07 Universal Dependencies

Universal Dependencies (UD) is an internationally harmonised annotation framework that aims to standardize the morphological and syntactic tagging of texts across languages in order to foster the development of multilingual language technologies and contrastive linguistics research. Grounded in dependency grammar principles, UD provides a universal set of grammatical categories – including parts of speech (POS), morphological features and syntactic dependency relations – and guidelines for their use. The framework is flexible, incorporating language-specific tags as needed. To date, over 240 corpora worldwide have been manually tagged using the framework, including the universal treebanks of both written and spoken Slovene.

- Introduction to Tags
- Annotation Guidelines
- References and Links

Introduction to Tags

The Universal Dependencies framework establishes a comprehensive and universal set of tags for parts of speech (POS), morphological features and syntactic dependencies that can be adopted in the treebanks of individual languages, or supplemented with new morphological features or derivations of core relations when necessary. For the Slovene language data, this includes the adoption of all 17 parts of speech (see Table 1), 22 morphological features spanning 62 distinct values (see Table 2), and 35 types of dependency relations (see Table 3).

Tag	Description
ADJ	adjective
ADP	adposition
ADV	adverb
AUX	auxiliary
CCONJ	coordinating conjunction
DET	determiner
INTJ	interjection
NOUN	noun
NUM	numeral
PART	particle
PRON	pronoun
PROPN	proper noun
PUNCT	punctuation
SCONJ	subordinating conjunction
SYM	symbol
VERB	verb
X	other

Table 1: Part-of-speech tags used in Slovene texts.

Feature	Value	Description
Abbr	Yes	abbreviation

Feature	Value	Description
Animacy	Anim, Inanim	animacy
Aspect	Imp, Perf	aspect
Case	Nom, Gen, Dat, Acc, Loc, Ins	case
Definite	Ind, Def	definiteness or state
Degree	Pos, Cmp, Sup	degree
Foreign	Yes	is this a foreign word?
Gender	Masc, Fem, Neut	gender
Gender[psor]	Masc. Fem, Neut	possessor’s gender
Mood	Ind, Imp, Cnd	mood
Number	Sing, Dual, Plur	number
Number[psor]	Sing, Dual, Plur	possessor’s number
NumForm	Word, Digit, Roman	numeral form
NumType	Card, Ord, Mult, Sets	numeral type
Person	1, 2, 3	person
Polarity	Neg, Pos	polarity
Poss	Yes	possessive
PronType	Prs, Int, Rel, Dem, Tot, Neg, Ind	pronominal type
Reflex	Yes	reflexive
Tense	Pres, Fut	tense
Variant	Bound, Short	alternative form of word
VerbForm	Fin, Inf, Sup, Part, Conv	form of verb or deverbative

Table 2: Tags for morphological features used in Slovene texts. In the corpus, these are listed in the form of feature and value pairs (e.g., Tense=Pres).

Tag	Description
acl	clausal modifier of noun
advcl	adverbial clause modifier
advmod	adverbial modifier
amod	adjectival modifier
appos	appositional modifier

Tag	Description
aux	auxiliary verb
case	case marking preposition
cc	coordinating conjunction
ccomp	clausal complement
conj	conjunct
cop	copula verb
csubj	clausal subject
dep	unspecified dependency
det	determiner
discourse	discourse element
dislocated	dislocated element
expl	expletive
fixed	fixed multi-word expression
flat	flat multi word-expression
goeswith	disjointed token
iobj	indirect object
list	list
mark	marker (subordinating conjunction)
nmod	nominal modifier
nsubj	nominal subject
nummod	numeric modifier
obj	(direct) object
obl	oblique nominal (adjunct)
orphan	dependent of missing parent
parataxis	parataxis
punct	punctuation symbol
reparandum	overriden disfluency
root	root element
vocative	vocative
xcomp	open clausal complement

Table 3: Tags for syntactic dependency relations (without subtypes) used in Slovene texts.

Annotation Guidelines

This chapter summarizes the annotation guidelines for the Universal Dependencies (UD) morphology and syntax as applied to Slovene texts. The guidelines are arranged from the latest, up-to-date version to the oldest version.

Version 1.7

Project SPOT
DOBROVOLJC, Kaja in TERČON, Luka: 2024. *Universal Dependencies: Smernice za označevanje besedil v slovenščini.* Različica 1.7. Rezultat projekta Na drevesnici temelječ pristop k raziskavam govorjene slovenščine. [DOCX] [PDF] - only in Slovene

Version 1.3

Project SPOT
DOBROVOLJC, Kaja in TERČON, Luka: 2023. *Universal Dependencies: Smernice za označevanje besedil v slovenščini.* Različica 1.3. Rezultat projekta Na drevesnici temelječ pristop k raziskavam govorjene slovenščine. [DOCX] [PDF]) - only in Slovene

Version 1.0

Project Development of Slovene in a Digital Environment
DOBROVOLJC, Kaja in TERČON, Luka: 2023. *Universal Dependencies: Smernice za označevanje besedil v slovenščini.* Rezultat projekta Razvoj slovenščine v digitalnem okolju. [DOCX] [PDF] - only in Slovene

Priloga k smernicam: Odprta vprašanja pri prenosu označevalne sheme Universal Dependencies na slovenska besedila

[DOCX] [PDF] - only in Slovene

References and Links

This chapter compiles relevant references and provides links to projects where the the Universal Dependencies (UD) morphology and syntax have been developed and applied to Slovene texts.

Main website of the Universal Dependencies project: https://universaldependencies.org/

General guidelines: https://universaldependencies.org/guidelines.html

Slovenian-specific guidelines (in English): https://universaldependencies.org/sl/index.html
Platform for discussing and proposing improvements to Slovenian-specific guidelines (in English):

https://github.com/UniversalDependencies/UD_Slovenian-SSJ/issues

Corpora containing manually revised UD tags

Slovenian-SSJ written language treebank: https://github.com/UniversalDependencies/UD_Slovenian-SSJ

Slovenian-SST spoken language treebank:

https://github.com/UniversalDependencies/UD_Slovenian-SST
Slovenian-SSJ treebank as part of the latest SUK 1.1 training corpus: Arhar Holdt, Špela; et al., 2024, Training corpus SUK 1.1, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042,

http://hdl.handle.net/11356/1959.

References

Dobrovoljc, K. (2024). Extending the Spoken Slovenian Treebank. Conference on Language Technologies and Digital Humanities (JT-DH-2024), Ljubljana, Slovenia.

https://doi.org/10.5281/zenodo.13936394

Dobrovoljc, K., Terčon, L., Ljubešić, N. (2023). Universal Dependencies za slovenščino: nove smernice, ročno označeni podatki in razčlenjevalni model. Slovenščina 2.0, 11(1): 218–246.

https://doi.org/10.4312/slo2.0.2023.1.218-246 [PDF] - only in Slovene

Dobrovoljc, K., Terčon, L., & Ljubešić, N. (2022). Universal Dependencies za slovenščino: nadgradnja smernic, učnih podatkov in razčlenjevalnega modela. In D. Fišer & T. Erjavec (Eds.), Jezikovne tehnologije in digitalna humanistika: zbornik konference (pp. 30–39). Inštitut za novejšo zgodovino. https://nl.ijs.si/jtdh22/pdf/JTDH2022_Dobrovoljc-et-al_Universal-Dependencies-za-slovenscino.pdf

de Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal Dependencies. Computational Linguistics, 47(2), 255–308. https://doi.org/10.1162/coli_a_00402

Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., & Zeman, D. (2020). Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. Proceedings of the Twelfth Language Resources and Evaluation Conference, 4034–4043.

https://aclanthology.org/2020.lrec-1.497

Dobrovoljc, K., Erjavec, T., & Krek, S. (2017). The Universal Dependencies Treebank for Slovenian. Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing, 33–38.

https://doi.org/10.18653/v1/W17-1406

Dobrovoljc, K., & Nivre, J. (2016). The Universal Dependencies Treebank of Spoken Slovenian. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 1566–1573. https://aclanthology.org/L16-1248