

Tutorial on Universal Dependencies

Adding a new language to UD

Marie-Catherine de Marneffe¹ Joakim Nivre² **Daniel Zeman**³

¹FNRS, Université catholique de Louvain, Belgium

²Department of Linguistics and Philology, Uppsala University, Sweden

³Institute of Formal and Applied Linguistics, Charles University, Prague, Czechia



Two Scenarios

You want your language in UD



Existing treebank
You have permission



Treebank conversion



No existing treebank
No permission/licence



Building from scratch



Common Steps

First steps

- ▶ Get an account in GitHub
 - ▶ All development goes on here

Get in contact

- ▶ Ask the release team (= Dan) to set up a repository
- ▶ Get in contact with any other teams working on your language, or a related one
- ▶ Register for the mailing list *
- ▶ All contributors given broad edit rights to all data, docs, and tools repositories
- ▶ Fully trust-based setup, git giving a safety net

* <https://lists.uu.se/sympa/info/lingfil-ud>

Release team



Data

- ▶ A GitHub repository for every treebank
 - ▶ UD_{Language}-{Treebank}
 - ▶ **master** branch holds the most recent official release
 - ▶ **dev** branch holds development data, not guaranteed to be valid
 - ▶ Prescribed file names in the repository (extra files under not-to-release)
 - ▶ Recommended: push data to dev early, watch the on-line validation page for errors
- ▶ **Official release:** LINDAT, May & November, all treebanks which contain valid data
 - ▶ **Data-freeze** period two weeks before the release
- ▶ **Release checklist:** https://universaldependencies.org/release_checklist.html



Validation

- ▶ Script(s) to validate treebank data
- ▶ Passing is compulsory
 - ▶ Format validation
 - ▶ Content (guidelines) validation
 - ▶ Language-specific documentation
 - ▶ File names and README contents
- ▶ Runs automatically every time a treebank is updated

Universal Dependencies Validation Report

This is the output of automatic on-line validation of UD data. Besides the official UD contents of the branch is modified. They are also rerun whenever the validation soft

The report on this page is an important indicator whether the current contents of the them. They were considered valid at the time of a previous release and the only error **NEGLECTED**. This is a warning that the error must be fixed as soon as possible (prior errors are fixed. If a treebank was valid in the previous release and now it contains 1 expired. Treebanks that were released in the past but excluded from the most recent

See the [release checklist](#) for more information on treebank requirements and validation

Click on the "report" link to see the full output of the validation software.

UD_Abaza-ATB: CURRENT VALID

UD_Afrikaans-AfriBooms: CURRENT VALID

UD_Akkadian-MCONG: SAPLING ERROR (TOTAL 251679; L0 Repo readme 1; L1 Unicode unicode-normalization 1; L2 Format invalid-head 42233; L2 Format skipper whitespace 2; L2 Morpho invalid-feature 703; L2 Morpho invalid-upos 300; L2 Morpho head 42234; L3 Syntax leaf-aux-cop 1; L3 Syntax leaf-cc 1; L3 Syntax leaf-fixed 5; L3 Syntax right-to-left-conj 10; L3 Syntax right-to-left-fixed 7; L3 Syntax too-many-subj Syntax cop-lemma 44) ([report](#))

UD_Akkadian-PISANDUB: CURRENT VALID

UD_Akkadian-RIAO: CURRENT VALID

UD_Akuntsu-TuDeT: CURRENT VALID

UD_Albanian-TSA: CURRENT VALID

UD_Amharic-ATT: CURRENT ERROR LEGACY; 2022-05-31 (TOTAL 77; L3 Syntax

UD_Amharic-Inku: SAPLING EMPTY

UD_Ancient_Greek-PROIEL: CURRENT VALID WARNING (TOTAL 18; L3 Warning

UD_Ancient_Greek-PTNK: SAPLING ERROR (TOTAL 397; L0 Repo readme 2; L1 Syntax too-many-subjects 14; L3 Warning orphan-parent 4; L4 Morpho feature-valu

UD_Ancient_Greek-Perseus: CURRENT ERROR LEGACY; 2022-05-31 (TOTAL

UD_Ancient_Hebrew-PTNK: CURRENT VALID

<https://quest.ms.mff.cuni.cz/udvalidator/>

Documentation

- ▶ GitHub *docs* repository, Markdown pages → HTML
 - ▶ Easy to add examples with tree visualizations
 - ▶ Automatically regenerated on every push and published on <https://universaldependencies.org> (takes a few minutes)
- ▶ One set of guidelines **per language** (not treebank)
 - ▶ **Mandatory** index page summarizing the language
 - ▶ Template pre-generated when adding a new language
 - ▶ Document as you annotate (or as you write conversion rules)
 - ▶ **Mandatory** pages for lang-spec features and relations
 - ▶ Optional other pages for features, relations and constructions
 - ▶ Automatically generated treebank pages with statistics
- ▶ https://universaldependencies.org/contributing_language_specific.html
- ▶ <http://spyysalo.github.io/annodoc/>



Language-specific Documentation

→ ↻ 🏠 🔒 universaldependencies.org/ml/index.html



[home](#) [edit](#) [page](#) [issue tracker](#)

This page pertains to UD version 2.

UD for Malayalam



Tokenization and Word Segmentation

- In general, words are delimited by whitespace characters or punctuations.
- Multiword tokens are relatively common in Malayalam. In the following situations, we understand orthographic tokens as corresponding to multiple syntactic words and split them:
 - The copula ആകു / *āk* "to be" is written as a suffix of the nominal/adjectival predicate. However, sometimes it is suffixed to another word in the clause, indicating that it is a clitic rather than a derivational morpheme that would derive a verb from a noun/adjective.
 - The quotative particle or the complementizer എന്ന് / *enn* "that" usually occurs as a suffix of the verb or the copula. Given that we split the copula as a syntactic word, we split the complementizer as well. (Also, it increases parallelism with languages where complementizers are independent words, and avoids having to define a language-specific feature for verb with complementizer.)
 - The coordinating clitics -ഉം / *-um* and -ഒ / *-o* are written together with conjuncts but analyzed as separate syntactic words.
 - In orthography sometimes the object and the verb of a sentence occur as a multiword token. For example, in the sentence പെൺകുട്ടി തന്റെ സുഹൃത്തിന് കത്തെഴുതി. / *penkutti tanre suhrttin katteluti*. "The girl wrote a letter to her friend.", കത്ത് / *katt* "letter" and എഴുതി / *eluti* "wrote" occur as a multiword token and are split.
- There are letters that can be encoded in multiple ways, even after standard Unicode normalization (NFC), which is required in UD.
 - The *viram* sign, used across Indic scripts to cancel the vowel (*a* or schwa) inherently present in a consonant character, may in Malayalam actually result in a half vowel *ĩ*, especially at the end of a word. To signal that there is really no vowel, some consonants in the Malayalam script have so-called *chillu* variants, which have their own Unicode position. However, there is an older alternative of



Language-specific Feature Documentation

→ ↻ 🏠 🔒 universaldependencies.org/nhi/feat/Animacy-obj.html



[home](#) [edit page](#) [issue tracker](#)

This page pertains to UD version 2.

Animacy [obj] : Whether the indefinite object is human or non-human.

The *Animacy* feature in Western Sierra Puebla Nahuatl is relevant for indefinite object prefixes on Verbs. Specifically, there is one indefinite object prefix for indefinite human objects (te-), and another for indefinite non-human objects (tla-). We use the layered features in order to be explicit that this feature only applies to the object of the Verb, not to any of the other arguments which can also be encoded in the Verb.

Hum: human

Examples

- *niteitstinimi* "Ando viendo personas."

Nhum: Non-human

Examples

- *nitlaoni* "Yo tomo (cosas/bebidas)."



Language-specific Relation Documentation

→ ↻ 🏠 🔒 universaldependencies.org/yue/dep/case-loc.html



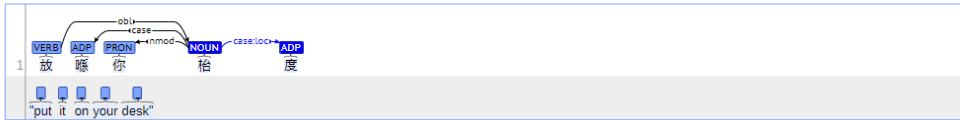
[home](#) [edit](#) [page](#) [issue tracker](#)

This page pertains to UD version 2.

case:loc: postpositional localizer

We treat localizers as postpositions which typically denote spatial locations analogous to adpositions or case markers in some languages, although a few localizers have further grammaticalized functions denoting temporal and other non-spatial concepts. (See [ADP](#) for a list of localizers.)

The head of the localizer is the noun or the main verb of the clause preceding it. Localizers are always tagged ADP (adposition). When it follows a noun, it receives the `case:loc` relation label. But if it follows a clause and acts as a subordinator, it receives the `mark` relation (but retains the tag ADP).



case:loc in other languages: [\[hy\]](#) [\[hyw\]](#) [\[yue\]](#) [\[zh\]](#)



Language-specific Rules for Validator

← → ↺ ↻ quest.ms.mff.cuni.cz/udvalidator/cgi-bin/unidep/langspec/specify_auxiliary.pl?code=cs



Specify auxiliaries for Czech

Remember: Not everything that a traditional grammar labels as auxiliary is necessarily an [auxiliary in UD](#). Just because a verb combines with another verb does not necessarily mean that one of the verbs is aux; other, or in some languages as a serial verb construction ([compound:svc](#)). Language-specific tests whether a verb is auxiliary are [grammatical](#) rather than [semantic](#): just because something has a modal or near-modal verb does not count as auxiliaries at all. Some verbs function as auxiliaries in some constructions and as full verbs in others (e.g., *to have* in English). There are also auxiliaries whose nature is completely different

Remember: A language typically has at most one lemma for [copula](#). Exceptions include deficient paradigms (different present and past copula, positive and negative, imperfect and iterative), and also the Roman become, to stay, to look like, to be called" etc. are not copulas in UD, even if a traditional grammar classifies them as such. In UD they should head an [xcomp](#) relation instead. Existential "to be" can be copula only two different verbs, then **the existential one is not a copula**. A copula is normally tagged [AUX](#). Exception: in some languages a personal or demonstrative pronoun / determiner can be used as a copula and then

Edit or add auxiliaries

[být bývat bývávat](#)

Known auxiliaries for this and other languages

Language	Total	Copula	Perfect	Past	Future	Passive	Conditional	Necessitative	Potential	Desiderative
Czech	cs 3	být bývat bývávat		být	být	být	být			
Upper Sorbian	hsb 1	być	być		być	być	być			
Polish	pl 10	bywać być to		być	być	być zostawać zostać	by			
Slovak	sk 3	byť bývať		byť	byť	byť	by			
Ukrainian	uk 4	бувати бути		бути	бути	бути	б би			
Belarusian	be 4	быць гэта		быць	быць	быць	б бы			
Russian	ru 5	быть это		быть	быть	быть бывать быть	б бы			
Old East Slavic	orv 4	быти		стати	быти	быти	бъ бы			
Old Church Slavonic	cu 1	быти	быти				быти			
Bulgarian	bg 6	бѣда съм	сѣм		ще	бивам бѣда съм	би			
Macedonian	mk 0									
Pomak	qpm 6	býdom da som še šom štom								
Serbian	sr 2	biti		biti	hteti	biti	biti			



Language-specific Rules for Validator

- ▶ **Morphological features**

- ▶ https://quest.ms.mff.cuni.cz/udvalidator/cgi-bin/unidep/langspec/specify_feature.pl?lcode=cs

- ▶ **Dependency relations**

- ▶ https://quest.ms.mff.cuni.cz/udvalidator/cgi-bin/unidep/langspec/specify_deprel.pl?lcode=cs

- ▶ **Auxiliaries and copulas**

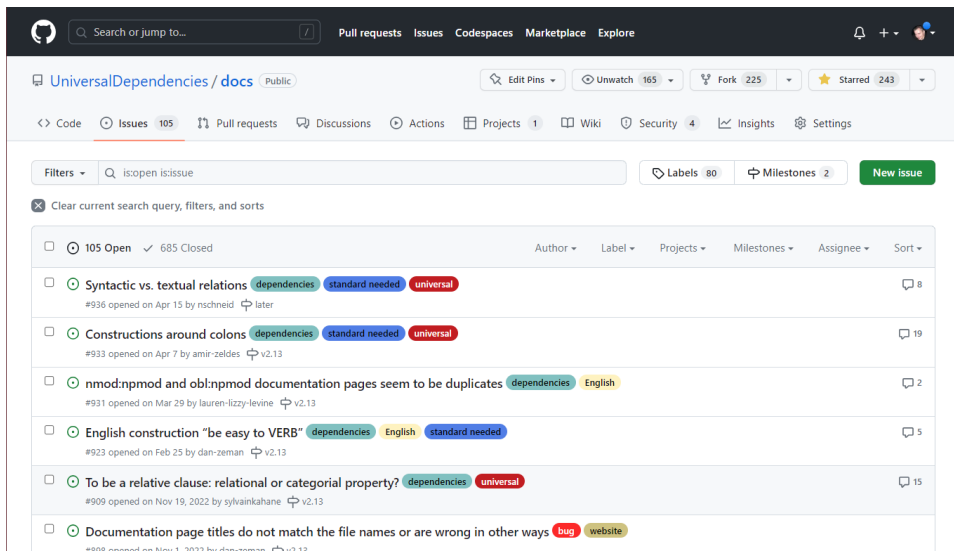
- ▶ https://quest.ms.mff.cuni.cz/udvalidator/cgi-bin/unidep/langspec/specify_auxiliary.pl?lcode=cs

- ▶ Do not make other people's treebanks invalid!



Linguistic Discussion

Linguistic discussion goes on under the *docs* repository



The screenshot shows the GitHub interface for the `UniversalDependencies/docs` repository. The 'Issues' tab is selected, displaying 105 open issues. The issues are listed with their titles, labels, and comment counts. The labels include 'dependencies', 'standard needed', 'universal', 'English', 'bug', and 'website'.

Filters: Labels: 80 Milestones: 2 [New issue](#)

☒ Clear current search query, filters, and sorts

<input type="checkbox"/>	105 Open ✓ 685 Closed	Author	Label	Projects	Milestones	Assignee	Sort
<input type="checkbox"/>	Syntactic vs. textual relations dependencies standard needed universal 8	#936 opened on Apr 15 by nschneid	later				
<input type="checkbox"/>	Constructions around colons dependencies standard needed universal 19	#933 opened on Apr 7 by amir-zeldes	v2.13				
<input type="checkbox"/>	nmod:npmmod and obl:npmmod documentation pages seem to be duplicates dependencies English 2	#931 opened on Mar 29 by lauren-lizzy-levine	v2.13				
<input type="checkbox"/>	English construction "be easy to VERB" dependencies English standard needed 5	#923 opened on Feb 25 by dan-zeman	v2.13				
<input type="checkbox"/>	To be a relative clause: relational or categorial property? dependencies universal 15	#909 opened on Nov 19, 2022 by sylvainkahane	v2.13				
<input type="checkbox"/>	Documentation page titles do not match the file names or are wrong in other ways bug website	#808 opened on Nov 1, 2023 by dan-zeman	v2.13				



Linguistic Discussion

- ▶ The issue tracker for the *docs* repository is where all the UD activity is happening
 - ▶ Universal guidelines \Rightarrow *docs* issues
 - ▶ Language-specific guidelines \Rightarrow **still *docs* issues** (and not treebank repository issues)
 - ▶ Even if the language has only one treebank
 - ▶ Bugs in a treebank \Rightarrow treebank repository issues



Where To Get The Data?

Free text:

- ▶ Plenty of options:
 - ▶ Wikimedia projects: Wikipedia, Wikinews, ...
 - ▶ Public domain texts (varies by country)
 - ▶ Out of copyright (e.g. old literature, folktales)
 - ▶ Laws/state administrative texts
 - ▶ **Cairo Cicling Corpus**
<https://github.com/UniversalDependencies/cairo/blob/master/translations.txt>

Borderline:

- ▶ Examples from linguistic literature
- ▶ Shuffled sentences from the web

Non-free text:

- ▶ Contact copyright holders early on



How Much Data?

- ▶ **Required minimum:** 20 sentences and 100 words
 - ▶ Useful sample: 100 sentences
 - ▶ Very good: 100K tokens
 - ▶ Biggest treebank: 3M tokens
- ▶ CoNLL-2006, smallest treebank: 29K tokens
- ▶ You can add more data for the next release!



How Long Will It Take?

- Some approximate numbers:

Language	Annotators	Tokens	Months
Kazakh	2	4,500	1
Buryat	1	10,000	3
Irish	1	23,600	12

In all the above cases, annotation guidelines were developed from scratch by people with no prior exposure to UD.



Annotation Tools

- ▶ No official annotation tool
- ▶ A list of tools: <http://universaldependencies.org/tools.html>



Bootstrapping

- ▶ Annotate 20 sentences
- ▶ <https://lindat.mff.cuni.cz/services/udpipe/>
- ▶ <https://stanfordnlp.github.io/stanza/>
- ▶ <https://trankit.readthedocs.io/en/latest/overview.html>



Bootstrapping

- ▶ Annotate 20 sentences
- ▶ Train a parser on that

- ▶ <https://lindat.mff.cuni.cz/services/udpipe/>
- ▶ <https://stanfordnlp.github.io/stanza/>
- ▶ <https://trankit.readthedocs.io/en/latest/overview.html>



Bootstrapping

- ▶ Annotate 20 sentences
- ▶ Train a parser on that
- ▶ Use it to parse next 100 sentences

- ▶ <https://lindat.mff.cuni.cz/services/udpipe/>
- ▶ <https://stanfordnlp.github.io/stanza/>
- ▶ <https://trankit.readthedocs.io/en/latest/overview.html>



Bootstrapping

- ▶ Annotate 20 sentences
 - ▶ Train a parser on that
 - ▶ Use it to parse next 100 sentences
 - ▶ Manually fix annotation
 - ▶ **UD Requirement: At least UPOS, HEAD and DEPREL must be manually verified**
 - ▶ Lemmas and features can be released even if predicted automatically (without manual verification of each word)
-
- ▶ <https://lindat.mff.cuni.cz/services/udpipe/>
 - ▶ <https://stanfordnlp.github.io/stanza/>
 - ▶ <https://trankit.readthedocs.io/en/latest/overview.html>



Bootstrapping

- ▶ Annotate 20 sentences
 - ▶ Train a parser on that
 - ▶ Use it to parse next 100 sentences
 - ▶ Manually fix annotation
 - ▶ **UD Requirement: At least UPOS, HEAD and DEPREL must be manually verified**
 - ▶ Lemmas and features can be released even if predicted automatically (without manual verification of each word)
 - ▶ Retrain the parser on 120 sentences
 - ▶ ...
-
- ▶ <https://lindat.mff.cuni.cz/services/udpipe/>
 - ▶ <https://stanfordnlp.github.io/stanza/>
 - ▶ <https://trankit.readthedocs.io/en/latest/overview.html>



Transliteration

https://github.com/dan-zeman/translit/blob/main/conllu_translit.pl

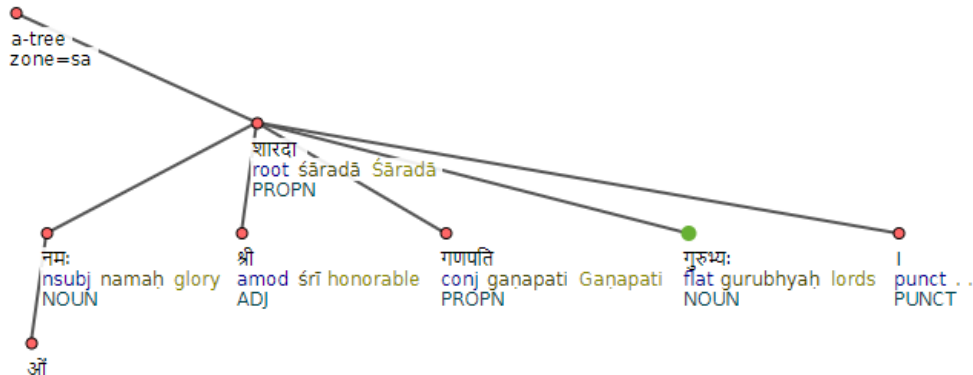
```
# newpar id = panc0.p2
```

```
# sent_id = panc0.s2
```

```
# text = ओ नमः श्रीशारदागणपतिगुरुभ्यः।
```

```
# translit = oh namaḥ śrīśāradāgaṇapatiḡurubhyaḥ.
```

1	ओ	ओ	INTJ	-	-	2	discourse	-	Translit=oh LTranslit=oh Gloss=oh
2	नमः	नमस्	NOUN	-	Case=Nom Gender=Neut Number=Sing	4	nsubj	-	Translit=namaḥ LTranslit=namas Gloss=glory
3-6	श्रीशारदागणपतिगुरुभ्यः			-	-	-		-	Translit=śrīśāradāgaṇapatiḡurubhyaḥ Gloss=(to-the-lor
3	श्री	श्री	ADJ	-	Compound=Yes	4	amod	-	Translit=śrī LTranslit=śrī Gloss=honorable
4	शारदा	शारदा	PROPN	-	Compound=Yes	0	root	-	Translit=śāradā LTranslit=śāradā Gloss=Śāradā
5	गणपति	गणपति	PROPN	-	Compound=Yes	4	conj	-	Translit=gaṇapati LTranslit=gaṇapati Gloss=Gaṇapati
6	गुरुभ्यः	गुरु	NOUN	-	Case=Dat Gender=Masc Number=Plur	4	flat	-	Translit=gurubhyaḥ LTranslit=guru Gloss=lords
7	।	।	PUNCT	-	-	4	punct	-	Translit=. LTranslit=. Gloss=.



Udapi

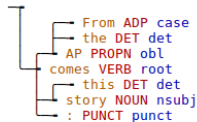
- ▶ A library and command line tool for processing UD data
 - ▶ **Python**, Java, Perl
- ▶ Format conversions
- ▶ Fixes of systematic errors
- ▶ Validation tests
- ▶ Evaluation, filtering, statistics
- ▶ Tree visualization
- ▶ <https://udapi.github.io>
 - ▶ <https://ufal.mff.cuni.cz/~zeman/vyuka/deptreebanks/NPFL075-working-with-UD.pdf>



Tree Visualization Tools

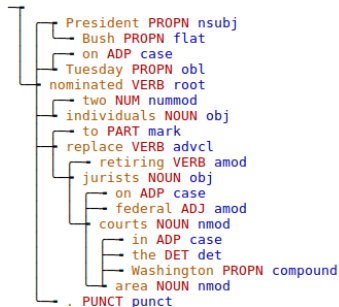
```
cat en_ewt-ud-dev.conllu | udapy -T | less -R
```

```
docname = weblog-blogspot.com_nominations_20041117172713_ENG_20041117_172713
# sent_id = weblog-blogspot.com_nominations_20041117172713_ENG_20041117_172713-0001
# text = From the AP comes this story :
```



```
# sent_id = weblog-blogspot.com_nominations_20041117172713_ENG_20041117_172713-0002
```

```
# text = President Bush on Tuesday nominated two individuals to replace retiring jurists on federal courts in the Washington area:
```

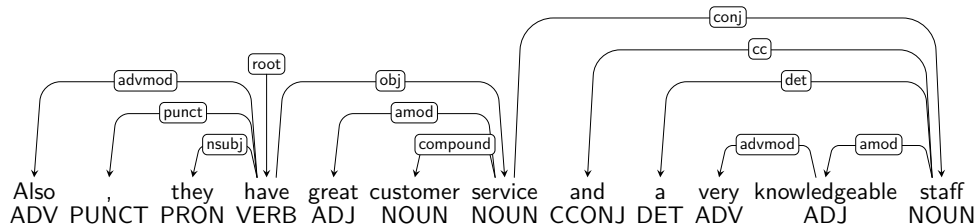


```
# sent_id = weblog-blogspot.com_nominations_20041117172713_ENG_20041117_172713-0003
```

```
# text = Bush nominated Jennifer M. Anderson for a 15-year term as associate judge of the Superior Court of the District of Colum
```

Tree Visualization Tools

cat en_ewt-ud-dev.conllu | udapy write.Tikz



Summary

What you need to do

- ▶ Join the project
- ▶ Start annotating or converting
- ▶ Ask if you get stuck!

