07 Universal Dependencies

Universal Dependencies (UD) je mednarodno usklajena označevalna shema, ki si prizadeva za poenoteno oblikoslovno in skladenjsko označevanje besedil v vseh svetovnih jezikih, da bi pospešila razvoj večjezičnih jezikovnih tehnologij in kontrastivnih jezikoslovnih raziskav. Znotraj sheme, ki temelji na načelih odvisnostne slovnice, je bil vzpostavljen univerzalni nabor slovničnih kategorij (besednih vrst, oblikoslovnih lastnosti in odvisnostnih skladenjskih relacij) in smernic za njihovo pripisovanje, obenem pa zasnova sheme dovoljuje tudi vpeljevanje jezikovnospecifičnih oznak, če je to potrebno. Doslej je bilo s to shemo ročno označenih že več kot 250 korpusov po vsem svetu, med katerimi sta tudi univerzalni odvisnostni drevesnici pisne (SSJ) in govorjene (SST) slovenščine.

- Predstavitev oznak
- Označevalne smernice
- Reference in povezave

Predstavitev oznak

Shema Universal Dependencies določa univerzalni nabor oznak za besedne vrste, oblikoslovne lastnosti in odvisnostne skladenjske relacije, ki se lahko prenesejo na drevesnice posameznih jezikov ali pa dopolnijo z novimi oblikoslovnimi oznakami oz. izpeljavami jedrnih relacij, če je to potrebno. Konkretno je bilo na slovenske podatke preneseno vseh 17 besednih vrst (Tabela 1), 22 oblikoslovnih lastnosti z 62 različnimi vrednostmi (Tabela 2) in 35 odvisnostnih relacij (Tabela 3).

Oznaka	Opis
ADJ	pridevnik
ADP	adpozicija (predlog)
ADV	prislov
AUX	pomožnik
CCONJ	priredni veznik
DET	določilnik
INTJ	medmet
NOUN	samostalnik
NUM	števnik
PART	členek
PRON	zaimek
PROPN	lastnoimenski samostalnik
PUNCT	ločilo
SCONJ	podredni veznik
SYM	simbol
VERB	glagol
X	drugo

Tabela 1: Na slovenskih besedilih uporabljene oznake za besedne vrste.

Lastnost	Vrednosti	Opis
Abbr	Yes	okrajšanost

Lastnost	Vrednosti	Opis
Animacy	Anim, Inanim	živost
Aspect	Imp, Perf	vid
Case	Nom, Gen, Dat, Acc, Loc, Ins	sklon
Definite	Ind, Def	določnost
Degree	Pos, Cmp, Sup	stopnja
Foreign	Yes	tujejezičnost
Gender	Masc, Fem, Neut	spol
Gender[psor]	Masc. Fem, Neut	spol svojine
Mood	Ind, Imp, Cnd	naklon
Number	Sing, Dual, Plur	število
Number[psor]	Sing, Dual, Plur	število svojine
NumForm	Word, Digit, Roman	oblika številnika
NumType	Card, Ord, Mult, Sets	vrste številnika
Person	1, 2, 3	oseba
Polarity	Neg, Pos	nikalnost
Poss	Yes	svojina
PronType	Prs, Int, Rel, Dem, Tot, Neg, Ind	vrsta zimka
Reflex	Yes	povratnost
Tense	Pres, Fut	čas
Variant	Bound, Short	oblika zaimka
VerbForm	Fin, Inf, Sup, Part, Conv	vrsta glagola

Tabela 2: Na slovenskih besedilih uporabljene oznake za oblikoslovne lastnosti. V korpusu so navedene v obliki parov lastnosti in vrednosti (npr. Tense=Pres).

Oznaka	Angleški opis	Slovenski opis
acl	clausal modifier of noun	stavčni prilastki
advcl	adverbial clause modifier	prislovni odvisniki
advmod	adverbial modifier	prislovna določila (v širšem smislu)
amod	adjectival modifier	pridevniški prilastki
appos	appositional modifier	pristavčna določila

Oznaka	Angleški opis	Slovenski opis
aux	auxiliary verb	pomožni glagoli
case	case marking preposition	predlogi
cc	coordinating conjunction	priredni vezniki
ccomp	clausal complement	stavčna dopolnila (predmetni odvisniki)
conj	conjunct	priredno zloženi elementi
cop	copula verb	vezni glagoli
csubj	clausal subject	osebkovi odvisniki
dep	unspecified dependency	nedoločena povezava
det	determiner	določilniki
discourse	discourse element	diskurzni členki
dislocated	dislocated element	dislocirani elementi
expl	expletive	ekspletivne besede
fixed	fixed multi-word expression	funkcijske zveze
flat	flat multi word-expression	eksocentrične zveze
goeswith	disjointed token	razdruženi deli besed
iobj	indirect object	nepremi predmeti
list	list	seznami
mark	marker (subordinating conjunction)	podredni vezniki
nmod	nominal modifier	samostalniški prilastki
nsubj	nominal subject	samostalniški osebki
nummod	numeric modifier	številčna določila
obj	(direct) object	premi predmeti
obl	oblique nominal (adjunct)	odvisne samostalniške zveze
orphan	dependent of missing parent	elementi v eliptičnih strukturah
parataxis	parataxis	stavčna soredja
punct	punctuation symbol	ločila
reparandum	overriden disfluency	samopopravljanja
root	root element	koren povedi
vocative	vocative	ogovori
xcomp	open clausal complement	odprta stavčna dopolnila

Tabela 3: Na slovenskih besedilih uporabljene oznake za odvisnostne skladenjske relacije (brez podoznak).

Označevalne smernice

V tem poglavju so zbrane označevalne smernice za oblikoslovje in skladnjo po sistemu Universal Dependencies (UD).

Različica 1.7

projekt 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]

Različica 1.3

projekt 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]

Različica 1.0

projekt Razvoj slovenščine v digitalnem okolju

DOBROVOLJC, Kaja in TERČON, Luka: 2023. *Universal Dependencies: Smernice za označevanje besedil v slovenščini.* Rezultat projekta Razvoj slovenščine v digitalnem okolju. [DOCX] [PDF]

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

[DOCX] [PDF]

Reference in povezave

V tem poglavju so zbrane relevantne reference in povezave na projekte, v katerih se je označevalni sistem razvijal ter uporabljal.

Krovna spletna stran projekta Universal Dependencies: https://universaldependencies.org/

Splošne smernice: https://universaldependencies.org/guidelines.html

Smernice za slovenščino (v angleščini): https://universaldependencies.org/sl/index.html
Platforma za diskusijo glede smernic za slovenščino in predloge izboljšav (v angleščini):

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

Korpusi z ročno pregledanimi oznakami UD

Drevesnica pisne slovenščine SSJ: https://github.com/UniversalDependencies/UD_Slovenian-SSJ

Drevesnica govorjene slovenščine SST: https://github.com/UniversalDependencies/UD_Slovenian-SST
Drevesnica SSJ kot del trenutne različice učnega korpusa SUK: 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.

Reference

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]

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