SKLADENJSKA DREVESNICA GOVORJENE SLOVENŠČINE: STANJE IN PERSPEKTIVE

KAJA DOBROVOLJC
Univerza v Ljubljani, Filozofska fakulteta, Ljubljana, Slovenija
kaja.dobrovoljc@ff.uni-lj.si
Institut Jožef Stefan, Odsek za umetno inteligenco, Ljubljana, Slovenija

V prispevku predstavljamo drevesnico SST (angl. Spoken Slovenian Treebank), prvi skladenjsko razčlenjeni korpus govorjene slovenščine, ki vsebuje uravnotežen in reprezentativni nabor besedil referenčnega korpusa govorjene slovenščine Gos z ročno pripisanimi podatki o lemah, besednih vrstah in oblikoslovnih lastnostih besed ter njihovimi odvisnostnimi skladenjskimi razmerji. Konkretno drevesnica temelji na označevalni shemi Universal Dependencies (UD), ki si prizadeva za mednarodno poenoteno oblikoskladenjsko označevanje besedil in se zaradi svoje interoperabilnosti, fleksibilnosti in naslavljanja širokega nabora slovničnih pojavov – tudi tipično govorjenih – vse pogosteje uporablja tudi za razčlenjevanje govorjenih besedil. Po predstavitvi zasnove, vsebine in dostopnosti obstoječe različice drevesnice SST v drugem delu prispevka predstavimo prve rezultate in načrte v povezavi z njenim nadaljnjim razvojem, kot sta razširitev z novimi besedili in nadgradnja smernic za označevanje tipično govorjenih pojavov.

DOI
https://doi.org/10.18690/um.ff.4.2024.3

ISBN
978-961-286-882-6

Ključne besede:
jezikoslovno označevanje, skladenjsko razčlenjeni korpusi, odvisnostna slovnica, govorjeni jezik, korpusno jezikoslovje

Univerzitetna založba Univerze v Mariboru

DOI
https://doi.org/
10.18690/um.ff.4.2024.3

ISBN
978-961-286-882-6

Keywords:
linguistic annotation,
syntactically parsed corpora,
dependency grammar,
spoken language,
corpus linguistics

University of Maribor Press

SPOKEN SLOVENIAN TREEBANK: CURRENT SITUATION AND PERSPECTIVES

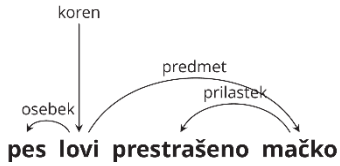
KAJA DOBROVOLJC
University of Ljubljana, Faculty of Arts, Ljubljana, Slovenia
kaja.dobrovoljc@ff.uni-lj.si
Jožef Stefan Institute, Department for Artificial Intelligence, Ljubljana, Slovenia

In this paper we present the Spoken Slovenian Treebank (SST), the first syntactically annotated corpus of spoken Slovene containing a balanced and representative set of transcriptions from the Gos reference corpus of spoken Slovene, with manually annotated lemmas, morphological features and syntactic dependencies. The treebank is based on the Universal Dependencies (UD) annotation scheme, which aims at harmonised corpus annotation across languages and is increasingly applied to spoken data due to its interoperability, flexibility and the coverage of a wide range of grammatical structures, including speech-specific phenomena. After summarising the design, content and accessibility of the existing version of the SST, the second part of this paper describes the first results of the ongoing development, which includes the extension of the corpus with new data and the improvement of speech-specific annotation guidelines.

1 Uvod[1]

Jezikoslovno označeni korpusi, tj. besedilne zbirke, v katerih so besedam, stavkom, besedilom ali drugim oblikam jezikovnih podatkov pripisane opisne oznake različnih vrst, poleg svoje pomembne vloge na področju računalniške obdelave jezikov, kjer se tovrstni viri uporabljajo za učenje statističnih modelov in evalvacijo na njih temelječih orodij, predstavljajo enega temeljnih virov za razvoj slovničnih teorij in na podatkih temelječe jezikoslovne raziskave (Ide in Pustojevsky 2017).

Med najbolj splošno razširjenimi oblikami jezikoslovno označenih korpusov so skladenjsko razčlenjeni korpusi, t. i. skladenjske drevesnice, ki zgradbo stavkov opisujejo v obliki hierarhično urejenih grafov (skladenjskih dreves), v katerih točke grafa označujejo posamezne besede, povezave med točkami pa skladenjska razmerja med njimi. Med raznolikimi teoretskimi pristopi (gl. pregled v Kübler et al. 2009 in Ledinek 2014) danes prevladujejo predvsem skladenjske drevesnice na temelju odvisnostne slovnice (Tesnière 1959, Mel′čuk 1988), ki skladenjsko strukturo opisuje skozi binarne asimetrične relacije (odvisnosti) med posameznimi besedami. Preprost primer take razčlembe prikazuje slika 1, v katerem je povedek *lovi* analiziran kot jedro stavka, osebek in predmet pa kot njegova odvisna oz. podrejena elementa; prav tako je samostalniškemu jedru predmeta podrejeno njegovo določilo v obliki levega prilastka.



Slika 1: Primer odvisnostnega drevesa
Vir: lasten

Čeprav imajo skladenjsko razčlenjeni korpusi pisne slovenščine že precej dolgo zgodovino (Džeroski et al. 2006, Erjavec et al. 2010, Krek et al. 2020) in danes največja drevesnica slovenskega jezika obsega že preko 13.000 razčlenjenih povedi (Arhar Holdt et al. 2023), so se maloštevilni poskusi jezikoslovnega označevanja

[1] Prispevek je nastal v okviru raziskovalnega projekta Na drevesnici temelječ pristop k raziskavam govorjene slovenščine (št. Z6-4617) in raziskovalnega programa Jezikovni viri in tehnologije za slovenski jezik (št. P6-0411), ki ju sofinancira Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije (ARIS).

govornega gradiva doslej osredotočali predvsem na organizacijo diskurza (npr. Verdonik 2020) in semantično analizo (npr. Antloga 2022), ne pa na slovnično analizo na nižjih ravneh, kot sta oblikoslovna in skladenjska analiza.

Da so tako označeni korpusi pomembni gradivni vir za raziskave govora, potrjujejo tudi številne govorne drevesnice za druge jezike, ki so nastajale od začetka 90. let prejšnjega stoletja, kot so korpusi Switchboard za angleščino (Godfrey et al. 1992), CGN za nizozemščino (van der Wouden et al. 2002), PDTSL za češčino (Hajič et al. 2008), NDC in LIA za norveščino (Øvrelid et al. 2018, Käsen et al. 2022), Rhapsodie za francoščino (Lacheret-Dujour et al. 2019) ter večjezični zbirki Verbmobil (Hinrichs et al. 2000) in CHILDES (MacWhinney 2014), če jih naštejemo le nekaj.

Kot odgovor na ta infrastrukturni manko v slovenskem prostoru je bila leta 2016 izdelana prva skladenjska drevesnica govorjene slovenščine, drevesnica SST (angl. *Spoken Slovenian Treebank*), ki še danes predstavlja edini tovrstni jezikovni vir za govorjeno slovenščino, a njegova vsebina v slovenskem prostoru doslej še ni bila podrobneje predstavljena. Da bi zapolnili to vrzel in osvetlili doslej le delno izkoriščen metodološki potencial tega jezikovnega vira za bodoče jezikoslovne in jezikovnotehnološke raziskave govorjene slovenščine, v nadaljevanju prispevka opišemo zasnovo, vsebino in dostopnost obstoječe različice drevesnice SST (Dobrovoljc in Nivre 2016), v drugem delu prispevka pa predstavimo tudi njeno aktualno nadgradnjo znotraj nacionalnega projekta, v okviru katerega bo drevesnica bistveno nadgrajena z vidika obsega in raznolikosti vsebovanih besedil.

2 Zasnova in vsebina drevesnice SST

Korpus, na katerem temelji drevesnica SST, je bil izdelan za potrebe označevanja in analize diskurznofunkcijskih stalnih besednih zvez v slovenskem govoru (Dobrovoljc 2018) in obsega nekaj manj kot 30.000 besed. Zasnovan je bil kot reprezentativni vzorec takratnega referenčnega korpusa govorjene slovenščine, korpusa Gos 1.0 (Verdonik in Zwitter Vitez 2011; Zwitter et al. 2013), ki si je prizadeval za ohranjanje raznolikosti govornih dogodkov in govorcev referenčnega korpusa.

Natančneje to pomeni, da je bil iz vsakega izmed 287 govornih dogodkov oz. besedil korpusa Gos s pomočjo računalniškega programa vzorčen sorazmerni delež pojavnic glede na delež pojavnic tega besedila v referenčnem korpusu nasploh. Vzorec vsakega besedila vsebuje niz ene ali več zaporednih govornih vlog, torej neprekinjen in zaključen govor enega ali več govorcev,[2] pri čemer je bil začetek vzorčenja znotraj besedila, torej izbira prve vloge v vzorčenem nizu, določen naključno.

Kot prikazuje tabela 1, vzorčeni korpus vključuje enako število različnih govornih dogodkov in zelo podobno razmerje posameznih podzvrsti kot referenčni korpus Gos, tj. 33,5 % besedil javnega informativno-izobraževalnega diskurza, 23 % besedil javnega razvedrilnega diskurza, 28 % besedil nejavnega zasebnega diskurza in 15,5 % besedil nejavnega nezasebnega diskurza, zaradi omejenosti na kratke zaključene segmente posameznih govornih dogodkov pa vzorčeni korpus vsebuje nekoliko manjše število različnih govorcev (606) kot referenčni korpus (1.561). Posamezno besedilo vzorčenega korpusa tako v povprečju vsebuje 102 pojavnici, 11 izjav, 8 izmenjanih vlog in 2 različna govorca.

Tabela 1: Velikost in sestava obstoječe različice korpusa SST

Tip diskurza	Besedila	Govorci	Vloge	Izjave	Pojavnice
javni informativno-izobraževalni	129	263	703	959	9.899
javni razvedrilni	42	78	499	726	6.833
nejavni nezasebni	45	102	425	497	4.535
nejavni zasebni	71	163	833	1.006	8.221
SKUPAJ	**287**	**606**	**2.460**	**3.188**	**29.488**

Vir: Dobrovoljc 2018: 120

Pri izdelavi vzorca smo sledili zapisovalnim načelom izvornega korpusa, kar pomeni, da so meje vlog, izjav in besed v vzorčenem korpusu enake tistim, ki so bile ročno določene ob nastanku referenčnega korpusa Gos, prav tako pa je bil podedovan tudi sam nabor (besednih in nebesednih) pojavnic.

[2] Edina izjema v algoritmu so bila besedila javnega informativno-izobraževalnega diskurza, pri katerih se je lahko vzorec po dosegu predvidenega števila pojavnic lahko končal zgolj z zaključeno izjavo in ne nujno tudi zaključkom celotne vloge govorca. Zaradi monološke narave tovrstnega diskurza so namreč vloge posameznih govorcev pogosto zelo dolge ali obsega celotni govorni dogodek. Take nezaključene vloge sicer predstavljajo zgolj 5,5 % vseh vzorčenih vlog korpus.

3 Označevalna shema

Poleg podedovanih ročnih transkripcij in segmentacij izvornega korpusa so bili (standardiziranim) pojavnicam korpusa SST nato ročno pripisani še podatki o osnovni obliki (lemi), besedni vrsti in drugih oblikoslovnih lastnostih, ter skladenjski vlogi. Konkretno te oznake sledijo dvema označevalnima shemama: označevalni shemi MULTEXT-East za leme in oblikoskladenjske lastnosti ter shemi Universal Dependencies (UD) za oblikoslovne lastnosti in odvisnostne skladenjske relacije, pri čemer slednja v slovenskih drevesnicah posredno vsebuje tudi prvo, kot podrobneje razložimo v nadaljevanju.

3.1 Shema Universal Dependencies

Universal Dependencies je mednarodno oz. medjezično usklajena shema za slovnično označevanje besedil na oblikoslovni in skladenjski ravni, ki je bila leta 2013 vzpostavljena z namenom, da z neposredno primerljivostjo označenih korpusov za čim več svetovnih jezikov omogoči napredek na področju razvoja jezikovnih tehnologij na eni strani ter kontrastivojezikoslovnih raziskav na drugi. Znotraj sheme UD je bil tako vzpostavljen univerzalni nabor jezikoslovnih kategorij (besednih vrst, oblikoslovnih lastnosti in odvisnostnih skladenjskih relacij) in smernic za njihovo pripisovanje, ki odslej omogoča enotno označevanje podobnih slovničnih pojavov v različnih svetovnih jezikih. Shema temelji na načelih odvisnostne slovnice, njena teoretska izhodišča pa so podrobneje pojasnjena v prispevku de Marneffe et al. (2021). Do danes je bila shema UD prenesena že na več kot 245 korpusov v več kot 140 svetovnih jezikih (Zeman et al. 2023), med njimi tudi na drevesnico pisne (SSJ; Dobrovoljc et al. 2017, Dobrovoljc et al. 2023) in govorjene slovenščine (SST; Dobrovoljc in Nivre 2016).

Čeprav se shema UD večinoma uporablja za razčlenjevanje pisnih besedil, se danes vse pogosteje uporablja tudi za slovnično označevanje transkripcij govorjenega jezika. Poleg že izpostavljenih prednosti sheme, kot sta široka uveljavljenost ter dobro razmerje med mednarodno standardizacijo na eni in jezikovnospecifično fleksibilnostjo na drugi strani, je shema z vidika raziskav govorjenega jezika zanimiva zlasti zaradi svoje visoke stopnje interoperabilnosti, saj omogoča neposredne konstrastivne raziskave med drevesnicami različnih jezikov ali jezikovnih zvrsti (npr. primerjave med govornimi drevesnicami v različnih jezikih ali primerjave med drevesnicami pisnega in govorjenega jezika). Širok nabor univerzalnih slovničnih

kategorij, kot so denimo skladenjske relacije za zvalnike, diskurzne členke in samopopravke, obenem omogoča celosten, enonivojski pristop k slovnični analizi jezika, v skladu s katerim se lahko oblikoslovne oz. skladenjske oznake pripišejo vsem izgovorjenim pojavom, brez kakršnegakoli predhodnega izključevanja netekočnosti in drugih strukturnih posebnosti govora, kot je bilo to pogosto praksa v preteklosti.

Shema UD je bila za razčlenjevanje govorjenega jezika prvič preizkušena prav na slovenski drevesnici SST, odtlej pa je temu vzoru sledilo že več kot 20 drugih drevesnic govorjenega jezika po vsem svetu, ki tako kot SST vključujejo zgolj transkripcije govora, ter 40 drevesnic, ki vsebujejo tako pisna kot govorjena besedila. Ta trend potrjuje, da je shema dovolj fleksibilna, da jo je mogoče uporabiti tudi za razčlenjevanje govornih korpusov.

3.2 Nabor oznak sheme UD

Shema UD obsega 17 splošnih, 'univerzalnih' oznak za besedne vrste (npr. *ADJ* za pridevnike), 24 univerzalnih oznak za oblikoslovne lastnosti (npr. *Gender* za spol) z več kot 200 različnimi vrednostmi (npr. *Fem* za ženski spol) ter 37 odvisnostnih skladenjskih relacij (npr. *obj* za predmet), pri čemer univerzalnost ne pomeni, da se ti slovnični pojavi pojavljajo v vseh jezikih, temveč da se pojavljajo v dovolj velikem številu jezikov, da so jezikoslovno relevantni. Medtem ko je nabor besednih vrst nespremenljiv, lahko avtorji drevesnic za posamezne jezike predlagajo tudi dodatne oblikoslovne lastnosti in/ali njihove vrednosti (kot npr. *Gender[psor]* za označevanje spola svojine v slovanskih jezikih) ter izpeljave posamičnih skladenjskih relacij v obliki z dvopičjem ločenih podoznak (kot npr. *cc:preconj* za prvi del dvodelnih veznikov).

Za slovenščino so podrobnejše smernice, ki s podrobnejšimi pojasnili in številnimi primeri opisujejo prenos oznak vseh treh tipov na konkretne jezikovne pojave v slovenščini, dokumentirane tako na krovni spletni strani projekta UD[3] (v angleščini) kot v obliki samostojnega priročnika v slovenščini (Dobrovoljc in Terčon 2023a).[4] Za lažjo predstavo Tabela 2 prikazuje nabor vseh univerzalnih relacij s splošnimi

[3] https://universaldependencies.org/
[4] https://wiki.cjvt.si/books/07-universal-dependencies/page/oznacevalne-smernice

opisi v slovenščini, podrobnejše pregledne tabele za druge ravni pa so poleg omenjenih priročnikov na voljo tudi na spletišču CJVT.[5]

Tabela 2: Seznam jedrnih odvisnostnih skladenjskih relacij sheme UD

Relacija	Kratek opis
acl	stavčni prilastki
advcl	prislovni odvisniki
advmod	prislovna določila (v širšem smislu)
amod	pridevniški prilastki
appos	pristavčna določila
aux	pomožni glagoli
case	predlogi
cc	priredni vezniki
ccomp	stavčna dopolnila (predmetni odvisniki)
conj	priredno zloženi elementi
cop	vezni glagoli
csubj	osebkovi odvisniki
dep	nedoločena povezava
det	določilniki
discourse	diskurzni členki
dislocated	dislocirani elementi
expl	ekspletivne besede
fixed	funkcijske zveze
flat	eksocentrične zveze
goeswith	razdruženi deli besed
iobj	nepremi predmeti
list	seznami
mark	podredni vezniki
nmod	samostalniški prilastki
nsubj	samostalniški osebki
nummod	številčna določila
obj	premi predmeti
obl	odvisne samostalniške zveze
orphan	elementi v eliptičnih strukturah
parataxis	stavčna soredja
punct	ločila
reparandum	samopopravljanja
root	koren povedi
vocative	ogovori
xcomp	odprta stavčna dopolnila

Vir: https://wiki.cjvt.si/books/07-universal-dependencies/page/predstavitev-oznak

Smernice UD se tako osredotočajo predvsem na raven oblikoslovnega in skladenjskega označevanja, manj specifične pa so glede načel segmentiranja, tokenizacije in lematizacije, kjer splošne smernice UD podajajo zgolj nekaj okvirnih

[5] https://wiki.cjvt.si/books/07-universal-dependencies/page/predstavitev-oznak

priporočil. Kot smo že omenili, drevesnica SST sledi zapisovalnim odločitvam izvornega korpusa Gos, leme pa so ji bile pripisane v skladu s smernicami sheme MULTEXT-East (Holozan et al. 2023),[6] na katerih temeljijo tudi drugi referenčni korpusi slovenskega jezika.

Tudi nasploh je shema MULTEXT-East neposredno vključena v obe slovenski drevesnici UD, saj uradni format sheme, CONLL-U (gl. razdelek 4.2), v enem izmed stolpcev omogoča ohranjanje lokalnih oz. jezikovnospecifičnih oznak, kar smo glede na razmeroma visoko stopnjo podobnosti med obema shema na ravni oblikoslovne analize v samem procesu označevanja tudi izkoristili.

4 Označevanje in objava

4.1 Ročno pregledovanje

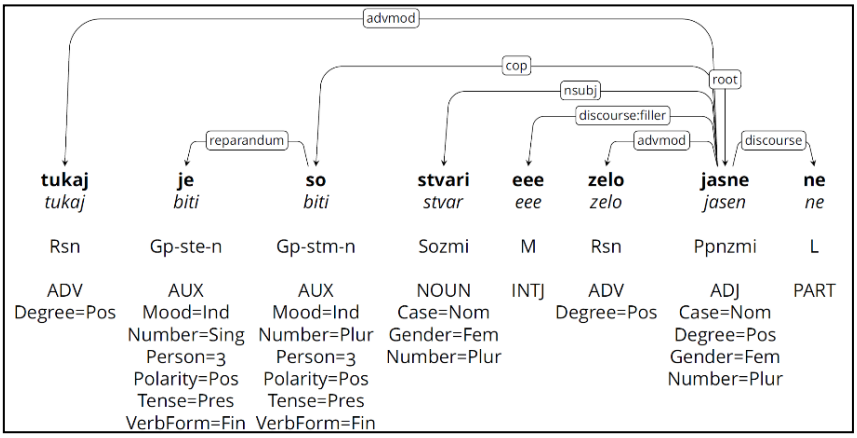
Po izdelavi vzorčenega korpusa, opisanega v 2. razdelku, so bila besedilom v prvi fazi označevanja tako najprej pripisane leme in oblikoskladenjske oznake po shemi MULTEXT-East (Erjavec 2012, Holozan et al. 2023), ki smo jih nato avtomatsko pretvorili še v besedne vrste in oblikoslovne oznake UD. Za to pretvorbo je bil namreč že ob nastanku drevesnice pisne slovenščine SSJ (Dobrovoljc et al. 2017) izdelan računalniški program,[7] ki temelji na številnih ročno zasnovanih pravilih za preslikavo med tema podobnima označevalnima sistemoma. Oblikoslovno označeni korpus je bil nato v drugi fazi strojno skladenjsko razčlenjen z orodjem MaltParser in naložen na spletno platformo WebAnno CLARIN.SI,[8] na kateri so bile vse relacije tudi ročno pregledane. Ta proces podrobneje popisuje prispevek Dobrovoljc in Nivre (2016), ki omenja tudi identificirane posebnosti govorjenega jezika, kakršne predstavimo v razdelku 5.3.

4.2 Primer razčlenjene izjave in format CONLL-U

Korpus SST torej za vsako zapisano besedo prinaša ročno pripisani podatek o osnovni obliki leme, oznako MTE ter oblikoslovne in skladenjske lastnosti (odvisnostne relacije) po shemi UD. Primer tako označene izjave v drevesnici SST je prikazan na sliki 2, na kateri oznake pod transkribiranimi besednimi oblikami

[6] https://wiki.cjvt.si/books/04-oblikoskladnja-multext-east/page/oznacevalne-smernice
[7] https://github.com/clarinsi/jos2ud
[8] https://clarin.si/webanno/

predstavljajo ročno pripisane leme (prva vrstica v poševnem), oznake MTE (druga vrstica) in oblikoslovne lastnosti po shemi UD (tretja in vse nadaljnje vrstice), puščice nad njimi pa odvisnostna razmerja med posameznimi besedami.



Slika 2: Primer označene izjave v drevesnici SST

Vir: lasten

Drevesnica SST je objavljena v standardnem formatu sheme UD, tj. tabelaričnem formatu CONLL-U, v vrsticah zapisane besede oz. pojavnice, v stolpcih pa njihove ročno pripisane lastnosti, kot prikazuje primer izjave v formatu CONLL-U na sliki 3. Konkretno format CONLL-U sestavlja 10 stolpcev:

1. ID: identifikator pojavnice
2. FORM: besedna oblika pojavnice
3. LEMMA: osnovna oblika pojavnice
4. UPOS: besedna vrsta po shemi UD
5. XPOS: oznaka po lokalni označevalni shemi (v primeru SST je to MTE)
6. FEATS: oblikoslovne lastnosti po shemi UD
7. HEAD: identifikator nadrejene pojavnice
8. DEPREL: vrsta skladenjskega razmerja do nadrejene pojavnice
9. DEPS: nadgrajeni odvisnostni graf (v primeru SST ne pripisujemo)
10. MISC: poljubna oznaka (v primeru SST se tukaj beleži podatek o obliki pojavnice v pogovornem zapisu)

Mejo med posameznimi povedmi oz. izjavami označuje ena prazna vrstica in vrstice z metapodatki, ki se začnejo z znakom #. Med slednjimi sta obvezni vrstici z unikatnim identifikatorjem povedi (# sent_id) in izpisanim golim besedilom (# text), dodajanje drugih metapodatkov pa je poljubno. Kot prikazuje primer na sliki 3, drevesnica SST trenutno vsebuje še povezavo do zvočnega posnetka izjave (# sound_url) in identifikacijsko številko govorca (# speaker_id). Tako identifikator povedi kot identifikator govorca sledita nomenklaturi izvornega korpusa Gos, v katerem je na ta način mogoče poiskati tudi podrobnejše podatke o dogodkih in govorcih, npr. da dogodek Gos119 označuje intervju v informativni TV oddaji na temo poslovanja Slovenskih železnic in da je govorec Bm-gost-07155 moškega spola.

# sent_id = Gos119.s72									
# speaker_id = Bm-gost-07155									
# sound_url = https://nl.ijs.si/project/gos20/Gos119/Gos119.s72.mp3									
1	tukaj	tukaj	ADV	Rgp	Degree=Pos	8	advmod	_	pron=tuki
2	je	biti	VERB	Va-r3s-n	Mood=Ind…	4	reparandum	_	pron=je
3	so	biti	AUX	Va-r3p-n	Mood=Ind...	8	cop	_	pron=so
4	stvari	stvar	NOUN	Ncfpn	Case=Nom…	8	nsubj	_	pron=stvari
5	eee	eee	INTJ	I	_	8	discourse:filler	_	pron=eee
6	zelo	zelo	ADV	Rgp	Degree=Pos	8	advmod	_	pron=zelo
7	jasne	jasen	ADJ	Agpfpn	Case=Nom…	0	root	_	pron=jasne
8	ne	ne	PART	Q	Polarity=Neg	8	discourse	_	pron=ne
9	tukaj	tukaj	ADV	Rgp	Degree=Pos	8	advmod	_	pron=tuki

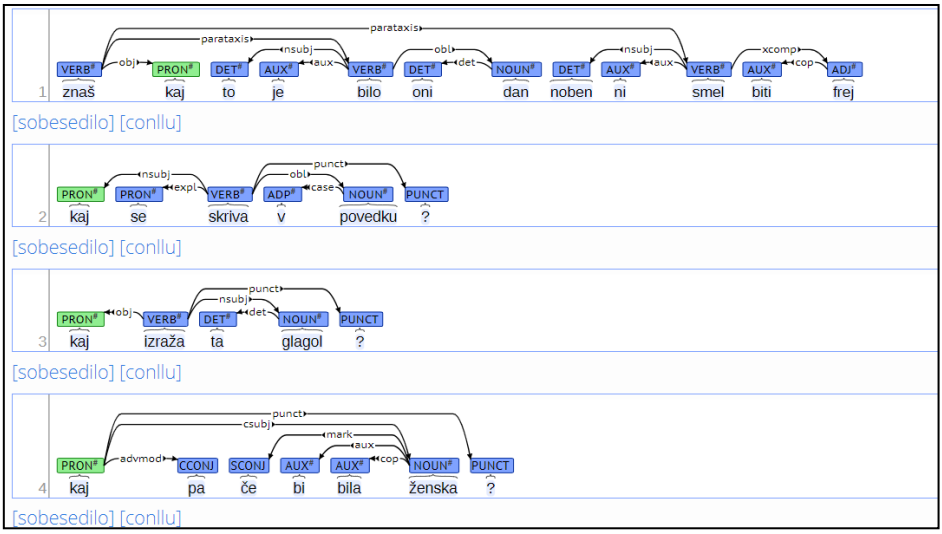
Slika 3: Primer zapisa označene izjave v formatu CONLL-U[9]
Vir: lasten

4.3 Dostopnost

Drevesnica SST je bila kot odprtodostopna podatkovna množica prvič objavljena leta 2015 pod enako licenco kot transkripcije izvornega korpusa (CC BY-NC-SA), in sicer kot del različice 1.3 uradne zbirke drevesnic UD. Drevesnice UD so namreč distribuirane kot del enotne, skupne korpusne zbirke, ki izhaja dvakrat letno (maja in novembra) in na ta način omogoča kontinuirano priključevanje novih drevesnic in redne izboljšave že obstoječih. V okviru teh polletnih posodobitev se je od prve objave redno izboljševala tudi vsebina drevesnice SST, denimo z odpravljanjem posamičnih napak, posodabljanjem glede na spremenjene označevalne smernice ali

[9] Zaradi omejenosti s prostorom je v šestem Stolpcu (FEATS) navedena zgolj prva oblikoslovna lastnost v nizu, v zadnjem stolpcu (MISC) pa je atribut 'pronunciation' okrajšan na 'pron'.

z dodajanjem novih metapodatkov. Zadnja različica drevesnice SST je tako izšla kot del zbirke UD v2.12 (Zeman et al. 2023).



Slika 4: Primer rezultatov iskanja po drevesnici SST na portalu Drevesnik
Vir: lasten

5 Nadaljnji razvoj

Da bi zagotovili še kvalitetnejšo gradivno osnovo za nadaljnje raziskave, je kot eden izmed ciljev nacionalnega projekta *Na drevesnici temelječ pristop k raziskavam govorjene slovenščine* (SPOT),[10] ki se v letih 2022–2024 izvaja na Filozofski fakulteti v Ljubljani, v teku temeljita nadgradnja drevesnice SST, ki jo skupaj z drugimi smernicami nadaljnjega razvoja predstavimo v nadaljevanju.

5.1 Povečanje korpusa

Obstoječa različica drevesnice SST obsega 30.000 pojavnic, kar jo umešča v zgornjo polovico drugih sorodnih govornih drevesnic za druge jezike, a obenem to predstavlja zgolj eno devetino drevesnice pisne slovenščine SSJ, ki trenutno obsega nekaj več kot 267.000 pojavnic oz. 13.000 razčlenjenih povedi. Poleg že omenjene potrebe po vzpostavitvi statistično močnejših empiričnih temeljev za bodoče

[10] https://spot.ff.uni-lj.si/

kvalitativne in kvantitativne raziskave te podatkovne množice je razširitev smiselna tudi zaradi dveh drugih aktualnih pomanjkljivosti.

Prva izvira iz dejstva, da je bila pred kratkim kot rezultat projekta RSDO objavljena nova, skoraj enkrat obsežnejša različica referenčnega korpusa govorjene slovenščine, korpus Gos 2 (Verdonik et al. 2023), ki v primerjavi s prvotno različico vsebuje nekoliko drugačno strukturo besedil, saj je bil korpus razširjen z novimi tipi govornih dogodkov, kot so posnetki znanstvenih srečanj, javnih dogodkov in parlamentarnih sej.

Druga pomembna omejitev obstoječe drevesnice SST pa je dejstvo, da je z besedilnega vidika zelo fragmentirana, saj glede na zasnovo vzorčenja, opisanega v 2. razdelku, vsebuje precej kratke segmente zelo širokega nabora govornih dogodkov. To ima seveda številne prednosti za raziskave, pri katerih sta pomembni raznolikost govorcev in govornih situacij, kot sta na primer leksikologija in dialektologija, omejuje pa uporabo v raziskavah jezikovnih pojavov nad nivojem izjave, kot so denimo raziskave strukturiranja diskurza in pragmatike.

5.1.1 Vzorčenje novih besedil

Ob upoštevanju vseh treh naštetih dejavnikov, tj. potrebe po povečanju obsega referenčnega slovnično označenega korpusa govorjene slovenščine, zagotavljanju njegove reprezentativnosti glede na Gos 2 in izboljšanju njegove uporabnosti za diskurznoanalitične raziskave, je bil v sodelovanju s projektom Mezzanine, v okviru katerega je prav tako predvidena izdelava ročno označenih korpusov za dialoška dejanja, netekočnosti in prozodično segmentacijo, pred kratkim izdelan vzorec korpusa Gos 2.0 v napovedanem obsegu 50.000 pojavnic, ki bo služil kot gradivna osnova za vse navedene kampanje.

Ker opis postopka vzorčenja presega namen in obseg tega prispevka, na tem mestu zgolj povzamemo, da je vzorčenje potekalo na podlagi ročnega predizbora specifičnih govornih dogodkov iz korpusa Gos 2 v dveh korakih. V prvem so bili povečani oz. podaljšani vzorci izbranih 22 dogodkov v obstoječi drevesnici SST (pribl. 450 novih besed na dogodek oz. skupno pribl. 10.000 pojavnic z delovnim imenom SPOG), v drugem pa so bili izdelani vzorci 57 povsem novih govornih dogodkov iz baze Artur (pribl. 800 novih besed na dogodek oz. skupno pribl. 40.000 besednih pojavnic z delovnim imenom IRISS).

5.1.2 Sestava nove različice drevesnice SST

Kot prikazuje povzetek vsebine vseh treh podkorpusov razširjene različice drevesnice SST (tj. prvotne različice drevesnice SST ter njene razširitve s podkorpusoma SPOG in IRISS), bo nova različica drevesnice bistveno večja (+199 % pojavnic), vsebovala bo še bolj raznolik nabor govorcev (+13 %) in dogodkov (+20 %), vzorčeni segmenti dogodkov pa bodo v povprečju tudi daljši. V primerjavi s prvotno različico, v kateri posamezni izsek povprečno obsega 103 besede, 11 izjav in 9 izmenjanih vlog, izseki v novi različici namreč v povprečju obsegajo 257 besed, 19 izjav oz. 12 izmenjanih vlog.

Tabela 3: Velikost in sestava nove, razširjene različice drevesnice SST (v označevanju)

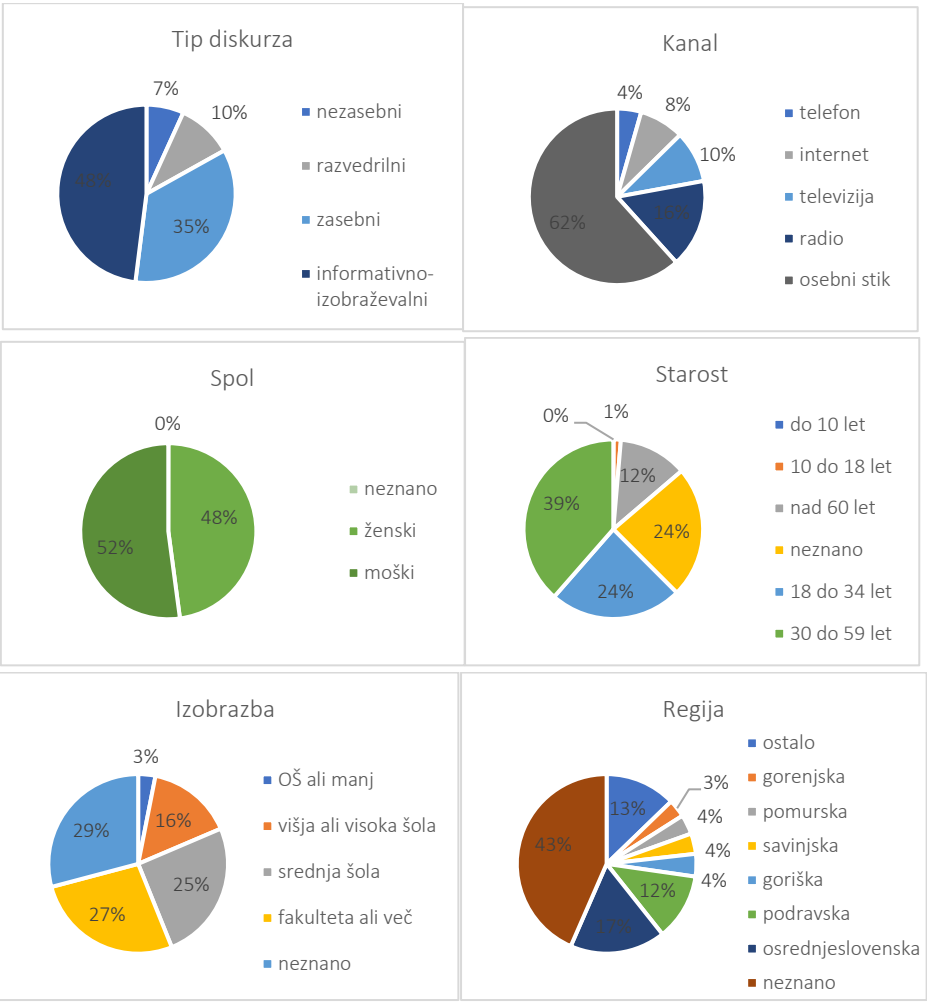
Tip diskurza	Besedila	Govorci	Vloge	Izjave	Pojavnice[11]
SST-izvorna	287	606	2.460	3.188	29.488
SPOG	22	63	1.224	1.374	10.184
IRISS	57	72	612	1.968[12]	48.624
SST-novi	**344**	**687**	**4.296**	**6.530**	**88.296**

Vir: lasten

Kot prikazujejo grafi na sliki 3, pa razlike v načinu vzorčenja prvotne različice drevesnice (tj. manjše število pojavnic velikega nabora dogodkov) in načinu vzorčenja njene razširitve (tj. večje število pojavnic manjšega nabora dogodkov) ne vplivajo na samo reprezentativnost nove različice drevesnice SST, saj tudi ta vsebuje raznolik in uravnotežen nabor govornih situacij ter demografskih lastnosti govorcev.

[11] Štetje pojavnic v tabeli 1 je bilo izvedeno na podlagi korpusov v formatu CONLL-U, v katerih so v nasprotju z izvornim zapisom korpusa v XML trenutno kot pojavnice obravnavana tudi anonimizirana imena (npr. *[ime]*, *[priimek]*), in oznake za premore in nerazumljivi govor, ki se v transkripcijah pojavljajo tudi kot samostojne izjave. Štetje pojavnic v korpusu IRISS vključuje tudi ločila (ki jih v drugih dveh korpusih ni). Morebitno poenotenje podkorpusov z vidika vključevanja ločil, nebesednih pojavnic in uporabe velikih začetnic je predmet širše diskusije o poenotenju podkorpusov Gos 2 (Verdonik et al. 2022) in bo implementirano ob koncu označevalne kampanje.

[12] Štetje izjav korpusa IRISS temelji na resegmentirani različici, v kateri so bile meje med izjavami postavljene na podlagi končnih ločil, kot so pika, vejica in vprašaj. Izvorna segmentacija, podedovana iz baze Artur, je bila namreč bistveno bolj fragmentirana in ni upoštevala skladenjsko-pomenske zaključenosti segmentov. Tako je bil na primer niz štirih izjav v bazi Artur oz. Gos 2 (1) /*Drage prijateljice, dragi prijatelji*/, (2) /*govorjene slovenščine.*/, (3) /*razmišljal sem, kako naj začnem ta*/ in (4) /*svoj nastop*/ avtomatsko resegmentiran v niz dveh izjav: (1) /*Drage prijateljice, dragi prijatelji govorjene slovenščine.*/ in (2) /*razmišljal sem, kako naj začnem ta svoj nastop.*/



n = 88.296

Slika 5: Sestava nove različice drevesnice SST (v označevanju) z vidika deleža pojavnic glede na tip govornega dogodka, komunikacijski kanal in demografske lastnosti govorcev[13]

Vir: lasten

[13] Graf regionalne pripadnosti govorcev prikazuje delež govorcev glede na statistično regijo prebivališča, pri čemer 'ostalo' označuje regije z manjšo zastopanostjo, tj. jugovzhodna Slovenija (3,0 %), koroška (2,5 %), posavska (2,2 %), obalno-kraška (2,1 %), primorsko-notranjska (1,2 %), tujina (0,8 %), Italija (0,4 %), Avstrija (0,3 %) in Madžarska (0,3 %). Pri zbiranju posnetkov za korpus Gos 1, iz katerega izhajajo besedila korpusa SST in SPOG, je bilo mogoče navesti več regionalnih pripadnosti. V primeru tovrstnih govorcev (pribl. 10 % vseh pojavnic) graf prikazuje zgolj prvo navedeno regijo.

5.2 Sistematična označevalna kampanja

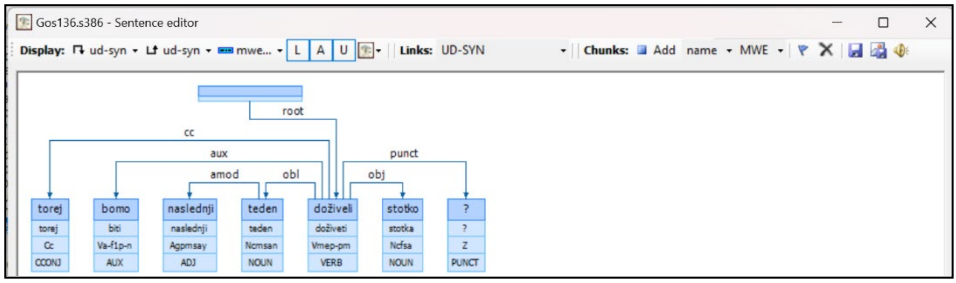
Na podlagi zgoraj opisanih novih korpusnih podatkov (tj. korpusov SPOG in IRISS) že poteka ročno pripisovanje slovničnih oznak v obliki dveh vzporednih označevalnih kampanj. V prvi se besedilom pripisujejo leme in oblikoskladenjske oznake po shemi MTE, ki bodo nato po že preizkušenem pretvorbenem postopku strojno preslikane v besedne vrste in oblikoslovne oznake sheme UD. Konkretno označevalci pregledujejo leme in oznake, ki so bile besedilom strojno pripisane z označevalnikom CLASSLA-Stanza,[14] pri čemer se pregledovanje po izkušnjah nedavne sorodne kampanje na podatkih pisne slovenščine (Pori et al. 2022) osredotoča zgolj na besedne oblike, pri katerih je glede na oblikoslovni leksikon Sloleks (Čibej et al. 2023) možnih več različnih interpretacij (npr. *škarje* kot samostalnik v imenovalniku ali tožilniku), ne pa oblikoslovno nedvoumne oblike (npr. *srajca* kot samostalnik v imenovalniku).

Vzporedno s to aktivnostjo trenutno poteka tudi kampanja skladenjskega razčlenjevanja po shemi UD, v kateri označevalci pregledujejo in popravljajo odvisnostne relacije, ki so bile tem besedilom pripisane s strojnim označevalnikom Trankit,[15] ki se je na skriti tesni množici platforme SloBench izkazal kot najuspešnejši pri nalogi skladenjskega razčlenjevanja.[16] Konkretno sta bila korpusa SPOG oz. IRISS razčlenjena z lokalno razvitim modelom (Krsnik in Dobrovoljc 2024), ki je bil glede na že znane prednosti združevanja različnih vrst učnih podatkov pri razčlenjevanju govorjene slovenščine (Dobrovoljc in Martinc 2018) naučen na kombinaciji obeh referenčnih drevesnic UD za slovenščino, drevesnice SSJ in SST.

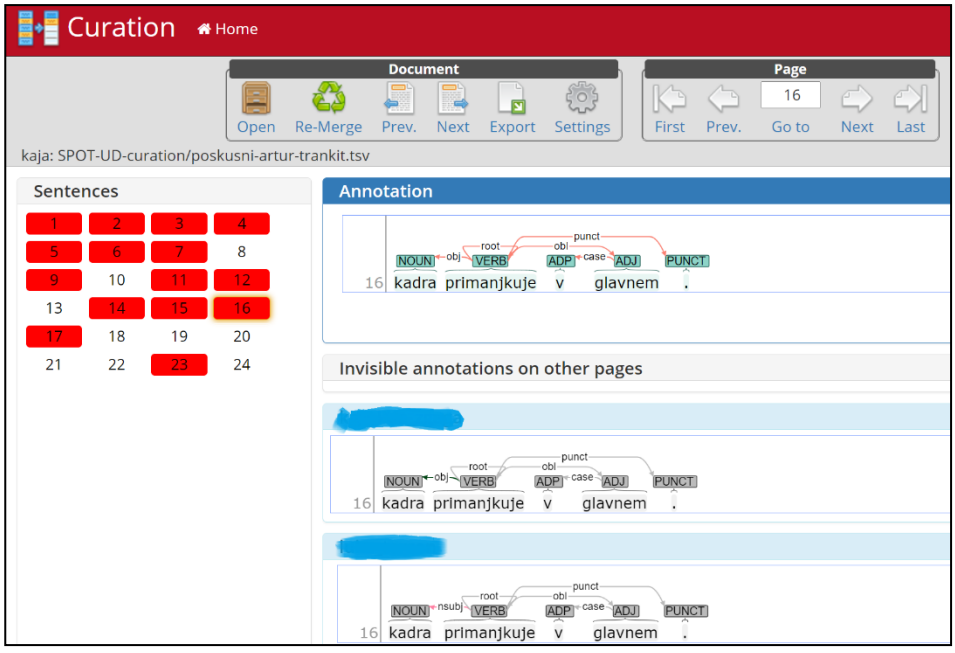
V nasprotju s prvotno drevesnico SST, ki je bila označena s strani ene same označevalke, bodo v tokratni kampanji besedila pregledali 2–3 neodvisni označevalci, kar po eni strani zagotavlja večjo zanesljivost pripisanih relacij, po drugi strani pa tudi lažjo identifikacijo težavnejših mest tovrstne jezikoslovne analize, ki opozarjajo na pomanjkljivosti izhodiščnih jezikoslovnih kategorij oz. smernic za njihovo pripisovanje. Označevanje poteka v orodju Q-CAT (Brank 2023), ki je bilo za potrebe te specifične kampanje nadgrajeno tudi z možnostjo hkratnega poslušanja zvočnih posnetkov izjav za korpuse v formatu CONLL-U, če so povezave do

[14] https://pypi.org/project/classla/
[15] https://github.com/nlp-uoregon/trankit
[16] Primerjava orodij Trankit, Stanza in CLASSLA-Stanza na skriti tesni množici platforme SloBench: https://slobench.cjvt.si/leaderboard/view/11.

posnetkov podane v vrsticah z atributom # sound_url (slika 6). Ker Q-CAT ne podpira primerjave oznak različnih označevalcev oz. njihovega poenotenja v obliki končnih odločitev, za ta korak (t. i. kuriranje) uporabljamo spletno označevalno platformo WebAnno (slika 7).



Slika 6: Razčlenjevanje izjav v orodju Q-CAT (gumb za poslušanje posnetka je v desnem zgornjem kotu)
Vir: lasten



Slika 7: Primerjava oznak dveh označevalcev na platformi WebAnno
Vir: lasten

5.3 Nadgradnja označevalnih smernic s posebnostmi govora

Kot smo omenili že v 3. razdelku, so splošne, jezikovno neodvisne smernice sheme UD objavljene na spletni strani projekta, podrobnejše smernice s pojasnili in primeri prenosa sheme na slovenske podatke pa so bile popisane v obliki priročnika (Dobrovoljc in Terčon 2023a), ki se je v svoji prvi različici osredotočal predvsem na izčrpen opis razčlenjevanja besedil pisne slovenščine, kakršna se pojavljajo v učnem korpusu pisne slovenščine SUK/SSJ.

Za potrebe razčlenjevanja transkripcij govorjene slovenščine smo nedavno ta priročnik dopolnili še z opisi označevanja skladenjskih posebnosti govorjenega jezika, ki služijo kot izhodišče za razčlenjevanje v okviru zgoraj opisane kampanje (Dobrovoljc in Terčon 2023b). Te nadgrajene smernice poleg prenosa (zelo ohlapnih) splošnih smernic UD, ki govor omenjajo zgolj pri najbolj izstopajočih skladenjskih pojavih, kot so samopopravljanja in diskurzni členki, popisujejo doslej nedokumentirano podrobnejšo obravnavo teh in številnih drugih pojavov v okviru označevanja prvotne drevesnice SST (Dobrovoljc in Nivre 2016), obenem pa upoštevajo tudi priporočila, ki so se oblikovala v poznejših diskusijah (Kahane et al. 2021, Dobrovoljc 2022) in skoti primere dobre prakse drugih sorodnih drevesnic govorjenega jezika.

Na oblikoslovni ravni so bili tako dopolnjeni ponazoritveni seznami nepregibnih in zaprtih besednih vrst (npr. prislovi *plus, kao, komot, direkt, tukajle, ene, prvo*; priredni vezniki *aber, ar*; podredni veznik *ka;* določilnik *ovi*) ter številni medmeti in členki, kot so *eee, eem, živjo, porkaš, vav, opala*, *alora, arki, evo, tipo*. Popisane so bile tudi odločitve glede dogovorne besednovrstne kategorizacije nedokončanih besed, nebesednih pojavnic in anonimiziranih imen.

Bistveno obsežnejše pa so dopolnitve na skladenjski ravni, kjer smernice prinašajo podrobnejše opise in ponazoritve relacije *discourse*, s katero se označujejo medmeti, diskurzni označevalci in druge oblike ustaljenih, skladenjsko manj vpetih izrazov, kot so *oh, ja, (a) ne, tako, hvala, škoda* in (s podoznako *discourse:filler*) tudi zapolnjeni premori tipa *eee*. Prav tako je bilo obsežno dopolnjeno poglavje s predstavitvijo relacije *reparandum*, s katero se označujejo samopopravljanja različnih tipov, od popravljanja napačno začetih besed (npr. *kako orož- orožje pa to*) ali napačnih besednih oblik (npr. *da so te eee ti stroški čim manjši*) do ponavljanj in popravljanj znotraj stavka (npr. *nekega dne sem se eee sem se skregal*).

Med drugimi izpostavljenimi posebnostmi govora smernice denimo naslavljajo tudi označevanje navideznih odvisnikov znotraj relacije *advcl* (npr. *če smem vprašati, kot rečeno*), nadaljevalnikov znotraj relacije *conj* (npr. *in tako naprej, ali pa kaj takega*), ekspletivne vloge kazalnega zaimka *to* znotraj relacije *expl* (npr. *tako da to pol nekega posebnega izobraževanja pa verjetno ni ne*), izpustov povedka znotraj relacije *orphan* (npr. *pri nas pa občasno, kam pa?, tudi Francozinja v težavah*), ponovne začetke izjav znotraj relacije *parataxis* (npr. *kaj si zdaj pravkar katero črto boš narisala*) in obravnavo netipičnega besednega reda (npr. *imamo pa tudi debelo ono uro jekleno*), če jih naštejemo le nekaj.

Nenazadnje smernice prinašajo tudi razširjeni seznam stalnih besednih zvez, pri katerih se notranja struktura členi z relacijo *fixed* (npr. *a la, a ne, hvala bogu, ker da, se pravi da*), ter znotraj relacije *flat* opisujejo skladenjsko členjenje zapisov besednih zvez, kakršne bi v izvorno pisnih besedilih pričakovali zapisane drugače, npr. izgovorjenih decimalnih števil (*dva cela pet*) ali naslovov spletnih strani (npr. *trikrat dvojni v pika radio capris pika si poševnica kikiriki*).

Kot smo že omenili v razdelku 5.2, nameravamo smernice pred uradno objavo dodatno dopolniti še na podlagi analize najpogostejših nestrinjanj med označevalci, ki utegnejo opozoriti na jezikoslovna vprašanja, ki so bila v smernicah ali literaturi nasploh naslovljena pomanjkljivo, kot je denimo določanje nadrejenega elementa diskurznih členkov in drugih 'šibko' vpetih struktur. Z namenom zagotavljanja dosledne in izčrpno dokumentirane označenosti nove, razširjene drevesnice SST bodo morebitne novosti ob koncu prenesene tudi na besedila prvotne različice drevesnice SST.

6 Zaključek

V prispevku smo predstavili zasnovo, vsebino, dostopnost in aktualno nadgradnjo drevesnice SST, odprto dostopnega skladenjsko razčlenjenega korpusa govorjene slovenščine, v katerem so vsaki transkribirani pojavnici ročno pripisane informacije o besednih vrsti, oblikoslovnih lastnostih in odvisnostnih skladenjskih relacijah po mednarodno uveljavljenih shemah MULTEXT-East in Universal Dependencies.

Kot taka drevesnica predstavlja pomembno podatkovno množico za nadaljnji razvoj in evalvacijo tehnologij za obdelavo slovenskega govora, kot so na primer na govor prilagojeni slovnični označevalniki, ter izjemno dragoceni gradivni vir za kvalitativne

in kvantitativne jezikoslovne analize slovničnih značilnosti govorjene slovenščine, tudi preko primerjave z drugimi sorodnimi drevesnicami pisnih in govorjenih besedil.

Za polni izkoristek tega potenciala v raziskavah slovenskega govora je seveda poleg nadaljnjega razvoja drevesnice z vidika obsega, zanesljivosti oznak in transparentne dokumentiranosti, ki smo ga nakazali v tem prispevku, smiselno tudi njeno kontinuirano dopolnjevanje z drugimi ravnmi jezikoslovnega opisa, ki bi omogočale celostne slovnične analize govora s hkratnim upoštevanjem prozodičnih, slovničnih in pragmatičnih vidikov govornega sporazumevanja.

Literatura

Špela ANTLOGA, 2022: Identifikacija metafore in metonimije v jezikovnih korpusih: Poskus kategorizacije označenih metonimičnih prenosov v korpusu g-KOMET. *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave*, 11/1, 91–117. https://doi.org/10.4312/slo2.0.2023.1.91-117.

Špela ARHAR HOLDT et al., 2022: Training corpus SUK 1.0, *Slovenian language resource repository CLARIN.SI*, http://hdl.handle.net/11356/1747.

Janez BRANK, 2023: Q-CAT Corpus Annotation Tool 1.5, *Slovenian language resource repository CLARIN.SI*, http://hdl.handle.net/11356/1844.

Brian MacWHINNEY, 2000: The CHILDES Project: Tools for Analyzing Talk, 3. izdaja. Psychology Press.

Marie-Catherine de MARNEFFE, Christopher D. MANNING, Joakim NIVRE, Daniel ZEMAN, 2021: Universal Dependencies. *Computational Linguistics*, 47/2, 255–308.

Jaka ČIBEJ et al., 2022: Morphological lexicon Sloleks 3.0, *Slovenian language resource repository CLARIN.SI*, http://hdl.handle.net/11356/1745.

Kaja DOBROVOLJC, 2018: *Leksikalne prvine govorjenega jezika v uporabniških spletnih vsebinah: primer večbesednih diskurznih označevalcev*. Doktorska disertacija. Ljubljana: Filozofska fakulteta UL.

Kaja DOBROVOLJC, 2022: Spoken Language Treebanks in Universal Dependencies: an Overview. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. 1798–1806.

Kaja DOBROVOLJC, Tomaž ERJAVEC, Simon KREK, 2017: The Universal Dependencies Treebank for Slovenian. *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing, BSNLP@EACL 2017*. 33–38.

Kaja DOBROVOLJC, Joakim NIVRE, 2016: The Universal Dependencies Treebank of Spoken Slovenian. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. 1566–1573.

Kaja DOBROVOLJC, Matej MARTINC, 2019: Er ... Well, it matters, right? On the Role of Data Representations in Spoken Language Dependency Parsing. *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*. 37–46.

Kaja DOBROVOLJC, Luka TERČON, 2023a: *Universal Dependencies: Smernice za označevanje besedil v slovenščini. Različica 1.0*.

Kaja DOBROVOLJC, Luka TERČON, 2023b: *Universal Dependencies: Smernice za označevanje besedil v slovenščini. Različica 1.3*. Ljubljana: Center za jezikovne vire in tehnologije Univerze v Ljubljani.

Kaja DOBROVOLJC, Luka TERČON, Nikola LJUBEŠIĆ, 2023: Universal Dependencies za slovenščino: nove smernice, ročno označeni podatki in razčlenjevalni model. *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave,* 11/1, 218-246.

Sašo DŽEROSKI, Tomaž ERJAVEC, Nina LEDINEK, Petr PAJAS, Zdenek ŽABOKRTSKY, Andreja ŽELE, 2006: Towards a Slovene Dependency Treebank. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. 1388–1391

Tomaž ERJAVEC, 2012: MULTEXT-East: morphosyntactic resources for Central and Eastern European languages. *Language Resources and Evaluation* 46, 131–142. https://doi.org/10.1007/s10579-011-9174-8

Tomaž ERJAVEC, Darja FIŠER, Simon KREK, Nina LEDINEK, 2010: The JOS Linguistically Tagged Corpus of Slovene. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*. 1806–1809.

John J. GODFREY, Edward C. HOLLIMAN, Jane McDANIEL, 1992: SWITCHBOARD: Telephone speech corpus for research and development. *Acoustics, Speech, and Signal Processing, IEEE International Conference*. 517–520.

Erhard W. HINRICHS, Julia BARTELS, Yasuhiro KAWATA, Valia KORDONI, Heike TELLJOHANN, 2000: The Tubingen treebanks for spoken German, English, and Japanese. *Verbmobil: Foundations of Speech-to-Speech Translation*. Ur. Wolfgang Wahlster. Springer Berlin Heidelberg. 550–574.

Peter HOLOZAN et al., 2023: *Specifikacije za učni korpus: lematizacija in MSD. Različica 2.0*.

Nany IDE, James PUSTEJOVSKY, 2017: *Handbook of linguistic annotation*. Berlin: Springer.

Sylvain KAHANE, Bernard CARON, Emmett STRICKLAND, Kim GERDES, 2021: Annotation guidelines of UD and SUD treebanks for spoken corpora: A proposal. *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, Syntaxfest 2021)*. 35–47.

Andre KÅSEN, Kristin HAGEN, Anders NØKLESTAD, Joel PRIESTLY, Per Erik SOLBERG, Dag Trygve Truslew HAUG, 2022: The Norwegian Dialect Corpus Treebank. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. 4827–4832.

Simon KREK et al., 2020: The ssj500k Training Corpus for Slovene Language Processing. *Zbornik Konference Jezikovne tehnologije in digitalna humanistika 2020*. 24-33.

Luka KRSNIK, Kaja DOBROVOLJC, 2024: Trankit model for linguistic processing of spoken Slovenian, *Slovenian language resource repository CLARIN.SI*, ISSN 2820-4042, http://hdl.handle.net/11356/1909.

Sandra KÜBLER, Ryan MCDONALD, Joakim NIVRE, 2009: *Dependency Parsing*. Morgan and Claypool Publishers.

Anne LACHERET-DUJOUR, Sylvain KAHANE, Paola PIETRANDREA, 2019: *Rhapsodie: A Prosodic and Syntactic Treebank for Spoken French*. John Benjamins Publishing Company.

Nina LEDINEK, 2014: *Slovenska skladnja v oblikoskladenjsko in skladenjsko označenih korpusih slovenščine*. Ljubljana: Založba ZRC.

Igor A. MEL'ČUK, 1988: *Dependency Syntax: Theory and Practice*. State University Press of New York.

Lilja ØVRELID, Andre KÅSEN, Kristin HAGEN, Anders NØKLESTAD, Per Erik SOLBERG, Janne Bondi JOHANNESSEN, 2018: The LIA Treebank of Spoken Norwegian Dialects. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. 4482–4488.

Eva PORI, Jaka ČIBEJ, Tina MUNDA, Luka TERČON and Špela ARHAR HOLDT, 2022: Lematizacija in oblikoskladenjsko označevanje korpusa SentiCoref. *Zbornik konference Jezikovne tehnologije in digitalna humanistika 2022*. Ur. Darja Fišer, Tomaž Erjavec. Ljubljana: Inštitut za novejšo zgodovino.

Miha ŠTRAVS, Kaja DOBROVOLJC, 2022: Service for querying dependency treebanks Drevesnik 1.0, *Slovenian language resource repository CLARIN.SI*, http://hdl.handle.net/11356/1715.

Lucien TESNIÈRE, 1959: *Éléments de Syntaxe Structurale*. Paris: Klincksieck.

Ton VAN DER WOUDEN et al., 2002: Harvesting Dutch trees: Syntactic properties of spoken Dutch. In *Computational Linguistics in the Netherlands: Selected Papers from the Thirteenth CLIN Meeting*. Ur. Tanja Gaustad. Brill. 129-141.

Darinka VERDONIK, 2020: Dialogue act annotated spoken corpus GORDAN 1.0 (transcription), *Slovenian language resource repository CLARIN.SI*, http://hdl.handle.net/11356/1291

Darinka VERDONIK, Ana ZWITTER VITEZ, 2011: *Slovenski govorni korpus GOS*. Ljubljana: Trojina, zavod za uporabno slovenistiko.

Darinka VERDONIK, Andreja BIZJAK, Andrej ŽGANK, Simon DOBRIŠEK, 2022: Metapodatki o posnetkih in govorcih v govornih virih: primer baze Artur. *Zbornik konference Jezikovne tehnologije in digitalna humanistika 2022*. Ur. Darja Fišer, Tomaž Erjavec. Ljubljana: Inštitut za novejšo zgodovino.

Darinka VERDONIK et al., 2023: Spoken corpus Gos 2.1 (transcriptions), *Slovenian language resource repository CLARIN.SI*, http://hdl.handle.net/11356/1863.

Daniel ZEMAN et al., 2023: Universal Dependencies 2.12, *LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University*, http://hdl.handle.net/11234/1-5150.

Ana ZWITTER VITEZ et al., 2013: Spoken corpus Gos 1.0, *Slovenian language resource repository CLARIN.SI*, http://hdl.handle.net/11356/1040.