*) in English, by a morphological inflection in Swedish and not at all in Czech; passive is expressed by a periphrastic construction involving an auxiliary and a participle in English, by a morphological inflection in

PoS Tags	Features: Inflectional	Features: Lexical	Syntactic Relations: Clausal: Core	Syntactic Relations: Clausal: Non-Core	Syntactic Relations: Nominal
ADJ	Animacy	Abbr	nsubj	advcl	acl
ADP	Aspect	Foreign	csubj	advmod	amod
ADV	Case	NumType	ccomp	aux	appos
AUX	Clusivity	Poss	iobj	cop	case
CCONJ	Definite	PronType	obj	discourse	clf
DET	Degree	Reflex	xcomp	dislocated	det
INTJ	Evident	Typo		expl	nmod
NOUN	Gender			mark	nummod
NUM	Mood			obl	
PART	NounClass			vocative	
PRON	Number		**Linking**	**MWE**	**Special**
PROPN	Person		cc	compound	dep
PUNCT	Polarity		conj	fixed	goeswith
SCONJ	Polite		list	flat	orphan
SYM	Tense		parataxis		punct
VERB	VerbForm				reparandum
X	Voice				root

Table 1: Universal part-of-speech tags (left), morphological features (middle) and syntactic relations (right).

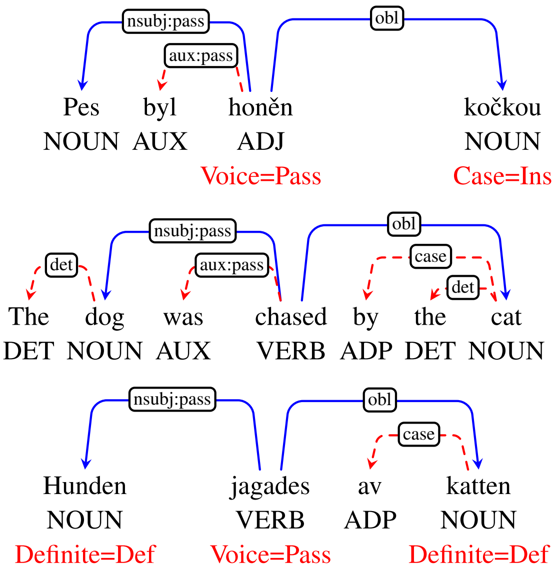


Figure 3: Parallel sentences in Czech, English and Swedish. Common syntactic relations in blue, differences in morphosyntactic encoding highlighted in red. The Czech passive participle has both adjectival and verbal features; it is tagged ADJ due to its similarity to adjectives.

Swedish, and by a combination of these strategies in Czech (because the participle is unique to the passive construction); and the oblique modifier is introduced by a preposition in English and Swedish but marked by instrumental case in Czech.

UD provides a taxonomy of 37 universal relation types to classify syntactic relations, as shown in Table 1. The taxonomy distinguishes between relations that occur at the clause level (linked to a predicate) and those that occur in noun phrases (linked to a nominal head). At the clause level, a distinction is made between core arguments (essentially subjects and objects) and all other dependents (Thompson, 1997; Andrews, 2007). It is important to note that not all relations in the taxonomy are syntactic dependency relations in the narrow sense. First, there are special relations for function words like determiners, classifiers, adpositions, auxiliaries, copulas and subordinators, whose dependency status is controversial. In addition, there are a number of special relations for linking relations (including coordination), certain types of multiword expressions, and special phenomena like ellipsis, disfluencies, punctuation and typographical errors. Many of these relations cannot plausibly be interpreted as syntactic head-dependent relations, and should rather be thought of as technical devices for encoding flat structures in the form of a tree.

The inventory of universal relation types is fixed, but subtypes can be added in individual languages to capture additional distinctions that are useful. This is illustrated in Figure 3, where the relations NSUBJ[3] (nominal subject) and AUX (auxiliary) are subtyped to NSUBJ:PASS and AUX:PASS to capture properties of passive constructions.

3. Changes from UD v1 to UD v2

We now discuss the most important changes from UD v1 to UD v2. More information about these changes can be found on the UD website.[4]

3.1. Tokenization and Word Segmentation

In UD v1, word-internal spaces were not allowed. This restriction has now been lifted in two circumstances:

1. For languages with writing systems that use spaces to mark units smaller than words (typically syllables),

[3]Syntactic relations in UD are normally written in all lowercase, as shown in Table 1, but in this paper we use small capitals in running text for clarity.

[4]https://universaldependencies.org/v2/summary.html

Feature		Value(s)	
Old	**New**	**Old**	**New**
	Clusivity		Ex, In
	Evident		Nfh
	NounClass		Bantu1–23, Wol1–12, …
	Polite		Infm, Form, Elev, Humb
	Abbr		Yes
	Foreign		Yes
	Typo		Yes
Animacy			Hum
Case			Equ, Cmp, Cns, Per
Degree			Equ
Definite			Spec
Number			Count, Tri, Pauc, Grpa, Grpl, Inv
VerbForm			Gdv, Vnoun
Mood			Prp, Adm
Aspect			Iter, Hab
Voice			Mid, Antip, Dir, Inv
PronType			Emp, Exc
Person			0, 4
Negative	Polarity		
Aspect		Pro	Prosp
VerbForm		Trans	Conv
Definite		Red	Cons

Table 2: Revisions to morphological features and values in UD v2: new features (group 1), new values (group 2), and renamed features and values (groups 3 and 4).

spaces are allowed in any word; the phenomenon has to be declared in the language-specific documentation.

2. For other languages, spaces are allowed only for a restricted list of exceptions like numbers (*100 000*) and abbreviations (*i. e.*); the latter have to be listed explicitly in the language-specific documentation.

The first case was deemed necessary, because in languages like Vietnamese all polysyllabic words would otherwise have to be annotated as fixed multiword expressions, which would seriously distort the syntactic representations compared to other languages. The second case is more a matter of convenience, but it seemed useful to allow *multitoken words* – a single (syntactic word) corresponding to multiple orthographic tokens – as well as multiword tokens, although this option should be used very restrictively.

3.2. Morphological Annotation

The universal part-of-speech tagset is essentially the same in UD v2 as in UD v1, but the tag for coordinating conjunctions has been renamed from CONJ to CCONJ[5] and the guidelines have been modified slightly for three tags:

1. The use of AUX is extended from auxiliary verbs in a narrow sense to also include copula verbs and nonverbal TAME particles (tense, aspect, mood, evidentiality, and, sometimes, voice or polarity particles).
2. The use of PART is limited to a small set of words that must be listed in the language-specific documentation.
3. The distinction between PRON and DET is made more flexible to accommodate cross-linguistic variation.

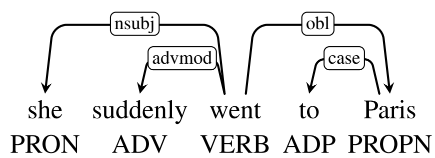
The inventory of universal morphological features has been extended with new features and new values for existing features. In addition, a few features and feature values have been renamed or removed. These changes, which are summarized in Table 2, are motivated by the addition of new languages to UD as well as an effort to harmonize UD with the UniMorph project (Sylak-Glassman et al., 2015).

3.3. Syntactic Annotation

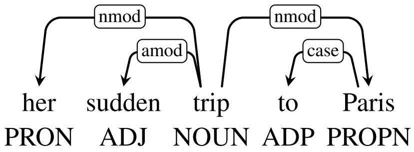
Although most syntactic relations are the same in UD v2 as in UD v1, the guidelines have often been improved by providing more explicit criteria and examples from multiple languages. Here we only list cases where relations have been removed, added or renamed, or where the use of an existing relation has changed significantly.

Clauses and Dependents of Predicates As explained earlier, UD assumes a distinction between core and non-core dependents of predicates. For nominal core arguments, UD v1 used the labels NSUBJ, DOBJ and IOBJ. These relations remain conceptually unchanged, but the second label has been changed from DOBJ to OBJ, because this seems to better convey the intended interpretation of "second core argument" or "P/O argument" (without connection to specific cases or semantic roles). In addition, the NSUBJPASS label for passive subjects is removed, and passive subjects are subsumed under the NSUBJ relation, but with a strong recommendation to use the subtype NSUBJ:PASS for languages where the distinction is relevant. Analogously, the relations CSUBJPASS (for clausal passive subject) and AUXPASS (for passive auxiliary) are now subsumed under CSUBJ and AUX (with possible subtypes CSUBJ:PASS and AUX:PASS).

The second change in this area concerns the analysis of *oblique* nominals at the clause level, that is, nominal expressions that are dependents of predicates but not core arguments, and which are typically accompanied by case marking in the form of adpositions or oblique morphological case. In UD v1, such expressions were subsumed under the NMOD relation (for nominal modifier), which also applies to nominal expressions that modify other nominals and are not dependents of predicates at the clause level. This violated a fundamental principle of UD, namely that distinct labels should be used for dependents of nominals and dependents of predicates, even if the overt form of the modifier is the same. In UD v2, the OBL relation is therefore used for oblique nominals at the clause level, while the NMOD relation is reserved for nominals modifying other nominal expressions. The distinction is illustrated in (1) and (2), which also show that the core/non-core distinction is only applied at the clause level. Hence, both the NSUBJ and the OBL relations in the clause example correspond to NMOD relations in the nominal example.

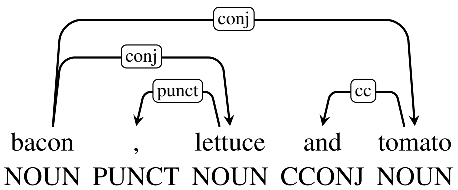
(1)


[5] The motivation is to make it parallel to SCONJ (for subordinating conjunctions), more similar to the syntactic relation CC with which it often cooccurs, and less similar to the relation CONJ with which it practically never cooccurs.

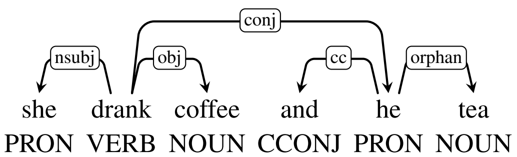
(2) 

The final modification in the annotation of clause structure is a more restricted application of the COP relation. In UD v2, the COP relation is restricted to function words (verbal or nonverbal) whose sole function is to link a nonverbal predicate to its subject and which does not add any meaning other than grammaticalized TAME categories. The range of constructions that are analyzed using the COP relation is subject to language-specific variation but can be identified using universal criteria described in the guidelines.

Coordination The question of whether and how coordination can be analyzed as a dependency structure is a vexed one (Popel et al., 2013; Gerdes and Kahane, 2015). UD treats coordination as an essentially symmetric relation, and uses the special CONJ relation to connect all non-first conjuncts to the first one. In this respect, UD v2 is exactly the same as UD v1, but UD v2 differs by attaching coordinating conjunctions (CC) and punctuation (PUNCT) inside coordinated structures to the immediately succeeding conjunct (instead of the first conjunct as in UD v1), following the approach of Ross (1967), as illustrated in (3).

(3) 

Ellipsis The analysis of elliptical constructions like gapping is completely different in UD v2 compared to UD v1. Let us first note that most cases of ellipsis are simply treated by "promoting" a dependent of the elided element to take its place in the syntactic structure. Thus, adjectival modifiers or even determiners can head nominals if the head noun is omitted. Similarly, auxiliary verbs can head clauses in constructions like VP ellipsis. However, in cases like gapping, this yields a rather unsatisfactory analysis where one core argument is typically attached to another. UD v2 therefore uses a special relation ORPHAN to indicate that this is an anomalous structure where the dependent is really a sibling of the word to which is it attached. As illustrated in (4), this gives an underspecified analysis of the predicate-argument structure, which can be fully resolved in the enhanced representation (see Section 3.4.).

(4) 

The choice of which dependent to promote is determined by an obliqueness hierarchy (where subjects precede objects) described in the guidelines. This new analysis of gapping is superior to the UD v1 analysis (which used a REMNANT relation), because it preserves the integrity of the two clauses and introduces fewer non-projective dependencies.

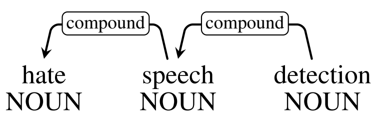
Functional Relations UD v2 also includes some changes in the annotation of functional relations, that is, relations holding between a function word or grammatical marker and its host (mostly a verb or noun). More specifically:

1. A new relation CLF is added for nominal classifiers.
2. The AUX relation is extended from auxiliary verbs in a narrow sense to also include nonverbal TAME particles in analogy with the extended use of the part-of-speech tag AUX (see Section 3.2.).
3. The AUXPASS relation is subsumed under the AUX relation (see above).
4. The COP relation is restricted to pure linking words (see above).
5. The NEG relation is removed from the set of universal relations, and polarity is instead encoded in a feature (see Section 3.2.).

3.3.1. Multiword Expressions

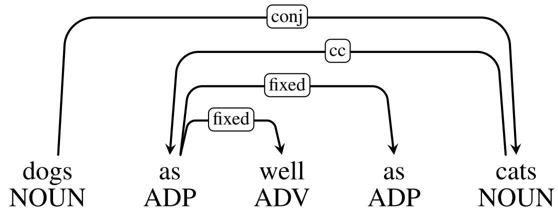
The guidelines for annotation of multiword expressions have been thoroughly revised in UD v2. Multiword expressions that are morphosyntactically regular (and only exhibit semantic non-compositionality) normally do not receive any special treatment at all. Hence, the UD guidelines in this area only apply to a few subtypes of the many phenomena that have been discussed in the literature on multiword expressions.

The first subtype is compounding. The relation COMPOUND is used for any kind of lexical compounds: noun compounds such as *phone book*, but also verb and adjective compounds, such as the serial verbs that occur in many languages, or a Japanese light verb construction such as *benkyō suru* ("to study"). The compound relation is also used for phrasal verbs, such as *put up*: COMPOUND(*put, up*). Despite operating at the lexical level, compounds are regular headed constructions, as illustrated in (5). This behavior distinguishes compounds from the other two types of multiword expressions.

(5) 

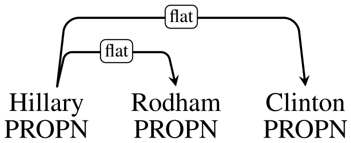
The second subtype is fixed expressions, highly grammaticalized expressions that typically behave as function words or short adverbials, for which the relation FIXED is used. The name and rough scope of usage is borrowed from the fixed expressions category of Sag et al. (2002).[6] Fixed multiword expressions are annotated with a flat structure. Since there is no clear basis for internal syntactic structure, we adopt the convention of always attaching subsequent words to the first one with the FIXED label, as shown in (6).

[6]This relation was called MWE in UD v1, but the name was found to be misleading as the relation only applies to a very small subset of multiword expressions.

(6) 

As with other clines of grammaticalization, it is not always clear where to draw the line between giving a regular syntactic analysis versus a fixed expression analysis of a conventionalized expression. In practice, the best solution is to be conservative and to prefer a regular syntactic analysis except when an expression is highly opaque and clearly does not have internal syntactic structure (except from a historical perspective).

The final subtype is headless multiword expressions analyzed with the relation FLAT. This class is less clearly recognized in most grammars of human languages, but in practice there are many linguistic constructions with a sequence of words that do not have any clear synchronic grammatical structure but are not fixed expressions. These include names, dates, and calqued expressions from other languages. We again adopt the convention that in these cases subsequent words are attached to the first word with the FLAT relation, as exemplified in (7).

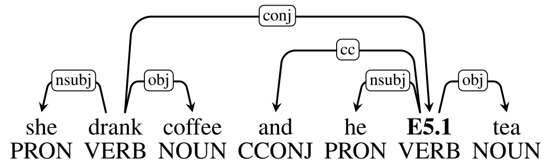
(7) 

This relation replaces two more specific relations from UD v1, NAME and FOREIGN. Subtypes like FLAT:NAME and FLAT:FOREIGN can be used in cases where a flat analysis is appropriate for complex names and foreign expressions.

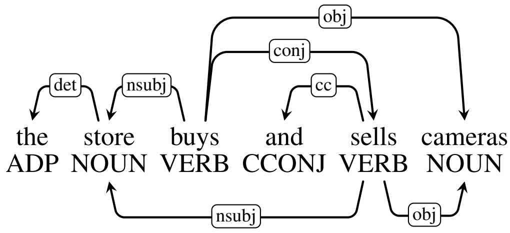
3.4. Enhanced Dependencies

UD v2 now also provides guidelines for *enhanced* dependency graphs. With a few exceptions, enhanced graphs consist of all the syntactic relations in the *basic* dependency tree and may contain additional relations and nodes that make otherwise implicit relations between tokens explicit, with the purpose of facilitating downstream natural language understanding tasks. The guidelines are based on the *CCprocessed* Stanford dependencies representation (de Marneffe et al., 2006) and a proposal for *enhanced* dependencies (Schuster and Manning, 2016), and define five types of enhancements. For more information, we refer to the documentation on the UD website.[7]

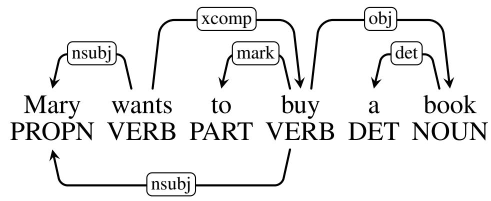
Null Nodes for Elided Predicates For sentences with elided predicates, in the basic representation, one word is promoted to be the head of the clause and all words that would have been a sibling of the promoted word if no predicate had been elided are attached with the ORPHAN relation (see Section 3.3.). The enhanced representation for sentences with gapping contains additional null nodes representing elided predicates. Arguments and modifiers of the elided predicate are attached to the null nodes, as illustrated in (8), which contains a null node (**E5.1**) and relations between the null node and the arguments in the second clause.

(8) 

Propagation of Conjuncts Conjoined predicates often share dependents (e.g., a subject) and conjoined dependents share a head. In (9), the two predicates (*buys* and *sells*) share the subject (*the store*) and object (*cameras*). The shared status of dependents and governors is made explicit in the enhanced representation through additional relations, such as the NSUBJ and OBJ relations below the sentence.[8]

(9) 

Controlled and Raised Subjects For sentences with control or raising predicates, in the basic representation, the argument that is shared between the matrix predicate and the embedded predicate is only attached to the matrix predicate. Thus, similarly as in the case of shared dependents in conjoined phrases, there is no explicit relation between the embedded predicate and its subject. In the enhanced representation, this implicit subject relation is made explicit with an additional relation, such as the NSUBJ relation[9] below the sentence in (10).

(10) 

Relative Pronouns In the enhanced representation, the coreferential status of relative pronouns is marked with the special REF relation. Further, to represent the implicit relation between the predicate of the relative clause and the antecedent of the relative pronoun, there is an additional relation between the predicate and the antecedent, such as the NSUBJ relation between *lived* and *boy* in (11).[10]

[7]https://universaldependencies.org/u/overview/enhanced-syntax.html

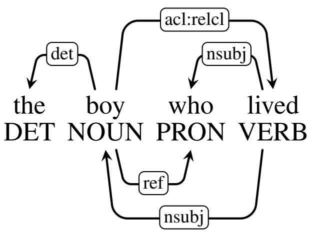
[8]The placement of arcs above and below the sentence, respectively, is only for perspicuity and does not imply any difference in status between different types of arcs.

[9]The fact that this relation is between an embedded predicate and an argument of the matrix verb can be optionally marked with the NSUBJ:XSUBJ subtype.

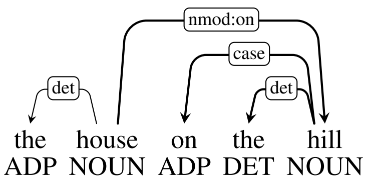
[10]The NSUBJ relation between *lived* and *who* is common to the basic and enhanced representation.

Language	#	Sents	Words	Language	#	Sents	Words	Language	#	Sents	Words
Afrikaans	1	1,934	49,276	German	4	208,440	3,753,947	Old Russian	2	17,548	168,522
Akkadian	1	101	1,852	Gothic	1	5,401	55,336	Persian	1	5,997	152,920
Amharic	1	1,074	10,010	Greek	1	2,521	63,441	Polish	3	40,398	499,392
Ancient Greek	2	30,999	416,988	Hebrew	1	6,216	161,417	Portuguese	3	22,443	570,543
Arabic	3	28,402	1,042,024	Hindi	2	17,647	375,533	Romanian	3	25,858	551,932
Armenian	1	2502	52630	Hindi English	1	1,898	26,909	Russian	4	71,183	1,262,206
Assyrian	1	57	453	Hungarian	1	1,800	42,032	Sanskrit	1	230	1,843
Bambara	1	1,026	13,823	Indonesian	2	6,593	141,823	Scottish Gaelic	1	2,193	42,848
Basque	1	8,993	121,443	Irish	1	1,763	40,572	Serbian	1	4,384	97,673
Belarusian	1	637	13,325	Italian	6	35,481	811,522	Skolt Sámi	1	36	321
Bhojpuri	1	254	4,881	Japanese	4	67,117	1,498,560	Slovak	1	10,604	106,043
Breton	1	888	10,054	Karelian	1	228	3,094	Slovenian	2	11,188	170,158
Bulgarian	1	11,138	156,149	Kazakh	1	1,078	10,536	Spanish	3	34,693	1,004,443
Buryat	1	927	10,185	Komi Permyak	1	49	399	Swedish	3	12,269	206,855
Cantonese	1	1,004	13,918	Komi Zyrian	2	327	3,463	Swedish Sign Language	1	203	1,610
Catalan	1	16,678	531,971	Korean	3	34,702	446,996	Swiss German	1	100	1,444
Chinese	5	12,449	285,127	Kurmanji	1	754	1,0260	Tagalog	1	55	292
Classical Chinese	1	15,115	74,770	Latin	3	41,695	582,336	Tamil	1	600	9,581
Coptic	1	1,575	40,034	Latvian	1	13,643	219,955	Telugu	1	1,328	6,465
Croatian	1	9,010	199,409	Lithuanian	2	3,905	75,403	Thai	1	1,000	22,322
Czech	5	127,507	2,222,163	Livvi	1	125	1,632	Turkish	3	9,437	91,626
Danish	1	5,512	100,733	Maltese	1	2,074	44,162	Ukrainian	1	7,060	122,091
Dutch	2	20,916	306,503	Marathi	1	466	3,849	Upper Sorbian	1	646	11,196
English	7	35,791	620,509	Mbyá Guaraní	2	1,144	13,089	Urdu	1	5,130	138,077
Erzya	1	1,550	15,790	Moksha	1	65	561	Uyghur	1	3,456	40,236
Estonian	2	32,634	465,015	Naija	1	948	12,863	Vietnamese	1	3,000	43,754
Faroese	1	1,208	10,002	North Sámi	1	3,122	26,845	Warlpiri	1	55	314
Finnish	3	34,859	377,619	Norwegian	3	42,869	666,984	Welsh	1	956	16,989
French	7	45,074	1,157,171	Old Church Slavonic	1	6,338	57,563	Wolof	1	2,107	44,258
Galician	2	4,993	164,385	Old French	1	17,678	170,741	Yoruba	1	100	2,664

Table 3: Languages in UD v2.5 with number of treebanks (#), sentences (Sents) and words (Words).

(11)


Case Information Finally, since many modifier relation types such as OBL or ACL are used for many different types of relations, and since adpositions or case information often disambiguate the semantic role, the enhanced representation provides augmented modifier relations that include adposition or case information in the relation name, such as the NMOD:ON relation in (12).

(12)


All enhancements are optional and users may decide to implement only a subset of these. As of UD release v2.5, only 24 treebanks include an enhanced representation, and even fewer treebanks implement all five enhancements (see also Droganova and Zeman (2019)). In many cases, the enhanced graphs can be computed automatically from a basic dependency tree (see Nivre et al. (2018) for a discussion and evaluation of a rule-based and a machine learning-based converter from basic to enhanced dependencies), and Droganova and Zeman (2019) recently used the Stanford Enhancer (Schuster and Manning, 2016) to automatically predict enhanced dependencies for all UD treebanks.

4. Available Treebanks

UD release v2.5[11] (Zeman et al., 2019) contains 157 treebanks representing 90 languages. Table 3 specifies for each language the number of treebanks available, as well as the total number of annotated sentences and words in that language. It is worth noting that the amount of data varies considerably between languages, from Skolt Sámi with 36 sentences and 321 words, to German with over 200,000 sentences and nearly 4 million words. The majority of treebanks are small but it should be kept in mind that many of these treebanks are new initiatives and can be expected to grow substantially in the future.

The languages in UD v2.5 represent 20 different language families (or equivalent), listed in Table 4. The selection is very heavily biased towards Indo-European languages (48 out of 90), and towards a few branches of this family – Germanic (10), Romance (8) and Slavic (13) – but it is worth noting that the bias is (slowly) becoming less extreme over time.[12] Another way of visualizing the gradual extension of UD to new language families and geographic areas can

[11] UD releases are numbered by letting the first digit (2) refer to the version of the guidelines and the second digit (5) to the number of releases under that version.

[12] The proportion of Indo-European languages has gone from 60% in v2.1 to 53% in v2.5.

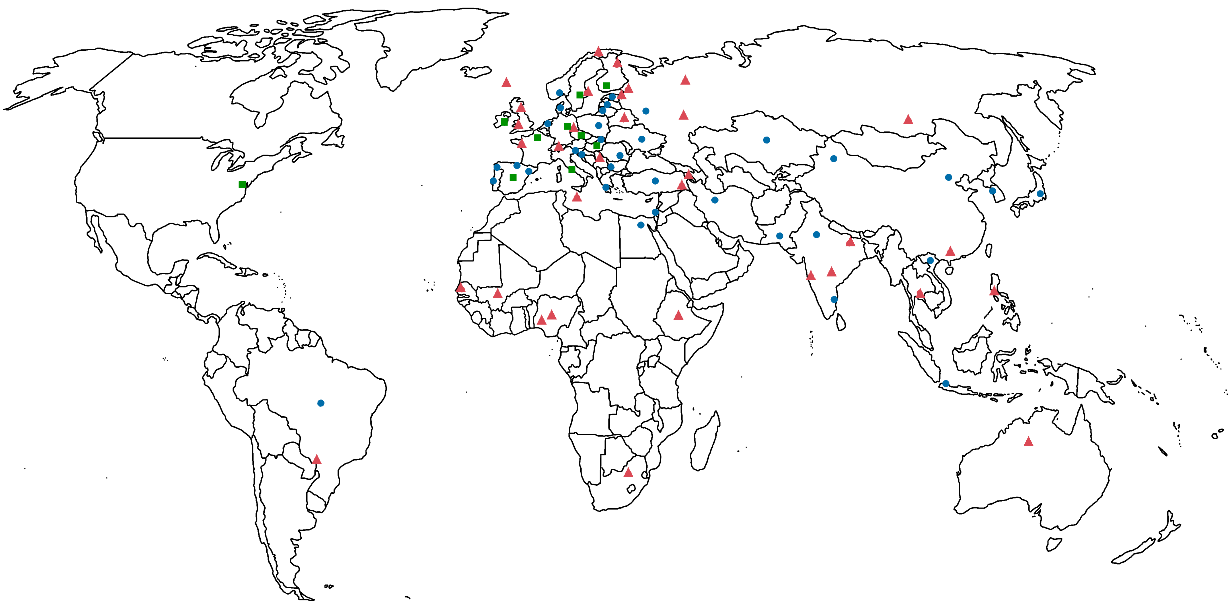


Figure 4: Map of the world with language coverage of UD. Locations are approximate. Languages released in v1.0 of the collection (2015) are in green ■, those released in v2.0 (2017) are in blue ●, and those released in v2.5 (2019) are in red ▲. Coordinates are approximate based on the capital city or centre of the country where either the largest population of speakers lives, or where the treebank was created.

be found in Figure 4, which shows the approximate geographic locations of languages added in UD v1.0 (green), UD v2.0 (blue) and UD v2.5 (red). It is clear that, whereas UD v1.0 was almost completely restricted to Europe, later versions have extended to other areas, and by v2.5 all inhabited continents are represented – although there are still large white areas on the map.

The treebanks in UD v2.5 are also heterogeneous with respect to the type of text (or spoken data) annotated. A very coarse-grained picture of this variation can be gathered from Table 5, which specifies the number of treebanks that contain some amount of data from different "genres", as reported by each treebank provider in the treebank documentation. The categories in this classification are neither mutually exclusive nor based on homogeneous criteria, but it is currently the best documentation that can be obtained.

Family	**Languages**
Afro-Asiatic	7
Austro-Asiatic	1
Austronesian	2
Basque	1
Dravidian	2
Indo-European	48
Japanese	1
Korean	1
Mande	1
Mongolic	1
Niger-Congo	2
Pama-Nyungan	1
Sino-Tibetan	3
Tai-Kadai	1
Tupian	1
Turkic	3
Uralic	11
Code-Switching	1
Creole	1
Sign Language	1

Table 4: Language families in UD v2.5.

5. Conclusion

The UD project has come a long way in only five years, and UD treebanks are now widely used in NLP as well as in linguistic research, especially with a typological orientation. Future priorities for the project include obtaining data from more languages – in order to achieve better coverage of major language families – but also obtaining more annotated data for existing languages – in order to make the data more useful for NLP as well as linguistic studies. Finally, the work on achieving cross-linguistic consistency needs to continue. Adopting a common set of categories and guidelines is a first step in this direction, but ensuring that these are applied consistently across a growing set of typologically diverse languages will continue to be a challenge for years to come. Fortunately, efforts in this direction are constantly being pursued in the active UD user community.

6. Acknowledgments

We want to thank our colleagues in the UD core guidelines group Yoav Goldberg, Ryan McDonald, Slav Petrov and Reut Tsarfaty for fruitful discussions and comments on a draft version of this paper, as well as all the 345 UD treebank contributors, listed in Zeman et al. (2019), without whom UD literally would not exist.

7. Bibliographical References

Andrews, A. D. (2007). The major functions of the noun phrase. In Timothy Shopen, editor, *Language Typology*

Genre	#	Genre	#
Academic	4	News	98
Bible	10	Non-fiction	57
Blog	17	Poetry	4
Email	2	Reviews	7
Fiction	42	Social	9
Grammar examples	13	Spoken	18
Learner essays	2	Web	9
Legal	22	Wiki	46
Medical	6		

Table 5: Genres in UD v2.5 with number of treebanks.

and Syntactic Description. Second Edition. Volume I: Clause Structure, pages 132–223. Cambridge University Press.

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

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

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)*, pages 4585–4592.

Marie-Catherine de Marneffe, et al., editors. (2017). *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*.

Marie-Catherine de Marneffe, et al., editors. (2018). *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*.

Droganova, K. and Zeman, D. (2019). Towards deep Universal Dependencies. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 144–152.

Gerdes, K. and Kahane, S. (2015). Non-constituent coordination and other coordinative constructions as dependency graphs. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 101–110.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*.

Nivre, J., Marongiu, P., Ginter, F., Kanerva, J., Montemagni, S., Schuster, S., and Simi, M. (2018). Enhancing Universal Dependency treebanks: A case study. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*.

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

Popel, M., Mareček, D., Štěpánek, J., Zeman, D., and Žabokrtský, Z. (2013). Coordination structures in dependency treebanks. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 517–527.

Alexandre Rademaker et al., editors. (2019). *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*.

Ross, J. R. (1967). *Constraints on Variables in Syntax*. Ph.D. thesis, Massachusetts Institute of Technology.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A. A., and Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, pages 1–15.

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 (LREC)*.

Sylak-Glassman, J., Kirov, C., Yarowsky, D., and Que, R. (2015). A language-independent feature schema for inflectional morphology. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 674–680.

Thompson, S. A. (1997). Discourse motivations for the core-oblique distinction as a language universal. In Akio Kamio, editor, *Directions in Functional Linguistics*, pages 59–82. John Benjamins.

Zeman, D., Popel, M., Straka, M., Hajic, J., Nivre, J., Ginter, F., Luotolahti, J., Pyysalo, S., Petrov, S., Potthast, M., Tyers, F., Badmaeva, E., Gokirmak, M., Nedoluzhko, A., Cinkova, S., Hajic jr., J., Hlavacova, J., Kettnerová, V., Uresova, Z., Kanerva, J., Ojala, S., Missilä, A., Manning, C. D., Schuster, S., Reddy, S., Taji, D., Habash, N., Leung, H., de Marneffe, M.-C., Sanguinetti, M., Simi, M., Kanayama, H., dePaiva, V., Droganova, K., Martínez Alonso, H., Çöltekin, c., Sulubacak, U., Uszkoreit, H., Macketanz, V., Burchardt, A., Harris, K., Marheinecke, K., Rehm, G., Kayadelen, T., Attia, M., Elkahky, A., Yu, Z., Pitler, E., Lertpradit, S., Mandl, M., Kirchner, J., Alcalde, H. F., Strnadová, J., Banerjee, E., Manurung, R., Stella, A., Shimada, A., Kwak, S., Mendonca, G., Lando, T., Nitisaroj, R., and Li, J. (2017). CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada.

Zeman, D., Hajič, J., Popel, M., Potthast, M., Straka, M., Ginter, F., Nivre, J., and Petrov, S. (2018). CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Brussels, Belgium.

Zeman, D., Nivre, J., Abrams, M., Aepli, N., Agić, Ž., Ahrenberg, L., Aleksandravičiūtė, G., Antonsen, L., Aplonova, K., Aranzabe, M. J., Arutie, G., Asahara, M., Ateyah, L., Attia, M., Atutxa, A., Augustinus, L., Badmaeva, E., Ballesteros, M., Banerjee, E., Bank, S., Barbu Mititelu, V., Basmov, V., Batchelor, C., Bauer, J., Bellato, S., Bengoetxea, K., Berzak, Y., Bhat, I. A., Bhat, R. A., Biagetti, E., Bick, E., Bielinskienė, A., Blokland, R., Bobicev, V., Boizou, L., Borges Völker, E., Börstell, C., Bosco, C., Bouma, G., Bowman, S., Boyd, A., Brokaitė, K., Burchardt, A., Candito, M., Caron, B., Caron, G., Cavalcanti, T., Cebiroğlu Eryiğit, G., Cecchini, F. M., Celano, G. G. A., Čéplö, S., Cetin, S., Chalub, F., Choi, J., Cho, Y., Chun, J., Cignarella, A. T., Cinková, S., Collomb, A., Çöltekin, Ç., Connor, M., Courtin, M., Davidson, E., de Marneffe, M.-C., de Paiva, V., de Souza, E., Diaz de Ilarraza, A., Dickerson, C., Dione, B., Dirix, P., Dobrovoljc, K., Dozat, T., Droganova, K., Dwivedi, P., Eckhoff, H., Eli, M., Elkahky, A., Ephrem, B., Erina, O., Erjavec, T., Etienne, A., Evelyn, W., Farkas, R., Fernandez Alcalde, H., Foster, J., Freitas, C., Fujita, K., Gajdošová, K., Galbraith, D., Garcia, M., Gärdenfors, M., Garza, S., Gerdes, K., Ginter, F., Goenaga, I., Gojenola, K., Gökırmak, M., Goldberg, Y., Gómez Guinovart, X., González Saavedra, B., Griciūtė, B., Grioni, M., Grūzītis, N., Guillaume, B., Guillot-Barbance, C., Habash, N., Hajič, J., Hajič jr., J., Hämäläinen, M., Hà Mỹ, L., Han, N.-R., Harris, K., Haug, D., Heinecke, J., Hennig, F., Hladká, B., Hlaváčová, J., Hociung, F., Hohle, P., Hwang, J., Ikeda, T., Ion, R., Irimia, E., Ishola, Ọ., Jelínek, T., Johannsen, A., Jørgensen, F., Juutinen, M., Kaşıkara, H., Kaasen, A., Kabaeva, N., Kahane, S., Kanayama, H., Kanerva, J., Katz, B., Kayadelen, T., Kenney, J., Kettnerová, V., Kirchner, J., Klementieva, E., Köhn, A., Kopacewicz, K., Kotsyba, N., Kovalevskaitė, J., Krek, S., Kwak, S., Laippala, V., Lambertino, L., Lam, L., Lando, T., Larasati, S. D., Lavrentiev, A., Lee, J., Lê H`ông, P., Lenci, A., Lertpradit, S., Leung, H., Li, C. Y., Li, J., Li, K., Lim, K., Liovina, M., Li, Y., Ljubešić, N., Loginova, O., Lyashevskaya, O., Lynn, T., Macketanz, V., Makazhanov, A., Mandl, M., Manning, C., Manurung, R., Mărănduc, C., Mareček, D., Marheinecke, K., Martínez Alonso, H., Martins, A., Mašek, J., Matsumoto, Y., McDonald, R., McGuinness, S., Mendonça, G., Miekka, N., Misirpashayeva, M., Missilä, A., Mititelu, C., Mitrofan, M., Miyao, Y., Montemagni, S., More, A., Moreno Romero, L., Mori, K. S., Morioka, T., Mori, S., Moro, S., Mortensen, B., Moskalevskyi, B., Muischnek, K., Munro, R., Murawaki, Y., Müürisep, K., Nainwani, P., Navarro Horñiacek, J. I., Nedoluzhko, A., Nešpore-Bērzkalne, G., Nguy˜ên Thị, L., Nguy˜ên Thị Minh, H., Nikaido, Y., Nikolaev, V., Nitisaroj, R., Nurmi, H., Ojala, S., Ojha, A. K., Olúòkun, A., Omura, M., Osenova, P., Östling, R., Øvrelid, L., Partanen, N., Pascual, E., Passarotti, M., Patejuk, A., Paulino-Passos, G., Peljak-Łapińska, A., Peng, S., Perez, C.-A., Perrier, G., Petrova, D., Petrov, S., Phelan, J., Piitulainen, J., Pirinen, T. A., Pitler, E., Plank, B., Poibeau, T., Ponomareva, L., Popel, M., Pretkalniņa, L., Prévost, S., Prokopidis, P., Przepiórkowski, A., Puolakainen, T., Pyysalo, S., Qi, P., Rääbis, A., Rademaker, A., Ramasamy, L., Rama, T., Ramisch, C., Ravishankar, V., Real, L., Reddy, S., Rehm, G., Riabov, I., Rießler, M., Rimkutė, E., Rinaldi, L., Rituma, L., Rocha, L., Romanenko, M., Rosa, R., Rovati, D., Roșca, V., Rudina, O., Rueter, J., Sadde, S., Sagot, B., Saleh, S., Salomoni, A., Samardžić, T., Samson, S., Sanguinetti, M., Särg, D., Saulīte, B., Sawanakunanon, Y., Schneider, N., Schuster, S., Seddah, D., Seeker, W., Seraji, M., Shen, M., Shimada, A., Shirasu, H., Shohibussirri, M., Sichinava, D., Silveira, A., Silveira, N., Simi, M., Simionescu, R., Simkó, K., Šimková, M., Simov, K., Smith, A., Soares-Bastos, I., Spadine, C., Stella, A., Straka, M., Strnadová, J., Suhr, A., Sulubacak, U., Suzuki, S., Szántó, Z., Taji, D., Takahashi, Y., Tamburini, F., Tanaka, T., Tellier, I., Thomas, G., Torga, L., Trosterud, T., Trukhina, A., Tsarfaty, R., Tyers, F., Uematsu, S., Urešová, Z., Uria, L., Uszkoreit, H., Utka, A., Vajjala, S., van Niekerk, D., van Noord, G., Varga, V., Villemonte de la Clergerie, E., Vincze, V., Wallin, L., Walsh, A., Wang, J. X., Washington, J. N., Wendt, M., Williams, S., Wirén, M., Wittern, C., Woldemariam, T., Wong, T.-s., Wróblewska, A., Yako, M., Yamazaki, N., Yan, C., Yasuoka, K., Yavrumyan, M. M., Yu, Z., Žabokrtský, Z., Zeldes, A., Zhang, M., and Zhu, H. (2019). Universal Dependencies 2.5. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. `http://hdl.handle.net/11234/1-3105`.

Zeman, D. (2008). Reusable tagset conversion using tagset drivers. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, pages 213–218.