Kaja Dobrovoljc
Univerza v Ljubljani
Filozofska fakulteta in
Institut »Jožef Stefan«

DOI: 10.4312/jis.69.4.187-209
1.01

UPORABA DREVESNICE SST V RAZISKAVAH GOVORJENE SLOVENŠČINE: PREDNOSTI IN OMEJITVE

Kljub porastu jezikoslovnih raziskav govorjene slovenščine, ki si prizadevajo za popis številnih doslej prezrtih posebnosti govorjenega jezika v primerjavi s pisnim, metodologija tovrstnih razprav večinoma temelji na kvalitativni analizi razmeroma majhnih ter zvrstno ali demografsko omejenih vzorcev jezikovne rabe, kar omejuje ponovljivost raziskav in možnost posploševanja spoznanj na govorjeno slovenščino kot celoto. Kot eno izmed možnosti za premostitev tega problema v prispevku predstavljamo drevesnico govorjene slovenščine SST (angl. *Spoken Slovenian Treebank*), prostodostopni oblikoslovno in skladenjsko označeni reprezentativni vzorec referenčnega korpusa govorjene slovenščine Gos, in ponazarjamo njen metodološki potencial za nadaljnje korpusne raziskave govorjene slovenščine. Na primeru treh tipično govorjenih pojavov (samopopravljanja, diskurzni členki in dodani ujemalni pridevniški prilastki) prikažemo uporabo drevesnice SST za enostaven priklic številnih avtentičnih primerov rabe, na primeru analize pogostosti samopopravljanj glede na različne sporazumevalne okoliščine pa ponazorimo tudi njeno uporabnost za raznolike statistične analize jezikovne rabe. Poleg najpomembnejših prednosti drevesnice SST, kot so uravnoteženost, odprta dostopnost, ročna slovnična označenost in neposredna primerljivost z drugimi tovrstnimi korpusi po svetu, v sklepnem delu izpostavimo tudi nekaj omejitev, kot sta razmeroma majhna velikost ter robustna, v pisni jezik usmerjena označevalna shema.

Ključne besede: korpusno jezikoslovje, govorjena slovenščina, jezikoslovno označevanje, odvisnostna drevesnica

Using the SST Treebank in Research on Spoken Slovenian: Advantages and Limitations

Despite the increase in linguistic research on spoken Slovenian, which strives to catalogue the many previously overlooked characteristics of the spoken language compared to the written form, the methodology of such discussions largely relies on the qualitative analysis of relatively small and demographically or genre-limited samples of language use, which limits the replicability of research and the ability to generalize findings to spoken Slovenian as a whole. To address this issue, this paper introduces the Spoken Slovene Treebank (SST), a freely accessible, morphologically and syntactically annotated

representative sample of the Gos spoken Slovene reference corpus, and illustrates its methodological potential for future corpus-based research of spoken Slovene. By examining three common spoken phenomena – self-repairs, discourse markers, and post-modifying adjectives – we showcase the SST Treebank's capability for straightforward retrieval of numerous authentic examples. Furthermore, by analysing the distribution of self-repairs across various communicative settings, we highlight its utility for diverse statistical analyses of language practices. In addition to highlighting the SST Treebank's major advantages, such as its balanced composition, open access, manual grammatical annotations, and direct comparability with other similar corpora worldwide, we also address some limitations in the concluding section, notably its relatively small size and the robust, written-language-oriented annotation scheme.

Keywords: corpus linguistics, spoken Slovenian, linguistic annotation, dependency treebank

1 Uvod

Pod vplivom funkcijskih jezikoslovnih teorij, ki v ospredje svojega zanimanja postavljajo analizo jezikovne rabe v najrazličnejših sporazumevalnih situacijah, smo v zadnjih treh desetletjih priča skokovitemu porastu raziskav govorjenega jezika (Leech 2000; Sinclair in Maurenen 2006; Carter in McCarthy 2017). To velja tudi za slovenski jezikoslovni prostor, v katerem se razprave o specifikah govorjenega jezika pojavljajo na različnih jezikoslovnih področjih: od slovničnih razprav (Smolej 2004; Krajnc Ivič 2004; Valh Lopert 2006; Marušič in Žaucer 2007; Zwitter Vitez 2018; Smolej 2022) in narečjeslovnih opisov (Zuljan Kumar 2007, 2019, 2022) do raziskav na področjih pragmatičnega (Verdonik 2007; Schlamberger Brezar 2007; Dobrovoljc 2018a), kognitivnega (Kranjc 1999, 2006) in računalniškega jezikoslovja (Dobrovoljc in Nivre 2016). V njih raziskovalci opozarjajo na številne posebnosti spontano govorjenega diskurza v primerjavi s pisnim, kot so oklevanja, samopopravljanja, elipse, vrivki in diskurznofunkcijska leksika, če naštejemo le nekaj najpogosteje obravnavanih (za pregled gl. Dobrovoljc, v pripravi).

Kljub aktualnosti in raznolikosti raziskav govorjene slovenščine pa njihova metodologija večinoma temelji na kvalitativni analizi razmeroma majhnih, zvrstno ali demografsko omejenih vzorcev jezikovne rabe, ki praviloma tudi niso prosto dostopni, kar omejuje ponovljivost raziskav ter možnost posploševanja dognanj na govorjeno slovenščino kot celoto. V poldrugem desetletju, kolikor mineva od prelomnega korpusnojezikoslovnega prispevka J. Zemljarič Miklavčič (2008), je bilo za premostitev tega problema izdelanih več uravnoteženih prosto dostopnih korpusov govorjene slovenščine. Referenčni korpus Gos (Verdonik in Zwitter Vitez 2011; Verdonik idr. 2024) vsebuje ročne prepise posnetkov javnega in zasebnega govora v najrazličnejših vsakodnevnih situacijah. Za podrobnejše slovnične analize pa je metodološko relevanten zlasti njegov reprezentativni vzorec z ročno pripisanimi oblikoslovnimi in skladenjskimi informacijami, drevesnica govorjene slovenščine SST (angl. *Spoken Slovenian Treebank*, Dobrovoljc in Nivre 2016).

V nasprotju z govornimi drevesnicami za druge jezike (za pregled gl. Dobrovoljc 2024a), ki se poleg jezikovnotehnoloških aplikacij vse bolj uporabljajo tudi za

jezikoslovne raziskave (van der Wouden idr. 2003; Hinrichs in Kübler 2005; Roland idr. 2007; Van Eynde 2009; Pietrandrea in Delsart 2019), metodološki potencial drevesnice SST za korpusnojezikoslovne raziskave govorjene slovenščine doslej še ni bil izkoriščen, kar sicer velja za slovenske slovnično označene korpuse nasploh (Ledinek 2018).

Da bi premostili to vrzel med porastom raziskav govorjene slovenščine na eni strani in razmeroma nepoznanim jezikovnim virom za tovrstne analize na drugi strani, je namen tega prispevka predstaviti uporabnost drevesnice SST za različne tipe korpusnojezikoslovnih raziskav slovenskega govora. V nadaljevanju tako drevesnico SST podrobneje predstavimo (2. razdelek) in ponazorimo način njene uporabe za analizo treh izbranih tipično govorjenih pojavov (3. razdelek), v sklepni diskusiji pa poleg najpomembnejših prednosti izpostavimo tudi nekaj omejitev (4. razdelek).

2 Drevesnica govorjene slovenščine SST

Drevesnica govorjene slovenščine SST je slovnično označeni reprezentativni vzorec referenčnega korpusa Gos. V nadaljevanju na kratko predstavimo zgradbo drevesnice z vidika vsebovanih besedil in slovničnih kategorij, ki so bile tem pripisane, več podrobnosti o zasnovi in izdelavi korpusa pa opisujeta prispevka K. Dobrovoljc s sodelavci (2016, 2024).

2.1 Vsebina

Korpus, na katerem temelji drevesnica SST, je bil v okviru doktorske raziskave leksikalnih prvin govorjenega jezika (Dobrovoljc 2018b) zasnovan kot reprezentativni vzorec takratnega referenčnega korpusa govorjene slovenščine, korpusa Gos 1.0 (Verdonik in Zwitter Vitez 2011; Zwitter Vitez idr. 2013) in obsega nekaj manj kot 30.000 besed. Z namenom ohranjanja raznolikosti in uravnoteženosti govornih situacij in demografskih značilnosti govorcev so v bili v drevesnico SST vključeni krajši izseki vseh 287 govornih dogodkov izvornega korpusa Gos, podedovani pa so bili tudi ročni prepisi (Verdonik idr. 2013), kar pomeni, da so meje vlog, izjav in besed v drevesnici SST enake tistim v korpusu Gos.

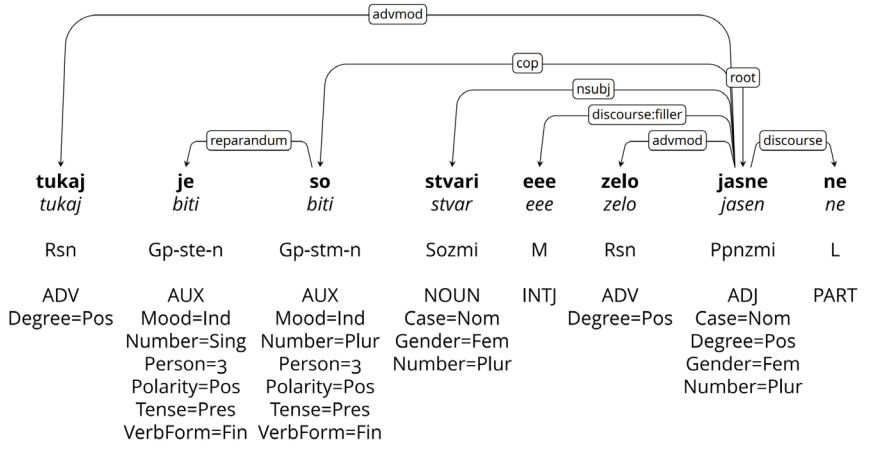
Kot povzema tabela 1, drevesnica SST tako temelji na raznolikem naboru besedil, ki zajemajo najširši spekter govornega sporazumevanja, od javnega informativno-izobraževalnega diskurza (npr. fakultetna predavanja, šolske učne ure, diskusije, intervjuji, informativne oddaje) ter javnega razvedrilnega diskurza (npr. jutranji radijski program, zabavne TV oddaje, resničnostni šovi, športni prenosi) do nejavne nezasebne komunikacije (npr. delovni sestanki, konzultacije, storitve, prodaje) in zasebne komunikacije, kot so pogovori med prijatelji in družinskimi člani.

Tip diskurza	Besedila	Govorci	Izjave	Pojavnice	Delež pojavnic
javni informativno-izobraževalni	129	263	959	9.899	33,5 %
javni razvedrilni	42	78	726	6.833	23 %
nejavni nezasebni	45	102	497	4.535	28 %
nejavni zasebni	71	163	1.006	8.221	15,5 %
SKUPAJ	**287**	**606**	**3.188**	**29.488**	**100 %**

Tabela 1: Velikost in sestava drevesnice SST

2.2 Označevalna shema

Kot prikazuje primer razčlenjene povedi na sliki 1, je vsaki besedi[1] v korpusu SST pripisanih več slovničnih lastnosti, kot so lema, besedna vrsta, oblikoslovne lastnosti in skladenjska vloga v povedi, ki sledijo naboru oznak in načelom dveh (medsebojno povezanih) označevalnih shem.



Slika 1: Primer oblikoslovno in skladenjsko označene izjave v drevesnici SST

Prvi dve vrstici pod zapisanim govorom (tj. odebeljenimi besedami na sliki 1) predstavljajo oznake, pripisane v skladu s shemo MULTEXT-East (Erjavec 2012; Holozan idr. 2023). V slovenskem prostoru je poznana tudi pod imenom JOS, uporablja pa se tudi pri drugih referenčnih jezikovnih virih slovenskega jezika, kot sta korpus Gigafida (Krek idr. 2020) in oblikoslovni leksikon Sloleks (Dobrovoljc idr. 2015). Poleg podatka o osnovni obliki besede oziroma lemi (npr. lema *stvar* za

[1] Med dvema oblikama zapisa govora v korpusu Gos, pogovornim (npr. *tko*) in standardiziranim (npr. *tako*), osnovo drevesnice SST predstavlja standardizirani zapis.

pregibno obliko *stvari*) so besedam pripisane tudi t. i. oblikoskladenjske oznake, ki predstavljajo strnjen opis oblikoslovnih lastnosti oblike (npr. oznaka *Sozmi* za občni samostalnik ženskega spola v imenovalniku množine).

Poleg zgoraj navedenih oznak po lokalno uveljavljeni shemi MTE-JOS vsebuje drevesnica SST še dodatne slovnične oznake, pripisane po mednarodno uveljavljeni shemi Universal Dependencies (de Marneffe idr. 2021), ki si prizadeva za medjezikovno poenoteno slovnično označevanje besedil, tj. enako označevanje enakih slovničnih pojavov v različnih jezikih. Shema UD tako obsega nabor »univerzalnih« jezikoslovnih kategorij (17 oznak za besedne vrste, 24 oznak za oblikoslovne lastnosti, 37 odvisnostnih skladenjskih relacij) in splošnih smernic za njihovo pripisovanje besedilnim korpusom, do danes pa je bila prenesena že na več kot 280 korpusov v več kot 160 jezikih po svetu (Zeman idr. 2023). Med njimi je poleg drevesnice SST tudi drevesnica pisne slovenščine SSJ (Dobrovoljc idr. 2017; Dobrovoljc idr. 2023).

Na ravni oblikoslovja je shema UD precej podobna shemi MTE-JOS, saj za vsako obliko predvideva pripis besedne vrste (npr. oznaka NOUN za občni samostalnik) in drugih oblikoslovnih lastnosti, ki so podane kot pari atributov in njihovih vrednosti (npr. Gender=Fem za ženski spol). Na ravni skladnje pa shema temelji na načelih odvisnostne slovnice (Tesnière 1959; Mel′čuk 1988), ki za vsako besedo v povedi oziroma izjavi določa njeno nadrejeno besedo (jedro zveze) in vrsto njunega skladenjskega razmerja (tip relacije). To lahko na grafični ravni ponazorimo s puščico, ki poteka od nadrejenega k podrejenemu elementu, kot na sliki 1 ponazarja relacija *nsubj* (samostalniški osebek), ki poteka od povedka *so jasne* do osebka *stvari*, pri tem pa je tudi notranja sestava povedka analizirana kot odvisnost veznega glagola od povedkovega določila z relacijo *cop* (kopula). Ko tako skladenjsko analiziramo vsako besedo v povedi, ustvarimo t. i. odvisnostno drevo, v katerem ima lahko vsaka beseda poljubno število odvisnih besed, a natanko eno nadrejeno besedo in z njo povezano skladenjsko vlogo.[2]

Nabor »univerzalnih« skladenjskih relacij sheme UD prikazuje tabela 2, pri čemer so njihove splošne, jezikovno neodvisne opredelitve na voljo na krovni strani projekta (v angleščini),[3] njihov prenos na konkretne skladenjske strukture v slovenščini pa je podrobneje opisan v samostojnem priročniku (Dobrovoljc in Terčon 2023), ki vsebuje tudi številne ponazoritve.

[2] Čeprav je odvisnostno skladenjsko razčlenjevanje v slovenskem prostoru že precej uveljavljen pristop, se shema UD od lokalno razvitega sistema JOS-SYN (Ledinek 2014; Arhar Holdt idr. 2023), ki se osredotoča predvsem na besednozvezno in stavčno skladnjo, razlikuje po tem, da vsebuje bistveno daljši seznam relacij, saj skuša zajeti najširši nabor skladenjskih pojavov v povedi (npr. tudi nepropozicijske pojave izven vezljivostnega vzorca povedka), pri njihovi kategorizaciji pa upošteva tudi strukturne lastnosti podrejenih elementov (npr. ločuje med samostalniškimi-*nsubj* in stavčnimi-*csubj* osebki). V tem vidiku je shema UD bolj sorodna shemi praške odvisnostne drevesnice PDT, na kateri je temeljila *Slovenska odvisnostna drevesnica* (Džeroski idr. 2006), prvi tovrstni korpus v našem prostoru, ki pa ni več aktivno vzdrževan.

[3] Povezava: https://universaldependencies.org/, dostop: 4. 4. 2024.

Nadrejeni \ Podrejeni	Samostalniške zveze	Stavki	Določila	Funkcijske besede
Jedrni stavčni argumenti	*nsubj* *obj* *iobj*	*csubj* *ccomp xcomp*		
Drugi stavčni argumenti	*obl* *vocative* *expl* *dislocated*	*advcl*	*advmod* *discourse*	*aux* *cop mark*
Določila samostalnikov	*nmod* *appos* *nummod*	*acl*	*amod*	*det* *clf* *case*
Priredja	**Večbesedne enote**	**Ohlapne relacije**	**Posebne relacije**	**Drugo**
conj ***cc***	*fixed* *compound* *flat*	*list* *parataxis*	*orphan* *goeswith repa-randun*	*punct root* *dep*

Tabela 2: Seznam odvisnostnih relacij po shemi Universal Dependencies[4] (Vir: de Marneffe idr. 2021)

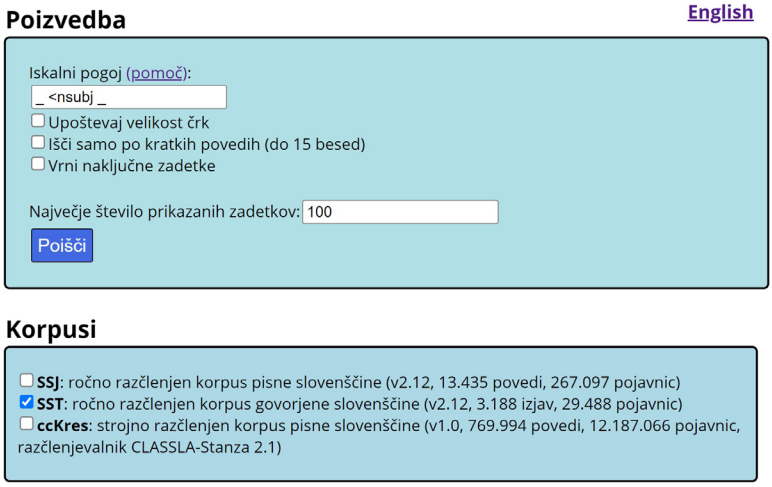
Poleg že izpostavljenih prednosti sheme UD, kot sta mednarodna uveljavljenost in visoka stopnja interoperabilnosti (tj. možnost neposrednih kontrastivnih analiz med drevesnicami različnih jezikov ali različnih jezikovnih zvrsti, kot sta pisni in govorjeni jezik), je bila ta shema za označevanje drevesnice SST izbrana predvsem zato, ker nabor »univerzalnih« skladenjskih oznak (tabela 2) že privzeto vključuje tudi oznake, ki se nanašajo zlasti na tipično govorjene pojave, kot so samopopravljanja (relacija *reparandum*), ogovori (*vocative*) ali diskurzni členki (*discourse*). To v praksi omogoča celosten, enonivojski pristop k slovnični analizi govornih transkripcij, brez kakršnegakoli predhodnega izključevanja netekočnosti in drugih strukturnih posebnosti govora, kot je bilo to pogosto praksa v preteklosti. Shema UD je bila na govorjenih podatkih prvič preizkušena prav na drevesnici SST, odtlej pa še na več kot 40 drugih drevesnicah po svetu, ki vsebujejo (tudi) govorjena besedila, kar potrjuje njeno širše prepoznano uporabnost za skladenjsko razčlenjevanje govorjenega jezika (Kahane idr. 2021; Dobrovoljc 2022).

[4] Približni prevodi relacij, ki se pojavljajo v slovenskih drevesnicah UD: **acl**: stavčni prilastki, **advcl**: prislovni odvisniki, **advmod**: prislovna določila (v širšem smislu, saj so relacija označuje tako prislove v vlogi določil povedka kot prislove v vlogi določil drugih besednih vrst, npr. pridevnikov), **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**: osiroteli argumenti v eliptičnih strukturah, **parataxis**: stavčna soredja, **punct**: ločila, **reparandum**: samopopravljanja, **root**: koren povedi, **vocative**: ogovori, **xcomp**: odprta stavčna dopolnila.

2.3 Dostopnost

Drevesnica SST je odprto dostopna podatkovna zbirka z licenco Creative Commons BY-SA. V standardnem tabelaričnem formatu CONLL-U[5] je distribuirana kot del skupne korpusne zbirke UD, kakršna z vsemi novimi in starimi drevesnicami vred izhaja dvakrat letno (Zeman idr. 2023). Za jezikoslovno analizo drevesnic so bila razvita tudi številna spletna orodja, ki omogočajo iskanje po tako označenih besedilih in vizualizacijo razčlenjenih povedi tudi tehnično manj podkovanim uporabnikom. Izpostavimo lahko orodje Grew-match (Guillaume 2021),[6] ki ga odlikujeta aktivno vzdrževanje in dobra dokumentiranost, lokalno pa je bil v okviru projekta CLARIN.SI s prilagoditvijo odprtokodnega orodja Dep_search (Luotolahti idr. 2017) za ta namen razvit spletni portal Drevesnik (Štravs in Dobrovoljc 2022).[7]

V primerjavi z drugimi obstoječimi portali za brskanje po drevesnicah UD je prednost orodja Drevesnik predvsem to, da ima razmeroma preprost in dobro dokumentiran iskalni jezik,[8] podpira iskanje po oblikoskladenjskih oznakah MTE-JOS, omogoča hkratno poizvedovanje po več korpusih, iskanje pa se lahko omeji na krajše povedi (npr. za didaktične potrebe). Po vnosu iskalnega pogoja (slika 2) se uporabniku prikažejo rezultati v obliki skladenjskih dreves oziroma skladenjsko razčlenjenih izjav (slika 3), v katerih se iskana beseda (npr. beseda z relacijo *nsubj*, ki opravlja vlogo samostalniškega osebka) obarva zeleno. Uporabniki lahko rezultate tudi shranijo, bodisi v obliki seznama zadetkov z besedami v okolici bodisi v obliki podkorpusa vseh prikazanih povedi.



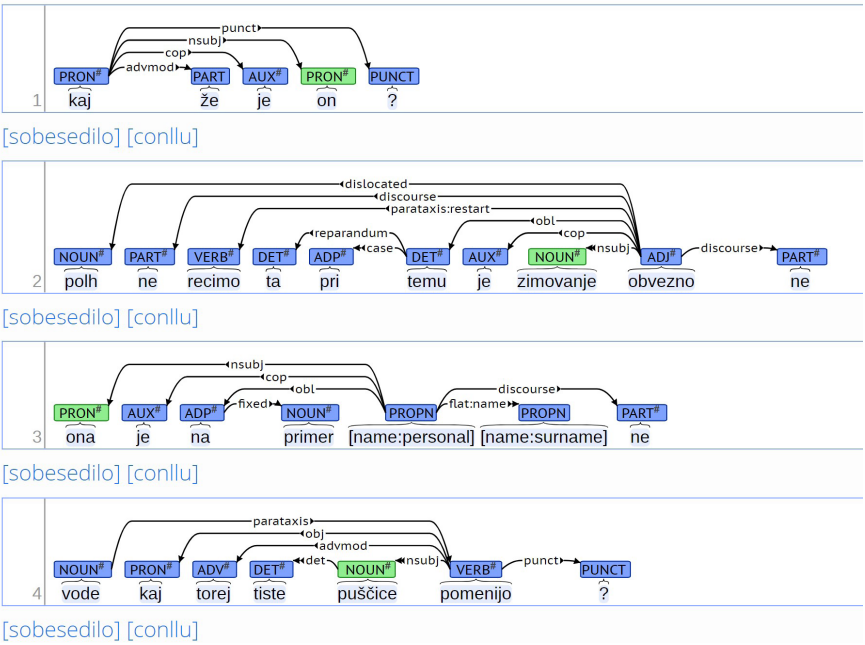
Slika 2: Vmesnik za oblikovanje iskalnega pogoja na portalu Drevesnik s primerom iskanja samostalniških osebkov (*nsubj*)

[5] Povezava: https://universaldependencies.org/format.html, dostop: 4. 4. 2024.

[6] Povezava: https://universal.grew.fr/, dostop: 4. 4. 2024.

[7] Povezava: https://orodja.cjvt.si/drevesnik/, dostop: 4. 4. 2024.

[8] Povezava: https://orodja.cjvt.si/drevesnik/help/, dostop: 4. 4. 2024.



Slika 3: Vmesnik za prikaz rezultatov iskanja po drevesnici SST na portalu Drevesnik s primerom prvih nekaj zadetkov za iskalni pogoj na sliki 2

3 Primer korpusne analize izbranih skladenjskih posebnosti govora

Da bi ponazorili metodološki potencial zgoraj predstavljene drevesnice za slovnične raziskave govorjene slovenščine, v nadaljevanju predstavimo primer korpusne poizvedbe po treh izbranih tipično govorjenih jezikovnih pojavih, ki se pogosto pojavljajo o razpravah o skladenjskih specifikah govorjene slovenščine (Dobrovoljc, v pripravi). Pri tem se osredotočimo tako na kvalitativni kot kvantitativni vidik, saj v prvem delu (razdelek 3.1) predstavimo uporabo korpusa za iskanje avtentičnih primerov rabe, v drugem delu (razdelek 3.2) pa na primeru analize distribucije izbranega pojava v različnih okoliščinah sporazumevanja ponazorimo še uporabo korpusa za različne statistične analize jezikovne rabe.

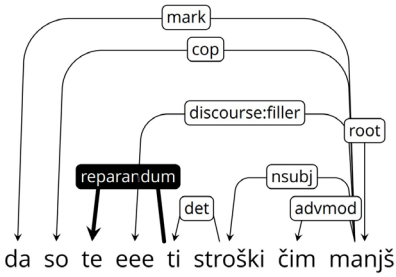
3.1 Priklic avtentičnih primerov rabe

V nadaljevanju predstavimo način iskanja po primerih treh izbranih skladenjskih posebnosti govorjene slovenščine, pri čemer smo iskanje izvedli na različici SST v2.12 na portalu Drevesnik.[9]

[9] V okviru nacionalnega projekta SPOT (*Na drevesnici temelječ pristop k raziskavam govorjene slovenščine*) je v izdelavi sicer nova, razširjena in izboljšana, različica drevesnice SST, ki bo predvidoma obsegala 80.000 pojavnic.

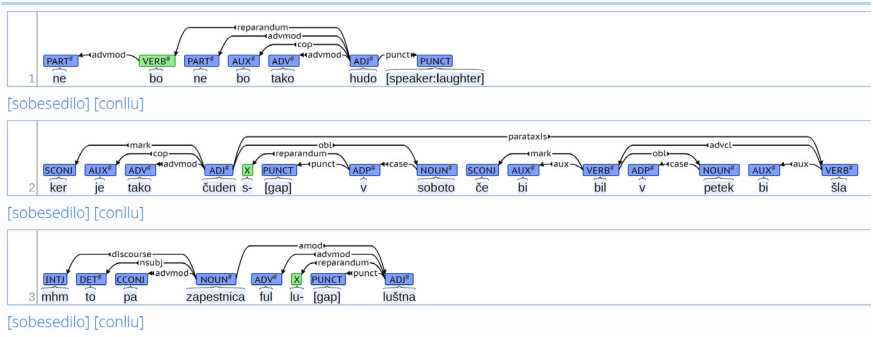
3.1.1 Samopopravljanja

Med najbolj izstopajočimi in v literaturi najpogosteje obravnavanimi jezikovnimi pojavi v govoru so nedvomno različne oblike (samo)popravljanj, s katerimi govorci že izrečeno nadomestijo s popravkom, ki se na sintagmatski osi umešča na isto mesto, pri čemer je lahko popravljena struktura izpeljana v celoti ali zgolj delno, kot v primeru nedokončanih besed, besednih zvez ali stavkov (slika 4). Za označevanje tega pojava se po shemi UD uporablja relacija *reparandum* (Dobrovoljc in Terčon 2023: 121–123), pri čemer je jedro želene ubeseditve oziroma popravka (angl. *repair*) nadrejeno jedru prve oziroma popravljene ubeseditve (angl. *reparandum*).



Slika 4: Označevanje samopopravljanj v drevesnici SST z relacijo *reparandum*

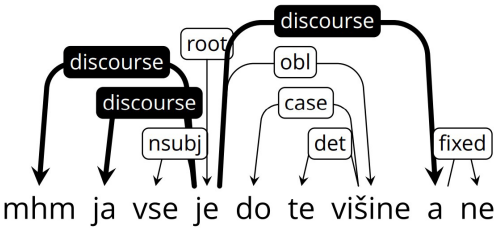
Z orodjem Drevesnik lahko primere samopopravljanj prikličemo s preprostim iskalnim pogojem, ki išče pare besed, povezanih s to relacijo (iskalni pogoj »_ <reparandum _«). Kot ponazarja slika 5, lahko na ta način v drevesnici SST prikličemo 680 primerov samopopravkov v 459 izjavah, primernih za nadaljnjo analizo.



Slika 5: Izsek rezultatov iskanja primerov samopopravljanja v drevesnici SST

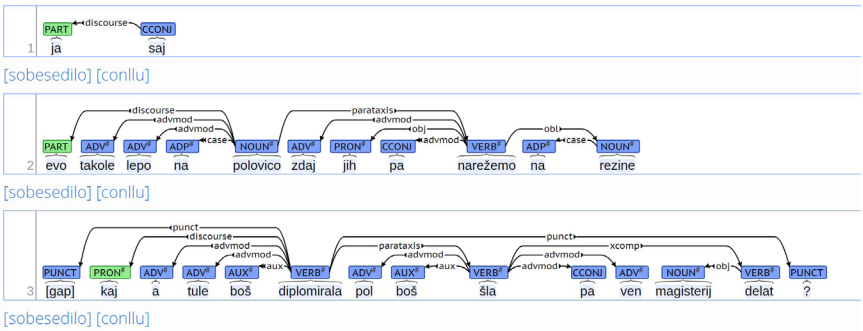
3.1.2 Diskurzni členki

Druga leksikalno-skladenjska značilnost spontanega govora je pogosta raba diskurznofunkcijskih sredstev, s katerimi govorci uravnavajo potek komunikacije in odnose z drugimi udeleženci in za katere se v literaturi najpogosteje pojavljajo poimenovanja, kot so diskurzni označevalci ali diskurzni členki. Med raznolikimi opredelitvami in poimenovanji tovrstnih izrazov, ki tudi v slovenskem prostoru segajo vse od ozkih zamejitev na pomensko najbolj izpraznjene pragmatične izraze (Verdonik 2007) do širokih zamejitev na vse tipe diskurznih organizatorjev (Dobrovoljc 2018b), so po shemi UD tovrstni izrazi opredeljeni predvsem skladenjsko. Gre torej za izraze, ki jih zaznamuje ohlapna vpetost v siceršnjo stavčno strukturo in pojavljanje na skladenjsko perifernih položajih, kot sta začetek ali konec izjave (Dobrovoljc in Terčon 2023: 89–90), v slovenščini pa se v tej vlogi najpogosteje pojavljajo različne medmetne in členkovne besede, kot so *ja*, *ne*, *no*, *tako*, *aha*, *pač*, *zdaj*, *dobro.* Po shemi UD je njihova skladenjska vloga označena z relacijo *discourse*, pri čemer se kot njihov nadrejeni element (dogovorno) izbere jedro najrelevantnejšega stavka (običajno povedek glavnega stavka), kot prikazuje primer na sliki 6 spodaj.



Slika 6: Označevanje diskurznih členkov v drevesnici SST z relacijo *discourse*

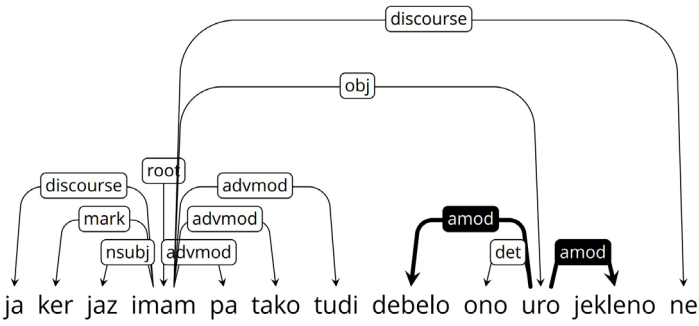
Z orodjem Drevesnik primere tovrstnih struktur poiščemo z iskanjem po parih besed, povezanih s to relacijo (iskalni pogoj »_ <discourse _«), s čimer v drevesnici SST prikličemo 1.587 primerov diskurznih členkov oziroma 1.015 izjav, ki jih vsebujejo (slika 7). Čeprav je korpusni pristop k preučevanju teh izrazov v slovenskem prostoru že precej uveljavljen, je pomembna prednost drevesnice SST to, da so v njej ti večfunkcijski izrazi, ki se lahko v jeziku pojavljajo bodisi kot skladenjsko neintegrirani diskurzni členki (npr. *jah zdaj kako kateri ne*) bodisi kot skladenjsko integrirani stavčni členi različnih tipov (*kaj delaš zdaj*), že razdvoumljeni. V prvem primeru so namreč označeni kot diskurzni členki (oznaka *discourse*), v drugem pa denimo kot prislovna določila (oznaka *advmod*). Priklic ustreznih primerov v drevesnici SST je tako veliko hitrejši kot v neoznačenih ali zgolj oblikoslovno označenih korpusih, v katerih imajo tovrstni izrazi običajno pripisano isto besedno vrsto (npr. prislov).



Slika 7: Izsek rezultatov iskanja primerov diskurznih členkov v drevesnici SST

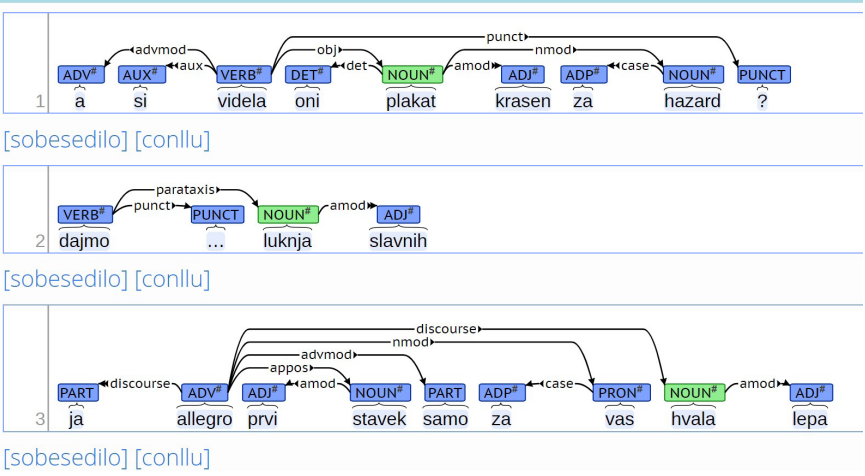
3.1.3 Dodajanje ujemalnih prilastkov

V skladu z dodajalno naravo govorjene komunikacije nasploh je ena pogosto izpostavljenih značilnosti spontanega govora tudi atipična stava prilastkov, zlasti levega ujemalnega pridevniškega prilastka, ki ga govorci v spontanem govoru pogosto dodajajo po izraženi samostalniški odnosnici (npr. *moja stoenka rdeča*). Za ta pojav v shemi UD ni predvidena posebna oznaka ali poseben način označevanja, saj je ena izmed pomembnih prednosti odvisnostnega pristopa k skladenjskemu razčlenjevanju jezika prav neobčutljivost na besednoredne posebnosti posameznih jezikov ali jezikovnih zvrsti. V skladu s tem po shemi UD prilastke, ki sledijo odnosnici, analiziramo enako, kot če bi ti stali pred njo, tj. z relacijo *amod* (pridevniški prilastek), ki poteka od samostalniškega jedra do pridevniškega prilastka. Kot prikazuje primer na sliki 8, sta oba pridevniška prilastka (*debelo* in *jekleno*) povezana z isto relacijo, razlika je zgolj v smeri povezave.



Slika 8: Označevanje pridevniških prilastkov v drevesnici SST z relacijo *amod*

Tudi orodje Drevesnik relacije privzeto išče ne glede na smer povezave (npr. v primeru iz razdelka 3.1.2 tako diskurzne členke pred povedkom kot tiste za njim), lahko pa želeno smer opredelimo z operatorjema @R oziroma @L, ki določa, da naj podrejeni element stoji desno oziroma levo od nadrejenega. Več primerov tovrstnih desnih ujemalnih pridevniških prilastkov lahko torej poiščemo z iskanjem po samostalnikih z desno stavljenim pridevniškim prilastkom (iskalni pogoj »NOUN >amod@R _«) in na ta način prikličemo 34 potencialno relevantnih primerov za nadaljnjo analizo, kot prikazuje slika 9 spodaj.



Slika 9: Izsek rezultatov iskanja primerov desnih pridevniških prilastkov v drevesnici SST

3.2 Statistična analiza jezikovne rabe

Poleg priklica relevantnih primerov rabe za nadaljnje kvalitativne jezikoslovne analize je pomembna prednost drevesnice SST oziroma korpusnih jezikovnih virov nasploh dejstvo, da omogoča tudi kvantitativne analize jezikovne rabe, kakršne so relevantne zlasti za funkcijsko usmerjene jezikoslovne discipline, ki v središče svojega zanimanja postavljajo analize jezikovne rabe in njene odvisnosti od okoliščin sporazumevanja (Stubbs in Halbe 2012; Adolphs in Carter 2013).

Kot primer uporabe drevesnice SST za tovrstne raziskave za prvo zgoraj izpostavljeno značilnost, samopopravljanja (razdelek 3.1.1), v tabeli 3 prikazujemo še pogostost njihovega pojavljanja glede na tip diskurza, sporazumevalni kanal ter starost in spol govorca, tj. število primerov, ki jih z enakim iskanjem najdemo v podkorpusih in ustrezajo posamezni okoliščini. Pri tem poleg absolutne pogostosti (tj. dejanskega števila pojavitev) navajamo tudi relativno pogostost (tj. število pojavitev na 1.000 pojavnic opazovanega podkorpusa), ki omogoča neposredno

primerjavo med posameznimi podkorpusi ne glede na njihov delež v celotnem korpusu.[10]

	Podkorpus	Vseh pojavnic	Samopopravljanja (Abs. pogostost)	Samopopravljanja (Rel. pogostost)
Tip diskurza	javni informativno-izobraževalni	9.899	191	19,3
	javni razvedrilni	6.833	126	18,4
	nejavni nezasebni	4.535	134	29,5
	nejavni zasebni	8.221	229	27,9
Kanal	osebni	13.884	346	24,9
	TV	6.480	122	18,8
	radio	6.126	120	19,6
	telefon	2.998	92	30,7
Spol	ženski	12.659	247	19,5
	moški	16.802	433	25,8
	neznano	27	0	0
Starost	do 10 let	59	2	33,9
	10 do 18 let	1.070	28	26,2
	18 do 34 let	8.536	203	23,8
	35 do 59 let	8.006	199	24,9
	nad 60 let	1.637	49	29,9
	neznano	10.180	199	19,5
	Skupaj	**29.488**	**680**	**23,1**

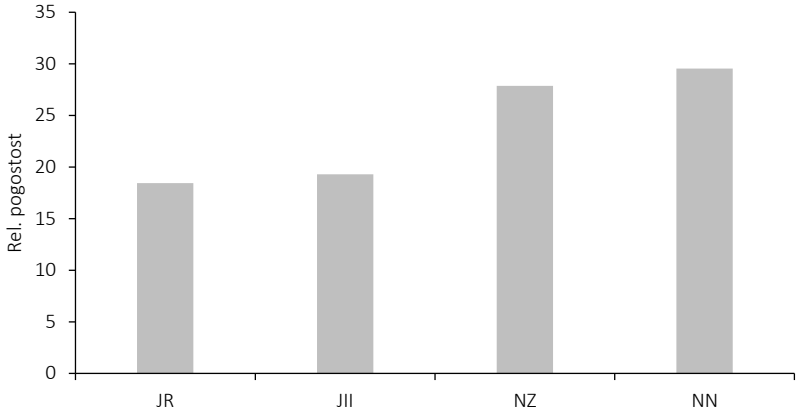
Tabela 3: Pogostost relacije *reparandum* v drevesnici SST glede na izbrane okoliščine in demografske lastnosti govorcev

Kot prikazujejo grafični povzetki v nadaljevanju,[11] se samopopravljanja razmeroma pogosto pojavljajo v vseh štirih tipih diskurza (slika 10), pri čemer je samopopravljanj v javnih govornih položajih manj kot v nejavnih. To ugotovitev potrjuje tudi analiza rabe glede na sporazumevalni kanal (slika 11), ki kaže, da je samopopravljanj

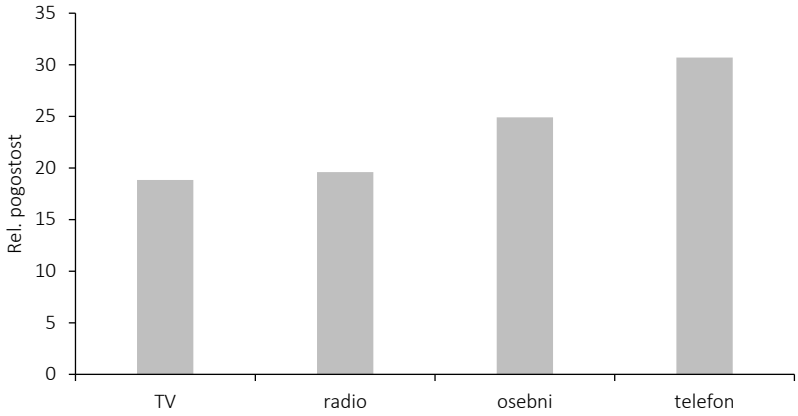
[10] Če ponazorimo: v drevesnici SST imamo podkorpus nejavnega zasebnega govora, ki obsega 8.221 pojavnic, med katerimi se pojavi 229 primerov samopopravljanj. Če bi ta podkorpus obsegal 1.000 pojavnic, bi bilo samopopravljanj 27,9 (229 / 8.221 * 1.000 = 27,9). To je denimo manj pogosto kot v nejavnem nezasebnem diskurzu (134 / 4.535 * 1.000 = 29,5), ki ima v absolutnem smislu sicer manj pojavitev samopopravljanj kot podkorpus zasebnega govora.

[11] V grafih ne prikazujemo kategorije *neznano*.

najmanj v radijskih in televizijskih govornih dogodkih, v katerih sodelujejo izkušenejši javni govorci, ki v komunikacijo vstopajo tudi bolje pripravljeni.



Slika 10: Pogostost samopopravljanja glede na tip diskurza[12]



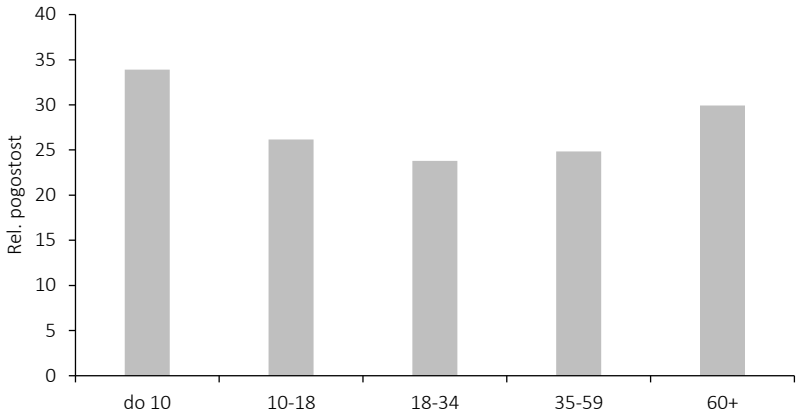
Slika 11: Pogostost samopopravljanja glede na sporazumevalni kanal

Druga zanimiva ugotovitev te ponazoritvene korpusne analize pa je, da je stopnja samopopravljanj morda odvisna tudi od demografski značilnosti govorca. Medtem ko se ta kognitivno-skladenjski mehanizem v drevesnici SST pojavlja približno enakomerno ne glede na starost govorca (slika 12),[13] analiza glede na

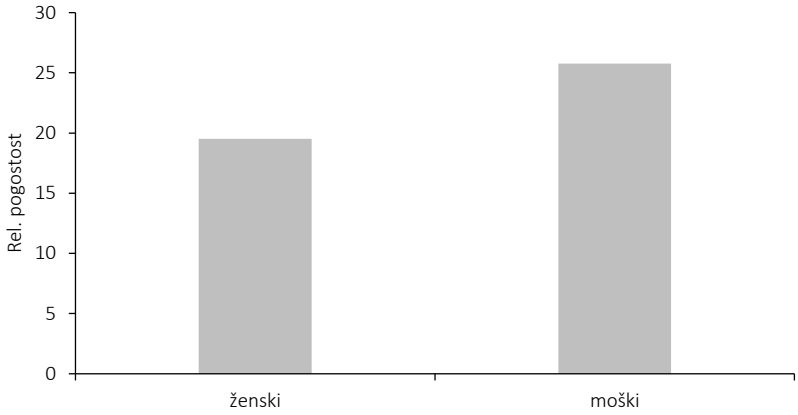
[12] JII = javni informativno-izobraževalni, JR = javni razvedrilni, NN = nejavni nezasebni, NZ = nejavni zasebni.

[13] Izstopajoče relativne pogostosti govorcev, mlajših od 10 let, v interpretaciji ne upoštevamo, saj gre zgolj za 2 pojavitvi samopopravljanj v že tako majhnem podkorpusu, ki v drevesnici SST obsega zgolj 59 besed (tabela 3).

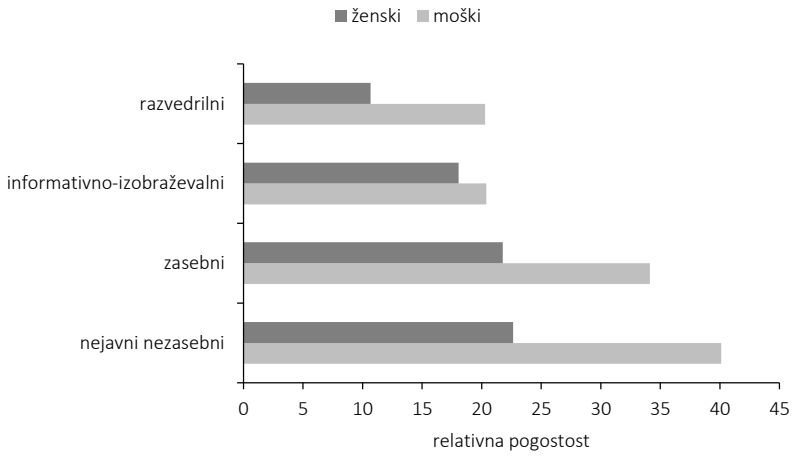
spol govorca (slika 13) kaže, da se moški (samo)popravljajo pogosteje kot ženske. Kot prikazuje slika 14, ta trend opazimo ne glede na tip govornega položaja, pri čemer je razlika med spoloma zlasti očitna v manj formalnih govornih položajih (tj. razvedrilnem in nejavnem govoru). Pri tem je seveda treba poudariti, da glede na omejeno velikost drevesnice SST te ugotovitve niso dokončne, nakazujejo pa zanimivo hipotezo, ki bi jo veljalo preveriti na gradivu večjega obsega.



Slika 12: Pogostost samopopravljanja glede na starost govorca



Slika 13: Pogostost samopopravljanja glede na spol govorca



Slika 14: Pogostost samopopravljanja glede na spol govorca in tip diskurza

4 Diskusija

V 2. in 3. razdelku smo predstavili več metodoloških prednosti govorne drevesnice SST za nadaljnje jezikoslovne raziskave te vseprisotne, a premalo raziskane oblike jezikovnega sporazumevanja. Poleg skrbno zasnovane vsebine, ki po vzoru izvornega korpusa Gos vključuje zapise govora demografsko raznolikih govorcev v najrazličnejših govornih situacijah, so njena najpomembnejša prednost predvsem ročno pripisane slovnične oznake različnih tipov, ki omogočajo enostaven priklic avtentičnih primerov opazovanih jezikovnih pojavov in analizo njihove distribucije v jezikovni rabi. Čeprav smo v prispevku ta potencial ponazorili na primeru nekaj tipično govorjenih pojavov, je korpus na enak način uporaben tudi za raziskave poljubnih drugih leksikalnih, oblikoslovnih ali skladenjskih lastnosti govorjene slovenščine. Poleg zgoraj predstavljenih metod, ki temeljijo na poizvedovanju po vnaprej opredeljenih strukturah, drevesnica SST omogoča tudi številne naprednejše korpusnojezikoslovne analize, kot so merjenje leksikalne raznolikosti ali skladenjske kompleksnosti govorjenega jezika ter avtomatsko odkrivanje tipično govorjenih pojavov. Prav za izvedbo slednjega je pomembna prednost drevesnice SST dejstvo, da temelji na medjezikovno in medžanrsko primerljivi označevalni shemi, s katero so poleg drevesnice pisne slovenščine SSJ označeni še številni drugi (pisni in govorni) korpusi po vsem svetu.

Kljub ponazorjenemu potencialu pa drevesnica SST izkazuje tudi nekatere omejitve. Z vidika vsebine korpusa sta njeni največji pomanjkljivosti majhna velikost in fragmentiranost, saj drevesnica SST trenutno obsega zgolj eno četrtino komplementarne drevesnice pisne slovenščine, zaradi želje po zajemu širokega nabora dogodkov in govorcev pa so izseki vsebovanih besedil zelo kratki in s

tem manj primerni za jezikoslovne analize pojavov, ki segajo nad ravnjo povedi oziroma izjave. Prav tako je zaradi prilagajanja standardnemu formatu sheme UD format korpusa z vidika analiz govorjenega jezika razmeroma osiromašen, saj ne omogoča neposrednega prikaza večplasnosti in kompleksnosti govorjenega jezika, npr. sopostavitve celotnih zvočnih posnetkov, metapodatkov o govorcih in dogodkih ali različnih ravni zapisovanja govora. Nekatere izmed izpostavljenih pomanjkljivosti se sicer aktivno odpravljajo v okviru nacionalnega projekta SPOT (*Na drevesnici temelječ pristop k raziskavam govorjene slovenščine*, ARIS št. Z6-4617), znotraj katerega je nastala tudi večja in izboljšana različica drevesnice SST (Dobrovoljc 2024a, 2024b).

Poleg vsebine in formata korpusa nekatere omejitve izvirajo tudi iz izbrane označevalne sheme. Shema UD je bila namreč zasnovana predvsem kot praktična rešitev za pereč metodološki problem neprimerljivosti slovnično označenih korpusov in se kot taka ne vzpostavlja kot izčrpno utemeljena slovnična teorija, še zlasti pa ne kot slovnična teorija govorjenega jezika. To v praksi pomeni, da shema UD govorjeni jezik opisuje predvsem skozi prizmo njegovega odstopanja od pravil pisnega jezika, ne pa kot avtonomni jezikovni sistem z lastnimi zakonitostmi. Tako kot je opazen trend v jezikoslovju nasploh, kjer so v porastu kognitivno utemeljeni pristopi k opisovanju skladenjskih mehanizmov govorjenega jezika, bi tudi shemo UD in druge podobne označevalne sheme v prihodnosti veljalo dopolniti s spoznanji in opredelitvami tovrstnih raziskav. Za dosego takih ciljev sicer tudi v našem prostoru manjka sodoben slovnični opis, ki bi kot enakovreden del sporazumevalnega kontinuuma vključeval analizo govorjene slovenščine.

Drugotna vloga govorjenega jezika znotraj sheme UD obenem v praksi pomeni, da je oznak, ki se nanašajo na posebnosti govorjenega jezika, malo oziroma so te razmeroma robustne, zato priklic relevantnih jezikovnih pojavov ni vedno enostaven. V literaturi obravnavane skladenjske posebnosti govorjene slovenščine (Dobrovoljc, v pripravi) lahko z vidika težavnosti njihove analize v drevesnici SST razvrstimo v tri temeljne skupine. Poleg že izpostavljenih samopopravljanj, diskurznih členkov in desnih ujemalnih prilastkov (razdelek 3.1) lahko v prvo skupino struktur, ki jih je v drevesnici SST s pomočjo orodij, kot je Drevesnik, mogoče priklicati relativno enostavno, umestimo še tihe in zapolnjene premore, dodajalne oziroma soredne stavke, različne vrste medstavčnih razmerij, brezosebkove stavke, ogovore, določni člen *ta*, nedoločni člen *en* ter deiktike oziroma druge leksikalne posebnosti na podlagi vnaprej določenega seznama. Nekoliko kompleksnejše poizvedbe z naprednejšimi orodji bi zahtevali pojavi v drugi skupini, kamor se umeščajo različne oblike ponavljanj, poročanega govora in elips, skladenjska neskladja, stavčni in drugi vrivki, skladenjski paralelizmi ter netipična stava naslonk in drugih stavčnih členov.

Kot tretjo skupino pa lahko izpostavimo primere skladenjsko nezaključenih izjav ali primere vzajemno grajenih skladenjskih dreves med različnimi udeleženci, saj trenutni način označevanja korpusa ne omogoča njihovega izčrpnega priklica. To

in druge pomanjkljivosti bi lahko v prihodnjih različicah korpusa naslovili z vpeljavo novih označb. Tudi nasploh velja poudariti, da je shema UD zasnovana kot odprtokodni kolaborativni projekt, ki se na podlagi diskusij uporabnikov nenehno razvija, fleksibilna zasnova nabora oznak pa avtorjem posameznih drevesnic omogoča vpeljavo poljubnih oblikoslovnih lastnosti ali skladenjskih pod(oznak) (t. i. izpeljanih relacij), če je potrebno. Tudi drevesnica SST je objavljena kot odprto dostopna podatkovna zbirka, zato je za nadaljnje izboljšave in prilagoditve na voljo najširši raziskovalni skupnosti.

Nenazadnje pa izpostavimo še dejstvo, ki se kot ovira za polni izkoristek metodološkega potenciala drevesnice SST in drugih sorodnih skladenjsko razčlenjenih korpusov kaže v praksi – zapletena zgradba jezikovnega vira. Odvisnostno razčlenjene povedi so namreč kompleksni grafi, po katerih je mogoče učinkovito poizvedovati zgolj z dobrim poznavanjem zaledne označevalne sheme in dovoljšno mero tehničnih veščin, ki jih zahtevajo orodja za korpusno analizo. V približevanje tovrstnih virov jezikoslovni skupnosti je bilo v zadnjem času vloženega kar nekaj truda, denimo z izčrpnim popisom označevalnih smernic (Dobrovoljc in Terčon 2023; Arhar Holdt idr. 2023; Holozan idr. 2023), razvojem specializiranih orodij za analizo skladenjsko razčlenjenih korpusov (Štravs in Dobrovoljc 2022; Dobrovoljc idr. 2024; Brank 2023; Krsnik idr. 2024) in povezanimi izobraževalnimi dogodki (Dobrovoljc 2019). Vendarle pa bi veljalo na dolgi rok okrepiti tudi interdisciplinarno povezovanje razvijalcev jezikovnih virov na eni strani in raziskovalcev posameznih slovničnih pojavov na drugi.

5 Zaključek

V prispevku smo predstavili drevesnico SST, uravnoteženi oblikoslovno in skladenjsko razčlenjeni korpus govorjene slovenščine, ter s ponazoritveno korpusno analizo izbranih tipično govorjenih skladenjskih pojavov skušali prikazati njegovo uporabnost za korpusne raziskave govorjene slovenščine, zlasti za potrebe enostavnega priklica velikega števila avtentičnih primerov in razne statistične analize jezikovne rabe. Kljub izpostavljenim omejitvam, kot sta majhna velikost korpusa in robustna, v pisni jezik usmerjena označevalna shema, drevesnica SST predstavlja pomembno metodološko novost v slovenskem prostoru, saj poleg naprednejših korpusnojezikoslovnih analiz slovenskega govora omogoča tudi neposredne kontrastivne analize z drevesnico pisne slovenščine ter z enako označenimi govornimi korpusi v številnih drugih jezikih. Da bi bil ta pomemben metodološki potencial vira kar najbolje izkoriščen, si je smiselno prizadevati za povečevanje interdisciplinarnega sodelovanja med raziskovalci različnih jezikoslovnih področij. Škoda bi namreč bilo, da kljub izjemnemu napredku na področju razvoja slovnično označenih korpusov na eni strani in orodij za njihovo jezikoslovno analizo na drugi metodologija slovenističnega korpusnega jezikoslovja ostane omejena na analize na podlagi neoznačenih in/ali nereprezentativnih vzorcev jezikovne rabe.

Zahvala

Delo, predstavljeno v prispevku, je sofinancirala Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije 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).

Viri

Brank, Janez, 2023: *Q-CAT Corpus Annotation Tool 1.5*. Repozitorij CLARIN.SI. http://hdl.handle.net/11356/1844. (Dostop 4. 4. 2024.)

Dobrovoljc, Kaja idr., 2024: *Spletni portal CJVT Označevalnik 2.1*. https://orodja.cjvt.si/oznacevalnik/slv/. (Dostop 4. 4. 2024.)

Dobrovoljc, Kaja, 2019: *Slovnične analize ročno označenega korpusa ssj500k z orodjem Q-CAT*. https://videolectures.net/novaSlovnicaLjubljana_dobrovoljc_slovnicne_analize/ (Dostop 4. 4. 2024.)

Krsnik, Luka, Dobrovoljc, Kaja in Robnik-Šikonja, Marko, 2023: *Dependency tree extraction tool STARK 2.0*. Repozitorij CLARIN.SI. http://hdl.handle.net/11356/1899. (Dostop 4. 4. 2024.)

Štravs, Miha in Dobrovoljc, Kaja, 2022: *Service for querying dependency treebanks Drevesnik 1.0*. Repozitorij CLARIN.SI. http://hdl.handle.net/11356/1715. (Dostop 4. 4. 2024.)

Zeman, Daniel idr., 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. (Dostop 4. 4. 2024.)

Zwitter Vitez, Ana idr., 2013: *Spoken corpus Gos 1.0*. Repozitorij CLARIN.SI. http://hdl.handle.net/11356/1040. (Dostop 4. 4. 2024.)

Literatura

Adolphs, Svenja in Carter, Ronald, 2013: *Spoken corpus linguistics: From monomodal to multimodal.* Oxon: Routledge.

Arhar Holdt, Špela, Terčon, Luka, Krek, Simon, Ledinek, Nina, Može, Sara, Saksida, Amanda in Holz, Nanika, 2023: *Navodila za skladenjsko označevanje slovenščine po sistemu JOS-SYN*. Različica 2.0. https://wiki.cjvt.si/books/06-odvisnostna-skladnja-jos-syn/page/oznacevalne-smernice. (Dostop 4. 4. 2024.)

Carter, Ronald in McCarthy, Michael, 2017: Spoken grammar: Where are we and where are we going? *Applied linguistics* 38/1. 1–20. DOI: https://doi.org/10.1093/applin/amu080.

de Marneffe, Marie-Catherine, Manning, Christopher D., Nivre, Joakim in Zeman, Daniel, 2021: Universal Dependencies. *Computational Linguistics* 47/2. 255–308. DOI: https://doi.org/10.1162/coli_a_00402.

Dobrovoljc, Kaja in Nivre, Joakim, 2016: The Universal Dependencies Treebank of Spoken Slovenian. Calzolari, Nicoletta, Choukri, Khalid, Declerck, Thierry, Goggi, Sara, Grobelnik, Marko, Bente, Maegaard, Mariani, Joseph, Mazo, Helene, Moreno, Asuncion, Odijk, Jan in Piperidis, Stelios (ur.): *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016).* Portorož: European Language Resources Association. 1566–1573. https://aclanthology.org/L16-1248. (Dostop 4. 4. 2024.)

Dobrovoljc, Kaja in Terčon, Luka, 2023: *Universal Dependencies: Smernice za označevanje besedil v slovenščini.* Različica 1.3. Ljubljana: Center za jezikovne vire in tehnologije Univerze v Ljubljani. https://wiki.cjvt.si/books/07-universal-dependencies/page/oznacevalne-smernice. (Dostop 4. 4. 2024.)

Dobrovoljc, Kaja, 2018a: Formulaičnost v slovenskem jeziku. *Slovenščina 2.0* 6/2. 67–95. DOI: https://doi.org/10.4312/slo2.0.2018.2.67-95.

Dobrovoljc, Kaja, 2018b: *Leksikalne prvine govorjenega jezika v uporabniških spletnih vsebinah: primer večbesednih diskurznih označevalcev*. Doktorska disertacija. Ljubljana: Filozofska fakulteta, Univerza v Ljubljani. https://repozitorij.uni-lj.si/IzpisGradiva.php?id=106400. (Dostop 4. 4. 2024.)

Dobrovoljc, Kaja, 2022: Spoken Language Treebanks in Universal Dependencies: an Overview. Calzolari, Nicoletta, Béchet, Frédéric, Blache, Philippe, Choukri, Khalid, Cieri, Christopher, Declerck, Thierry, Goggi, Sara, Isahara, Hitoshi, Maegaard, Bente, Mariani, Joseph, Mazo, Hélène, Odijk, Jan in Piperidis, Stelios (ur.): *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille: European Language Resources Association. 1798–1806. https://aclanthology.org/2022.lrec-1.191. (Dostop 4. 4. 2024.)

Dobrovoljc, Kaja, 2024a: Skladenjska drevesnica govorjene slovenščine: stanje in perspektive. Krajnc Ivič, Mira (ur.): *Stanje in perspektive uporabe govornih virov v raziskavah govora.* Maribor: Univerza v Mariboru, Univerzitetna založba. DOI: https://doi.org/10.18690/um.ff.4.2024.

Dobrovoljc, Kaja, 2024b: Extending the Spoken Slovenian Treebank. Arhar Holdt, Špela in Erjavec, Tomaž (ur.): *Zbornik konference Jezikovne tehnologije in digitalna humanistika.* Ljubljana: Inštitut za novejšo zgodovino. 116–146. https://www.sdjt.si/wp/wp-content/uploads/2024/09/JT-DH_2024_Dobrovoljc.pdf. (Dostop 5. 11. 2024.)

Dobrovoljc, Kaja, Erjavec, Tomaž in Krek, Simon, 2017: The Universal Dependencies Treebank for Slovenian. Erjavec, Tomaž, Piskorski, Jakub, Pivovarova, Lidia, Šnajder, Jan, Steinberger, Josef in Yangarber, Roman (ur.): *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing, BSNLP@EACL 2017*. Valencia: Association for Computational Linguistics. 33–38. DOI: https://doi.org/10.18653/v1/W17-1406.

Dobrovoljc, Kaja, Krek, Simon in Erjavec, Tomaž, 2015: Leksikon besednih oblik Sloleks in smernice njegovega razvoja. Gorjanc, Vojko, Gantar, Polona, Kosem, Iztok in Krek, Simon (ur.): *Slovar sodobne slovenščine: problemi in rešitve.* Ljubljana: Znanstvena založba Filozofske fakultete. 80–105. DOI: https://doi.org/10.4312/9789612379759.

Dobrovoljc, Kaja, Terčon, Luka in Ljubešić, Nikola, 2023: Universal Dependencies za slovenščino: nove smernice, ročno označeni podatki in razčlenjevalni model. *Slovenščina 2.0* 11/1. 218–246. DOI: https://doi.org/10.4312/slo2.0.2023.1.218-246.

Dobrovoljc, Kaja, v pripravi: Pregled raziskav skladenjskih posebnosti govorjene slovenščine.

Džeroski, Sašo, Erjavec, Tomaž, Ledinek, Nina, Pajas, Petr, Žabokrtrtsky, Zdenek in Žele, Andreja, 2006: Towards a Slovene Dependency Treebank. Calzolari, Nicoletta, Choukri,

Khalid, Gangemi, Aldo, Maegaard, Bente, Mariani, Joseph, Odijk, Jan in Tapias, Daniel (ur.): *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*. Genoa: European Language Resources Association. 1388–1391. https://aclanthology.org/L06-1068/. (Dostop 4. 4. 2024.)

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

Guillaume, Bruno, 2021: Graph Matching and Graph Rewriting: GREW tools for corpus exploration, maintenance and conversion. Gkatzia, Dimitra in Seddah, Djamé (ur.): *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. [Online]: Association for Computational Linguistics. 168–175. https://aclanthology.org/2021.eacl-demos.21/. (Dostop 4. 4. 2024.)

Hinrichs, Erhard in Kübler, Sandra, 2005: Treebank profiling of spoken and written German. *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories*. Barcelona. 65–76. http://www.sfs.uni-tuebingen.de/~kuebler/papers/GermanEstimation.pdf. (Dostop 4. 4. 2024.)

Holozan, Peter, Krek, Simon, Pivec, Matej, Rigač, Simon, Rozman, Simon, Velušček, Aleš, Pori, Eva in Arhar Holdt, Špela, 2023: *Specifikacije za učni korpus: lematizacija in MSD*. Različica 2.0. https://wiki.cjvt.si/books/04-oblikoskladnja-multext-east/page/oznacevalne-smernice. (Dostop 4. 4. 2024.)

Kahane, Sylvain, Caron, Bernard, Strickland, Emmett in Gerdes, Kim, 2021: Annotation guidelines of UD and SUD treebanks for spoken corpora: A proposal. Dakota, Daniel, Evang, Kilian in Kübler, Sandra (ur.): *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, Syntaxfest 2021)*. Sofia: Association for Computational Linguistics. 35–47. https://aclanthology.org/2021.tlt-1.4/. (Dostop 4. 4. 2024.)

Krajnc Ivič, Mira, 2004: Besedilnoskladenjske značilnosti javne govorjene besede (na gradivu mariborščine). *Slavistična revija* 52/4. 475–498.

Kranjc, Simona, 1999: *Razvoj govora predšolskih otrok*. Ljubljana: Znanstveni inštitut Filozofske fakultete.

Kranjc, Simona, 2006: *Poglavja iz skladnje otroškega govora*. Domžale: Izolit.

Krek, Simon, Arhar Holdt, Špela, Erjavec, Tomaž, Čibej, Jaka, Repar, Andraž, Gantar, Polona, Ljubešić, Nikola, Kosem, Iztok in Dobrovoljc, Kaja, 2020: Gigafida 2.0: The Reference Corpus of Written Standard Slovene. Calzolari, Nicoletta, Béchet, Frédéric, Blache, Philippe, Choukri, Khalid, Cieri, Christopher, Declerck, Thierry, Goggi, Sara, Isahara, Hitoshi, Maegaard, Bente, Mariani, Joseph, Mazo, Hélène, Moreno, Asuncion, Odijk, Jan in Piperidis, Stelios (ur.): *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille: European Language Resources Association. 3340–3345. https://aclanthology.org/2020.lrec-1.409. (Dostop 4. 4. 2024.)

Ledinek, Nina, 2014: *Slovenska skladnja v oblikoskladenjsko in skladenjsko označenih korpusih slovenščine.* Ljubljana: Založba ZRC. DOI: https://doi.org/10.3986/9789612547479.

Ledinek, Nina, 2018: Skladenjska analiza slovenščine in slovenski jezikoslovno označeni korpusi. *Jezik in slovstvo* 63/2–3. 103–116. DOI: https://doi.org/10.4312/jis.63.2-3.103-116.

Leech, Geoffrey, 2000: Grammars of spoken English: new outcomes of corpus-oriented research. *Language Learning* 50/4. 675–724. DOI: https://doi.org/10.1111/0023-8333.00143.

Luotolahti, Juhani, Kanerva, Jenna in Ginter, Filip, 2017: Dep_search: Efficient Search Tool for Large Dependency Parsebanks. Tiedemann, Jörg in Tahmasebi, Nina (ur.): *Proceedings of the 21st Nordic Conference on Computational Linguistics*. Gothenburg: Association for Computational Linguistics. 255–258. https://aclanthology.org/W17-0233/. (Dostop 4. 4. 2024.)

Marušič, Franc in Žaucer, Rok, 2007: O določnem ta v pogovorni slovenščini (z navezavo na določno obliko pridevnika). *Slavistična revija* 55/1–2. 223–247.

Mel′čuk, Igor A., 1988: *Dependency Syntax: Theory and Practice*. New York: State University Press of New York.

Pietrandrea, Paola in Delsart, Aline, 2019: Macrosyntax at work. Lacheret-Dujour, Anne, Kahane, Sylvain in Pietrandrea, Paola (ur.): *Rhapsodie: A Prosodic and Syntactic Treebank for Spoken French.* John Benjamins. 285–314. DOI: https://doi.org/10.1075/scl.89.

Roland, Douglas, Dick, Frederic in Elman, Jefferey L., 2007: Frequency of basic English grammatical structures: A corpus analysis. *Journal of memory and language* 57/3. 348–379. DOI: https://doi.org/10.1016/j.jml.2007.03.002.

Schlamberger Brezar, Mojca, 2007: Vloga povezovalcev v govorjenem diskurzu. *Jezik in slovstvo* 52/3–4. 21–32. DOI: https://doi.org/10.4312/jis.52.3-4.21-32.

Sinclair, Mch. John in Mauranen, Anna, 2006: *Linear Unit Grammar: Integrating speech and writing.* John Benjamins. DOI: https://doi.org/10.1075/scl.25.

Smolej, Mojca, 2004: Načini tvorjenja govorjenega diskurza – paradigmatska in sintagmatska os. Kržišnik, Erika (ur.): *Aktualizacija jezikovnovrstne teorije na Slovenskem*. Obdobja 22. Ljubljana: Filozofska fakulteta. 423–436. https://centerslo.si/wp-content/uploads/2015/10/22-Smolej.pdf. (Dostop 4. 4. 2024.)

Smolej, Mojca, 2022: *Skladanje: izbrana poglavja iz skladnje slovenskega jezika*. Ljubljana: Založba Univerze v Ljubljani. DOI: https://doi.org/10.4312/9789610606000.

Stubbs, Michael in Halbe, Dorothea, 2012: Corpus Linguistics: Overview. Chapell, A. (ur.): *The Encyclopedia of Applied Linguistics*. Oxford: Blackwell. DOI: https://doi.org/10.1002/9781405198431.wbeal0033.

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

Valh Lopert, Alenka, 2006: Skladenjski elementi govorjenega jezika v jutranjem programu komercialnega radia (Radio City). *Jezikoslovni zapiski* 12/2. 51–62. https://www.dlib.si/details/URN:NBN:SI:DOC-QY5LTB3J. (Dostop 4. 4. 2024.)

van der Wouden, Ton, Schuurman, Ineke, Schouppe, Machteld in Hoekstra, Heleen, 2003: Harvesting Dutch Trees: Syntactic Properties of Spoken Dutch. Gaustad, Tanja (ur.): *Computational Linguistics in the Netherlands 2002*. Leiden: Brill. 129–141. DOI: https://doi.org/10.1163/9789004334441_011.

Van Eynde, Frank, 2009: A Treebank-driven investigation of predicative complements in Dutch. *LOT Occasional Series* 14. 131–145. https://dspace.library.uu.nl/bitstream/handle/1874/297144/bookpart.pdf?sequence=2&isAllowed=y. (Dostop 4. 4. 2024.)

Verdonik, Darinka in Zwitter Vitez, Ana, 2011: *Slovenski govorni korpus GOS*. Ljubljana: Trojina, zavod za uporabno slovenistiko. DOI: https://doi.org/10.4312/9789610603528.

Verdonik, Darinka, 2007: *Jezikovni elementi spontanosti v pogovoru: diskurzni označevalci in popravljanja*. Maribor: Univerzitetna založba Univerze v Mariboru.

Verdonik, Darinka, Dobrovoljc, Kaja, Erjavec, Tomaž in Ljubešić, Nikola, 2024: Gos 2: A New Reference Corpus of Spoken Slovenian. Calzolari, Nicoletta, Kan, Min-Yen, Hoste, Veronique, Lenci, Alessandro, Sakti, Sakriani in Xue, Nianwen (ur.): *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024).* Torino: ELRA and ICCL. 7825–7830. https://aclanthology.org/2024.lrec-main.691/. (Dostop 5. 6. 2024.)

Verdonik, Darinka, Kosem, Iztok, Vitez Zwitter, Ana, Krek, Simon in Stabej, Marko, 2013: Compilation, transcription and usage of a reference speech corpus: The case of the Slovene corpus GOS. *Language resources and evaluation* 47. 1031–1048. DOI: https://doi.org/10.1007/s10579-013-9216-5.

Zemljarič Miklavčič, Jana, 2008: *Govorni korpusi*. Ljubljana: Znanstvena založba Filozofske fakultete. DOI: https://doi.org/10.4312/9789612379902.

Zuljan Kumar, Danila, 2007: *Narečni diskurz: Diskurzivna analiza briških pogovorov*. Ljubljana: Založba ZRC SAZU. DOI: https://doi.org/10.3986/9789612540050.

Zuljan Kumar, Danila, 2019: Besedni red v slovenskem narečnem diskurzu. *Slovenski jezik / Slovene Linguistic Studies* 12. 53–74. DOI: https://doi.org/10.3986/sjsls.12.1.04.

Zuljan Kumar, Danila, 2022: *Skladnja nadiškega in briškega narečja*. Ljubljana: Založba ZRC SAZU. DOI: https://doi.org/10.3986/9789610506195.

Zwitter Vitez, Ana, 2018: Enota analize spontanega govora: interakcija prozodije, pragmatike in skladnje. *Jezik in slovstvo* 63/2–3. 157–175. DOI: https://doi.org/10.4312/jis.63.2-3.157-175.